Information Theory, Learning, and Statistics

John Duchi

October 9, 2026

Contents

1 Introduction and setting **12**
- 1.1 Information theory . . . 12
- 1.2 Moving to statistics and machine learning . . . 13
- 1.3 Outline and chapter discussion . . . 14
- 1.4 How to read or teach from this book . . . 16
- 1.5 A remark about measure theory . . . 18

2 An information theory primer **19**
- 2.1 Basics of information theory . . . 19
  - 2.1.1 Definitions . . . 19
  - 2.1.2 Chain rules . . . 24
  - 2.1.3 Data processing inequalities . . . 27
- 2.2 General divergence measures and definitions . . . 28
  - 2.2.1 Partitions, algebras, and quantizers . . . 28
  - 2.2.2 KL-divergence . . . 28
  - 2.2.3 f-divergences . . . 30
  - 2.2.4 Inequalities and relationships between divergences . . . 33
  - 2.2.5 Convexity and data processing for divergence measures . . . 36
- 2.3 First steps into optimal procedures: testing inequalities . . . 37
  - 2.3.1 Le Cam's inequality and binary hypothesis testing . . . 37
  - 2.3.2 Fano's inequality and multiple hypothesis testing . . . 38
- 2.4 A first operational result: entropy and source coding . . . 40
  - 2.4.1 The source coding problem . . . 41
  - 2.4.2 The Kraft-McMillan inequalities . . . 42
  - 2.4.3 Entropy rates and longer codes . . . 44
- 2.5 Bibliography and further reading . . . 45
- 2.6 Exercises . . . 46

3 Exponential families and statistical modeling **52**
- 3.1 Exponential family models . . . 52
- 3.2 Why exponential families? . . . 54
  - 3.2.1 Analytical properties . . . 54
  - 3.2.2 Fitting an exponential family model . . . 56
- 3.3 Divergence measures and information for exponential families . . . 57
- 3.4 Generalized linear models and regression . . . 59
  - 3.4.1 Fitting a generalized linear model from a sample . . . 61

3.4.2 The information in a generalized linear model . . . 61
3.5 Lower bounds on testing a parameter's value . . . 63
3.6* Deferred proofs . . . 65
3.6.1 Proof of Proposition 3.2.2 . . . 65
3.7 Bibliography and further reading . . . 67
3.8 Exercises . . . 67

I Concentration, information, stability, and generalization 69

4 Concentration inequalities 71
4.1 Basic tail inequalities . . . 71
4.1.1 Sub-Gaussian random variables . . . 73
4.1.2 Sub-exponential random variables . . . 77
4.1.3* Orlicz norms . . . 82
4.1.4 An application of concentration: random projections . . . 84
4.2 Martingale methods . . . 85
4.2.1 Sub-Gaussian martingales and Azuma-Hoeffding inequalities . . . 86
4.2.2 Examples and bounded differences . . . 87
4.3* Matrix concentration . . . 89
4.4 Technical proofs . . . 93
4.4.1 Proof of Theorem 4.1.11 . . . 93
4.4.2 Proof of Theorem 4.1.16 . . . 94
4.4.3 Proof of Proposition 4.3.2 . . . 95
4.5 Bibliography and further reading . . . 96
4.6 Exercises . . . 97

5 Estimation and generalization 103
5.1 Uniformity and metric entropy . . . 103
5.1.1 Symmetrization and uniform laws . . . 104
5.1.2 Rademacher complexity bounds . . . 106
5.1.3 Metric entropy, coverings, and packings . . . 108
5.1.4 Application: concentration of covariance matrices . . . 110
5.2 Generalization bounds . . . 111
5.2.1 Finite and countable classes of functions . . . 112
5.2.2 Large classes . . . 114
5.3 M-estimators and estimation . . . 117
5.3.1 Standard conditions and convex optimization . . . 118
5.3.2 Some growth properties of convex functions . . . 119
5.3.3 Convergence analysis for convex M-estimators . . . 121
5.3.4 Consequences for exponential families and generalized linear models . . . 123
5.3.5 Proof of Theorem 5.3.8 . . . 125
5.4 Bibliography and further reading . . . 126
5.5 Exercises . . . 126

6 Generalization and stability **132**
6.1 The variational representation of Kullback-Leibler divergence . . . 133
6.2 PAC-Bayes bounds . . . 134
6.2.1 Relative bounds . . . 136
6.2.2 A large-margin guarantee . . . 138
6.2.3 A mutual information bound . . . 140
6.3 Interactive and adaptive data analysis . . . 141
6.3.1 The interactive setting . . . 141
6.3.2 Second moment errors and mutual information . . . 142
6.3.3 Limiting interaction in interactive analyses . . . 143
6.3.4 Error bounds for a simple noise addition scheme . . . 147
6.4 Bibliography and further reading . . . 149
6.5 Exercises . . . 150

7 Advanced concentration inequalities **154**
7.1 From divergences to concentration and back . . . 154
7.1.1 Concentration of covariance matrices via the variational representation . . . 156
7.1.2* The connection between moment generating functions and KL-divergence . . . 158
7.2* Transportation inequalities . . . 160
7.2.1 Transportation between finitely supported distributions . . . 160
7.2.2 Transportation between general distributions . . . 162
7.2.3 A tensorized transportation inequality . . . 163
7.2.4 A heuristic proof of Theorem 7.2.2 . . . 165
7.2.5 Proof of Corollary 7.2.3 . . . 166
7.3* Two applications of transportation . . . 166
7.3.1 Metric Gaussianity, transport inequalities, and expansion of sets . . . 166
7.3.2 A weak and strong converse for hypothesis testing . . . 170
7.4 Discussion and bibliographic remarks . . . 173
7.5 Exercises . . . 174

8 Privacy and disclosure limitation **178**
8.1 Disclosure limitation, privacy, and definitions . . . 178
8.1.1 Basic mechanisms . . . 180
8.1.2 Resilience to side information, Bayesian perspectives, and data processing . . . 183
8.2 Weakenings of differential privacy . . . 184
8.2.1 Basic mechanisms . . . 185
8.2.2 Connections between privacy measures . . . 186
8.2.3 Side information protections under weakened notions of privacy . . . 189
8.3 Composition and privacy based on divergence . . . 191
8.3.1 Composition of Rényi-private channels . . . 191
8.3.2 Privacy games and composition . . . 192
8.4 Additional mechanisms and privacy-preserving algorithms . . . 194
8.4.1 The exponential mechanism . . . 194
8.4.2 Local sensitivities and the inverse sensitivity mechanism . . . 197
8.5 Deferred proofs . . . 201
8.5.1 Proof of Lemma 8.2.10 . . . 201
8.6 Bibliography and further reading . . . 203

8.7 Exercises . . . 205

II Fundamental limits and optimality **212**

9 Minimax lower bounds: the Le Cam, Fano, and Assouad methods **214**
9.1 Basic framework and minimax risk . . . 214
9.2 Preliminaries on methods for lower bounds . . . 216
9.2.1 From estimation to testing . . . 216
9.2.2 Inequalities between divergences and product distributions . . . 217
9.2.3 Metric entropy and packing numbers . . . 219
9.3 Le Cam's method . . . 220
9.4 Fano's method . . . 221
9.4.1 The classical (local) Fano method . . . 221
9.4.2 A distance-based Fano method . . . 225
9.5 Assouad's method . . . 229
9.5.1 Well-separated problems . . . 229
9.5.2 From estimation to multiple binary tests . . . 229
9.5.3 Example applications of Assouad's method . . . 231
9.6 Deferred proofs . . . 233
9.6.1 Proof of Proposition 9.4.6 . . . 233
9.6.2 Proof of Corollary 9.4.7 . . . 233
9.6.3 Proof of Theorem 9.5.2 . . . 234
9.7 Bibliography and further reading . . . 234
9.8 Exercises . . . 235

10 Constrained risk inequalities **240**
10.1 Strong data processing inequalities . . . 240
10.2 Local privacy . . . 242
10.3 Communication complexity . . . 245
10.3.1 Classical communication complexity problems . . . 246
10.3.2 Deterministic communication: lower bounds and structure . . . 248
10.3.3 Randomization, conditional information complexity, and direct sums . . . 250
10.3.4* The structure of randomized communication and communication complexity of primitives . . . 254
10.4 Communication complexity in estimation . . . 257
10.4.1 Direct sum communication bounds . . . 258
10.4.2 Communication data processing . . . 259
10.4.3 Applications: communication and privacy lower bounds . . . 261
10.5* Proof of Theorem 10.4.4 . . . 265
10.5.1 Proof of Lemma 10.5.3 . . . 269
10.6 Bibliography and further reading . . . 270
10.7 Exercises . . . 271

11 Asymptotically exact optimality guarantees and score correlation 275
11.1 The Cramér-Rao inequality . . . 276
11.1.1 Compact sets and the failure of the Cramér-Rao bound . . . 277
11.1.2 Regularization and the failure of the Cramér-Rao bound . . . 277
11.2 The Van Trees inequality: a Bayesian Cramér-Rao bound . . . 278
11.2.1 The Van Trees inequality in one dimension . . . 278
11.2.2 The Van Trees inequality in d-dimensions . . . 280
11.2.3 The Van Trees inequality for a function of the parameter . . . 282
11.2.4 The full multivariate Van Trees inequality . . . 283
11.3 Beyond parametric problems . . . 285
11.3.1 Influence function lower bounds . . . 286
11.3.2 An extended example: M-estimation lower bounds . . . 288
11.4 Super-efficiency and instance optimality . . . 289
11.5 Case study: a lower bound for differential privacy . . . 291
11.5.1 The alignment (or score attack) . . . 292
11.5.2* The identifiability inequality . . . 294
11.5.3* The privacy inequality . . . 297
11.5.4* Putting the bounds together and finalizing Theorem 11.5.2 . . . 298
11.6 Bibliography and further reading . . . 299
11.7 Exercises . . . 301

12 Beyond local minimax techniques: a miscellany 303
12.1 Nonparametric regression: minimax upper and lower bounds . . . 303
12.1.1 Kernel estimates of the function . . . 303
12.1.2 Minimax lower bounds on estimation with Assouad's method . . . 306
12.2 Global Fano Method . . . 309
12.2.1 A mutual information bound based on metric entropy . . . 309
12.2.2 Minimax bounds using global packings . . . 310
12.2.3 Example: nonparametric regression . . . 311
12.3 Multiple hypothesis testing and detection . . . 312
12.3.1 The multiple testing problem . . . 313
12.3.2 Optimality results in multiple testing . . . 315
12.3.3 The signal detection problem . . . 318
12.3.4 The sparse signal detection problem . . . 320
12.4* Strong converses and high-probability lower bounds . . . 323
12.4.1 Refined Fano inequalities . . . 324
12.4.2 High probability estimation lower bounds . . . 326
12.4.3 Proof of Theorem 12.4.4 . . . 328
12.5 Bibliography and further reading . . . 329
12.6 Exercises . . . 330

13 Testing and functional estimation 338
13.1 Geometrizing rates of convergence . . . 338
13.1.1 Fisher information and divergence measures . . . 342
13.2 Le Cam's convex hull method . . . 344
13.2.1 The χ^2-mixture bound . . . 345
13.2.2 Estimating the norm of a Gaussian vector and multiple hypothesis testing . . . 347

13.2.3* Lower bounds on estimating integral functionals . . . 349
13.3 Instance-optimal bounds, super-efficiency, and risk transfer . . . 351
13.3.1 Risk transfer inequalities . . . 352
13.3.2 A general risk transfer bound . . . 355
13.3.3* Risk transfer with mixtures . . . 356
13.4 Minimax hypothesis testing . . . 359
13.4.1 Detecting a difference in populations . . . 360
13.4.2 Signal detection and testing a Gaussian mean . . . 361
13.4.3 Goodness of fit and two-sample tests for multinomials . . . 363
13.5* Deferred and technical proofs . . . 366
13.5.1 Valid asymptotic information expansions of divergences . . . 366
13.5.2 Proof of Lemma 13.5.1 . . . 368
13.5.3 Proof of Lemma 13.5.5 . . . 368
13.5.4 Proof of Lemma 13.2.7 . . . 369
13.6 Bibliography and further reading . . . 370
13.7 Exercises . . . 371

III Entropy, predictions, divergences, and information 375

14 Predictions, loss functions, and entropies 377
14.1 Proper losses, scoring rules, and generalized entropies . . . 377
14.1.1 A convexity primer . . . 379
14.1.2 From a proper loss to an entropy . . . 380
14.1.3 The information in an experiment . . . 381
14.2 Characterizing proper losses and Bregman divergences . . . 382
14.2.1 Characterizing proper losses for Y taking finitely many vales . . . 382
14.2.2 General proper losses . . . 384
14.2.3 Proper losses for vector-valued Y . . . 388
14.3 From entropies to convex losses, arbitrary predictions, and link functions . . . 390
14.3.1 Convex conjugate linkages . . . 390
14.3.2 Convex conjugate linkages with affine constraints . . . 393
14.4 Exponential families, maximum entropy, and log loss . . . 395
14.4.1 Maximizing entropy . . . 398
14.5* Technical and deferred proofs . . . 400
14.5.1 Proof of Theorem 14.2.15 . . . 400
14.5.2 Proof of Lemma 14.5.1 . . . 402
14.5.3 Proof of Proposition 14.4.1 . . . 403
14.5.4 Proof of Proposition 14.4.2 . . . 404
14.6 Bibliography and further reading . . . 404
14.7 Exercises . . . 405

15 Calibration and proper losses 408
15.1 Proper losses and calibration error . . . 408
15.2 Measuring calibration . . . 410
15.2.1 The impossibility of measuring calibration . . . 411
15.2.2 Alternative calibration measures . . . 413

15.3 Auditing calibration with proper losses . . . 415
15.3.1 The post-processing gap and calibration audits for squared error . . . 415
15.3.2 Calibration audits for losses based on conjugate linkages . . . 416
15.4 The power of working in quantile space . . . 418
15.4.1 Conformal prediction . . . 418
15.4.2 Calibeating: improving squared error by calibration . . . 420
15.5 Continuous and equivalent calibration measures . . . 422
15.5.1 Calibration measures . . . 423
15.5.2* Equivalent calibration measures . . . 425
15.6* Deferred technical proofs . . . 431
15.6.1 Proof of Lemma 15.2.1 . . . 431
15.6.2 Proof of Proposition 15.5.2 . . . 431
15.6.3 Proof of Lemma 15.5.4 . . . 431
15.6.4 Proof of Theorem 15.5.6 . . . 432
15.7 Bibliography and further reading . . . 433
15.8 Exercises . . . 434

16 Classification, surrogate risk, and loss equivalence 437
16.1 Surrogate risk consistency in binary classification . . . 438
16.1.1 A general classification calibration result . . . 441
16.1.2 Convex losses for binary classification . . . 442
16.1.3 Proof of Theorem 16.1.5 . . . 443
16.2 General surrogate risk consistency . . . 445
16.2.1 Uniform calibration . . . 445
16.2.2 Pointwise calibration . . . 446
16.2.3 Examples: multiclass surrogate risk consistency . . . 447
16.3 Generalized entropies and surrogate risk consistency . . . 449
16.3.1 Multiclass classification and strictly concave entropies . . . 449
16.3.2 Conjugate surrogates for a discrete loss . . . 451
16.3.3 Proof of Theorem 16.3.2 . . . 454
16.4 Universal loss equivalence and entropy . . . 455
16.5* Deferred and technical proofs . . . 457
16.5.1 Proof of Lemma 16.1.8 . . . 458
16.5.2 Proof of Theorem 16.3.8 . . . 458
16.5.3 Proof of Theorem 16.4.1 for binary labels . . . 460
16.5.4 Proof of Lemma 16.5.3 . . . 462
16.5.5 Proof of Lemma 16.5.5 . . . 463
16.6 Bibliography and further reading . . . 463
16.7 Exercises . . . 464

IV Optimization, games, learning, and redundancy 469

17 Stochastic and online convex optimization 471
17.1 Preliminaries on convex optimization . . . 472
17.2 Online convex optimization methods . . . 472
17.2.1 Projected subgradient methods . . . 473

17.2.2 Mirror descent methods . . . 474
17.2.3 Convergence analysis of mirror descent . . . 476
17.2.4 Instantiations of the regret guarantee . . . 477
17.2.5 Proof of Theorem 17.2.4 . . . 479
17.3 Optimality guarantees and fundamental limits . . . 481
17.3.1 From optimization to testing . . . 481
17.3.2 Constructing the optimization lower bound . . . 484
17.3.3 Instantiations and optimality . . . 486
17.3.4 A lower bound for high-dimensional stochastic optimization . . . 488
17.4 A miscellany of online optimization guarantees . . . 489
17.4.1 An online-to-batch conversion . . . 490
17.4.2 A more refined convergence guarantee for entropic mirror descent . . . 491
17.4.3 Proof of Proposition 17.4.2 . . . 492
17.5 Bibliography and further reading . . . 493
17.6 Exercises . . . 494

18 Exploration, exploitation, and bandit problems 499
18.1 The multi-armed bandit problem . . . 499
18.2 Confidence-based algorithms . . . 500
18.3 General losses and information-based bounds . . . 504
18.3.1 An information-based regret bound . . . 505
18.3.2 Posterior (Thompson) sampling . . . 508
18.3.3 Information-based exploration . . . 511
18.3.4 An extended example: linear bandits . . . 513
18.4 Online gradient descent approaches . . . 515
18.4.1 Some empirical comparisons . . . 518
18.5 Minimax lower bounds . . . 518
18.5.1 Action separation and a modulus of continuity . . . 520
18.5.2 Assouad's method for lower bounds . . . 522
18.5.3 Proof of Theorem 18.5.1 . . . 525
18.6 A technical proof . . . 526
18.6.1 Proof of Lemma 18.2.1 . . . 526
18.7 Bibliography and further reading . . . 527
18.8 Exercises . . . 528

19 Minimax games and Bayesian estimation 531
19.1 Robust Bayesian procedures and maximum entropy . . . 532
19.1.1 Saddle points for maximum entropy . . . 533
19.1.2 Exponential family models as robust Bayesian procedures . . . 533
19.2 The coding game and sequential prediction . . . 534
19.3 Expected regret, information capacity, and redundancy . . . 536
19.3.1 Information capacity and regret duality . . . 537
19.3.2 Instantiations and corollaries of regret/capacity duality . . . 539
19.3.3 Maximum generalized entropy and Robust Bayesian procedures . . . 540
19.4 Minimax strategies for regret . . . 541
19.5 Mixture (Bayesian) strategies and redundancy . . . 543
19.5.1 Bayesian redundancy and objective, reference, and Jeffreys priors . . . 546

19.5.2 Heuristic calculations: normality and Theorem 19.5.1 . . . 548
19.6* Regret/capacity dualities and Theorem 19.3.1 with rigor . . . 549
19.6.1 Lower semicontinuity and capacity-achieving priors . . . 549
19.6.2 Duality when the domain is finite . . . 551
19.6.3 Proof of Corollary 19.3.3 . . . 552
19.6.4 Regret/capacity duality for arbitrary domains . . . 554
19.6.5 A formal statement of regret/capacity duality . . . 557
19.7 Bibliography and further reading . . . 558
19.8 Exercises . . . 558

V Appendices **562**

A Miscellaneous mathematical results **563**
A.1 Integral convergence and approximation . . . 563
A.2 Probabilistic convergence . . . 565
A.2.1 Classical results on convergence in distribution . . . 565
A.2.2 Assorted convergence results for probability distributions . . . 566

B Convex analysis: sets and functions **569**
B.1 Convex sets . . . 569
B.1.1 Operations preserving convexity . . . 572
B.1.2 Relative interiors . . . 574
B.1.3 Representation and separation of convex sets . . . 576
B.1.4 Cones attached to a convex set . . . 581
B.2 Convex functions . . . 582
B.2.1 Equivalent definitions of convex functions . . . 583
B.2.2 Continuity properties of convex functions . . . 586
B.2.3 Closure properties of convex functions . . . 589
B.2.4 Sublinear and support functions . . . 593
B.2.5 Operations preserving convexity . . . 595
B.2.6 Smoothness properties, first-order developments for convex functions, and subdifferentiability . . . 597
B.2.7 Calculus rules of subgradients . . . 603
B.3 Exercises . . . 604

C Convex analysis: optimality, stability, and duality **607**
C.1 Optimality conditions and stability properties . . . 608
C.1.1 Subgradient characterizations for optimality . . . 608
C.1.2 Stability properties of minimizers . . . 609
C.2 Conjugacy and duality properties . . . 613
C.2.1 Gradient dualities and the Fenchel-Young inequality . . . 614
C.2.2 Smoothness and strict convexity of conjugates . . . 615
C.2.3 Jensen's inequality . . . 617
C.3 Limits at infinity of convex functions and sets . . . 618
C.3.1 Boundedness and closedness of convex sets . . . 618
C.3.2 Asymptotic growth and existence of minimizers . . . 620

C.3.3 Essentially smooth convex functions 622
C.4 Bibliography and further reading 624
C.5 Exercises 625

Acknowledgements

I have been fortunate to have exceptional mentors, first as an undergraduate and graduate student, and later as a faculty member at Stanford. I must therefore begin with my PhD advisors, Michael Jordan and Martin Wainwright, who set me on the path toward connecting information, optimization, and statistics; they both remain an inspiration. My unofficial PhD advisor, Yoram Singer, introduced me to online and stochastic optimization, and while these topics appear only near the end of the book, the perspectives on convexity and optimality throughout it grew out of our collaboration.

Without the support, discussions, insights, and intellectual stimulation of many colleagues and friends, I would never have begun this book, much less completed it. I would like particularly like to thank Alekh Agarwal, Rina Barber, Peter Bartlett, Emmanuel Candès, Persi Diaconis, Dylan Foster, Elad Hazan, John Lafferty, Percy Liang, Lester Mackey, Sasha Rakhlin, Philippe Rigollet, Aaron Sidford, Adam Smith, Kunal Talwar, Ryan Tibshirani, and Ben Van Roy. I must also thank the faculty, students, and staff in Stanford's Statistics and Electrical Engineering departments, all of whom conspire to provide a delightfully supportive place to work, think, and collaborate.

I have also been lucky to ride the coattails of many students and postdoctoral researchers. Many of them have taken the classes out of which this book grew, provided feedback on it, identified and corrected mistakes in exercises, and changed my perspective on the field. Accordingly, I would like to thank my current and former students and postdoctoral associates, including Alnur Ali, Felipe Areces, Hilal Asi, Yu Bai, Yair Carmon, Maxime Cauchois, Karan Chadha, Chen Cheng, Gary Cheng, John Cherian, Noah Cowan, Isaac Gibbs, Suyash Gupta, Saminul Haque, Tatsunori Hashimoto, Sungyoon Kim, Rohith Kuditipudi, Daniel Levy, Annie Marsden, Chris Mohri, Hongseok Namkoong, Feng Ruan, Aman Sinha, Audrey Xie, Fanny Yang, and Yifan Zhu. It has been a delight to work with them all, once as pupils and now, for many, as colleagues.

Finally, I thank my parents, Mary and Joe, who inspire a commitment to seeking the truth and doing my best to serve students and neighbors. And my family: Emily, Annie, and Jack, who delight me with their joy, endless curiosity, and love.

Chapter 1

Introduction and setting

This book explores the connections between information theory, statistics, computation, and learning. Signal processing, machine learning, and statistics all revolve around extracting useful information from signals and data. In signal processing and information theory, a central question is how to best *design* signals—and the channels over which they are transmitted—to maximally communicate and store information, and to enable the most effective decoding. In machine learning and statistics, by contrast, nature provides a fixed data distribution, and it is the learner's or statistician's goal to recover information about this (unknown) distribution. We will show how information-theoretic perspectives can provide clean techniques to perform this recovery.

The discovery of fundamental limits—results that demonstrate that certain procedures are optimal—forms a central aspect of information theory. Thus, information-theoretic tools allow a characterization of the attainable results in a variety of communication and statistical settings. As we explore in the coming chapters in the context of statistical, inferential, and machine learning tasks, this allows us to develop procedures whose optimality we can certify, so that in various senses, no better procedure is possible. Such results are useful for a myriad of reasons; we would like to avoid making bad decisions or false inferences, we may realize a task is impossible, we can explicitly calculate the amount of data necessary for solving different statistical problems, and by identifying characteristics making a problem difficult, we can design ways to circumvent these challenges.

1.1 Information theory

Information theory addresses several fundamental questions: What is information? How much information content do various signals and data hold? How much information can be reliably transmitted over a noisy communication channel? Instead of comprehensively delineating the discipline, we will instead grossly oversimplify information theory into two main inquiries, with corresponding chains of tasks.

1. How much information does a signal contain?
2. How much information can a noisy channel reliably transmit?

In this context, we provide two main high-level examples, one for each of these tasks.

Example 1.1.1 (Source coding)**:** The source coding (or data compression) problem is to take information from a source, encode it, decode it, and recover the original message. Graphically, we have

$$\mathsf{Source} \;\to\; \mathsf{Encoder} \;\to\; \mathsf{Decoder} \;\to\; \mathsf{Receiver}$$

The question, then, is how to design an encoder and decoder that use as few bits as possible to describe a source (or a message) while preserving the information necessary for the receiver to reconstruct the message with high probability. This fewest number of bits is then the information content of the source (signal). $\diamond$

Example 1.1.2: The channel coding, or data transmission, problem is similar to the source coding problem of Example 1.1.1, except that a noisy *channel* intervenes between the encoder and decoder. The graphical representation becomes

$$\text{Source} \rightarrow \text{Encoder} \rightarrow \text{Channel} \rightarrow \text{Decoder} \rightarrow \text{Receiver}$$

Here one investigates the maximum number of bits that may be sent per each channel use subject to the constraint that the receiver can reconstruct the desired message with small error probability. Because the channel introduces noise, we require some redundancy, and information theory studies the exact amount of redundancy—in the form of additional bits—that must be sent to allow such reconstruction. $\diamond$

1.2 Moving to statistics and machine learning

We advocate a study of statistics and machine learning that keeps in mind these same views. Let us therefore shoehorn statistics and machine learning into source-coding and channel-coding analogies, which will help illuminate the perspective that information-theoretic techniques afford.

In the analogy with source coding, we observe a sequence of data points $X_1, \ldots, X_n$ drawn from some (unknown) distribution P on a space $\mathcal{X}$. For example, we might observe species that biologists collect. Then by analogy, we construct a model (often a generative model) that encodes the data using relatively few bits:

$$\text{Source } (P) \stackrel{X_1,\ldots,X_n}{\longrightarrow} \text{Encoder} \stackrel{\widehat{P}}{\rightarrow} \text{Decoder} \rightarrow \text{Receiver}.$$

We estimate $\widehat{P}$—an empirical version of the distribution P that is easier to communicate than the raw signal $X_1, \ldots, X_n$—with the hope that we learn information about the generating distribution P, or at least describe it efficiently.

We can connect estimation and inference to problems of channel coding; though the analogies are not perfect, they suggest the connections we develop. Consider a statistical problem in which we wish to estimate an unknown function f on a space $\mathcal{X}$, and we observe noisy versions of $f(X_i)$ at points X_i, either chosen by an experimenter or drawn from a population. Recalling the graphical description of Example 1.1.2, we now have a channel $P(Y \mid f(X))$ that gives noisy observations of $f(X)$ for each X_i, and we think of the encoder as choosing those X_i at which to evaluate $f(X_i)$. That is, we have

$$\text{Source } (f) \rightarrow \text{Encoder} \stackrel{f(X_1),\ldots,f(X_n)}{\longrightarrow} \text{Channel } P(Y \mid f(X)) \stackrel{Y_1,\ldots,Y_n}{\longrightarrow} \text{Decoder}.$$

The estimation—decoding—problem is to estimate f. In many statistical problems, we have no choice in the design of the "encoding" $X_1, \ldots, X_n$ as the observations X_i are sampled from a population, and the evaluations $(f(X_1), \ldots, f(X_n))$ provide only a partial view of f, making it different from classical coding in information theory. In some cases that we explore later—such as experimental design, randomized controlled trials, reinforcement learning and bandits (and associated exploration/exploitation tradeoffs)—we can influence the encoding part of the above scheme.

Example 1.2.1: A classical example of the statistical paradigm in this lens is linear regression. Here X_i belong to $\mathbb{R}^d$, and the source $f(x) = \theta^\top x$ for some vector $\theta \in \mathbb{R}^d$. Then the channel is often of the form

$$Y_i = \underbrace{\theta^\top X_i}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}},$$

where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ are independent mean zero normal perturbations. Given a sequence of pairs (X_i, Y_i), we wish to recover the true θ in the linear model.

In *active learning* or *active sensing* scenarios, also known as sequential experimental design, we may choose the sequence X_i to better explore properties of θ. As one concrete idea, without a constraint on measurement power or design norm, choosing vectors with $\|X_i\| \to \infty$ makes the signal $\theta^\top X_i$ dominate the noise. Meaningful formulations therefore impose constraints on the design, such as norm, energy, or sampling constraints. $\diamond$

The remainder of book explores these ideas.

1.3 Outline and chapter discussion

The book separates into four distinct parts, each interacting with the others, but it is possible to read an individual part as a reasonably self-contained unit. The book begins with a review (Chapter 2) that introduces the basic information-theoretic quantities that we discuss: mutual information, entropy, and divergence measures. It is required reading for all the chapters that follow. Chapter 3 provides an overview of exponential family models, which form a core tool in the statistical learning toolbox and provide a ready example in which to interpret our results. Readers familiar with this material can skip it, but it grounds examples and applications in subsequent chapters, and we will return to it throughout the book.

Part I of the book covers what we might term "stability" based results. At a high level, we ask what can be gained by considering situations where individual observations in a sequence of random variables $X_1, \ldots, X_n$ have little effect on functions of the sequence. We begin in Chapter 4 with concentration inequalities, discussing how sums and related quantities can converge quickly. While this material is essential for the remaining chapters, it does not depend on particular information-theoretic techniques. We discuss some applications to problems in statistical learning, especially empirical risk minimization, in this part of the book. Chapter 5 provides results on uniform concentration, with applications to both "generalization"—the standard theoretical tool in machine learning, most typically applying to the accuracy of prediction models—and to estimation problems, where we will give guarantees on estimation of model parameters, which constitute core statistical problems.

We turn in Chapter 6 to investigate generalization and convergence guarantees—arguing that functions of a sample $X_1, \ldots, X_n$ are representative of the full population P from which the sample is drawn—by controlling different information-theoretic quantities. In this context, we develop PAC-Bayesian bounds, and we also use the same framework to present tools to control generalization and convergence in *interactive* data analyses. These types of analyses reflect modern statistics, where one performs some type of data exploration before committing to a fuller analysis, though this adaptivity can invalidate classical statistical guarantees, because the analysis now depends on the sample. Chapter 7 treats more advanced ideas and sophisticated concentration results, such as on random matrices. Finally, Chapter 8 studies disclosure limitation and privacy techniques, all of which repose on different notions of stability in distribution.

Part II studies fundamental limits, using information-theoretic techniques to derive *lower bounds* on the possible rates of convergence for estimation, learning, and other statistical problems. Chapter 9 kicks this off by developing the Assouad, Fano, and Le Cam methods, the three major methods for lower bounds, and it contains the building blocks from which most other lower bound techniques follow. At a high level, we might consider it, along with Part I, as exhibiting the central object of study of this book: how do distributions get close to one another, and how can we leverage that closeness? We give a brief treatment of some lower bounding techniques beyond these approaches in Chapter 12, including applications to certain nonparametric problems, as well as a few results that move beyond the typical lower bounds, which apply in expectation, to some that mimic "strong converses" in information theory, meaning that with exceedingly high probability, one cannot hope to achieve anything better than average case error guarantees.

In modern statistical learning problems, one frequently has concerns beyond statistical risk, such as communication or computational cost, or the privacy of study participants. Accordingly, Chapter 10 develops recent techniques for obtaining optimality guarantees simultaneously along several dimensions, connecting to communication-complexity ideas from information theory. Chapter 11 provides a closer look at optimality results for squared error loss, which admits specialized classical techniques yielding sharp, even constant-level, lower bounds; modern applications of these ideas reappear frequently, as the chapter illustrates. Finally, we conclude the discussion of fundamental limits with testing problems and functional estimation, where one wishes to estimate only a single parameter of a larger model (Chapter 13). While estimating a single scalar may seem, *a priori*, simpler than estimating an entire model, adequately addressing its complexity requires a nuanced treatment and new uses of our information-theoretic tools.

Part III revisits all of our information-theoretic notions from Chapter 2, but instead of simply giving definitions and a few consequences, provides operational interpretations of the different information-theoretic quantities, such as entropy. This includes Shannon's original results on the relationship between coding and entropy (which we overview in Section 2.4.1), but it also interprets entropy and information as measures of uncertainty in statistical experiments and statistical learning, a perspective typically missing from information-theoretic treatments of entropy (Chapter 14). Our treatment shows a deep connection between entropy and loss functions used for prediction, where a duality allows moving back and forth between them.

We connect these ideas to the problem of *calibration* in Chapter 15, where we ask that a prediction model be valid in that, e.g., on 75% of the days the model provides a prediction of 75% of rain, it rains. We also use these information-theoretic notions of risk, entropy, and losses to connect to problems in optimization and machine learning. In particular, Chapter 16 explores when replacing a "true" target loss by an easier-to-optimize surrogate loses essentially nothing. This allows us to delineate when (at least asymptotically) it is possible to computationally efficiently learn good predictors and design good experiments in statistical machine learning problems. Because of the connections with optimization and convex duality, these chapters rest on a nontrivial foundation of convex analysis; Appendices B and C provide a fairly comprehensive review of the results we require. For readers unfamiliar with convex optimization and analysis, these chapters may be tough going; we therefore distinguish the big-picture ideas from the technical conditions necessary for the most general results.

Part IV finishes the book with a treatment of stochastic optimization, online game playing, and minimax problems. Chapter 17 takes a modern perspective on stochastic optimization as minimizing random models of functions, and it includes the "book" proofs of convergence for the workhorses of modern machine-learning optimization. It also leverages the earlier results on fundamental limits to develop optimality theory for convex optimization in the same framework.

Chapter 18 explores online decision-making problems and, more broadly, problems that require exploration and exploitation. This includes bandit problems and basic questions in causal estimation, where information-theoretic tools allow a clean treatment. The concluding Chapter 19 revisits Chapter 14 on loss functions and predictions, but considers it in the context of games between nature and a statistician/learner. Once again leveraging the perspective on entropy and loss functions we have developed, we provide a generalization of the celebrated redundancy–capacity theorem from information theory, recast as a loss minimization game.

1.4 How to read or teach from this book

This book has more material than one could realistically cover in a quarter or semester class, though I have in some iteration of the classes from which this book derives lectured on all of the material in the book. Many of the parts of the book comprise a base chapter—typically the first in the part—off of which the remaining parts can be read nearly independently. As such, the book provides opportunities for a modular approach, where one may choose the chapters and sections of most interest to a given instructor or class.

Nonetheless, I include here both a suggested syllabus for a ten week quarter—roughly twenty 80 minute lectures—offered to advanced undergraduates and beginning graduate students.

1. Lectures 1–2: A blitz through the basics of information theory, covering Sections 2.1–2.3.

2. Lectures 3–4: Material on concentration inequalities and the basics of generalization, which provides motivation for optimality results as well as more sophisticated concentration inequalities (Sections 4.1, 4.2, and 5.1).

3. Lectures 5–6: These lectures continue to cover somewhat classical topics in generalization, but introduce the information-theoretic tools (such as the variational representation of the KL-divergence) to provide stronger concentration and convergence guarantees (Sections 5.1–5.3, and Sections 6.1 and 6.2 of Chapter 6).

4. Lectures 7–10: This is the first of what I would call the modular elements of the course. These two weeks cover interactive and private data analysis (finishing Chapter 6, and then Chapter 8), as they rely on the same types of arguments—limiting information release of procedures—and allow students to connect to real-world concerns.

5. Lectures 11–12: In the sixth week of the class, I shift gears to develop the optimality theory in estimation that inforation-theory permits. These lectures cover Chapter 9.

6. Lectures 13–14: These lectures continue the study of optimal estimators, but focus on estimation with constraints (e.g., privacy), covering Sections 10.1, 10.2, as well as moving beyond minimax to more instance-optimal estimation (Sections 11.1–11.2).

7. Lectures 15–18: I have presented a unit on loss functions, proper losses, forecasting, and scoring. In my courses, this usually revolves around Sections 14.1–14.3, the last of which students typically find quite beautiful, and then proceeding to Sections 15.1–15.3.

8. Lectures 19–20: I have typically concluded the course by a discussion of some part of online and stochastic convex optimization; in recent years, renewed interest in causal inference has motivated covering Chapter 18, particularly Sections 18.3 and 18.5.

Any lecturer should, of course, feel free to substitute different chapters within the book. As the coming chapter dependency graphs illustrate, each part of the book is nearly independent of the others, given sufficient information theory background and preparation, which can make teaching to student interest quite feasible. In some iterations of the course, I have extended the unit on loss functions to include material on surrogate risk consistency, in Chapter 16; for a mathematically sophisticated group of students given an entire semester, I would include Chapter 19, because the generalization of capacity/redundancy duality to arbitrary losses and prediction problems (Theorem 19.3.1) has particular mathematical elegance.

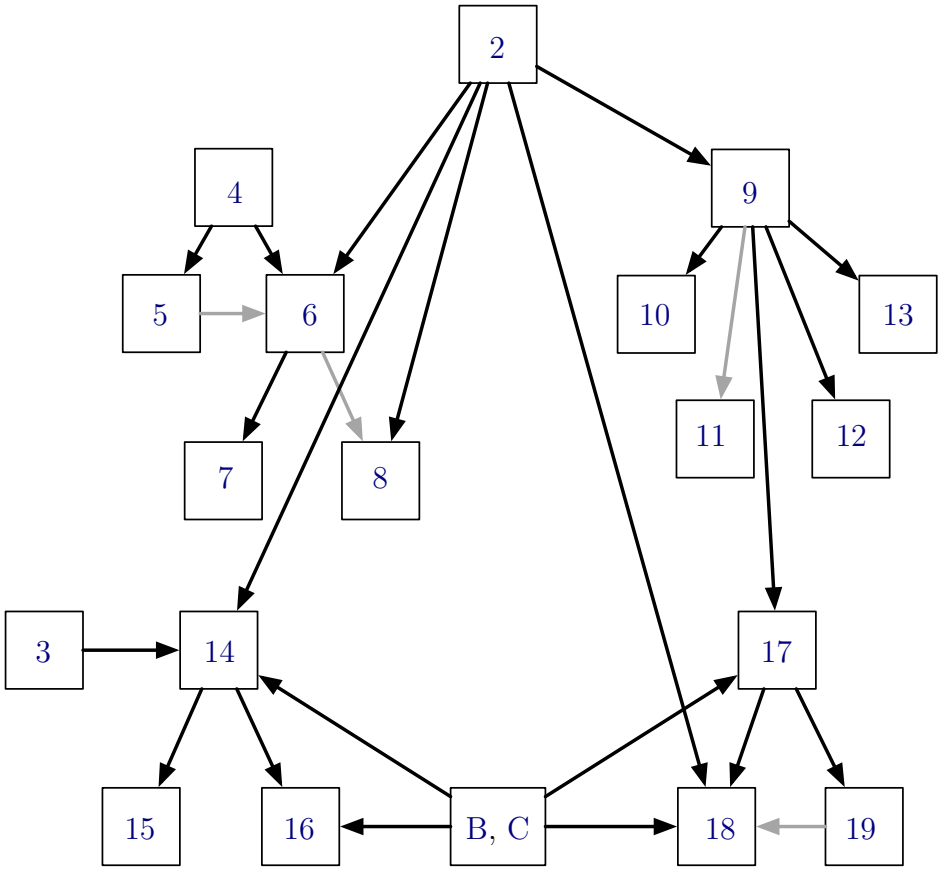


Figure 1.1. A dependency graph of the chapters in the book. Lighter lines indicate a weak dependence: the material in Chapter 5 serves as a strong comparator and baseline for some of the results in Chapter 6, but is not strictly necessary for the material.

I include two illustrations of the chapter dependencies in the book. Figure 1.1 illustrates the strict dependencies: certain chapters (such as Chapter 2, covering information theoretic foundations) necessarily precede other chapters. On the other hand, Figure 1.2 provides a network of motivations. Each dashed line in the figure indicates that the preceding chapter gives motivating problems and examples for, or helps to situate, the following chapter. This sparser graph—it fails to cover each chapter—is by design, as some chapters we trust to be interesting simply on their own. In either case, reading the book linearly provides sufficient structure to follow the paths the figures lay out.

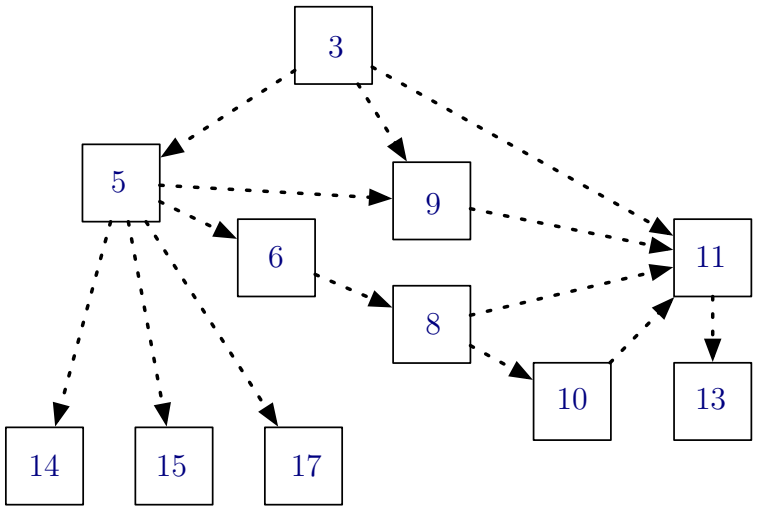


Figure 1.2. A network of motivation: several chapters, particularly Chapter 3, provide motivation and help to situate many of the problems we choose to cover in the book.

1.5 A remark about measure theory

As this book focuses on a number of fundamental questions in statistics, machine learning, and information theory, fully general statements of the results often require measure theory. Thus, formulae such as $\int f(x)dP(x)$ or $\int f(x)d\mu(x)$ appear. While knowledge of measure theory is useful, it is not essential for understanding the proofs and main results. Readers unfamiliar with measure theory may read many of our integrals in the familiar continuous form $\int f(x)p(x)dx$. The most frequent cases we encounter will involve expectations of a function f of random variable X following distribution P, that is, $\mathbb{E}_P[f(X)]$. Normally, we write $\mathbb{E}_P[f(X)] = \int f(x)dP(x)$, or sometimes $\mathbb{E}_P[f(X)] = \int f(x)p(x)d\mu(x)$, saying that "$P$ has density p with respect to the underlying measure μ." In the common case where μ is Lebesgue measure on Euclidean space, one may read the integrals

$$\mathbb{E}_P[f(X)] = \int f(x)dP(x) \quad \text{or} \quad \mathbb{E}_P[f(X)] = \int f(x)p(x)d\mu(x)$$

as

$$\mathbb{E}_P[f(X)] = \int f(x)p(x)dx.$$

Nothing essential will be lost.

Chapter 2

An information theory primer

This introductory chapter reviews the information-theoretic tools used throughout the book: entropy, KL-divergence, mutual information, f-divergences, and the basic testing inequalities that connect them to statistical optimality. As our main goal is to develop the machinery for later applications, our presentation is brisk. A recurring theme is our use of information theory to prove *converse results*: fundamental limits showing that no procedure can improve over a particular benchmark. We give the first of these here to preview more of what is to come, as these fundamental limits form one of the core connections between statistics and information theory. And though many of these quantities appear here as converse tools, later chapters also use them to certify stability, generalization, and convergence. The same information-theoretic quantities also enjoy strong operational interpretations. We illustrate one of them here through the connection between source coding, or lossless compression, and Shannon entropy, and the remainder of the book explores more.

2.1 Basics of information theory

In this section, we review the basic definitions in information theory, including (Shannon) entropy, KL-divergence, mutual information, and their conditional versions. Before beginning, we make one notational convention clear (especially to an information theorist reading this book): any time we use a log, it will be base-e. This is more convenient for our analyses, and it makes differentiating cleaner.

To begin, we will assume that all distributions are discrete; this makes the quantities somewhat easier to manipulate and allows us to avoid complicated measure-theoretic quantities. In Section 2.2, we extend the important definitions (for our purposes)—those of KL-divergence and mutual information—to general distributions, where basic ideas such as entropy no longer make sense. However, for the identities we use most often, quantization lets us reduce many arguments to the discrete case.

2.1.1 Definitions

We provide the basic definitions of entropy, information, and divergence, first for discrete random variables and (later) for variables with densities.

Entropy: We begin with a central concept in information theory: the entropy. Let P be a distribution on a finite (or countable) set $\mathcal{X}$, and let p denote the probability mass function associated with P. That is, if X is a random variable distributed according to P, then $P(X = x) = p(x)$. The *entropy of* X (or of P) in *nats* is

$$H(X) := -\sum_x p(x) \log p(x).$$

Because $p(x) \le 1$ for all x, it is clear that this quantity is positive. We will show later that if $\mathcal{X}$ is finite, the uniform distribution $X \sim \mathsf{Uniform}(\mathcal{X})$ maximizes the entropy, with $H(X) = \log |\mathcal{X}|$.

We will also provide several operational interpretations of the entropy. The most common interpretation—which forms the beginning of Shannon's information theory [377]—is via the source-coding theorem. We present this in Section 2.4.1, where we show that if we wish to encode a random variable X, distributed according to P, with a k-ary string (i.e. each entry of the string takes on one of k values), then the minimal expected length of the encoding is given by $H(X) = -\sum_x p(x) \log_k p(x)$. Moreover, this is achievable (to within a length of at most 1 symbol) by using Huffman codes (among many other types of codes). As an example, consider encoding a random variable X with equi-probable distribution on m items, which has $H(X) = \log(m)$. In base-2, this makes sense: we simply assign an integer to each item and encode each integer with the natural (binary) integer encoding of length $\lceil \log m \rceil$.

We can also define the *conditional entropy*, which is the amount of information left in a random variable after observing another. In particular, we define

$$H(X \mid Y = y) = -\sum_x p(x \mid y) \log p(x \mid y) \quad \text{and} \quad H(X \mid Y) = \sum_y p(y) H(X \mid Y = y),$$

where $p(x \mid y)$ is the p.m.f. of X given that $Y = y$.

Example 2.1.1 (Uniform random variables)**:** If a random variable X is uniform on a set of size m, then $H(X) = \log m$. $\diamond$

Example 2.1.2 (Bernoulli random variables)**:** A $\mathsf{Bernoulli}(p)$ random variable X takes the value $X = 1$ with probability p and $X = 0$ with probability $1 - p$ and has entropy $H(X) = h_2(p) := -p \log p - (1 - p) \log(1 - p)$. This *binary entropy* h_2 will frequently arise. $\diamond$

Example 2.1.3 (Geometric random variables)**:** A random variable X is $\mathsf{Geometric}(p)$, for some $p \in (0, 1]$, if has support $\{1, 2, \ldots\}$, and $P(X = k) = (1 - p)^{k-1} p$; this is the probability distribution of the number X of $\mathsf{Bernoulli}(p)$ trials until a single success. Exercise 2.4 asks you to prove that the entropy of such a random variable is

$$H(X) = -\frac{1 - p}{p} \log(1 - p) - \log p,$$

which satisfies that $H(X) \uparrow \infty$ as $p \downarrow 0$. $\diamond$

Example 2.1.4 (A discrete random variable with infinite entropy)**:** While most "reasonable" discrete random variables have finite entropy, it is possible to construct distributions with infinite entropy. For example, the random variable X with p.m.f.

$$p(k) = \frac{A}{k \log^2 k} \quad \text{where} \quad A^{-1} = \sum_{k=2}^{\infty} \frac{1}{k \log^2 k} < \infty,$$

on $\{2, 3, \ldots\}$ has $H(X) = +\infty$. (Exercise 2.5 steps through a proof of this result.) $\diamond$

Kullback-Leibler (KL)-divergence: Now we define two additional quantities, both more fundamental than entropy: they can always be defined for any distributions and any random variables, as they measure distance between distributions. Entropy simply makes no sense for non-discrete random variables, let alone random variables with continuous and discrete components, though it proves useful for some of our arguments and interpretations.

Before defining these quantities, we recall the definition of a convex function $f : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ as any bowl-shaped function, that is, one satisfying

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{2.1.1}$$

for $\lambda \in [0,1]$ and all x, y. (Here, we adopt the standard practice that $\infty \le \infty$, and say that $f(x) = +\infty$ if $x \notin \operatorname{dom} f$.) The function f is *strictly* convex if the convexity inequality (2.1.1) is strict for $\lambda \in (0,1)$ and $x \ne y$. We recall a standard result:

Proposition 2.1.5 (Jensen's inequality)**.** *Let f be convex and X be a discrete random variable. Then*

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)].$$

If f is strictly convex*, then $f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$ unless X is constant.*

Proof An induction on the size k of the support of X gives the result. The base case $k = 2$ follows immediately from inequality (2.1.1), so we assume the result holds for random variables supported on k values. Let X take values $x_1, \ldots, x_{k+1}$, where $P(X = x_i) = p_i > 0$. Then

$$f(\mathbb{E}[X]) = f\left(\sum_{i=1}^k p_i x_i + p_{k+1} x_{k+1}\right) \le \left(\sum_{i=1}^k p_i\right) f\left(\frac{1}{\sum_{j=1}^k p_j} \sum_{i=1}^k p_i x_i\right) + p_{k+1} f(x_{k+1}),$$

where we use that $\sum_{i=1}^k p_i = 1 - p_{k+1}$. The inequality is strict if f is strictly convex, and by the inductive assumption,

$$f\left(\frac{1}{\sum_{j=1}^k p_j} \sum_{i=1}^k p_i x_i\right) \le \sum_{i=1}^k \frac{1}{\sum_{j=1}^k p_j} p_i f(x_i)$$

because $\sum_{i=1}^k p_i / \sum_{j=1}^k p_j = 1$. □

One may generalize Jensen's inequality; we state one standard result, deferring proof and a few other statements to Appendix C.2.3 as it relies on convex analysis we will develop.

Proposition 2.1.6 (Jensen's inequality, the vector-valued case)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex and X be a random variable with $\mathbb{E}[\|X\|] < \infty$ and $\mathbb{E}[X] \in \operatorname{int} \operatorname{dom} f$. Then*

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)],$$

and moreover, if f is strictly convex, then $f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$ unless X is constant.

Now we may define and provide a few properties of the KL-divergence. Let P and Q be distributions defined on a discrete set $\mathcal{X}$. The *KL-divergence* between them is

$$D_{\rm kl}\left(P \| Q\right) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

We observe immediately that $D_{\rm kl}\left(P\|Q\right) \geq 0$. To see this, we apply Jensen's inequality (Proposition 2.1.5) to the function $-\log$ and the random variable $q(X)/p(X)$, where X is distributed according to P (with the convention that $p \log \frac{p}{0} = +\infty$ for $p > 0$ and $0 \log \frac{0}{q} = 0$):

$$D_{\rm kl}\left(P\|Q\right) = -\mathbb{E}\left[\log \frac{q(X)}{p(X)}\right] \geq -\log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] = -\log\left(\sum_{x:p(x)>0} q(x)\right) \geq 0.$$

Moreover, as $-\log$ is strictly convex, equality holds if and only if $P = Q$. Another consequence of the positivity of the KL-divergence is that whenever the set $\mathcal{X}$ is finite with cardinality $|\mathcal{X}| < \infty$, for any random variable X supported on $\mathcal{X}$ we have $H(X) \leq \log |\mathcal{X}|$. Indeed, letting $m = |\mathcal{X}|$, Q be the uniform distribution on $\mathcal{X}$ so that $q(x) = \frac{1}{m}$, and X have distribution P on $\mathcal{X}$, we have

$$0 \leq D_{\rm kl}\left(P\|Q\right) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -H(X) - \sum_x p(x) \log q(x) = -H(X) + \log m, \tag{2.1.2}$$

so that $H(X) \leq \log m$. Thus, the uniform distribution has the highest entropy over all distributions on the set $\mathcal{X}$.

Mutual information: Having defined KL-divergence, we may now describe the information content between two random variables X and Y. The *mutual information* $I(X;Y)$ between X and Y is the KL-divergence between their joint distribution and their products (marginal) distributions:

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.1.3}$$

We can rewrite this in several ways. First, using Bayes' rule, we have $p(x,y)/p(y) = p(x \mid y)$, so

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(y)p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \\ &= -\sum_x \sum_y p(y)p(x \mid y) \log p(x) + \sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y) \\ &= H(X) - H(X \mid Y). \end{aligned}$$

Similarly, we have $I(X;Y) = H(Y) - H(Y \mid X)$, so mutual information can be thought of as the amount of entropy removed (on average) in X by observing Y. We may also think of mutual information as measuring the difference between the joint distribution of X and Y and the product of their marginals—when they are treated as independent.

Comparing the definition (2.1.3) to that for KL-divergence, we see that if P_{XY} is the joint distribution of X and Y, while P_X and P_Y are their marginal distributions (distributions when X and Y are treated independently), then

$$I(X;Y) = D_{\rm kl}\left(P_{XY}\|P_X \times P_Y\right) \geq 0.$$

Moreover, $I(X;Y) = 0$ if and only if X and Y are independent.

As with entropy, we may also define the *conditional information between X and Y given Z*, which is the mutual information between X and Y when Z is observed (on average). That is,

$$I(X;Y \mid Z) := \sum_z I(X;Y \mid Z = z)p(z) = H(X \mid Z) - H(X \mid Y, Z) = H(Y \mid Z) - H(Y \mid X, Z).$$

Entropies of continuous random variables For continuous random variables, we may define an analogue of the entropy known as *differential entropy*, which for a random variable X with density p is

$$h(X) := -\int p(x) \log p(x) dx. \tag{2.1.4}$$

The differential entropy may be negative—it is no longer directly a measure of the number of bits required to describe a random variable X (on average), as was the case for the entropy. We can similarly define the conditional entropy

$$h(X \mid Y) = -\int p(y) \int p(x \mid y) \log p(x \mid y) dx dy.$$

We remark that the conditional differential entropy of X given Y for Y with arbitrary distribution—so long as X has a density—is

$$h(X \mid Y) = \mathbb{E}\left[-\int p(x \mid Y) \log p(x \mid Y) dx\right],$$

where $p(x \mid y)$ denotes the conditional density of X when $Y = y$. The KL divergence between distributions P and Q with densities p and q becomes

$$D_{\rm kl}\left(P \| Q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx,$$

and similarly, we have the analogues of mutual information as

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X).$$

As we show in the next subsection, we can define the KL-divergence between arbitrary distributions (and mutual information between arbitrary random variables) more generally without requiring discrete or continuous distributions, and the preceding inequalities will follow these definitions. Before investigating these issues, however, we present a few examples.

Example 2.1.7 (Entropy of a uniform random variable)**:** For X uniform on a set $[a, b]$, we have $h(X) = \log(b - a)$. $\diamond$

Example 2.1.8 (Entropy of normal random variables)**:** A random variable has normal distribution with mean μ and variance σ^2, denoted $X \sim \mathsf{N}(\mu, \sigma^2)$, if it has density $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$. To compute the differential entropy (2.1.4), recognize that

$$h(X) = -\int p(x) \left[\frac{1}{2} \log \frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x - \mu)^2\right] = \frac{1}{2} \log(2\pi\sigma^2) + \frac{\mathbb{E}[(X - \mu)^2]}{2\sigma^2} = \frac{1}{2} \log(2\pi e\sigma^2).$$

For a multivariate Gaussian in $\mathbb{R}^n$, where $X \sim \mathsf{N}(\mu, \Sigma)$ for a vector $\mu \in \mathbb{R}^n$ and $\Sigma \succ 0$ with density $p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu))$, a similar calculation yields

$$\begin{aligned} h(X) &= \frac{1}{2}\mathbb{E}\left[n \log(2\pi) + \log \det(\Sigma) + (X - \mu)^\top \Sigma^{-1}(X - \mu)\right] \\ &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma) + \frac{1}{2} \operatorname{tr}(\Sigma\Sigma^{-1}) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log \det(\Sigma). \end{aligned}$$

These entropies are independent of the mean μ, and they grow as Σ increases (at least in the positive semidefinite order). $\diamond$

Continuing our examples with normal distributions, we may compute the divergence between two multivariate Gaussian distributions:

Example 2.1.9 (Divergence between Gaussian distributions)**:** Let P be the multivariate normal $\mathsf{N}(\mu_1, \Sigma)$, and Q be the multivariate normal distribution with mean μ_2 and identical covariance $\Sigma \succ 0$. Then we have that

$$D_{\rm kl}\left(P\|Q\right) = \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2). \tag{2.1.5}$$

We leave the computation of the identity (2.1.5) as Exercise 2.2. $\diamond$

An interesting consequence of Example 2.1.9 is that if a random vector X with a density has covariance $\Sigma \in \mathbb{R}^{n \times n}$, then the multivariate Gaussian with identical covariance has larger differential entropy. Put another way, the Gaussian distribution always maximizes differential entropy for random variables with second moments.

Proposition 2.1.10. *Let X be a random vector on $\mathbb{R}^n$ with a density, and assume that $\mathrm{Cov}(X) = \Sigma \succ 0$. Then for $Z \sim \mathsf{N}(0, \Sigma)$,*

$$h(X) \le h(Z).$$

Proof Without loss of generality, we assume that X has mean 0. Let P be the distribution of X with density p, and let Q be multivariate normal with mean 0 and covariance Σ; let Z be this random variable. Then

$$\begin{aligned} D_{\rm kl}\left(P\|Q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx &= -h(X) + \int p(x) \left[\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det\Sigma + \frac{1}{2}x^\top \Sigma^{-1} x\right] dx \\ &= -h(X) + h(Z), \end{aligned}$$

because Z has the same covariance as X. As $0 \le D_{\rm kl}\left(P\|Q\right)$, we have $h(Z) \ge h(X)$ as desired. $\square$

Among the interesting consequences of entropy maximization—which are beyond our scope—are quantitative central limit theorems; we remark on these in the bibliographic details at the end of the section.

2.1.2 Chain rules

One of our major undertakings throughout this book will be to control the behavior of the vector

$$X_1^n := (X_1, \ldots, X_n)$$

of random variables. (We will use the notation $X_j^i = (X_j, X_{j+1}, \ldots, X_i)$ when convenient.) The most convenient way to do so will be via *chain rules*, where we describe behavior of the vector by the behavior of its constituent parts, that is, the individual X_i. We now illustrate this idea by developing it in the contexts of entropy, KL divergence, and mutual information.

Chain rules for entropy: We begin by describing relationships between the entropy of collections of random variables $X_1, \ldots, X_n$ and individual members of the collection.

Proposition 2.1.11. *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1^{n-1}).$$

This result follows essentially immediately from the standard decomposition of a probability distribution $p(x, y) = p(x)p(y \mid x)$. To see the chain rule, then, note that

$$\begin{aligned} H(X, Y) &= -\sum_{x,y} p(x)p(y \mid x) \log p(x)p(y \mid x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(x) - \sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) = H(X) + H(Y \mid X). \end{aligned}$$

Now set $X = X_1^{n-1}$, $Y = X_n$, and simply induct.

A related corollary of the definitions of mutual information is the well-known result that *conditioning reduces entropy*:

$$H(X \mid Y) \le H(X) \text{ because } I(X; Y) = H(X) - H(X \mid Y) \ge 0.$$

So on average, knowing about a variable Y can only decrease your uncertainty about X. That conditioning reduces entropy for continuous random variables is also immediate, as for X continuous we have $I(X; Y) = h(X) - h(X \mid Y) \ge 0$, so that $h(X) \ge h(X \mid Y)$.

Chain rules for information and divergence: As another immediate corollary to the chain rule for entropy, we see that mutual information also obeys a chain rule:

Proposition 2.1.12. *Let X and $Y_1^n = (Y_1, \ldots, Y_n)$ be random variables. Then*

$$I(X; Y_1^n) = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

This result will hold for any random variables once we provide the general definitions of information and divergence; conveniently, as we will see, we need only address the discrete case, allowing a fairly easy proof of this result. Indeed, we have

$$I(X; Y_1^n) = H(Y_1^n) - H(Y_1^n \mid X) = \sum_{i=1}^n \left[H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})\right] = \sum_{i=1}^n I(X; Y_i \mid Y_1^{i-1}).$$

The KL-divergence obeys similar chain rules, making mutual information and KL-divergence measures useful tools for evaluation of distances and relationships between groups of random variables. We begin with the simplest case, supposing that the distribution $P = P_1 \times P_2 \times \cdots \times P_n$, and $Q = Q_1 \times \cdots \times Q_n$, that is, that P and Q are product distributions over independent random variables $X_i \sim P_i$ or $X_i \sim Q_i$. Then we immediately have the tensorization identity

$$D_{\rm kl}\left(P \| Q\right) = D_{\rm kl}\left(P_1 \times \cdots \times P_n \| Q_1 \times \cdots \times Q_n\right) = \sum_{i=1}^n D_{\rm kl}\left(P_i \| Q_i\right).$$

These identities hold for arbitrary distributions P_i and Q_i or random variables X, Y.

As a final tensorization identity, we consider a more general chain rule for KL-divergences, which will frequently be useful. We abuse notation temporarily, and for random variables X and Y with distributions P and Q, respectively, we denote

$$D_{\rm kl}\left(X \| Y\right) := D_{\rm kl}\left(P \| Q\right).$$

In analogy to the entropy, we can also define the *conditional KL divergence*. Let X and Y have distributions $P_{X|z}$ and $P_{Y|z}$ conditioned on $Z = z$, respectively. Then we define

$$D_{\rm kl}\left(X\|Y \mid Z\right) := \mathbb{E}_Z\left[D_{\rm kl}\left(P_{X|Z}\|P_{Y|Z}\right)\right],$$

so that if Z is discrete we have $D_{\rm kl}\left(X\|Y \mid Z\right) = \sum_z p(z) D_{\rm kl}\left(P_{X|z}\|P_{Y|z}\right)$. With this notation, we have the following chain rule:

Proposition 2.1.13. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be random variables. Then*

$$D_{\rm kl}\left(X_1, \ldots, X_n\|Y_1, \ldots, Y_n\right) = \sum_{i=1}^n D_{\rm kl}\left(X_i\|Y_i \mid X_1^{i-1}\right), \tag{2.1.6}$$

Proof We prove the result in the discrete case, which—as we discuss presently—is fully general for this purpose. For distributions P_{XY} and Q_{XY} we have

$$\begin{aligned} D_{\rm kl}\left(P_{XY}\|Q_{XY}\right) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x)p(y \mid x) \left[\log \frac{p(y \mid x)}{q(y \mid x)} + \log \frac{p(x)}{q(x)}\right] \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \sum_y p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}, \end{aligned}$$

where the final equality uses that $\sum_y p(y \mid x) = 1$ for all x. □

In different notation, for distributions P and Q on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, defining define $P_i(A \mid x_1^{i-1}) = P(X_i \in A \mid X_1^{i-1} = x_1^{i-1})$, and similarly for Q_i, to be the conditional laws of the ith coordinate on the same history, we have the following:

Corollary 2.1.14 (KL chain rule)**.** *Let P, Q be distributions on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Then*

$$D_{\rm kl}\left(P\|Q\right) = \sum_{i=1}^n \mathbb{E}_P[D_{\rm kl}\left(P_i(\cdot \mid X_1^{i-1})\|Q_i(\cdot \mid X_1^{i-1})\right)].$$

Frequently we will find it convenient to bound joint information and entropies by simpler quantities involving single variables. To that end, we give several *tensorization* identities, showing how to transform questions about the joint distribution of many random variables to simpler questions about their marginals. As a first example, because conditioning decreases entropy, for any sequence of (discrete or continuous, as appropriate) random variables, we have

$$H(X_1, \ldots, X_n) \le H(X_1) + \cdots + H(X_n) \quad \text{and} \quad h(X_1, \ldots, X_n) \le h(X_1) + \ldots + h(X_n).$$

The inequalities hold with equality if and only if $X_1, \ldots, X_n$ are mutually independent (see Exercise 2.6).

We return to information and divergence. Frequently, random variables Y_i are independent conditional on a vector X; for example, in a (Bayesian) linear regression problem with fixed design, we draw a random vector T, and conditional on $T = \theta$, draw $Y_i = x_i^\top \theta + \varepsilon_i$, where x_i are known *a priori*, making $Y_i \mid T$ independent. Letting the variables Y_i be independent conditional on X,

$$P(Y_1 = y_1, \ldots, Y_n = y_n \mid X = x) = P(Y_1 = y_1 \mid X = x) \cdots P(Y_n = y_n \mid X = x),$$

we have the inequality

$$\begin{aligned} I(X; Y_1, \ldots, Y_n) &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X, Y_1^{i-1})] \\ &= \sum_{i=1}^n [H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X)] \le \sum_{i=1}^n [H(Y_i) - H(Y_i \mid X)] = \sum_{i=1}^n I(X; Y_i) \end{aligned} \tag{2.1.7}$$

because conditioning reduces entropy. (See Exercise 2.7 for an example of this result.)

2.1.3 Data processing inequalities

A standard problem in information theory (and statistical inference) is to understand the degradation of a signal after it is passed through some noisy channel (or observation process). The simplest of such results, which we will use frequently, is that we can only lose information by adding noise. In particular, assume we have the Markov chain

$$X \to Y \to Z, \tag{2.1.8}$$

where the graphical structure indicates Z is independent of X given Y. Then we obtain the classical *data processing inequality*.

Proposition 2.1.15. *Let (X, Y, Z) have the dependence structure (2.1.8). Then $I(X; Z) \le I(X; Y)$.*

Proof We expand the mutual information $I(X; Y, Z)$ in two ways:

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y \mid Z) \\ &= I(X; Y) + \underbrace{I(X; Z \mid Y)}_{=0}, \end{aligned}$$

where we note that the final equality follows because X is independent of Z given Y:

$$I(X; Z \mid Y) = H(X \mid Y) - H(X \mid Y, Z) = H(X \mid Y) - H(X \mid Y) = 0.$$

Since $I(X; Y \mid Z) \ge 0$, this gives the result. □

There are related data processing inequalities for the KL-divergence—which we generalize in the next section—as well. In this case, we may consider a relationship $X \to Z$, meaning that given $X = x$, we draw $Z \sim Q(\cdot \mid x)$. If we let P_1 and P_2 be distributions on X and M_1 and M_2 be the induced distributions on Z, that is, $M_i(A) = \int Q(Z \in A \mid X = x) dP_i(x)$, then we have

$$D_{\rm kl}\left(M_1 \| M_2\right) \le D_{\rm kl}\left(P_1 \| P_2\right),$$

the basic KL-divergence data processing inequality. A consequence of this is that, for any function ϕ and random variables X and Y on the same space, we have

$$D_{\rm kl}\left(\phi(X) \| \phi(Y)\right) \le D_{\rm kl}\left(X \| Y\right).$$

We explore these data processing inequalities more when we generalize KL-divergences in the next section and in the exercises.

2.2 General divergence measures and definitions

Having given our basic definitions of mutual information and divergence, we show how the definitions of KL-divergence and mutual information extend to arbitrary distributions P and Q and arbitrary sets $\mathcal{X}$. This requires a bit of setup, including defining set algebras (which, we will see, simply correspond to quantization of the set $\mathcal{X}$), but allows us to define divergences in full generality.

2.2.1 Partitions, algebras, and quantizers

Let $\mathcal{X}$ be an arbitrary space. A *quantizer* on $\mathcal{X}$ is any function q that maps $\mathcal{X}$ to a finite collection of integers. In particular, a quantizer q partitions the space $\mathcal{X}$ into the subsets of $x \in \mathcal{X}$ for which $\mathsf{q}(x) = i$. A related notion is an *algebra* of subsets of $\mathcal{X}$, which is any collection $\mathcal{A}$ satisfying

1. The set $\mathcal{X} \in \mathcal{A}$.
2. The collection of sets $\mathcal{A}$ is closed under finite set operations: union, intersection, and complementation. That is, $A, B \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$, $A \cap B \in \mathcal{A}$, and $A \cup B \in \mathcal{A}$.

There is a 1-to-1 correspondence between quantizers—and their associated partitions of the set $\mathcal{X}$—and finite algebras on a set $\mathcal{X}$.[1] It should be clear that there is a one-to-one correspondence between finite *partitions* of the set $\mathcal{X}$ and quantizers q, so we must argue that finite partitions of $\mathcal{X}$ are in one-to-one correspondence with finite algebras defined over $\mathcal{X}$.

In one direction, consider a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$ and let the sets $A_i = \mathsf{q}^{-1}(\{i\})$ be the associated partition. We define the algebra $\mathcal{A}_{\mathsf{q}}$ as the collection of all finite set operations performed on $A_1, \ldots, A_m$ (which is finite, as finite set operations performed on the partition $A_1, \ldots, A_m$ induce only a finite collection of sets).

For the other direction, consider a finite algebra $\mathcal{A}$ over the set $\mathcal{X}$. To construct a quantizer $\mathsf{q}_{\mathcal{A}}$ corresponding to this algebra, define an *atom* of $\mathcal{A}$ as any non-empty set $A \in \mathcal{A}$ such that if $B \subset A$ and $B \in \mathcal{A}$, then $B = A$ or $B = \emptyset$. We claim there is a unique partition of $\mathcal{X}$ with atomic sets from $\mathcal{A}$; we prove this inductively. For the base case, observe that there is at least 1 atomic set, as $\mathcal{A}$ is finite; call it A_1. For the induction, assume we have atomic sets $A_1, \ldots, A_k \in \mathcal{A}$. Let $B = (A_1 \cup \cdots \cup A_k)^c$ be their complement, which we assume is non-empty (otherwise we have a partition of $\mathcal{X}$ into atomic sets). The complement B is either atomic, in which case the sets $\{A_1, A_2, \ldots, A_k, B\}$ consist of atoms of $\mathcal{A}$ and partition $\mathcal{X}$, or B is not atomic. If B is not atomic, one set of the form $A \cap B$ for $A \in \mathcal{A}$ must be atomic: each belongs to $\mathcal{A}$ and there are a finite number of them. This means there is a non-empty atomic set $A_{k+1} \subset B$. This induction must stop at a finite index m as $\mathcal{A}$ is finite, yielding a collection $A_1, \ldots, A_m$ of disjoint atomic sets in $\mathcal{A}$ for which $\cup_i A_i = \mathcal{X}$. (The uniqueness is an exercise.) We then define the quantizer

$$\mathsf{q}_{\mathcal{A}}(x) = i \quad \text{when } x \in A_i.$$

2.2.2 KL-divergence

We may now present the general definition of a KL-divergence between distributions P and Q as the limit of divergences between finer and finer partitions of their supports; this will hold for *any* pair of distributions. Let P and Q be distributions on a space $\mathcal{X}$. Now, let $\mathcal{A}$ be a finite algebra on

[1] Pedantically, this one-to-one correspondence holds up to permutations of the partition the quantizer induces.

$\mathcal{X}$ (as in the previous section, this is equivalent to picking a partition of $\mathcal{X}$ and then constructing the associated algebra), and let $\mathsf{atoms}(\mathcal{A})$ be its atoms. The KL-divergence between P and Q *conditioned on* $\mathcal{A}$ is

$$D_{\rm kl}\left(P\|Q \mid \mathcal{A}\right) := \sum_{A \in \mathsf{atoms}(\mathcal{A})} P(A) \log \frac{P(A)}{Q(A)}.$$

That is, we simply sum over the partition of $\mathcal{X}$. Another way to write this is as follows. Let $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$ be a quantizer, and define the sets $A_i = \mathsf{q}^{-1}(\{i\})$ to be the pre-images of each i (i.e., the different quantization regions, or the partition of $\mathcal{X}$ that q induces). Then the *quantized* KL-divergence between P and Q is

$$D_{\rm kl}\left(P\|Q \mid \mathsf{q}\right) := \sum_{i=1}^{m} P(A_i) \log \frac{P(A_i)}{Q(A_i)}.$$

We may now give the fully general definition of KL-divergence between P and Q:

$$\begin{aligned} D_{\rm kl}\left(P\|Q\right) &:= \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathcal{A}\right) \quad \text{such that } \mathcal{A} \text{ is a finite algebra on } \mathcal{X}\right\} \\ &= \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathsf{q}\right) \quad \text{such that } \mathsf{q} \text{ quantizes } \mathcal{X}\right\}. \end{aligned} \tag{2.2.1}$$

This also gives a rigorous definition of mutual information. Indeed, if X and Y are random variables with joint distribution P_{XY} and marginal distributions P_X and P_Y, we simply define

$$I(X;Y) = D_{\rm kl}\left(P_{XY}\|P_X \times P_Y\right).$$

When P and Q have densities p and q, we can show the definition (2.2.1) reduces to

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx,$$

while if P and Q both have probability mass functions p and q, then—as we see in Exercise 2.10—the definition (2.2.1) is equivalent to

$$D_{\rm kl}\left(P\|Q\right) = \sum_{x} p(x) \log \frac{p(x)}{q(x)},$$

precisely as in the discrete case. The basic reason for these equalities is that finer partitions can only increase divergence measures (see Exercises 2.10–2.11), so that taking limits of finer and finer partitions is more or less sensible.

Importantly, if the set $\mathcal{X}$ is a product space, meaning that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ for some $n < \infty$ (this is the case for mutual information, for example), then we may assume our quantizer *always* quantizes sets of the form $A = A_1 \times A_2 \times \cdots \times A_n$, that is, Cartesian products. Written differently, when we consider algebras on $\mathcal{X}$, the atoms of the algebra may be assumed to be Cartesian products of sets, and our partitions of $\mathcal{X}$ can always be taken as Cartesian products. Written slightly differently, if P and Q are distributions on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and q^i is a quantizer for the set $\mathcal{X}_i$ (inducing the partition $A_1^i, \ldots, A_{m_i}^i$ of $\mathcal{X}_i$) we may define

$$D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) = \sum_{j_1, \ldots, j_n} P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n) \log \frac{P(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}{Q(A_{j_1}^1 \times A_{j_2}^2 \times \cdots \times A_{j_n}^n)}.$$

Then the general definition (2.2.1) of KL-divergence specializes to

$$D_{\rm kl}\left(P\|Q\right) = \sup\left\{D_{\rm kl}\left(P\|Q \mid \mathsf{q}^1, \ldots, \mathsf{q}^n\right) \quad \text{such that } \mathsf{q}^i \text{ quantizes } \mathcal{X}_i\right\}.$$

So, excepting pathologies (see the bibliographic section), we only need consider "rectangular" sets in the definitions of KL-divergence.

As a consequence of this, we can define general mutual information measures in a quite convenient way, justifying our slapdash approach to chain rules, as in Proposition 2.1.12, where we simply assumed all random variables were discrete. In analogy with the definition (2.2.1) and the rectangular partitions above, however, for any random variables X_0 and X_1^n taking values in abstract spaces $\mathcal{X}_i$, we define

$$I(X_0; X_1^n) := \sup_{m\in\mathbb{N}} \sup_{\mathsf{q}_i:\mathcal{X}_i\to[m]} I\left(\mathsf{q}_0(X_0); \mathsf{q}_1(X_1), \ldots, \mathsf{q}_n(X_n)\right).$$

There is some subtlety here in developing chain rules, because while we may have a conditional independence relationship $X \to Y \to Z$, this does not mean that Z is independent of X given $\mathsf{q}(Y)$ for some quantized value of Y (think of the extreme case that $\mathsf{q}(Y) = 0$). Nonetheless, chain rules extend in these cases [e.g. 218, Corollary 7.14], and in a sense, then, there is no loss of generality in assuming our random variables are discrete in developing the chain rules. (See the bibliographic details for more.)

Measure-theoretic definition of KL-divergence A cleaner integral form makes the general definition (2.2.1) of KL-divergence more useful. Let P and Q be absolutely continuous with respect to a measure μ on $\mathcal{X}$ with densities p and q, respectively. (For example, take $\mu = P + Q$.) Then

$$D_{\rm kl}\left(P\|Q\right) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x). \tag{2.2.2}$$

The proof of equality (2.2.2) is somewhat involved, requiring the technology of Lebesgue integration; see the bibliographic section.

For readers who have not seen measure theory, the interpretation of the equality (2.2.2) should be as follows. When integrating a function $f(x)$, replace $\int f(x)d\mu(x)$ with one of two pairs of symbols: one may simply think of $d\mu(x)$ as dx, so that we are performing standard integration $\int f(x)dx$, or one should think of the integral operation $\int f(x)d\mu(x)$ as summing the argument of the integral, so $d\mu(x) = 1$ and $\int f(x)d\mu(x) = \sum_x f(x)$. (This corresponds to μ being "counting measure" on $\mathcal{X}$.)

2.2.3 f-divergences

A more general notion of divergence is the so-called f-divergence, or Ali-Silvey divergence [16, 122] (see also the alternate interpretations in the article by Liese and Vajda [289]). Here, the definition is as follows. Let P and Q be probability distributions on the set $\mathcal{X}$, and let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. If $\mathcal{X}$ is a discrete set, then the *f-divergence* between P and Q is

$$D_f\left(P\|Q\right) := \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right).$$

More generally, for a set $\mathcal{X}$ and a quantizer $\mathsf{q} : \mathcal{X} \to \{1, \ldots, m\}$, for the partition $A_i = \mathsf{q}^{-1}(\{i\}) = \{x \in \mathcal{X} \mid \mathsf{q}(x) = i\}$ we can define the quantized divergence

$$D_f\left(P \| Q \mid \mathsf{q}\right) = \sum_{i=1}^{m} Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right),$$

giving the general definition of an f divergence (in analogy with the definition (2.2.1) of KL divergence)

$$D_f\left(P \| Q\right) := \sup\left\{D_f\left(P \| Q \mid \mathsf{q}\right) \quad \text{such that } \mathsf{q} \text{ quantizes } \mathcal{X}\right\}. \tag{2.2.3}$$

The definition (2.2.3) often lets us prove identities for f-divergences—including KL-divergence or mutual information—by assuming that all distributions have finite discrete support (i.e., taking a finite quantization and passing to a supremum).

We make a few additional remarks on the definition (2.2.3). First, it is clear that $D_f\left(P \| Q\right) \geq 0$ by Jensen's inequality (Proposition 2.1.5). Indeed, let q quantize $\mathcal{X}$ into sets $\{A_i\}_{i=1}^m$ where $Q(A_i) > 0$ for each i. Then as f is convex and $\sum_i Q(A_i) = 1$,

$$D_f\left(P \| Q \mid \mathsf{q}\right) = \sum_{i=1}^{m} Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right) \geq f\left(\sum_{i=1}^{m} Q(A_i) \frac{P(A_i)}{Q(A_i)}\right) = f(1) = 0.$$

Finer quantization also increases the associated f-divergence. We say that a quantizer q_1 is a *finer* quantizer than q_2, which we write as $\mathsf{q}_1 \preceq \mathsf{q}_2$, if q_1 induces a partition $A_1, \ldots, A_n$ and q_2 induces a partition $B_1, \ldots, B_m$, where for each B_i, there exists a k and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We obtain the following corollary of the definition (2.2.3):

Corollary 2.2.1. *If* $\mathsf{q}_1 \preceq \mathsf{q}_2$, *then*

$$D_f\left(P \| Q \mid \mathsf{q}_1\right) \geq D_f\left(P \| Q \mid \mathsf{q}_2\right).$$

See Exercises 2.10 and 2.12.

There is a measure-theoretic version of the definition (2.2.3), which is frequently easier to use, though defining it properly requires us to use the mathematical appendices on closed convex functions and perspective transforms. Assume without loss of generality that P and Q are absolutely continuous with respect to the base measure μ. The f-divergence between P and Q is then

$$D_f\left(P \| Q\right) := \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x). \tag{2.2.4}$$

This "definition" requires us to address points x such that $q(x) = 0$, which we do by making the integral (2.2.4) depend on how quickly $qf(p/q)$ grows as $q \downarrow 0$. We rely on the *recession function* $f'_\infty(t) := \lim_{\alpha \downarrow 0} \alpha f(1 + t/\alpha)$, a type of asymptotic derivative of f (see Appendix C.3.2; we may replace 1 with any $t_0 \in \operatorname{int} \operatorname{dom} f$ without changing f'_∞). Then we take

$$D_f\left(P \| Q\right) := \int_{q(x)>0} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x) + \int_{q(x)=0} f'_\infty(p(x)) d\mu(x). \tag{2.2.5}$$

(Exercise 2.15 shows an alternative way to write this integral, which can be more convenient.)

In the exercises, we explore several properties of f-divergences, including the quantized representation (2.2.3), showing different data processing inequalities and orderings of quantizers based on the fineness of their induced partitions. Broadly, f-divergences satisfy essentially the same properties as KL-divergence, such as data-processing inequalities, and they provide a generalization of mutual information. We explore f-divergences from additional perspectives later—they are important both for optimality in estimation and related to consistency and prediction problems, as we discuss in Section 16.4.

Examples: Four key f-divergences—the KL-divergence, Hellinger distance (or divergence), variation distance, and χ^2-divergence—will form the basis of many of our results. We give these examples here; in Section 2.2.4 we provide a few examples of their uses as well as providing a few natural inequalities between them.

Example 2.2.2 (KL-divergence)**:** By taking $f(t) = t \log t$, which is convex and satisfies $f(1) = 0$, we obtain $D_f(P\|Q) = D_{\rm kl}(P\|Q)$. $\diamond$

Example 2.2.3 (KL-divergence, reversed)**:** By taking $f(t) = -\log t$, we obtain $D_f(P\|Q) = D_{\rm kl}(Q\|P)$. $\diamond$

Example 2.2.4 (Total variation distance)**:** The total variation distance between probability distributions P and Q defined on a set $\mathcal{X}$ is the maximum difference between probabilities they assign on subsets of $\mathcal{X}$:

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \sup_{A \subset \mathcal{X}} (P(A) - Q(A)), \tag{2.2.6}$$

where the second equality follows by considering complements $P(A^c) = 1 - P(A)$. The total variation distance, as we shall see later, is important for verifying the optimality of different tests, and appears in the measurement of difficulty of solving hypothesis testing problems. With the choice $f(t) = \frac{1}{2}|t - 1|$, we obtain the total variation distance, that is, $\|P - Q\|_{\rm TV} = D_f(P\|Q)$. There are several alternative characterizations, which we provide as Lemma 2.2.5 next; it will be useful in the sequel when we develop inequalities relating the divergences. $\diamond$

Lemma 2.2.5. *Let P and Q be probability distributions with densities p, q with respect to a base measure μ and $f(t) = \frac{1}{2}|t - 1|$. Then*

$$\begin{aligned}
\|P - Q\|_{\rm TV} = D_f(P\|Q) &= \frac{1}{2} \int |p(x) - q(x)| d\mu(x) \\
&= \int [p(x) - q(x)]_+ \, d\mu(x) = \int [q(x) - p(x)]_+ \, d\mu(x) \\
&= P(dP/dQ > 1) - Q(dP/dQ > 1) = Q(dQ/dP > 1) - P(dQ/dP > 1).
\end{aligned}$$

In particular, the set $A = \{x \mid p(x)/q(x) \geq 1\}$ maximizes $P(B) - Q(B)$ over $B \subset \mathcal{X}$ and so achieves $\|P - Q\|_{\rm TV} = P(A) - Q(A)$.

Proof Eliding the measure-theoretic details,[2] we immediately have

$$\begin{aligned}
D_f(P\|Q) &= \frac{1}{2} \int \left| \frac{p(x)}{q(x)} - 1 \right| q(x) d\mu(x) = \frac{1}{2} \int |p(x) - q(x)| d\mu(x) \\
&= \frac{1}{2} \int_{x: p(x) > q(x)} [p(x) - q(x)] \, d\mu(x) + \frac{1}{2} \int_{x: q(x) > p(x)} [q(x) - p(x)] \, d\mu(x) \\
&= \frac{1}{2} \int [p(x) - q(x)]_+ \, d\mu(x) + \frac{1}{2} \int [q(x) - p(x)]_+ \, d\mu(x).
\end{aligned}$$

[2] To make this fully rigorous, we would use the Hahn decomposition of the signed measure $P - Q$ to recognize that $\int f(dP - dQ) = \int f [dP - dQ]_+ - \int f [dQ - dP]_+$ for any integrable f.

Considering the last integral $\int [q(x) - p(x)]_+ d\mu(x)$, taking the set $A = \{x \mid q(x) > p(x)\}$ yields

$$Q(A) - P(A) = \int_A (q(x) - p(x)) d\mu(x) \geq \int_B (q(x) - p(x)) d\mu(x) = Q(B) - P(B)$$

for any set B, as any $x \in B \setminus A$ clearly satisfies $q(x) - p(x) \leq 0$. □

Example 2.2.6 (Hellinger distance)**:** The *Hellinger distance* between probability distributions P and Q defined on a set $\mathcal{X}$ is generated by the function $f(t) = \frac{1}{2}(\sqrt{t}-1)^2 = \frac{1}{2}t - \sqrt{t} + \frac{1}{2}$. The Hellinger distance is then

$$d_{\rm hel}(P, Q)^2 := \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x). \tag{2.2.7}$$

The non-squared version $d_{\rm hel}(P, Q)$ is indeed a distance between probability measures P and Q. It is sometimes convenient to rewrite the Hellinger distance in terms of the *affinity* between P and Q, as

$$d_{\rm hel}(P, Q)^2 = \frac{1}{2} \int (p(x) + q(x) - 2\sqrt{p(x)q(x)}) d\mu(x) = 1 - \int \sqrt{p(x)q(x)} d\mu(x), \tag{2.2.8}$$

which makes clear that $d_{\rm hel}(P, Q) \in [0, 1]$ is on roughly the same scale as the variation distance; we will say more when we provide inequalities relating them in the next section. ◇

Example 2.2.7 (χ^2 divergence)**:** The *χ^2-divergence* is generated by taking $f(t) = (t-1)^2$, so that

$$D_{\chi^2}(P \| Q) := \int \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x) d\mu(x) = \int \frac{p(x)^2}{q(x)} d\mu(x) - 1, \tag{2.2.9}$$

where the equality is immediate because $\int p d\mu = \int q d\mu = 1$. ◇

2.2.4 Inequalities and relationships between divergences

Because many of the results connecting information theory and statistics repose on ways in which probability distributions get close to one another, inequalities between divergences provide a central tool for understanding and developing bounds on statistical estimation and learning. These inequalities will be particularly important because, in some cases, different distributions admit easy calculations with some divergences, such as KL or χ^2 divergence, but it can be challenging to work with others that may be more "natural" for a particular problem. Most importantly, replacing a variation distance by bounding it with an alternative divergence admitting chain rules or some other type of tensorization is often convenient for analyzing the properties of product distributions (as will become apparent in Chapter 9). We record several of these results here, making a passing connection to mutual information as well.

The first inequality shows that the Hellinger distance and variation distance roughly generate the same topology on collections of distributions, as they upper and lower bound the other (if we tolerate polynomial losses).

Proposition 2.2.8. *The total variation distance and Hellinger distance satisfy*

$$d_{\rm hel}^2(P, Q) \leq \|P - Q\|_{\rm TV} \leq d_{\rm hel}(P, Q)\sqrt{2 - d_{\rm hel}^2(P, Q)}.$$

Proof We begin with the upper bound. We have by the Cauchy-Schwartz inequality that

$$\begin{aligned}\frac{1}{2}\int |p(x) - q(x)|d\mu(x) &= \frac{1}{2}\int |\sqrt{p(x)} - \sqrt{q(x)}| \cdot |\sqrt{p(x)} + \sqrt{q(x)}|d\mu(x) \\ &\le \left(\frac{1}{2}\int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x)\right)^{\frac{1}{2}} \left(\frac{1}{2}\int (\sqrt{p(x)} + \sqrt{q(x)})^2 d\mu(x)\right)^{\frac{1}{2}} \\ &= d_{\rm hel}(P, Q)\left(1 + \int \sqrt{p(x)q(x)}d\mu(x)\right)^{\frac{1}{2}}.\end{aligned}$$

As in Example 2.2.6, we have $\int \sqrt{p(x)q(x)}d\mu(x) = 1 - d_{\rm hel}(P, Q)^2$, so this (along with the representation Lemma 2.2.5 for variation distance) implies

$$\|P - Q\|_{\rm TV} = \frac{1}{2}\int |p(x) - q(x)|d\mu(x) \le d_{\rm hel}(P, Q)(2 - d_{\rm hel}^2(P, Q))^{\frac{1}{2}}.$$

For the lower bound on total variation, note that for any $a, b \in \mathbb{R}_+$, we have $a + b - 2\sqrt{ab} \le |a - b|$ (check the cases $a > b$ and $a < b$ separately); thus

$$d_{\rm hel}^2(P, Q) = \frac{1}{2}\int \left[p(x) + q(x) - 2\sqrt{p(x)q(x)}\right] d\mu(x) \le \frac{1}{2}\int |p(x) - q(x)|d\mu(x),$$

as desired. □

Several important inequalities relate the variation distance to the KL-divergence. We state two important inequalities in the next proposition, both of which are important enough to justify their own names.

Proposition 2.2.9. *The total variation distance satisfies the following relationships.*

(a) Pinsker's inequality*: for any distributions P and Q,*

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P\|Q\right). \tag{2.2.10}$$

(b) The Bretagnolle-Huber *inequality: for any distributions P and Q,*

$$\|P - Q\|_{\rm TV} \le \sqrt{1 - \exp(-D_{\rm kl}\left(P\|Q\right))} \le 1 - \frac{1}{2}\exp(-D_{\rm kl}\left(P\|Q\right)).$$

Proof Exercise 2.22 outlines one proof of Pinsker's inequality using the data processing inequality (Proposition 2.2.13). We present an alternative via the Cauchy-Schwarz inequality. Using the definition (2.2.1) of the KL-divergence, we may assume without loss of generality that P and Q are finitely supported, with p.m.f.s $p = (p_1, \ldots, p_m)$ and $q = (q_1, \ldots, q_m)$. Define the negative entropy function $h(p) = \sum_{i=1}^m p_i \log p_i$. Then showing that $D_{\rm kl}\left(P\|Q\right) \ge 2\|P - Q\|_{\rm TV}^2 = \frac{1}{2}\|p - q\|_1^2$ is equivalent to showing that

$$h(p) \ge h(q) + \langle \nabla h(q), p - q\rangle + \frac{1}{2}\|p - q\|_1^2, \tag{2.2.11}$$

because by inspection $h(p) - h(q) - \langle \nabla h(q), p - q\rangle = \sum_i p_i \log \frac{p_i}{q_i}$. We do this via a Taylor expansion: we have

$$\nabla h(p) = [\log p_i + 1]_{i=1}^m \quad \text{and} \quad \nabla^2 h(p) = \operatorname{diag}([1/p_i]_{i=1}^m).$$

By Taylor's theorem, there is some $\tilde{p} = (1-t)p + tq$, where $t \in [0,1]$, such that

$$h(p) = h(q) + \langle \nabla h(q), p - q \rangle + \frac{1}{2}\langle p - q, \nabla^2 h(\tilde{p})(p-q) \rangle.$$

Looking at the final quadratic, for any vector v and probability vector r, Cauchy-Schwarz gives

$$\langle v, \nabla^2 h(r) v \rangle = \sum_{i=1}^m \frac{v_i^2}{r_i} = \|r\|_1 \sum_{i=1}^m \frac{v_i^2}{r_i} \geq \left(\sum_{i=1}^m \sqrt{r_i} \frac{|v_i|}{\sqrt{r_i}} \right)^2 = \|v\|_1^2 .$$

Applying this with $r = \tilde{p}$ yields inequality (2.2.11).

For the claim (b), we use Proposition 2.2.8. Let $a = \int \sqrt{p(x)q(x)} d\mu(x)$ be a shorthand for the affinity, so that $d_{\rm hel}^2(P,Q) = 1 - a$. Then Proposition 2.2.8 gives $\|P - Q\|_{\rm TV} \leq \sqrt{1-a}\sqrt{1+a} = \sqrt{1-a^2}$. Now apply Jensen's inequality to the exponential: we have

$$\begin{aligned} \int \sqrt{p(x)q(x)} d\mu(x) &= \int \sqrt{\frac{q(x)}{p(x)}} p(x) d\mu(x) = \int \exp\left(\frac{1}{2} \log \frac{q(x)}{p(x)} \right) p(x) d\mu(x) \\ &\geq \exp\left(\frac{1}{2} \int p(x) \log \frac{q(x)}{p(x)} d\mu(x) \right) = \exp\left(-\frac{1}{2} D_{\rm kl}\left(P \| Q\right) \right). \end{aligned}$$

In particular, $\sqrt{1-a^2} \leq \sqrt{1 - \exp(-\frac{1}{2} D_{\rm kl}\left(P\|Q\right))^2}$, which is the first claim of part (b). For the second, note that $\sqrt{1-c} \leq 1 - \frac{1}{2}c$ for $c \in [0,1]$ by concavity of the square root. □

We also have the following bounds on the Hellinger distance in terms of the KL-divergence, and that in terms of the χ^2-divergence.

Proposition 2.2.10. *For any distributions P, Q,*

$$2 d_{\rm hel}^2(P,Q) \leq D_{\rm kl}\left(P\|Q\right) \leq \log(1 + D_{\chi^2}(P\|Q)) \leq D_{\chi^2}(P\|Q) .$$

Proof For the first inequality, note that $\log x \leq x - 1$ by concavity, or $1 - x \leq -\log x$, so that

$$\begin{aligned} 2 d_{\rm hel}^2(P,Q) &= 2 - 2 \int \sqrt{p(x)q(x)} d\mu(x) \\ &= 2 \int p(x) \left(1 - \sqrt{\frac{q(x)}{p(x)}} \right) d\mu(x) \leq 2 \int p(x) \log \sqrt{\frac{p(x)}{q(x)}} d\mu(x) = D_{\rm kl}\left(P\|Q\right). \end{aligned}$$

The last two inequalitites are simple: by Jensen's inequality,

$$D_{\rm kl}\left(P\|Q\right) \leq \log \int \frac{dP^2}{dQ} = \log(1 + D_{\chi^2}\left(P\|Q\right)).$$

The last inequality is immediate as $\log(1+t) \leq t$ for all $t > -1$. □

2.2.5 Convexity and data processing for divergence measures

f-divergences satisfy a number of very useful properties, which we use repeatedly. As the KL-divergence is an f-divergence, it of course satisfies these conditions; however, we state them in fuller generality, treating the KL-divergence results as special cases and corollaries.

We begin by exhibiting the general data processing properties and convexity properties of f-divergences, each of which specializes to KL divergence. We leave the proof of each of these as exercises. First, we show that f-divergences are jointly convex in their arguments.

Proposition 2.2.11. *Let P_1, P_2, Q_1, Q_2 be distributions on a set $\mathcal{X}$ and $f : \mathbb{R}_+ \to \mathbb{R}$ be convex with $f(1) = 0$. Then for any $\lambda \in [0, 1]$,*

$$D_f\left(\lambda P_1 + (1-\lambda)P_2 \| \lambda Q_1 + (1-\lambda)Q_2\right) \le \lambda D_f\left(P_1 \| Q_1\right) + (1-\lambda) D_f\left(P_2 \| Q_2\right).$$

The proof of this proposition we leave as Exercise 2.14, which we treat as a consequence of the more general "log-sum" like inequalities of Exercise 2.11. It is, however, an immediate consequence of the fully specified definition (2.2.5) of an f-divergence, because $\mathrm{pers}(f)$ is jointly convex. As an immediate corollary, we see that the same result is true for KL-divergence as well.

Corollary 2.2.12. *The KL-divergence $D_{\rm kl}\left(P \| Q\right)$ is jointly convex in its arguments P and Q.*

We can also provide more general data processing inequalities for f-divergences, paralleling those for the KL-divergence. In this case, we consider random variables X and Z on spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively, and a Markov transition kernel K giving the Markov chain $X \to Z$. That is, $K(\cdot \mid x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$, and conditioned on $X = x$, Z has distribution $K(\cdot \mid x)$ so that $K(A \mid x) = \mathbb{P}(Z \in A \mid X = x)$. Certainly, this includes the situation when $Z = \phi(X)$ for some function ϕ, and more generally when $Z = \phi(X, U)$ for a function ϕ and some additional randomness U. For a distribution P on X, we then define the marginals

$$K_P(A) := \int_{\mathcal{X}} K(A \mid x) dP(x).$$

We then have the following proposition.

Proposition 2.2.13 (The data processing inequality)**.** *Let P and Q be distributions on X and let K be any Markov kernel. Then*

$$D_f\left(K_P \| K_Q\right) \le D_f\left(P \| Q\right).$$

See Exercise 2.13 for a proof.

As a corollary, we obtain the following data processing inequality for KL-divergences, where we abuse notation to write $D_{\rm kl}\left(X \| Y\right) = D_{\rm kl}\left(P \| Q\right)$ for random variables $X \sim P$ and $Y \sim Q$.

Corollary 2.2.14. *Let $X, Y \in \mathcal{X}$ be random variables, let $U \in \mathcal{U}$ be independent of X and Y, and let $\phi : \mathcal{X} \times \mathcal{U} \to \mathcal{Z}$ for some spaces $\mathcal{X}$,$\mathcal{U}, \mathcal{Z}$. Then*

$$D_{\rm kl}\left(\phi(X, U) \| \phi(Y, U)\right) \le D_{\rm kl}\left(X \| Y\right).$$

Thus, further processing of random variables can only bring them "closer" in the space of distributions; downstream processing of signals cannot make them further apart as distributions.

2.3 First steps into optimal procedures: testing inequalities

A central benefit of the information theoretic tools we explore is that they allow us to certify the optimality of procedures—that no other procedure could (substantially) improve upon the one at hand. These certifications often involve inequalities governing the best possible behavior of statistical tests. Roughly, we put ourselves in the following scenario: nature chooses one of a possible set of (say) k worlds, indexed by probability distributions $P_1, P_2, \ldots, P_k$, and conditional on nature's choice of the world—the distribution $P^\star \in \{P_1, \ldots, P_k\}$ chosen—we observe data X drawn from $P^\star$. Intuitively, it will be difficult to decide which distribution P_i is the true $P^\star$ if all the distributions are similar—the divergence between the P_i is small, or the information between X and $P^\star$ is negligible—and easy if the distances between the distributions P_i are large. With this outline in mind, we present two inequalities, and first examples of their application, to make concrete these connections to the notions of information and divergence defined in this section.

2.3.1 Le Cam's inequality and binary hypothesis testing

The simplest instantiation of the above setting is the case when there are only two possible distributions, P_1 and P_2, and our goal is to make a decision on whether P_1 or P_2 is the distribution generating data we observe. Concretely, suppose that nature chooses one of the distributions P_1 or P_2 at random, and let $V \in \{1, 2\}$ index this choice. Conditional on $V = v$, we then observe a sample X drawn from P_v. Denoting by $\mathbb{P}$ the joint distribution of V and X, we have for any test $\Psi : \mathcal{X} \to \{1, 2\}$ that the probability of error is then

$$\mathbb{P}(\Psi(X) \neq V) = \frac{1}{2} P_1(\Psi(X) \neq 1) + \frac{1}{2} P_2(\Psi(X) \neq 2).$$

We can give an exact expression for the minimal possible error in the above hypothesis test. Indeed, Le Cam (see [282, 435, Lemma 1]) provides the following now standard variational representation of the total variation distance (2.2.6), which is the f-divergence associated with $f(t) = \frac{1}{2}|t-1|$, as a function of testing error.

Proposition 2.3.1. *Let $\mathcal{X}$ be an arbitrary set. For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_\Psi \{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\} = 1 - \|P_1 - P_2\|_{\mathrm{TV}},$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Proof Any test $\Psi : \mathcal{X} \to \{1, 2\}$ has an acceptance region, call it $A \subset \mathcal{X}$, where it outputs 1 and a region A^c where it outputs 2.

$$P_1(\Psi \neq 1) + P_2(\Psi \neq 2) = P_1(A^c) + P_2(A) = 1 - P_1(A) + P_2(A).$$

Taking an infimum over such acceptance regions, we have

$$\inf_\Psi \{P_1(\Psi \neq 1) + P_2(\Psi \neq 2)\} = \inf_{A \subset \mathcal{X}} \{1 - (P_1(A) - P_2(A))\} = 1 - \sup_{A \subset \mathcal{X}} (P_1(A) - P_2(A)),$$

which yields the total variation distance as desired. □

In the two-hypothesis case, we also know that the optimal test, by the Neyman-Pearson lemma [e.g. 287], is a likelihood ratio test. That is, assuming that P_1 and P_2 have densities p_1 and p_2, the optimal test is of the form

$$\Psi(X) = \begin{cases} 1 & \text{if } \frac{p_1(X)}{p_2(X)} \ge t \\ 2 & \text{if } \frac{p_1(X)}{p_2(X)} < t \end{cases}$$

for some threshold $t \ge 0$. In the case that the prior probabilities on P_1 and P_2 are each $\frac{1}{2}$, then $t = 1$ is optimal.

We give one example application of Proposition 2.3.1 to the problem of testing a normal mean.

Example 2.3.2 (Testing a normal mean)**:** Suppose we observe $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$ for $P = P_1$ or $P = P_2$, where P_v is the normal distribution $\mathsf{N}(\mu_v, \sigma^2)$, where $\mu_1 \neq \mu_2$. We would like to understand the sample size n necessary to guarantee that no test can have small error, that is, say, that

$$\inf_{\Psi} \left\{P_1(\Psi(X_1, \ldots, X_n) \neq 1) + P_2(\Psi(X_1, \ldots, X_n) \neq 2)\right\} \ge \frac{1}{2}.$$

By Proposition 2.3.1, we have that

$$\inf_{\Psi} \left\{P_1(\Psi(X_1, \ldots, X_n) \neq 1) + P_2(\Psi(X_1, \ldots, X_n) \neq 2)\right\} \ge 1 - \|P_1^n - P_2^n\|_{\mathrm{TV}},$$

where P_v^n denotes the n-fold product of P_v, that is, the distribution of $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_v$.

The interaction between total variation distance and product distributions is somewhat subtle, so it is often advisable to use a divergence measure more attuned to the i.i.d. nature of the sampling scheme. Two such measures are the KL-divergence and Hellinger distance. With that in mind, we apply Pinsker's inequality (2.2.10) to see that $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \le \frac{1}{2} D_{\mathrm{kl}}\left(P_1^n \| P_2^n\right) = \frac{n}{2} D_{\mathrm{kl}}\left(P_1 \| P_2\right)$, which implies that

$$1 - \|P_1^n - P_2^n\|_{\mathrm{TV}} \ge 1 - \sqrt{\frac{n}{2} D_{\mathrm{kl}}\left(P_1 \| P_2\right)}^{\frac{1}{2}} = 1 - \sqrt{\frac{n}{2}} \left(\frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2\right)^{\frac{1}{2}} = 1 - \frac{\sqrt{n}}{2} \frac{|\mu_1 - \mu_2|}{\sigma}.$$

In particular, if $n \le \frac{\sigma^2}{(\mu_1 - \mu_2)^2}$, then we have our desired lower bound of $\frac{1}{2}$.

Conversely, a calculation yields that $n \ge \frac{C\sigma^2}{(\mu_1 - \mu_2)^2}$, for some numerical constant $C \ge 1$, implies small probability of error. We leave this calculation to the reader. $\diamond$

2.3.2 Fano's inequality and multiple hypothesis testing

There are of course situations in which we wish to test between multiple competing hypotheses. In such situations, Fano's inequality, which we present shortly, is the most common tool for proving fundamental limits, lower bounds on probability of error, and converses (to results on achievability of some performance level) in information theory. We write this section in terms of general random variables, ignoring the precise setting of selecting an index in a family of distributions, though that is implicit in what we do.

Let X be a random variable taking values in a finite set $\mathcal{X}$, and assume that we observe a (different) random variable Y, and then must estimate or guess the true value of $\widehat{X}$. That is, we have the Markov chain

$$X \to Y \to \widehat{X},$$

and we wish to provide lower bounds on the probability of error that $\widehat{X} \neq X$. If we let the function $h_2(p) = -p \log p - (1-p) \log(1-p)$ denote the binary entropy (entropy of a Bernoulli random variable with parameter p), Fano's inequality takes the following form [e.g. 118, Chapter 2]:

Proposition 2.3.3 (Fano inequality)**.** *For any pair of random variables X and $\widehat{X}$,*

$$h_2(\mathbb{P}(\widehat{X} \neq X)) + \mathbb{P}(\widehat{X} \neq X) \log(|\mathcal{X}| - 1) \geq H(X \mid \widehat{X}). \tag{2.3.1}$$

Proof This proof follows by expanding an entropy functional in two different ways. Let E be the indicator for the event that $\widehat{X} \neq X$, that is, $E = 1$ if $\widehat{X} \neq X$ and is 0 otherwise. Then we have

$$\begin{aligned} H(X, E \mid \widehat{X}) &= H(X \mid E, \widehat{X}) + H(E \mid \widehat{X}) \\ &= \mathbb{P}(E = 1) H(X \mid E = 1, \widehat{X}) + \mathbb{P}(E = 0) \underbrace{H(X \mid E = 0, \widehat{X})}_{=0} + H(E \mid \widehat{X}), \end{aligned}$$

where the zero follows because given there is no error, X has no variability given $\widehat{X}$. Expanding the entropy by the chain rule in a different order, we have

$$H(X, E \mid \widehat{X}) = H(X \mid \widehat{X}) + \underbrace{H(E \mid \widehat{X}, X)}_{=0},$$

because E is perfectly predicted by $\widehat{X}$ and X. Combining these equalities, we have

$$H(X \mid \widehat{X}) = H(X, E \mid \widehat{X}) = \mathbb{P}(E = 1) H(X \mid E = 1, \widehat{X}) + H(E \mid \widehat{X}).$$

Noting that $H(E \mid \widehat{X}) \leq H(E) = h_2(\mathbb{P}(E = 1))$, as conditioning reduces entropy, and that $H(X \mid E = 1, \widehat{X}) \leq \log(|\mathcal{X}| - 1)$, as X can take on at most $|\mathcal{X}| - 1$ values when there is an error, completes the proof. □

We can rewrite Proposition 2.3.3 in a convenient way when X is uniform in $\mathcal{X}$. Indeed, $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$ by definition of the mutual information, so Proposition 2.3.3 implies

Corollary 2.3.4. *Assume that X is uniform on $\mathcal{X}$. For any Markov chain $X \to Y \to \widehat{X}$,*

$$\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{I(X; Y) + \log 2}{\log(|\mathcal{X}|)}. \tag{2.3.2}$$

Proof Let $P_{\mathsf{error}} = \mathbb{P}(X \neq \widehat{X})$ denote the probability of error. Noting that $h_2(p) \leq \log 2$ for any $p \in [0, 1]$ (recall inequality (2.1.2), that is, that uniform random variables maximize entropy), then using Proposition 2.3.3, we have

$$\log 2 + P_{\mathsf{error}} \log(|\mathcal{X}|) \geq h_2(P_{\mathsf{error}}) + P_{\mathsf{error}} \log(|\mathcal{X}| - 1) \overset{(i)}{\geq} H(X \mid \widehat{X}) \overset{(ii)}{=} H(X) - I(X; \widehat{X}).$$

Here step (i) uses Proposition 2.3.3 and step (ii) uses the definition $I(X; \widehat{X}) = H(X) - H(X \mid \widehat{X})$ of mutual information. The data processing inequality implies that $I(X; \widehat{X}) \leq I(X; Y)$, and using $H(X) = \log(|\mathcal{X}|)$ completes the proof. □

Corollary 2.3.4 shows that when X is uniform and we observe Y, we have

$$\inf_{\Psi} \mathbb{P}(\Psi(Y) \neq X) \geq 1 - \frac{I(X; Y) + \log 2}{\log |\mathcal{X}|},$$

where the infimum is taken over all testing procedures Ψ. Some interpretation of this quantity is helpful. If we think roughly of the number of bits it takes to describe a variable X uniformly chosen from $\mathcal{X}$, then we expect that $\log_2 |\mathcal{X}|$ bits are necessary (and sufficient). Thus, until we collect enough information that $I(X;Y) \approx \log |\mathcal{X}|$, so that $I(X;Y)/\log |\mathcal{X}| \approx 1$, we are unlikely to identify the variable X with nontrivial success probability. We must collect enough bits to actually discover X.

Example 2.3.5 (20 Questions)**:** In the familiar 20 Questions game, there are two players, the "chooser" and the "guesser," and an agreed upon universe $\mathcal{X}$. The chooser picks an element $x \in \mathcal{X}$, and the guesser's goal is to find x by using a series of yes/no questions about x. We consider optimal strategies for each player in this game, assuming $\mathcal{X}$ is finite and letting $m = |\mathcal{X}|$ be the universe size for shorthand.

For the guesser, it is clear that at most $\lceil \log_2 m \rceil$ questions are necessary to guess the item X that the chooser has picked—at each round of the game, the guesser asks a question that eliminates half of the remaining possible items. Indeed, let us assume that $m = 2^l$ for some $l \in \mathbb{N}$; if not, the guesser can always make her task more difficult by increasing the size of $\mathcal{X}$ until it is a power of 2. Thus, after k rounds, there are $m2^{-k}$ items left, and we have

$$m \left(\frac{1}{2}\right)^k \le 1 \text{ if and only if } k \ge \log_2 m.$$

For the converse—the chooser's strategy—let $Y_1, Y_2, \ldots, Y_k$ be the sequence of yes/no answers given to the guesser. Assume that the chooser picks X uniformly at random in $\mathcal{X}$. Then Fano's inequality (2.3.2) implies that for the guess $\widehat{X}$ the guesser makes,

$$\mathbb{P}(\widehat{X} \ne X) \ge 1 - \frac{I(X; Y_1, \ldots, Y_k) + \log 2}{\log m}.$$

By the chain rule for mutual information, we have

$$I(X; Y_1, \ldots, Y_k) = \sum_{i=1}^k I(X; Y_i \mid Y_1^{i-1}) = \sum_{i=1}^k H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid Y_1^{i-1}, X) \le \sum_{i=1}^k H(Y_i).$$

As the answers Y_i are yes/no, we have $H(Y_i) \le \log 2$, so $I(X; Y_1^k) \le k \log 2$, and

$$\mathbb{P}(\widehat{X} \ne X) \ge 1 - \frac{(k+1)\log 2}{\log m} = \frac{\log_2 m - 1}{\log_2 m} - \frac{k}{\log_2 m}.$$

The guesser must have $k \ge \log_2(m/2) = \log_2 m - 1$ to be guaranteed no mistakes. $\diamond$

2.4 A first operational result: entropy and source coding

The final section of this chapter explores the basic results in source coding. Source coding—in its simplest form—tells us the number of bits are necessary to perfectly encode a sequence of random variables $X_1, X_2, \ldots$ drawn according to a known distribution P.

2.4.1 The source coding problem

Assume we receive data consisting of a sequence of symbols $X_1, X_2, \ldots$, drawn from a known distribution P on a finite or countable space $\mathcal{X}$. We wish to choose an encoding, represented by a *d-ary code function* C that maps $\mathcal{X}$ to finite strings consisting of the symbols $\{0, 1, \ldots, d-1\}$. We denote this by $\mathsf{C} : \mathcal{X} \to \{0, 1, \ldots, d-1\}^*$, where the superscript * denotes that the length may change from input to input, and use $\ell_{\mathsf{C}}(x)$ to denote the length of the string $\mathsf{C}(x)$.

We consider a variety of types of codes, defining each in order of complexity of their decoding.

Definition 2.1. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* non-singular *if for each* $x, x' \in \mathcal{X}$ *we have*

$$\mathsf{C}(x) \neq \mathsf{C}(x') \quad \text{if } x \neq x'.$$

Generally speaking, we wish to transmit or encode a variety of codewords simultaneously, that is, we wish to encode a sequence $X_1, X_2, \ldots$ using the natural *extension* of the code C as the string $\mathsf{C}(X_1)\mathsf{C}(X_2)\mathsf{C}(X_3)\cdots$, where $\mathsf{C}(x_1)\mathsf{C}(x_2)$ denotes the concatenation of the strings $\mathsf{C}(x_1)$ and $\mathsf{C}(x_2)$. In this case, we require that the code be uniquely decodable:

Definition 2.2. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* uniquely decodable *if for all* $n, m \in \mathbb{N}$ *and sequences* $x_1, \ldots, x_n \in \mathcal{X}$ *and* $x'_1, \ldots, x'_m \in \mathcal{X}$ *we have*

$$\mathsf{C}(x_1)\mathsf{C}(x_2)\cdots\mathsf{C}(x_n) = \mathsf{C}(x'_1)\mathsf{C}(x'_2)\cdots\mathsf{C}(x'_m) \quad \text{if and only if } m = n \text{ and } x_1 = x'_1, \ldots, x_n = x'_n.$$

That is, the extension of the code C *to sequences is non-singular.*

Uniquely decodable codes may require inspection of an entire string before recovering the first element. With that in mind, we consider codes that can always be decoded instantaneously.

Definition 2.3. *A d-ary code* $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$ *is* prefix-free *or* instantaneous *if no codeword is the prefix to another codeword.*

As is hopefully apparent from the definitions, all prefix/instantaneous codes are uniquely decodable. The converse is not true, though we will see—as long as we care only about encoding sequences—using a prefix instead of uniquely decodable code has negligible consequences.

For example, written English, with periods (.) and spaces ( ) included at the ends of words (among other punctuation) is an instantaneous encoding of English into the symbols of the alphabet and punctuation, as punctuation symbols enforce that no "codeword" is a prefix of any other. More concrete examples may make things clearer.

Example 2.4.1 (Encoding strategies)**:** Consider the encoding schemes below, which encode the letters a, b, c, and d.

Symbol	$\mathsf{C}_1(x)$	$\mathsf{C}_2(x)$	$\mathsf{C}_3(x)$
a	0	00	0
b	00	10	10
c	000	11	110
d	0000	110	111

By inspection, C_1 is non-singular but certainly not uniquely decodable (does the sequence 0000 correspond to aaaa, bb, aab, aba, baa, ca, ac, or d?), while C_3 is a prefix code. We leave showing that C_2 is uniquely decodable as an exercise. $\diamond$

2.4.2 The Kraft-McMillan inequalities

We now turn to a few results on the connections between source-coding and entropy. Our first result, the *Kraft-McMillan inequality*, says that there is no difference in code-lengths attainable by prefix codes and uniquely decodable codes.

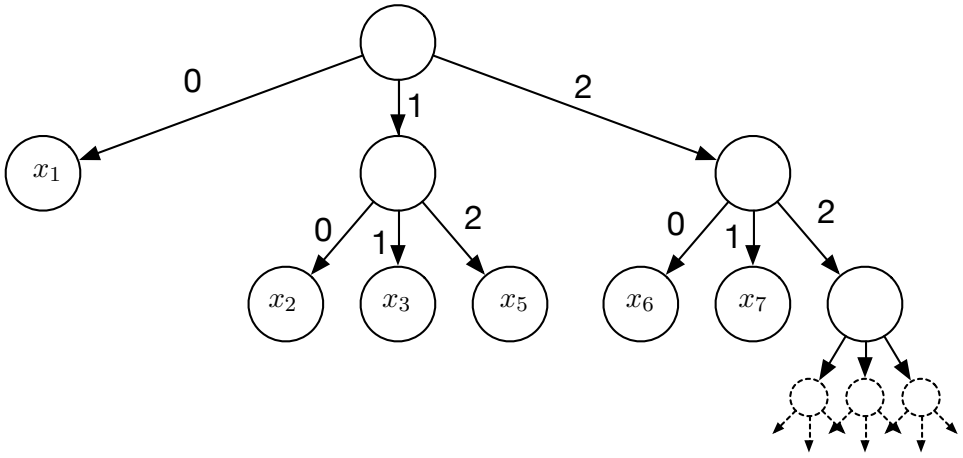


Figure 2.1. Prefix-tree encoding of a set of symbols. The encoding for x_1 is `0`, for x_2 is `10`, for x_3 is `11`, for x_4 is `12`, for x_5 is `20`, for x_6 is `21`, and nothing is encoded as `1`, `2`, or `22`.

Theorem 2.4.2. *Let $\mathcal{X}$ be a finite or countable set, and let $\ell : \mathcal{X} \to \mathbb{N}$ be a function. If $\ell(x)$ is the length of the encoding of the symbol x in a uniquely decodable d-ary code, then*

$$\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1. \tag{2.4.1}$$

Conversely, given any function $\ell : \mathcal{X} \to \mathbb{N}$ satisfying inequality (2.4.1), there is a prefix code whose codewords have length $\ell(x)$ for each $x \in \mathcal{X}$.

Proof We prove the first statement of the theorem first by a counting and asymptotic argument.

We begin by assuming that $\mathcal{X}$ is finite. As a consequence, there is some maximum length $\ell_{\max}$ such that $\ell(x) \le \ell_{\max}$ for all $x \in \mathcal{X}$. For a sequence $x_1, \ldots, x_n \in \mathcal{X}$, we have by the definition of our encoding strategy that $\ell(x_1, \ldots, x_n) = \sum_{i=1}^n \ell(x_i)$. In addition, for each m we let

$$E_n(m) := \{x_1^n \in \mathcal{X}^n \text{ such that } \ell(x_1^n) = m\}$$

denote the symbols x encoded with codewords of length m in our code. Then as the code is uniquely decodable we certainly have $\mathrm{card}(E_n(m)) \le d^m$ for all n and m. Moreover, for all $x_1^n \in \mathcal{X}^n$ we have $\ell(x_1^n) \le n\ell_{\max}$. We thus re-index the sum $\sum_x d^{-\ell(x)}$ and compute

$$\begin{aligned}\sum_{x_1, \ldots, x_n \in \mathcal{X}^n} d^{-\ell(x_1, \ldots, x_n)} &= \sum_{m=1}^{n\ell_{\max}} \mathrm{card}(E_n(m)) d^{-m} \\ &\le \sum_{m=1}^{n\ell_{\max}} d^{m-m} = n\ell_{\max}.\end{aligned}$$

The preceding relation is true for all $n \in \mathbb{N}$, so that

$$\left(\sum_{x_1^n \in \mathcal{X}^n} d^{-\ell(x_1^n)} \right)^{1/n} \le n^{1/n} \ell_{\max}^{1/n} \to 1$$

as $n \to \infty$. In particular, using that

$$\sum_{x_1^n \in \mathcal{X}^n} d^{-\ell(x_1^n)} = \sum_{x_1, \ldots, x_n \in \mathcal{X}^n} d^{-\ell(x_1)} \cdots d^{-\ell(x_n)} = \left(\sum_{x \in \mathcal{X}} d^{-\ell(x)} \right)^n,$$

we obtain $\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1$.

Returning to the case that $\mathrm{card}(\mathcal{X}) = \infty$, by defining the sequence

$$D_k := \sum_{x \in \mathcal{X}, \ell(x) \le k} d^{-\ell(x)},$$

as each subset $\{x \in \mathcal{X} : \ell(x) \le k\}$ is uniquely decodable, we have $D_k \le 1$ for all k. Then $1 \ge \lim_{k \to \infty} D_k = \sum_{x \in \mathcal{X}} d^{-\ell(x)}$.

The achievability of such a code intuitively follows by a pictorial argument (recall Figure 2.1), so we first sketch the result non-rigorously. Indeed, let $\mathcal{T}_d$ be an (infinite) d-ary tree. Then, at each level m of the tree, assign one of the nodes at that level to each symbol $x \in \mathcal{X}$ such that $\ell(x) = m$. Eliminate the subtree below that node, and repeat with the remaining symbols. The codeword corresponding to symbol x is then the path to the symbol in the tree.

A more formal version implementing this sketch follows. Let ℓ be a length function satisfying $\sum_{x \in \mathcal{X}} d^{-\ell(x)} \le 1$. Identify $\mathcal{X}$ with $\mathbb{N}$ (or a subset thereof) in such a way that $1 \le \ell(1) \le \ell(2) \le \ldots$, i.e., $\ell(x) \le \ell(y)$ whenever $x < y$, and let $\mathcal{X}_m = \{x \in \mathcal{X} \mid \ell(x) = m\}$ be the set of inputs with encoding length m. For each $x \in \mathbb{N}$, define the value

$$v(x) = \sum_{i < x} d^{-\ell(i)}.$$

We let the codeword $C(x)$ for x be the first $\ell(x)$ terms in the d-ary expansion of $v(x)$. Certainly the length of this encoding satisfies $|C(x)| = \ell(x)$. To see that it is prefix-free, take two symbols $x < y$, and assume for the sake of contradiction that $C(x)$ is a prefix of $C(y)$. Then $v(y) \ge v(x)$, while $v(y) - v(x) \le d^{-\ell(x)}$ because the two representations agree on the first $\ell(x)$ terms in the expansion. But

$$v(y) - v(x) = \sum_{i < y} d^{-\ell(i)} - \sum_{i < x} d^{-\ell(i)} = \sum_{x \le i < y} d^{-\ell(i)} = d^{-\ell(x)} + \sum_{x < i < y} d^{-\ell(i)} > d^{-\ell(x)},$$

a contradiction. □

With the Kraft-McMillan theorem in place, we may directly relate the entropy of a random variable to the length of possible encodings for the variable; in particular, we show that the entropy is essentially *the best* possible code length of a uniquely decodable source code. Throughout, we use the shorthand

$$H_d(X) := -\sum_{x \in \mathcal{X}} p(x) \log_d p(x).$$

Theorem 2.4.3. *Let $X \in \mathcal{X}$ be a discrete random variable distributed according to P and let ℓ_C be the length function associated with a d-ary encoding $\mathsf{C} : \mathcal{X} \to \{0, \ldots, d-1\}^*$. In addition, let $\mathcal{C}$ be the set of all uniquely decodable d-ary codes for $\mathcal{X}$. Then*

$$H_d(X) \le \inf \left\{ \mathbb{E}_P[\ell_{\mathsf{C}}(X)] \ : \ \mathsf{C} \in \mathcal{C} \right\} \le H_d(X) + 1.$$

Proof For the lower bound, we assume for simplicity that $\mathcal{X}$ is finite, and we identify $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ (let $m = |\mathcal{X}|$ for shorthand). Then as $\mathcal{C}$ consists of uniquely decodable codes, all the associated length functions must satisfy the Kraft-McMillan inequality (2.4.1). Letting $\ell_i = \ell(i)$, the minimal encoding length is at least

$$\inf_{\ell \in \mathbb{R}^m} \left\{ \sum_{i=1}^m p_i \ell_i : \sum_{i=1}^m d^{-\ell_i} \le 1 \right\}.$$

By introducing the Lagrange multiplier $\lambda \ge 0$ for the inequality constraint, we may write the Lagrangian for the preceding minimization problem as

$$\mathcal{L}(\ell, \lambda) = p^\top \ell + \lambda \left(\sum_{i=1}^m d^{-\ell_i} - 1 \right) \quad \text{with} \quad \nabla_\ell \mathcal{L}(\ell, \lambda) = p - \lambda \left[d^{-\ell_i} \log d \right]_{i=1}^m.$$

In particular, the optimal ℓ satisfies $\ell_i = \log_d \frac{\theta}{p_i}$ for some constant θ, and solving $\sum_{i=1}^m d^{-\log_d \frac{\theta}{p_i}} = 1$ gives $\theta = 1$ and $\ell(i) = \log_d \frac{1}{p_i}$.

For the attainment claim, simply set the encoding to be $\ell(x) = \lceil \log_d \frac{1}{P(X=x)} \rceil$, which satisfies the Kraft-McMillan inequality and thus yields a valid prefix code with

$$\mathbb{E}_P[\ell(X)] = \sum_{x \in \mathcal{X}} p(x) \left\lceil \log_d \frac{1}{p(x)} \right\rceil \le - \sum_{x \in \mathcal{X}} p(x) \log_d p(x) + 1 = H_d(X) + 1$$

as desired. □

Theorem 2.4.3 thus shows that, at least to within an additive constant of 1, the entropy both upper and lower bounds the expected length of a uniquely decodable code for the random variable X. This is the first of our promised "operational interpretations" of the entropy. Exercise 2.23 provides a related operational interpretation that among all codes $\mathsf{C} : \mathcal{X}^n \to \{0,1\}^k$ mapping length n sequences to k bits and achieving error $\epsilon \in (0,1)$, the best (asymptotic) rate of k/n is exactly $H(P)$.

2.4.3 Entropy rates and longer codes

Theorem 2.4.3 is a bit unsatisfying in that the additive constant 1 may be large relative to the entropy. A standard approach in information theory is to employ block-coding (encoding longer sequences), to (asymptotically) eliminate this error factor. While these results are important, we only sketch them here, as they are not central to the remainder of the book.

Definition 2.4. *The* entropy rate *of a sequence $X_1, X_2, \ldots$ of random variables is*

$$H(\{X_i\}) := \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{2.4.2}$$

whenever the limit exists.

The limit (2.4.2) will exist when we consider stationary random variables:

Definition 2.5. *We say a sequence $X_1, X_2, \ldots$ of random variables is stationary if for all n and all $k \in \mathbb{N}$ and all measurable sets $A_1, \ldots, A_k \subset \mathcal{X}$ we have*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_k \in A_k) = \mathbb{P}(X_{n+1} \in A_1, \ldots, X_{n+k} \in A_k).$$

With this definition, we have the following result.

Proposition 2.4.4. *Let the sequence of random variables $\{X_i\}$ taking values in the discrete space $\mathcal{X}$ be stationary. Then*

$$H(\{X_i\}) = \lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1})$$

and the limits (2.4.2) and above exist.

Because we will not use this result in the remainder of the book, we leave its proof as Exercise 2.24.

Finally, we present a result showing that it is possible to achieve average code length of at most the entropy rate, which for stationary sequences is smaller than the entropy of any single random variable X_i. To do so, we use a block code, which treats sets of random variables $(X_1, \ldots, X_m) \in \mathcal{X}^m$ as a single symbol to be jointly encoded. Then Exercise 2.25 outlines the proof of the next result.

Proposition 2.4.5. *Let $X_1, X_2, \ldots$ be a discrete stationary sequence. Then for any $\epsilon > 0$, there exists an $m \in \mathbb{N}$ and a d-ary prefix block encoder $\mathsf{C} : \mathcal{X}^m \to \{0, \ldots, d-1\}^*$ such that*

$$\limsup_n \frac{1}{n}\mathbb{E}_P[\ell_{\mathsf{C}}(X_1^n)] \le H_d(\{X_i\}) + \epsilon = \lim_n H_d(X_n \mid X_1, \ldots, X_{n-1}) + \epsilon.$$

2.5 Bibliography and further reading

The material in this chapter is classical in information theory. For our treatment of mutual information, entropy, and KL-divergence in the discrete case, Cover and Thomas provide a concise overview in Chapter 2 of their book [118]. Gray [217] provides a measure-theoretic version of these results, with Chapter 5 covering most of the material we present (or Chapter 7 in the newer edition [218] of the same book). In particular, Gray demonstrates that one may take the partitions in the (general) definition (2.2.1) of the KL-divergence to be Cartesian products of sets, and he develops the measure-theoretic version (2.2.2) of KL-divergences. We have been somewhat loose in the discussion of rectangular sets following the definition (2.2.1); to be precise, we must assume that the measures P and Q on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ are defined on the σ-algebra generated by rectangles $A_1 \times \cdots \times A_n$ for $A_i \subset \mathcal{X}_i$ (cf. [218, Ch. 7.4]); it is sufficient that P and Q be Borel measures, as rectangles generate the Borel product σ-field. Polyanskiy and Wu [346] provide a modern treatment of the measure-theoretic questions, and Duchi et al. [162, Section 2] also show how these ideas apply in the context of f-divergences. Csiszár and Körner [124] is the classic reference for coding theorems and results on communication, including strong converse results.

The fact that Gaussian random variables have the largest entropy (Proposition 2.1.10) has been used to prove stronger variants of the central limit theorem. Barron [44] first develops this, and Artstein et al. [26] and Madiman and Barron [299] provide quantitative results.

The f-divergence was independently discovered by Ali and Silvey [16] and Csiszár [122], and is consequently sometimes called an Ali-Silvey divergence or Csiszár divergence. Liese and Vajda [289] survey f-divergences and their relationships with different statistical concepts (taking a Bayesian

point of view), and various authors have extended the pairwise divergence measures to divergence measures between multiple distributions [224], making connections to experimental design [205, 162], which we investigate in Chapter 16. The inequalities relating divergences in Section 2.2.4 are classical, and standard references present them [282, 401]. For a proof that equality (2.2.4) is equivalent to the definition (2.2.3) with the appropriate closure operations, see the paper [162, Proposition 1]. Our proof of the Kraft-McMillan inequality (Theorem 2.4.2) follows Cover and Thomas [118] as well as Polyanskiy and Wu [346]; sometimes the converse inequality (that it is achievable) is called Chaitin's inequality, after the paper [107].

2.6 Exercises

Exercise 2.1: Prove the following identities about total variation. Throughout, let P_1 and P_2 have densities p_1 and p_2 on a (common) set $\mathcal{X} \subset \mathbb{R}$.

(a) $2 \|P_1 - P_2\|_{\rm TV} = \int |p_1(x) - p_2(x)| dx$.

(b) For functions $f : \mathcal{X} \to \mathbb{R}$, define the supremum norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Show that $2 \|P_1 - P_2\|_{\rm TV} = \sup_{\|f\|_\infty \le 1} \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x)) dx$.

(c) $\|P_1 - P_2\|_{\rm TV} = \int \max\{p_1(x), p_2(x)\} dx - 1$.

(d) $\|P_1 - P_2\|_{\rm TV} = 1 - \int \min\{p_1(x), p_2(x)\} dx$.

(e) For functions $f, g : \mathcal{X} \to \mathbb{R}$,

$$\inf \left\{ \int f(x) p_1(x) dx + \int g(x) p_2(x) dx : f + g \ge 1, f \ge 0, g \ge 0 \right\} = 1 - \|P_1 - P_2\|_{\rm TV} .$$

Exercise 2.2 (Divergence between multivariate normal distributions)**:** Let P_1 be $\mathsf{N}(\theta_1, \Sigma)$ and P_2 be $\mathsf{N}(\theta_2, \Sigma)$, where $\Sigma \succ 0$ is a positive definite matrix.

(a) Give $D_{\rm kl}(P_1 \| P_2)$.

(b) Show that $d_{\rm hel}^2(P_1, P_2) = 1 - \exp(-\frac{1}{8}(\theta_1 - \theta_2)^\top \Sigma^{-1} (\theta_1 - \theta_2))$.

Exercise 2.3: Let P_1 be $\mathsf{N}(\theta_1, \Sigma_1)$ and P_2 be $\mathsf{N}(\theta_2, \Sigma_2)$, where $\Sigma_i \succ 0$ are positive definite matrices. Give $D_{\rm kl}(P_1 \| P_2)$.

Exercise 2.4: Complete the calculation in Example 2.1.3.

Exercise 2.5: Let $X \in \{2, 3, \ldots\}$ have the distribution in Example 2.1.4.

(a) Show that for $\alpha > 1$, $\int_e^\infty \frac{1}{t \log^\alpha t} dt = \frac{1}{\alpha - 1}$, so the distribution is well-defined.

(b) Show that $H(X) = +\infty$.

Exercise 2.6 (Tensorization of entropy)**:** Let $X_1, \ldots, X_n$ be random variables.

(a) Show that if X_i are discrete, then $H(X_1^n) \le \sum_{i=1}^n H(X_i)$, with equality if and only if $X_1, \ldots, X_n$ are mutually independent.

(b) Show that if X_i have a joint density, then $h(X_1^n) \le \sum_{i=1}^n h(X_i)$, with equality if and only if $X_1, \ldots, X_n$ are mutually independent.

Exercise 2.7: Let $T \sim \mathsf{N}(0, \tau^2 I_d)$ and let $Y \mid T = \theta \sim \mathsf{N}(X\theta, \sigma^2 I_n)$, where $X \in \mathbb{R}^{n \times d}$ is a known matrix. Show that

$$I(T; Y_1^n) \le \sum_{i=1}^n I(T; Y_i)$$

and that

$$\lim_{\sigma^2 \uparrow \infty} \sigma^2 \sum_{i=1}^n I(T; Y_i) = \lim_{\sigma^2 \uparrow \infty} \sigma^2 I(T; Y_1^n).$$

That is, in the large noise limit, the tensorized information and information between T and Y are identical.

Exercise 2.8: A random variable X has $\mathsf{Laplace}(\lambda, \mu)$ distribution if it has density $p(x) = \frac{\lambda}{2} \exp(-\lambda |x - \mu|)$. Consider the test of P_1 versus P_2, where X has distribution $\mathsf{Laplace}(\lambda, \mu_1)$ under P_1 and distribution $\mathsf{Laplace}(\lambda, \mu_2)$ under P_2, where $\mu_1 < \mu_2$. Show that the minimal value over all tests ψ of P_1 versus P_2 is

$$\inf_\psi \left\{ P_1(\psi(X) \neq 1) + P_2(\psi(X) \neq 2) \right\} = \exp\left(-\frac{\lambda}{2} |\mu_1 - \mu_2| \right).$$

Exercise 2.9 (Log-sum inequality)**:** Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be non-negative reals. Show that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \ge \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}.$$

Hint. Use the convexity of the function $x \mapsto -\log(x)$.

Exercise 2.10: Given quantizers g_1 and g_2, we say that g_1 is a *finer* quantizer than g_2 under the following condition: assume that g_1 induces the partition $A_1, \ldots, A_n$ and g_2 induces the partition $B_1, \ldots, B_m$; then for any of the sets B_i, there exists $k \in \mathbb{N}$ and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We let $g_1 \prec g_2$ denote that g_1 is a finer quantizer than g_2. Prove

(a) Finer partitions increase the KL divergence: if $g_1 \prec g_2$,

$$D_{\rm kl}\left(P \| Q \mid g_2\right) \le D_{\rm kl}\left(P \| Q \mid g_1\right).$$

(b) If $\mathcal{X}$ is discrete (so P and Q have p.m.f.s p and q) then

$$D_{\rm kl}\left(P \| Q\right) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Exercise 2.11 (Generalized "log-sum" inequalities)**:** Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an arbitrary convex function, and for $b \ge 0$ let $0 \cdot f(b/0) := f'_\infty(b) = \lim_{t \to \infty} \frac{1}{t} f(tb)$.

(a) Let $a_i, b_i, i = 1, \ldots, n$ be non-negative reals. Prove that

$$\left(\sum_{i=1}^n a_i \right) f\left(\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i} \right) \le \sum_{i=1}^n a_i f\left(\frac{b_i}{a_i} \right).$$

(b) Generalizing the preceding result, let $a : \mathcal{X} \to \mathbb{R}_+$ and $b : \mathcal{X} \to \mathbb{R}_+$, and let μ be a finite measure on $\mathcal{X}$ with respect to which a and b are integrable. Show that

$$\int a(x)d\mu(x) f\left(\frac{\int b(x)d\mu(x)}{\int a(x)d\mu(x)}\right) \le \int a(x) f\left(\frac{b(x)}{a(x)}\right) d\mu(x).$$

If you are unfamiliar with measure theory, prove the following essentially equivalent result: let $u : \mathcal{X} \to \mathbb{R}_+$ satisfy $\int u(x)dx < \infty$. Show that

$$\int a(x)u(x)dx f\left(\frac{\int b(x)u(x)dx}{\int a(x)u(x)dx}\right) \le \int a(x) f\left(\frac{b(x)}{a(x)}\right) u(x)dx$$

whenever $\int a(x)u(x)dx < \infty$ and $\int b(x)u(x)dx < \infty$.

Hint. Use that the perspective $f_{\rm per}(x,t) = tf(x/t)$ for $t > 0$ is jointly convex in x and t, as in Proposition B.2.24; to give the result in full generality, use the closure of the perspective as in the representation (2.2.5).

Exercise 2.12 (Data processing and f-divergences I)**:** As with the KL-divergence, given a quantizer g of the set $\mathcal{X}$, where g induces a partition $A_1, \ldots, A_m$ of $\mathcal{X}$, we define the f-divergence between P and Q conditioned on g as

$$D_f\left(P\|Q \mid g\right) := \sum_{i=1}^m Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right) = \sum_{i=1}^m Q(g^{-1}(\{i\})) f\left(\frac{P(g^{-1}(\{i\}))}{Q(g^{-1}(\{i\}))}\right).$$

Given quantizers g_1 and g_2, we say that g_1 is a *finer* quantizer than g_2 under the following condition: assume that g_1 induces the partition $A_1, \ldots, A_n$ and g_2 induces the partition $B_1, \ldots, B_m$; then for any of the sets B_i, there exists $k \in \mathbb{N}$ and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We let $g_1 \prec g_2$ denote that g_1 is a finer quantizer than g_2.

(a) Let g_1 and g_2 be quantizers of the set $\mathcal{X}$, and let $g_1 \prec g_2$, meaning that g_1 is a finer quantization than g_2. Prove Corollary 2.2.1, that is, that

$$D_f\left(P\|Q \mid g_2\right) \le D_f\left(P\|Q \mid g_1\right).$$

Equivalently, show that whenever $\mathcal{A}$ and $\mathcal{B}$ are collections of sets partitioning $\mathcal{X}$, but $\mathcal{A}$ is a finer partition of $\mathcal{X}$ than $\mathcal{B}$, that

$$\sum_{B \in \mathcal{B}} Q(B) f\left(\frac{P(B)}{Q(B)}\right) \le \sum_{A \in \mathcal{A}} Q(A) f\left(\frac{P(A)}{Q(A)}\right).$$

(*Hint:* Use the result of Question 2.11(a)).

(b) Suppose that $\mathcal{X}$ is countable (or finite) so that P and Q have p.m.f.s p and q. Show that

$$D_f\left(P\|Q\right) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right),$$

where on the left we are using the partition definition (2.2.3); you should show that the partition into discrete parts of $\mathcal{X}$ achieves the supremum. You may assume that $\mathcal{X}$ is finite. (Though feel free to prove the result in the case that $\mathcal{X}$ is infinite.)

Exercise 2.13 (General data processing inequalities)**:** Let f be a convex function satisfying $f(1) = 0$. Let K be a Markov transition kernel from $\mathcal{X}$ to $\mathcal{Z}$, that is, $K(\cdot \mid x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$. (Written differently, we have $X \to Z$, and conditioned on $X = x$, Z has distribution $K(\cdot \mid x)$, so that $K(A \mid x)$ is the probability that $Z \in A$ given $X = x$.)

(a) Define the marginals $K_P(A) = \int K(A \mid x)p(x)dx$ and $K_Q(A) = \int K(A \mid x)q(x)dx$. Show that

$$D_f\left(K_P \| K_Q\right) \le D_f\left(P \| Q\right).$$

Hint: by equation (2.2.3), w.l.o.g. we may assume that $\mathcal{Z}$ is finite and $\mathcal{Z} = \{1, \ldots, m\}$; also recall Question 2.11.

(b) Let X and Y be random variables with joint distribution P_{XY} and marginals P_X and P_Y. Define the f-information between X and Y as

$$I_f(X; Y) := D_f\left(P_{XY} \| P_X \times P_Y\right).$$

Use part (a) to show the following general data processing inequality: if we have the Markov chain $X \to Y \to Z$, then

$$I_f(X; Z) \le I_f(X; Y).$$

Exercise 2.14 (Convexity of f-divergences)**:** Prove Proposition 2.2.11. *Hint:* Use Question 2.11.

Exercise 2.15 (The f-divergence as a perspective transform)**:** The *perspective transform* of a convex function f,

$$f_{\rm per}(t, u) := \begin{cases} uf(t/u) & \text{if } u > 0 \\ +\infty & \text{otherwise,} \end{cases}$$

is jointly convex in its arguments (see Proposition B.2.24). The *closure*, which remains jointly convex in its arguments and for any $t_0 \in \operatorname{int} \operatorname{dom} f$ takes, the value

$$\operatorname{cl} f_{\rm per}(t, u) = \begin{cases} uf(t/u) & \text{if } u > 0 \\ \lim_{\alpha \downarrow 0} \alpha f(t_0 - t + t/\alpha) & \text{if } u = 0 \\ +\infty & \text{if } u < 0. \end{cases}$$

Use Corollary C.3.4 to show that the formula (2.2.5) coincides with the formulation

$$D_f\left(P \| Q\right) = \int_{\mathcal{X}} \operatorname{cl} f_{\rm per}(p(x), q(x)) d\mu(x).$$

Exercise 2.16 (Variational forms of KL divergence)**:** Let P and Q be arbitrary distributions on a common space $\mathcal{X}$. Prove the following variational representation, known as the Donsker-Varadhan theorem, of the KL divergence:

$$D_{\rm kl}\left(P \| Q\right) = \sup_{f : \mathbb{E}_Q[e^{f(X)}] < \infty} \left\{\mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp(f(X))]\right\}.$$

You may assume that P and Q have densities.

Exercise 2.17: Let P and Q have densities p and q with respect to a base measure μ over the set $\mathcal{X}$. (This is no loss of generality, as we may take $\mu = P + Q$.) Define the support $\operatorname{supp} P := \{x \in \mathcal{X} : p(x) > 0\}$. Show that

$$D_{\rm kl}\left(P\|Q\right) \geq \log \frac{1}{Q(\operatorname{supp} P)}.$$

Exercise 2.18: Let $\{P_v\}_{v \in \mathcal{V}}$ be an arbitrary collection of distributions on a space $\mathcal{X}$ and μ be a probability measure on $\mathcal{V}$. Show that if $V \sim \mu$ and conditional on $V = v$, we draw $X \sim P_v$, then

(a) $I(X; V) = \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v)$, where $\overline{P} = \int P_v d\mu(v)$ is the (weighted) average of the P_v. You may assume that $\mathcal{V}$ is discrete if you like.

(b) For any distribution Q on $\mathcal{X}$, $I(X; V) = \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) - D_{\rm kl}\left(\overline{P} \| Q\right)$. Conclude that $I(X; V) \leq \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v)$, or, equivalently, $\overline{P}$ minimizes $\int D_{\rm kl}\left(P_v \| Q\right) d\mu(v)$ over all probabilities Q.

Exercise 2.19 (The triangle inequality for variation distance)**:** Let P and Q be distributions on $X_1^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$, and let $P_i(\cdot \mid x_1^{i-1})$ be the conditional distribution of X_i given $X_1^{i-1} = x_1^{i-1}$ (and similarly for Q_i). Show that

$$\|P - Q\|_{\rm TV} \leq \sum_{i=1}^n \mathbb{E}_P\Big[\left\|P_i(\cdot \mid X_1^{i-1}) - Q_i(\cdot \mid X_1^{i-1})\right\|_{\rm TV}\Big],$$

where the expectation is taken over X_1^{i-1} distributed according to P.

Exercise 2.20: Let $h(p) = -p \log p - (1-p) \log(1-p)$. Show that $h(p) \geq 2 \log 2 \cdot \min\{p, 1-p\}$.

Exercise 2.21 (Lin [290], Theorem 8)**:** Let $h(p) = -p \log p - (1-p) \log(1-p)$. Show that $h(p) \leq 2 \log 2 \cdot \sqrt{p(1-p)}$.

Exercise 2.22 (Proving Pinsker's inequality via data processing)**:** We work through a proof of Proposition 2.2.9.(a) using the data processing inequality for f-divergences (Proposition 2.2.13).

(a) Define $D_{\rm kl}\left(p\|q\right) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. Argue that to prove Pinsker's inequality (2.2.10), it is enough to show that $(p-q)^2 \leq \frac{1}{2} D_{\rm kl}\left(p\|q\right)$.

(b) Define the negative binary entropy $h(p) = p \log p + (1-p) \log(1-p)$. Show that

$$h(p) \geq h(q) + h'(q)(p-q) + 2(p-q)^2$$

for any $p, q \in [0, 1]$.

(c) Conclude Pinsker's inequality (2.2.10).

Exercise 2.23 (A strong converse in source coding)**:** A code $\mathsf{C} : \mathcal{X}^n \to \{0,1\}^k$ is an (n, k) binary *block code* if it takes sequences of length n and outputs binary sequences of length k. Let $\mathsf{dec} : \{0,1\}^k \to \mathcal{X}^n$ be a decoder. Then $(\mathsf{C}, \mathsf{dec})$ has error probability ϵ if

$$\mathbb{P}(\mathsf{dec} \circ \mathsf{C}(X_1^n) \neq X_1^n) \leq \epsilon.$$

Let X_i be i.i.d. P, where P is a distribution on a finite set $\mathcal{X}$ with $P(X = x) > 0$ for each $x \in \mathcal{X}$, and let $k(n, \epsilon)$ be the minimal block length k for which there exists an (n, k) binary block code with error probability at most $\epsilon > 0$. Justify the following steps in showing that

$$\lim_{n\to\infty} \frac{k(n,\epsilon)}{n} = H_2(P), \tag{2.6.1}$$

where $H_2(P) = -\sum_x p(x) \log_2 p(x)$. This says that among all error $\epsilon > 0$ codes, the entropy is the minimal code length, and increasing the allowed probability of error never increases this length.

(a) The existence of an (n, k) binary block code with error ϵ is equivalent to the existence of a set $A \subset \mathcal{X}^n$ with $|A| \le 2^k$ and $P^n(A) \ge 1 - \epsilon$. Letting $m(n, \epsilon)$ be the minimum cardinality of sets $A \subset \mathcal{X}^n$ with $P^n(A) \ge 1 - \epsilon$, the limit (2.6.1) is equivalent to

$$\frac{\log_2 m(n,\epsilon)}{n} \to H_2(P).$$

(b) Define $L_i = \log_2 \frac{1}{p(X_i)}$ and let $\mathbb{E}_{P_n}[L] = \frac{1}{n}\sum_{i=1}^n L_i$, and let $\delta > 0$. Let $B_n = \{x_1^n \in \mathcal{X}^n \mid |\mathbb{E}_{P_n}[L] - H_2(P)| \le \delta\}$. Then $P^n(B_n) \to 1$, and $|B_n| \le 2^{n(H_2(P)+\delta)}$.

(c) We have $\limsup_n \frac{\log_2 m(n,\epsilon)}{n} \le \limsup_n \frac{1}{n}\log_2 |B_n|$ and so $\limsup_n \frac{k(n,\epsilon)}{n} \le H_2(P)$.

(d) For large enough n, for all sets A from part (a), $|A| \ge \frac{1-\epsilon}{2} 2^{n(H_2(P)-\delta)}$, which implies the result (2.6.1).

Exercise 2.24: This exercise outlines a proof of Proposition 2.4.4.

(a) Demonstrate the Cesàro mean limit that if $c_n = \frac{1}{n}\sum_{i=1}^n a_i$ and $a_i \to a$, then $c_n \to a$.

(b) Prove Proposition 2.4.4.

Exercise 2.25 (Proving Proposition 2.4.5)**:**

(a) Argue that for any m, we may define a prefix code $\mathsf{C} : \mathcal{X}^m \to \{0, \ldots, d-1\}^*$ with

$$\ell_{\mathsf{C}}(x_1^m) \le \left\lceil \log_d \frac{1}{P(X_1^m = x_1^m)} \right\rceil.$$

(b) Argue that if n/m is an integer, $\mathbb{E}_P\left[\ell_{\mathsf{C}}(X_1^n)\right] \le \frac{n}{m} + \frac{n}{m} H_d(X_1, \ldots, X_m)$.

(c) Show that by encoding the length of the sequence of encoded words in each block of length m, we can construct a prefix code with

$$\frac{1}{n}\mathbb{E}[\ell_{\mathsf{C}}(X_1^m)] \le \frac{O(1) \cdot \log_2 m}{m} + \frac{1}{m} H_d(X_1^m).$$

(d) Conclude Proposition 2.4.5.

Chapter 3

Exponential families and statistical modeling

Our second introductory chapter aims at readers less familiar with statistical modeling, and, as did the preceding chapter on information theory, this chapter blitzes through the main concepts. Nonetheless, the distributions here should give us something on which to hang our hats, as exponential family models undergird much of statistics, providing the first natural step away from Gaussian distributions to a flexible set of models that retain analytical elegance. The information-theoretic quantities of the previous chapter thus become especially explicit and computationally tractable in exponential families. A key property is that fitting exponential family models reduces to the minimization of convex functions—convex optimization problems—an operation we treat as a computational primitive. This (accurate enough) perspective informs the philosophy we adopt throughout that once we formulate a problem as convex, it is solved.

3.1 Exponential family models

We begin by defining exponential family distributions, giving several examples to illustrate a few of their properties. Three key objects define a d-dimensional exponential family distribution on an underlying space $\mathcal{X}$: the *sufficient statistic* $\phi : \mathcal{X} \to \mathbb{R}^d$ representing what we model, a *canonical parameter* vector $\theta \in \mathbb{R}^d$, and a *carrier* $h : \mathcal{X} \to \mathbb{R}_+$.

In the discrete case, where $\mathcal{X}$ is a discrete set, the exponential family associated with the sufficient statistic ϕ and carrier h has probability mass function

$$p_\theta(x) = h(x) \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right), \qquad A(\theta) := \log \sum_{x \in \mathcal{X}} h(x) \exp(\langle \theta, \phi(x) \rangle),$$

where A is the *log partition function*, sometimes called the *cumulant generating function*. In the continuous case, p_θ is instead a density on $\mathcal{X} \subset \mathbb{R}^k$, taking identical form while $A(\theta) = \log \int_{\mathcal{X}} h(x) \exp(\langle \theta, \phi(x) \rangle) dx$. We can abstract away from this distinction between discrete and continuous distributions by making the definition measure-theoretic, which we do here for completeness. (But recall the remarks in Section 1.5.)

Definition 3.1. *The* exponential family *associated with the function ϕ and base measure μ is the set of distributions with densities p_θ with respect to μ, where*

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right), \tag{3.1.1}$$

and A is the log partition function *(or cumulant generating function)*

$$A(\theta) := \log \int_{\mathcal{X}} \exp\left(\langle \theta, \phi(x) \rangle\right) d\mu(x) \tag{3.1.2}$$

whenever A is finite (and is $+\infty$ otherwise). The family is regular *if the domain*

$$\Theta := \{\theta \mid A(\theta) < \infty\}$$

is open.

Definition 3.1 includes the carrier h in the base measure μ, but frequently we use the notation

$$p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)).$$

In some scenarios, it may be convenient to re-parameterize the problem in terms of some function $\eta(\theta)$ instead of θ itself; we will simply use the most convenient formulae.

Example 3.1.1 (Bernoulli distribution)**:** We have $X \in \{0, 1\}$ and $P(X = 1) = p$ for $p \in (0, 1)$; the end pointcases arise as limits. We take μ to be counting measure on $\{0, 1\}$, and $\theta = \log \frac{p}{1-p}$ yields canonical representation

$$\begin{aligned} P(X = x) = p(x) = p^x (1 - p)^{1-x} &= \exp(x \log p - x \log(1 - p)) \\ &= \exp\left(x \log \frac{p}{1-p} + \log(1 - p)\right) = \exp\left(x\theta - \log(1 + e^\theta)\right). \end{aligned}$$

The Bernoulli family thus has log partition function $A(\theta) = \log(1 + e^\theta)$. $\diamond$

Example 3.1.2 (Poisson distribution)**:** The Poisson distribution (for count data) is usually parameterized by $\lambda > 0$, and for $x \in \mathbb{N}$ has distribution $P_\lambda(X = x) = (1/x!)\lambda^x e^{-\lambda}$. Thus by taking μ to be counting (discrete) measure on $\{0, 1, \ldots\}$ and setting $\theta = \log \lambda$, we find the density (probability mass function in this case)

$$p(x) = \frac{1}{x!} \lambda^x e^{-\lambda} = \exp(x \log \lambda - \lambda) \frac{1}{x!} = \exp(x\theta - e^\theta) \frac{1}{x!}.$$

Notably, taking carrier $h(x) = (x!)^{-1}$ and log partition $A(\theta) = e^\theta$, we have probability mass function $p_\theta(x) = h(x) \exp(\theta x - A(\theta))$. $\diamond$

Example 3.1.3 (Normal distribution, mean parameterization)**:** For the d-dimensional normal distribution, we take μ to be Lebesgue measure on $\mathbb{R}^d$. If we fix the covariance and vary only the mean μ in the family $\mathsf{N}(\mu, \Sigma)$, then $X \sim \mathsf{N}(\mu, \Sigma)$ has density

$$p_\mu(x) = \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) - \frac{1}{2} \log \det(2\pi\Sigma)\right).$$

Reparameterizing $\theta = \Sigma^{-1}\mu$, we obtain

$$p_\theta(x) = \underbrace{\exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x - \frac{1}{2} \log \det(2\pi\Sigma)\right)}_{=:h(x)} \exp\left(x^\top \theta - \frac{1}{2} \theta^\top \Sigma \theta\right).$$

In particular, we have carrier $h(x) = \exp(-\frac{1}{2} x^\top \Sigma^{-1} x)/((2\pi)^{d/2} \det(\Sigma)^{1/2})$, sufficient statistic $\phi(x) = x$, and log partition $A(\theta) = \frac{1}{2} \theta^\top \Sigma \theta$. $\diamond$

Example 3.1.4 (Normal distribution)**:** Let $X \sim \mathsf{N}(\mu, \Sigma)$. We may re-parameterize this as $K = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, and we have density

$$p_{\theta,K}(x) \propto \exp\left(\langle \theta, x\rangle - \frac{1}{2}\langle xx^\top, K\rangle\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean or trace inner product, as appropriate. See Exercise 3.1 for the calculation of the log partition function and associated carrier. $\diamond$

We sometimes include conditions making the exponential family parameterization unique.

Definition 3.2. *Let $\{P_\theta\}_{\theta\in\Theta}$ be an exponential family as in Definition 3.1. The sufficient statistic ϕ is* minimal *if $\Theta = \operatorname{dom} A \subset \mathbb{R}^d$ has interior and there exists no nonzero vector u such that*

$$\langle u, \phi(x)\rangle \quad \textit{is constant } \mu\textit{-almost surely.}$$

Definition 3.2 says equivalently that the functions $1, \phi_1, \ldots, \phi_d$ are linearly independent modulo μ-null sets. While we do not prove this, via a suitable linear transformation—a variant of Gram-Schmidt orthonormalization—one may modify any non-minimal exponential family $\{P_\theta\}$ into an equivalent minimal exponential family $\{Q_\eta\}$, meaning that the two collections satisfy the equality $\{P_\theta\} = \{Q_\eta\}$ (see Brown [90, Chapter 1]).

3.2 Why exponential families?

We study exponential families because they combine modeling flexibility with analytic and computational tractability. As with the normal distribution, they often admit explicit calculations, but they allow discrete, bounded, or other non-Gaussian data. Their analytic tractability has made them the objects of substantial study; Brown [90] provides a deep and elegant treatment. Moreover, they arise as the solutions to natural optimization problems over probability distributions, and they enjoy robustness properties related to optimal Bayes procedures (see Section 14.4).

3.2.1 Analytical properties

Here, we enumerate a few of their key analytical properties, focusing on the log partition function $A(\theta) = \log \int e^{\langle\theta,\phi(x)\rangle} d\mu(x)$ and its relationship to expectations $\mathbb{E}_\theta$ under the distribution P_θ. We begin with a heuristic calculation, where we assume that we may exchange differentiation and integration. We then obtain the important expectation and covariance relationships that

$$\begin{aligned}\nabla A(\theta) &= \frac{1}{\int e^{\langle\theta,\phi(x)\rangle} d\mu(x)} \int \nabla_\theta e^{\langle\theta,\phi(x)\rangle} d\mu(x) \\ &= e^{-A(\theta)} \int \nabla_\theta e^{\langle\theta,\phi(x)\rangle} d\mu(x) = \int \phi(x) e^{\langle\theta,\phi(x)\rangle - A(\theta)} d\mu(x) = \mathbb{E}_\theta[\phi(X)] \end{aligned} \tag{3.2.1}$$

because $e^{\langle\theta,\phi(x)\rangle - A(\theta)} = p_\theta(x)$. A similar (and still heuristic for now) calculation gives

$$\nabla^2 A(\theta) = \mathbb{E}_\theta[\phi(X)\phi(X)^\top] - \mathbb{E}_\theta[\phi(X)]\mathbb{E}_\theta[\phi(X)]^\top = \operatorname{Cov}_\theta(\phi(X)). \tag{3.2.2}$$

That these identities hold is no accident and is central to the appeal of exponential family models.

The first and, from our perspective, most important result about exponential family models is their convexity. While (assuming the differentiation relationships above hold) the differentiation identity that $\nabla^2 A(\theta) = \operatorname{Cov}_\theta(\phi(X)) \succeq 0$ makes convexity of A immediate, one can also provide a direct argument without appealing to differentiation.

Proposition 3.2.1. *The cumulant-generating function $\theta \mapsto A(\theta)$ is convex.*

Proof Let $0 < \lambda < 1$ and $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$, where $\theta_1, \theta_2 \in \Theta$. Then $1/\lambda \geq 1$ and $1/(1-\lambda) \geq 1$, so Hölder's inequality implies

$$\begin{aligned}
\log \int \exp(\langle \theta_\lambda, \phi(x) \rangle) d\mu(x) &= \log \int \exp(\langle \theta_1, \phi(x) \rangle)^\lambda \exp(\langle \theta_2, \phi(x) \rangle)^{1-\lambda} d\mu(x) \\
&\leq \log \left(\int \exp(\langle \theta_1, \phi(x) \rangle)^{\frac{\lambda}{\lambda}} d\mu(x) \right)^\lambda \left(\int \exp(\langle \theta_2, \phi(x) \rangle)^{\frac{1-\lambda}{1-\lambda}} d\mu(x) \right)^{1-\lambda} \\
&= \lambda \log \int \exp(\langle \theta_1, \phi(x) \rangle) d\mu(x) + (1-\lambda) \log \int \exp(\langle \theta_2, \phi(x) \rangle) d\mu(x).
\end{aligned}$$

The case $\lambda \in \{0, 1\}$ is trivial. □

We now show that $A(\theta)$ is indeed infinitely differentiable and how it generates the moments of the sufficient statistics $\phi(x)$. To describe the properties, we provide a bit of tensor notation: for $x \in \mathbb{R}^d$, let $x^{\otimes k}$ denote the kth order tensor, which as a multilinear operator on $v_1, \ldots, v_k \in \mathbb{R}^d$ has definition

$$x^{\otimes k}(v_1, \ldots, v_k) := \langle x, v_1 \rangle \cdots \langle x, v_k \rangle = \prod_{i=1}^k \langle x, v_i \rangle.$$

When $k = 2$, this is the familiar outer product $x^{\otimes 2} = xx^\top$. With this notation, we demonstrate that we differentiate under the integral sign to compute any derivative of $e^{A(\theta)}$.

Proposition 3.2.2. *The cumulant-generating function $\theta \mapsto A(\theta)$ is infinitely differentiable on the interior of its domain $\Theta := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. The moment-generating function*

$$M(\theta) := \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x)$$

is analytic on the set $\Theta_\mathbb{C} := \{z \in \mathbb{C}^d \mid \operatorname{Re} z \in \operatorname{int} \Theta\}$. Additionally, the derivatives of M are computed by passing through the integral, that is,

$$\begin{aligned}
\nabla_\theta^k M(\theta) = \nabla_\theta^k \int e^{\langle \theta, \phi(x) \rangle} d\mu(x) &= \int \nabla_\theta^k e^{\langle \theta, \phi(x) \rangle} d\mu(x) \\
&= \int \phi(x)^{\otimes k} \exp(\langle \theta, \phi(x) \rangle) d\mu(x).
\end{aligned}$$

The proof of the proposition requires complex analysis, so we defer it to Section 3.6.1.

As particular consequences of Proposition 3.2.2, we have rigorously demonstrated the expectation and covariance relationships (3.2.1) and (3.2.2). Minimal exponential families (Definition 3.2) also enjoy a few additional regularity properties. Recall that A is *strictly convex* if

$$A(\lambda\theta_0 + (1-\lambda)\theta_1) < \lambda A(\theta_0) + (1-\lambda) A(\theta_1)$$

whenever $\lambda \in (0, 1)$ and $\theta_0, \theta_1 \in \operatorname{dom} A$. We have the following proposition.

Proposition 3.2.3. *Let $\{P_\theta\}$ be a regular exponential family. The log partition function A is strictly convex if and only if $\{P_\theta\}$ is minimal.*

Proof If the family is minimal, then for each $\theta \in \Theta$ and nonzero u, $\mathrm{Var}_\theta(u^\top \phi(X)) > 0$, while $\mathrm{Var}_\theta(u^\top \phi(X)) = u^\top \nabla^2 A(\theta) u$. This implies the strict positive definiteness $\nabla^2 A(\theta) \succ 0$ on the open domain Θ, which is equivalent to strict convexity (see Corollary B.2.3 in Appendix B.2.1). Conversely, if $\nabla^2 A(\theta) \succ 0$ for all $\theta \in \Theta$, then $\mathrm{Var}_\theta(u^\top \phi(X)) > 0$ for all $u \neq 0$ and so $u^\top \phi(x)$ is non-constant in x. □

3.2.2 Fitting an exponential family model

Their convexity and differentiability properties make exponential family models especially attractive from a computational perspective. A major focus in statistics is the convergence of estimates of different properties of a population distribution P and whether these estimates are computable. We will develop tools to address the first of these questions, and attendant optimality guarantees, throughout this book. To set the stage for what follows, let us consider what this entails in the context of exponential family models.

Suppose we have a population P (where, for simplicity, we assume P has a density p), and for a given exponential family $\mathcal{P}$ with densities $\{p_\theta\}$, we wish to find the model closest to P. Then it is natural (if we take on faith that the information-theoretic measures we have developed are the "right" ones) to find the distribution $P_\theta \in \mathcal{P}$ closest to P in KL-divergence, that is, to solve

$$\underset{\theta}{\text{minimize}} \; D_{\rm kl}\left(P \| P_\theta\right) = \int p(x) \log \frac{p(x)}{p_\theta(x)} dx. \tag{3.2.3}$$

This is evidently equivalent to minimizing

$$-\int p(x) \log p_\theta(x) dx = \int p(x) \left[-\langle \theta, \phi(x) \rangle + A(\theta)\right] dx = -\langle \theta, \mathbb{E}_P[\phi(X)] \rangle + A(\theta).$$

This is always a convex optimization problem (see Appendices B and C for much more on this), as A is convex and the first term is linear, and so has no non-global local optima. Here and throughout, we treat convex optimization as a technology: provided the problem has moderate dimension and evaluable objective, such problems are computationally straightforward.

Of course, we never have access to the population P fully; instead, we receive a sample $X_1, \ldots, X_n$ from P. In this case, a natural approach is to replace the expected (negative) log likelihood above with its empirical version and solve

$$\underset{\theta}{\text{minimize}} \; -\sum_{i=1}^n \log p_\theta(X_i) = \sum_{i=1}^n [-\langle \theta, \phi(X_i) \rangle + A(\theta)], \tag{3.2.4}$$

which is still a convex optimization problem (as the objective is convex in θ). The maximum likelihood estimate is any vector $\widehat{\theta}_n$ minimizing the negative log likelihood (3.2.4), and setting gradients to 0 any interior minimizer evidently matches moments, satisfying

$$\nabla A(\widehat{\theta}_n) = \mathbb{E}_{\widehat{\theta}_n}[\phi(X)] = \frac{1}{n} \sum_{i=1}^n \phi(X_i). \tag{3.2.5}$$

By convexity, any moment matching point (3.2.5) is a minimizer. (Such an estimate need not exist; for example, no Gaussian can fit a distribution with low rank covariance.) Assuming it exists, this $\widehat{\theta}_n$ is unique by strict convexity whenever $\mathrm{Cov}_\theta(\phi(X)) \succ 0$ for all θ.

Let us proceed heuristically to develop a rough convergence guarantee for the estimator $\widehat{\theta}_n$; the next paragraph assumes comfort with classical asymptotic statistics (and the central limit theorem) and is not essential for what comes later. Assume that the data X_i are i.i.d. from an exponential family model $P_{\theta^\star}$. Then the maximum likelihood estimate $\widehat{\theta}_n$ should converge to $\theta^\star$, and so

$$\frac{1}{n}\sum_{i=1}^n \phi(X_i) = \nabla A(\widehat{\theta}_n) = \nabla A(\theta^\star) + (\nabla^2 A(\theta^\star) + o(1))(\widehat{\theta}_n - \theta^\star).$$

But of course, $\nabla A(\theta^\star) = \mathbb{E}_{\theta^\star}[\phi(X)]$, and so the central limit theorem gives that

$$\frac{1}{n}\sum_{i=1}^n (\phi(X_i) - \nabla A(\theta^\star)) \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\mathrm{Cov}_{\theta^\star}(\phi(X))\right) = \mathsf{N}\left(0, n^{-1}\nabla^2 A(\theta^\star)\right),$$

where $\stackrel{\cdot}{\sim}$ means "is approximately distributed as." Multiplying by $(\nabla^2 A(\theta^\star)+o(1))^{-1} \approx \nabla^2 A(\theta^\star)^{-1}$, we thus see (still working in our heuristic)

$$\widehat{\theta}_n - \theta^\star = \left(\nabla^2 A(\theta^\star) + o(1)\right)^{-1} \frac{1}{n}\sum_{i=1}^n (\phi(X_i) - \nabla A(\theta^\star)) \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1} \cdot \nabla^2 A(\theta^\star)^{-1}\right), \tag{3.2.6}$$

where we use that $BZ \sim \mathsf{N}(0, B\Sigma B^\top)$ if $Z \sim \mathsf{N}(0, \Sigma)$. (It is possible to make each of these steps fully rigorous.) Thus the cumulant generating function A governs the error we expect in $\widehat{\theta}_n - \theta^\star$.

Much of the rest of this book explores properties of these minimization problems: at what rates do we expect $\widehat{\theta}_n$ to converge to a global minimizer of problem (3.2.3)? Can we show that these rates are optimal? Is this the "right" strategy for choosing a parameter? Exponential families form a particular working example to motivate this development.

3.3 Divergence measures and information for exponential families

Their elegant analytic properties mean that exponential family models also play nicely with the information-theoretic tools we develop. Indeed, consider the KL-divergence between two exponential family distributions P_θ and $P_{\theta+\Delta}$, where $\Delta \in \mathbb{R}^d$. Then we have

$$\begin{aligned} D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) &= \mathbb{E}_\theta\left[\langle \theta, \phi(X)\rangle - A(\theta) - \langle \theta+\Delta, \phi(X)\rangle + A(\theta+\Delta)\right] \\ &= A(\theta+\Delta) - A(\theta) - \mathbb{E}_\theta[\langle \Delta, \phi(X)\rangle] \\ &= A(\theta+\Delta) - A(\theta) - \nabla A(\theta)^\top \Delta. \end{aligned}$$

Similarly, we have

$$\begin{aligned} D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) &= \mathbb{E}_{\theta+\Delta}\left[\langle \theta+\Delta, \phi(X)\rangle - A(\theta+\Delta) - \langle \theta, \phi(X)\rangle + A(\theta)\right] \\ &= A(\theta) - A(\theta+\Delta) + \mathbb{E}_{\theta+\Delta}[\langle \Delta, \phi(X)\rangle] \\ &= A(\theta) - A(\theta+\Delta) - \nabla A(\theta+\Delta)^\top(-\Delta). \end{aligned}$$

These identities give an immediate connection with convexity. For a differentiable convex function h, the *Bregman divergence* associated with h is

$$D_h(u, v) = h(u) - h(v) - \langle \nabla h(v), u - v\rangle, \tag{3.3.1}$$

which is always nonnegative, as it is the gap between the linear approximation to the (convex) function h and its actual value. (See Figure 3.1.)

We catalog these results as the following proposition.

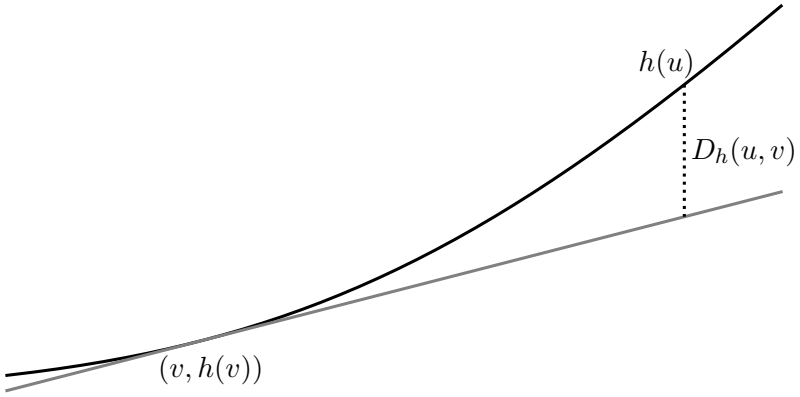


Figure 3.1. Illustration of the Bregman divergence $D_h(u, v)$ between u and v as the gap between h and its tangent.

Proposition 3.3.1. *Let $\{P_\theta\}$ be an exponential family model with cumulant generating function $A(\theta)$. Then*

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = D_A(\theta + \Delta, \theta) \quad \text{and} \quad D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) = D_A(\theta, \theta + \Delta).$$

Additionally, there exists a $t \in [0, 1]$ such that

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta + t\Delta)\Delta,$$

and similarly, there exists a $t \in [0, 1]$ such that

$$D_{\rm kl}\left(P_{\theta+\Delta} \| P_\theta\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta + t\Delta)\Delta.$$

Proof We have already shown the first two statements; the second two are applications of Taylor's theorem. □

When the perturbation Δ is small, that A is infinitely differentiable then gives that

$$D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta)\Delta + O(\|\Delta\|^3),$$

so that the Hessian $\nabla^2 A(\theta)$ tells quite precisely how the KL divergence changes as θ varies (locally). As we saw already in Example 2.3.2 (and see the next section), when the KL-divergence between two distributions is small, it is hard to test between them, and in the sequel, we will show converses to this. The Hessian $\nabla^2 A(\theta^\star)$ also governs the error in the estimate $\widehat{\theta}_n - \theta^\star$ in our heuristic (3.2.6). When the Hessian $\nabla^2 A(\theta)$ is large in the semidefinite order, the KL divergence $D_{\rm kl}\left(P_\theta \| P_{\theta+\Delta}\right)$ is large, and the asymptotic covariance (3.2.6) is small. For this—and other reasons we address later—for exponential family models, we call

$$\nabla^2 A(\theta) = {\rm Cov}_\theta(\phi(X)) = \mathbb{E}_\theta[\nabla \log p_\theta(X) \nabla \log p_\theta(X)^\top] \tag{3.3.2}$$

the *Fisher information* of the parameter θ in the model $\{P_\theta\}$. The objects we introduce here—Bregman divergences and Fisher information—will recur as central ingredients in our study of estimation, loss functions, and fundamental limits.

3.4 Generalized linear models and regression

We can specialize the general modeling strategies that exponential families provide to more directly address prediction problems, where we wish to predict a target $Y \in \mathcal{Y}$ given covariates $X \in \mathcal{X}$. Here, we almost always have that Y is either discrete or continuous with $\mathcal{Y} \subset \mathbb{R}$. In this case, we have a sufficient statistic $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, and we model $Y \mid X = x$ via the *generalized linear model* (GLM, or conditional exponential family model) with density or probability mass function

$$p_\theta(y \mid x) = \exp\left(\phi(x, y)^\top \theta - A(\theta \mid x)\right) h(y), \tag{3.4.1}$$

where as before h is the carrier and $A(\theta \mid x)$ denotes the conditional log partition, so

$$A(\theta \mid x) = \log \int \exp(\phi(x, y)^\top \theta) h(y) dy$$

in the case that $\mathcal{Y} \subset \mathbb{R}^k$ or, in the discrete case,

$$A(\theta \mid x) = \log \sum_y \exp(\phi(x, y)^\top \theta) h(y).$$

The log partition function $A(\cdot \mid x)$ provides the same insights for the conditional models (3.4.1) as it does for the unconditional exponential family models in the preceding sections. Indeed, as in Propositions 3.2.1 and 3.2.2, the log partition $A(\cdot \mid x)$ is always $\mathcal{C}^\infty$ on its domain and convex. Moreover, it gives the expected moments of the sufficient statistic ϕ conditional on x, as

$$\nabla A(\theta \mid x) = \mathbb{E}_\theta[\phi(X, Y) \mid X = x],$$

and

$$\nabla^2 A(\theta \mid x) = \mathrm{Cov}_\theta(\phi(X, Y) \mid X = x),$$

from which we can (typically) extract the mean or other statistics of Y conditional on x.

Three standard examples will be our most frequent motivators throughout this book: linear regression, binary logistic regression, and multiclass logistic regression. We give these, together with Poisson regression for count data.

Example 3.4.1 (Linear regression)**:** In linear regression, we wish to predict $Y \in \mathbb{R}$ from a vector $X \in \mathbb{R}^d$, and assume that $Y \mid X = x$ follows the normal distribution $\mathsf{N}(\theta^\top x, \sigma^2)$. In this case, we have

$$\begin{aligned} p_\theta(y \mid x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - x^\top\theta)^2\right) \\ &= \exp\left(\frac{1}{\sigma^2} y x^\top \theta - \frac{1}{2\sigma^2}\theta^\top x x^\top \theta\right) \exp\left(-\frac{1}{2\sigma^2} y^2 - \frac{1}{2}\log(2\pi\sigma^2)\right). \end{aligned}$$

Thus we have the exponential family representation (3.4.1) with sufficient statistic $\phi(x, y) = \frac{1}{\sigma^2} xy$, carrier $h(y) = \exp(-\frac{1}{2\sigma^2} y^2 - \frac{1}{2}\log(2\pi\sigma^2))$, and log partition $A(\theta \mid x) = \frac{1}{2\sigma^2}\theta^\top x x^\top \theta$. As $\nabla A(\theta \mid x) = \mathbb{E}_\theta[\phi(X, Y) \mid X = x] = \frac{1}{\sigma^2} x \mathbb{E}_\theta[Y \mid X = x]$, we recover $\mathbb{E}_\theta[Y \mid X = x] = \theta^\top x$. $\diamond$

Frequently, we wish to predict binary or multiclass random variables Y. For example, we may wish to assess the probability that an individual will have a heart attack in the next 5 years based on a set of covariates $x \in \mathbb{R}^d$, where $Y = 1$ indicates heart attack and $Y = -1$ indicates none.

Example 3.4.2 (Binary logistic regression)**:** If $Y \in \{-1, 1\}$, we model

$$p_\theta(y \mid x) = \frac{\exp(yx^\top\theta)}{1 + \exp(yx^\top\theta)},$$

where the model assigns high probability to the event $Y = y$ when $yx^\top\theta \gg 0$. We always have $p_\theta(y \mid x) + p_\theta(-y \mid x) = 1$, and the identity

$$yx^\top\theta - \log(1 + \exp(yx^\top\theta)) = \frac{y+1}{2}x^\top\theta - \log(1 + \exp(x^\top\theta))$$

implies the GLM representation $\phi(x, y) = \frac{y+1}{2}x$ and $A(\theta \mid x) = \log(1 + \exp(x^\top\theta))$.

As an alternative, we could represent $Y \in \{0, 1\}$ by

$$p_\theta(y \mid x) = \frac{\exp(yx^\top\theta)}{1 + \exp(x^\top\theta)} = \exp\left(yx^\top\theta - \log(1 + e^{x^\top\theta})\right),$$

which has the simpler sufficient statistic $\phi(x, y) = xy$. $\diamond$

Instead of a binary prediction problem, in many cases we have a *multiclass* prediction problem, where we seek to predict a label Y for an object x belonging to one of k different classes. For example, in image recognition, we are given an image x and wish to identify the subject Y of the image, where Y ranges over k distinct classes. This too we can model using exponential families.

Example 3.4.3 (Multiclass logistic regression)**:** In the case that we have a k-class prediction problem in which we wish to predict $Y \in \{1, \ldots, k\}$ from $X \in \mathbb{R}^d$, we assign parameters $\theta_y \in \mathbb{R}^d$ to each of the classes $y = 1, \ldots, k$. We then model

$$p_\theta(y \mid x) = \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} = \exp\left(\theta_y^\top x - \log\left(\sum_{j=1}^k e^{\theta_j^\top x}\right)\right).$$

Here, the idea is that if $\theta_y^\top x > \theta_j^\top x$ for all $j \neq y$, then the model assigns higher probability to class y than any other class; the larger the gap between $\theta_y^\top x$ and $\theta_j^\top x$, the larger the difference in assigned probabilities. This family is not minimal: adding a constant vector v to each parameter θ_y leaves the probabilities unchanged; to make the family minimal, we may assign $\theta_k = \mathbf{0}$ or enforce the normalization $\sum_{j=1}^k \theta_j = \mathbf{0}$. $\diamond$

Similar approaches allow us to model other situations. Poisson regression models are frequent choices for modeling count data, such as adverse events in issuing insurance premiums. To model such counts $Y \in \{0, 1, 2, \ldots\}$, we turn to Poisson regression.

Example 3.4.4 (Poisson regression)**:** When $Y \in \mathbb{N}$ is a count, the Poisson distribution with rate $\lambda > 0$ gives $P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$. Poisson regression models λ via $e^{\theta^\top x}$, giving model

$$p_\theta(y \mid x) = \frac{1}{y!}\exp\left(yx^\top\theta - e^{\theta^\top x}\right),$$

so that we have carrier $h(y) = 1/y!$ and the simple sufficient statistic $\phi(x, y) = yx$. The log partition function is $A(\theta \mid x) = e^{\theta^\top x}$. $\diamond$

The exercises explore additional examples.

3.4.1 Fitting a generalized linear model from a sample

We briefly revisit the approach in Section 3.2.2 for fitting exponential family models in the context of generalized linear models. In this case, the analogue of the maximum likelihood problem (3.2.4) is to solve

$$\minimize_\theta \quad -\sum_{i=1}^n \log p_\theta(Y_i \mid X_i) = \sum_{i=1}^n \left[-\phi(X_i, Y_i)^\top \theta + A(\theta \mid X_i)\right].$$

This is a convex optimization problem with $\mathcal{C}^\infty$ objective. Assuming a (finite and interior) minimizer exists, by convexity, equality (3.2.5) extends, so a necessary and sufficient condition for $\widehat{\theta}_n$ to minimize the above objective is that it achieves 0 gradient, that is,

$$\frac{1}{n}\sum_{i=1}^n \nabla A(\widehat{\theta}_n \mid X_i) = \frac{1}{n}\sum_{i=1}^n \phi(X_i, Y_i).$$

Once again, to find $\widehat{\theta}_n$ amounts to matching moments, as $\nabla A(\theta \mid X_i) = \mathbb{E}[\phi(X, Y) \mid X = X_i]$.

No exponential family or generalized linear model (GLM) has perfect fidelity: all models are approximations, though many capture important phenomena. Nonetheless, we can still *fit* any of the GLM models in Examples 3.4.1–3.4.4 to data of the appropriate type. In particular, for the logarithmic loss $\ell(\theta; x, y) = -\log p_\theta(y \mid x)$, we can define the empirical loss

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i).$$

Then, as $n \to \infty$, we expect that $L_n(\theta) \to \mathbb{E}[\ell(\theta; X, Y)]$, so that minimizing L_n approximates minimizing the population prediction loss. We shall therefore often study deviations of sample quantities (such as L_n) from their population counterparts.

3.4.2 The information in a generalized linear model

As in Section 3.3, we can compute the "information" about a parameter θ in a generalized linear model. In this case, P_θ specifies only the conditional distribution of Y given $X = x$, so when we compute information, we assume X follows some marginal distribution Q. We abuse composition notation, writing $(Y, X) \sim P_\theta \circ Q$ to mean $\mathbb{P}((X, Y) \in A) = \int_{(x,y)\in A} p_\theta(y \mid x) q(x) dy dx$. Then

$$\begin{aligned} D_{\rm kl}\left(P_\theta \circ Q \| P_{\theta+\Delta} \circ Q\right) = \mathbb{E}_\theta\left[\log \frac{p_\theta(Y \mid X)}{p_{\theta+\Delta}(Y \mid X)}\right] &= \mathbb{E}_\theta\left[\Delta^\top \phi(X, Y) - A(\theta \mid X) + A(\theta + \Delta \mid X)\right] \\ &= \mathbb{E}_\theta\left[D_{A(\cdot|X)}(\theta + \Delta, \theta)\right], \end{aligned}$$

and similarly

$$D_{\rm kl}\left(P_{\theta+\Delta} \circ Q \| P_\theta \circ Q\right) = \mathbb{E}_{\theta+\Delta}\left[D_{A(\cdot|X)}(\theta, \theta + \Delta)\right],$$

where we recall the Bregman divergence (3.3.1) and have used that $\mathbb{E}_\theta[\phi(X, Y) \mid X] = \nabla A(\theta \mid X)$. Performing a Taylor expansion, we have

$$A(\theta + \Delta \mid x) = A(\theta \mid x) + \langle \nabla A(\theta \mid x), \Delta\rangle + \frac{1}{2}\Delta^\top \nabla^2 A(\theta \mid x)\Delta + O(\|\Delta\|^3),$$

where the big-O hides dependence on x and θ. We thus (somewhat heuristically) obtain the following corollary:

Corollary 3.4.5. *Assume appropriate regularity conditions. Then as* $\Delta \to 0$,

$$D_{\rm kl}\left(P_\theta \circ Q \| P_{\theta+\Delta} \circ Q\right) = \frac{1}{2}\Delta^\top \mathbb{E}_Q[\nabla^2 A(\theta \mid X)]\Delta + O(\|\Delta\|^3)$$

and

$$D_{\rm kl}\left(P_{\theta+\Delta} \circ Q \| P_\theta \circ Q\right) = \frac{1}{2}\Delta^\top \mathbb{E}_Q[\nabla^2 A(\theta \mid X)]\Delta + O(\|\Delta\|^3).$$

In analogy with Proposition 3.3.1, we see again that the expected Hessian $\mathbb{E}_Q[\nabla^2 A(\theta \mid X)]$ tells quite precisely how the KL divergence changes as θ varies locally, but now, the distribution Q on X also enters the picture. So when A and the distribution Q are such that $\mathbb{E}_Q[\nabla^2 A(\theta \mid X)]$ is large in the semidefinite order, then it is easy to distinguish data coming from $P_\theta \circ Q$ from that drawn from $P_{\theta'} \circ Q$, and otherwise, it is not. We therefore call

$$\mathbb{E}[\nabla^2 A(\theta \mid X)] = \mathbb{E}[{\rm Cov}_\theta(\phi(X,Y) \mid X)] = \mathbb{E}[\nabla \log p_\theta(Y \mid X) \nabla \log p_\theta(Y \mid X)^\top] \tag{3.4.2}$$

the *Fisher information* of the parameter θ in the model P_θ.

Example 3.4.6 (The information in logistic regression)**:** For the binary logistic regression model (Example 3.4.2) with $Y \in \{0, 1\}$, we have

$$\nabla \log p_\theta(y \mid x) = yx - \frac{e^{x^\top\theta}}{1 + e^{x^\top\theta}} x = (y - p_\theta(1 \mid x))x$$

and $\nabla^2 A(\theta \mid X) = p_\theta(1 \mid x)(1 - p_\theta(1 \mid x))xx^\top$. Thus for $X \sim Q$ the Fisher information is

$$\mathbb{E}_Q[{\rm Var}_\theta(Y \mid X)XX^\top] \preceq \frac{1}{4}\mathbb{E}_Q[XX^\top].$$

When $\theta = 0$, we have identically ${\rm Var}_\theta(Y \mid X) = \frac{1}{4}$, which is the "maximal" information. Additionally, we see that when $X \sim Q$ has larger covariance, we expect $XX^\top$ to be larger in the semidefinite order, meaning the observations (X, Y) contain more information about θ (of course, this is mitigated by the fact that $p_\theta(y \mid x)$ becomes more extreme as $\|x\|$ grows). $\diamond$

Example 3.4.7 (The KL-divergence in logistic regression)**:** The binary logistic regression model with $Y \in \{0, 1\}$ also admits simple bounds on its KL-divergence. For these, we first make the simple observation that for the log-sum-exp function $f(t) = \log(1 + e^t)$, we have $f'(t) = \frac{e^t}{1+e^t}$ and $f''(t) = f'(t)(1 - f'(t))$. Since $0 \le f'' \le \frac{1}{4}$, Taylor's theorem gives

$$|f(t + \Delta) - f(t) + f'(t)\Delta| \le \frac{\Delta^2}{8}$$

for all $\Delta \in \mathbb{R}$. Computing the divergence directly, we have for any parameters θ_0, θ_1 that

$$D_{\rm kl}\left(P_{\theta_0}(\cdot \mid x) \| P_{\theta_1}(\cdot \mid x)\right) = f(\theta_1^\top x) - f(\theta_0^\top x) - f'(\theta_0^\top x)x^\top(\theta_1 - \theta_0) \le \frac{1}{8}\left(x^\top(\theta_0 - \theta_1)\right)^2, \tag{3.4.3}$$

so the divergence is at most quadratic. $\diamond$

3.5 Lower bounds on testing a parameter's value

We give a bit of a preview here of the tools we will develop to prove fundamental limits in Part II of the book, an *hors d'oeuvre* previewing the techniques we develop. In Section 2.3.1, we presented Le Cam's method and used it in Example 2.3.2 to give a lower bound on the probability of error in a hypothesis test comparing two normal means. This approach extends beyond this simple case, and here we give another example applying it to exponential family models.

We give a stylized version of the problem. Let $\{P_\theta\}$ be an exponential family model with parameter $\theta \in \mathbb{R}^d$. Suppose for some vector $v \in \mathbb{R}^d$, we wish to test whether $v^\top \theta > 0$ or $v^\top \theta < 0$ in the model. For example, in the regression settings in Section 3.4, we may be interested in the effect of a treatment on health outcomes. Then the covariates x contain information about an individual, with the first covariate x_1 indicating individual treatmen, while Y measures the outcome. Setting $v = e_1$, we then wish to test whether there is a positive or negative treatment effect $\theta_1 = e_1^\top \theta$.

Abstracting away the specifics of the scenario, we ask the following question: given an exponential family $\{P_\theta\}$ and a threshold t of interest, at what separation $\delta > 0$ does it become essentially impossible to test

$$v^\top \theta \le t \quad \text{versus} \quad v^\top \theta \ge t + \delta?$$

We give one approach to this using two-point hypothesis testing lower bounds. In this case, we consider testing sequences of alternatives

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_{1,n} : \theta = \theta_n$$

as n grows, where we observe a sample X_1^n drawn i.i.d. either according to P_{θ_0} (i.e., H_0) or P_{θ_n} (i.e., $H_{1,n}$). By choosing θ_n in a way that makes the separation $v^\top(\theta_n - \theta_0)$ large but testing H_0 against $H_{1,n}$ challenging, we can then (roughly) identify the separation δ at which testing becomes impossible.

Proposition 3.5.1. *Assume* $\operatorname{dom} A$ *has interior, and let* $\theta_0 \in \operatorname{int} \operatorname{dom} A$ *with* $\nabla^2 A(\theta_0)$ *positive definite. Then there exists a sequence of parameters* θ_n *with separation*

$$v^\top(\theta_n - \theta_0) = \frac{1}{\sqrt{n}} \sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v}$$

and for which

$$\inf_\Psi \left\{ P_{\theta_0}(\Psi(X_1^n) \neq 0) + P_{\theta_n}(\Psi(X_1^n) \neq 1) \right\} \ge \frac{1}{2} - O(n^{-1/2}).$$

Proof Let $\Delta \in \mathbb{R}^d$ be a potential perturbation of θ_0 to $\theta_1 = \theta_0 + \Delta$, which gives separation $\delta = v^\top \theta_1 - v^\top \theta_0 = v^\top \Delta$. Let $P_0 = P_{\theta_0}$ and $P_1 = P_{\theta_1}$. Then the smallest summed probability of error in testing between P_0 and P_1 based on n observations X_1^n is

$$\inf_\Psi \left\{ P_0(\Psi(X_1, \ldots, X_n) \neq 0) + P_1(\Psi(X_1, \ldots, X_n) \neq 1) \right\} = 1 - \|P_0^n - P_1^n\|_{\mathrm{TV}}$$

by Proposition 2.3.1. Following the approach of Example 2.3.2, we apply Pinsker's inequality (2.2.10) and use that the KL-divergence tensorizes to find

$$2 \|P_0^n - P_1^n\|_{\mathrm{TV}}^2 \le n D_{\mathrm{kl}}\left(P_0 \| P_1\right) = n D_{\mathrm{kl}}\left(P_{\theta_0} \| P_{\theta_0 + \Delta}\right) = n D_A(\theta_0 + \Delta, \theta_0),$$

where the final equality follows from the equivalence between KL and Bregman divergences for exponential families (Proposition 3.3.1).

To guarantee that the summed probability of error is at least $\frac{1}{2}$, that is, $\|P_0^n - P_1^n\|_{\rm TV} \le \frac{1}{2}$, it suffices to choose Δ satisfying $nD_A(\theta_0 + \Delta, \theta_0) \le \frac{1}{2}$. So to maximize the separation $v^\top \Delta$ while guaranteeing a constant probability of error, we (approximately) solve

$$\begin{array}{ll} \text{maximize} & v^\top \Delta \\ \text{subject to} & D_A(\theta_0 + \Delta, \theta_0) \le \frac{1}{2n}. \end{array}$$

Now, consider that $D_A(\theta_0 + \Delta, \theta_0) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta_0)\Delta + O(\|\Delta\|^3)$. Ignoring the higher order term, we consider maximizing $v^\top \Delta$ subject to $\Delta^\top \nabla^2 A(\theta_0)\Delta \le \frac{1}{n}$. A Lagrangian calculation shows that this has solution

$$\Delta = \frac{1}{\sqrt{n}} \frac{1}{\sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v}} \nabla^2 A(\theta_0)^{-1} v.$$

With this choice, we have separation $\delta = v^\top \Delta = \sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v / n}$, and $D_A(\theta_0 + \Delta, \theta_0) = \frac{1}{2n} + O(1/n^{3/2})$. The summed probability of error is at least

$$1 - \|P_0^n - P_1^n\|_{\rm TV} \ge 1 - \sqrt{\frac{n}{4n} - O(n^{-1/2})} = 1 - \sqrt{\frac{1}{4}} - O(n^{-1/2}) = \frac{1}{2} - O(n^{-1/2})$$

as desired. □

Let us briefly sketch why Proposition 3.5.1 is the "right" answer using the heuristics in Section 3.2.2. For an unknown parameter θ in the exponential family model P_θ, we observe $X_1, \ldots, X_n$, and wish to test whether $v^\top \theta \ge t$ for a given threshold t. Call the null $H_0 : v^\top \theta \le t$, and assume we wish to test at an asymptotic level $\alpha > 0$, meaning that the probability of falsely rejecting H_0 is asymptotically at most α. Assuming the heuristic (3.2.6), we have the approximate distributional equality

$$v^\top \widehat{\theta}_n \stackrel{\cdot}{\sim} \mathsf{N}\left(v^\top \theta, \frac{1}{n} v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v\right).$$

Note that we have $\widehat{\theta}_n$ on the right side of the distribution; it is possible to make this rigorous, but here we target only intuition building. A natural asymptotically level α test is then

$$T_n := \begin{cases} \text{Reject} & \text{if } v^\top \widehat{\theta}_n \ge t + z_{1-\alpha} \sqrt{v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v / n} \\ \text{Accept} & \text{otherwise,} \end{cases}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal, $\mathbb{P}(Z \ge z_{1-\alpha}) = \alpha$ for $Z \sim \mathsf{N}(0,1)$. Let θ_0 be such that $v^\top \theta_0 = t$, so H_0 holds. Then

$$P_{\theta_0}(T_n \text{ rejects}) = P_{\theta_0}\left(\sqrt{n} \cdot v^\top (\widehat{\theta}_n - \theta_0) \ge z_{1-\alpha} \sqrt{v^\top \nabla^2 A(\widehat{\theta}_n)^{-1} v}\right) \to \alpha.$$

At least heuristically, then, this separation $\delta = \sqrt{v^\top \nabla^2 A(\theta_0)^{-1} v}/\sqrt{n}$ is the fundamental separation in parameter values at which testing becomes possible (or below which it is impossible).

As a brief and suggestive aside, the precise growth of the KL-divergence $D_{\rm kl}\left(P_{\theta_0 + \Delta} \| P_{\theta_0}\right) = \frac{1}{2}\Delta^\top \nabla^2 A(\theta_0)\Delta + O(\|\Delta\|^3)$ near θ_0 plays the fundamental role in both the lower bound and upper bound on testing. When the Hessian $\nabla^2 A(\theta_0)$ is "large," meaning it is very positive definite, distributions with small parameter distances are still well-separated in KL-divergence, making testing easy, while when $\nabla^2 A(\theta_0)$ is small (nearly singular), the KL-divergence can be small even for large parameter separations Δ and testing is hard. As a consequence, at least for exponential family models, the Fisher information (3.3.2), which we defined as $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X))$, plays a central role in testing and, as we see later, estimation.

3.6* Deferred proofs

3.6.1 Proof of Proposition 3.2.2

We follow Brown [90]. We demonstrate only the first-order differentiability using Lebesgue's dominated convergence theorem, as higher orders and the interchange of integration and differentiation are essentially identical, and we assume without loss of generality (by working with a push-forward measure if necessary) that $\phi(x) = x$. Demonstrating first-order complex differentiability is enough to show analyticity of M.[1] By Proposition 3.2.1, A is convex and $\Theta = \operatorname{dom} A \subset \mathbb{R}^d$ is a convex set. We assume Θ has non-empty interior (if the interior is empty, then the convexity of Θ means that it must lie in a lower dimensional subspace (Proposition B.1.12); we simply take the interior relative to that subspace and may proceed). We claim the following lemma, which is the key to applying dominated convergence; we state it first for $\mathbb{R}^d$.

Lemma 3.6.1. *Consider a collection $\{\theta_i\} = \{\theta_1, \ldots, \theta_m\} \subset \Theta$. Let $\Theta_0 = \operatorname{Conv}\{\theta_i\}_{i=1}^m$ and $C \subset \operatorname{int} \Theta_0$ be closed. Then for any $k \in \mathbb{N}$, there exists a constant $K = K(C, k, \{\theta_i\})$ such that for all $\theta_0 \in C$,*

$$\|x\|^k \exp(\langle \theta_0, x \rangle) \le K \max_{j \le m} \exp(\langle \theta_j, x \rangle).$$

Proof Let $\mathbb{B} = \{u \in \mathbb{R}^d \mid \|u\| \le 1\}$ be the unit ball in $\mathbb{R}^d$. For any $\epsilon > 0$, there exists a $K = K(\epsilon)$ such that $\|x\|^k \le K e^{\epsilon \|x\|}$ for all $x \in \mathbb{R}^d$. As $C \subset \operatorname{int} \Theta_0$ is closed (and hence compact), there exists an $\epsilon > 0$ such that for all $\theta_0 \in C$, $\theta_0 + \epsilon \mathbb{B} \subset \Theta_0$, and by construction, for any $u \in \mathbb{B}$ we can write $\theta_0 + \epsilon u = \sum_{j=1}^m \lambda_j \theta_j$ for some $\lambda = \lambda(u) \in \mathbb{R}_+^m$ with $\langle \mathbf{1}, \lambda \rangle = 1$. We therefore have

$$\begin{aligned} \|x\|^k \exp(\langle \theta_0, x \rangle) \le K \exp\left(\langle \theta_0, x \rangle + \epsilon \|x\|\right) &= K \sup_{u \in \mathbb{B}} \exp(\langle \theta_0 + \epsilon u, x \rangle) \\ &\le K \sup_{\lambda \in \mathbb{R}_+^m, \langle \mathbf{1}, \lambda \rangle = 1} \exp\left(\sum_{j=1}^m \lambda_j \langle \theta_j, x \rangle\right). \end{aligned}$$

The convexity of $\exp(\cdot)$ shows $\exp(\sum_j \lambda_j \langle \theta_j, x \rangle) \le \sum_j \lambda_j e^{\langle \theta_j, x \rangle} \le \max_j e^{\langle \theta_j, x \rangle}$. □

A similar result is possible with differences of exponentials:

Lemma 3.6.2. *Under the conditions of Lemma 3.6.1, there exists a K such that for any distinct $\theta, \theta_0 \in C$*

$$\left| \frac{e^{\langle \theta, x \rangle} - e^{\langle \theta_0, x \rangle}}{\|\theta - \theta_0\|} \right| \le K \max_{j \le m} e^{\langle \theta_j, x \rangle}.$$

Proof We write

$$\frac{\exp(\langle \theta, x \rangle) - \exp(\langle \theta_0, x \rangle)}{\|\theta - \theta_0\|} = \frac{\exp(\langle \theta - \theta_0, x \rangle) - 1}{\|\theta - \theta_0\|} \exp(\langle \theta_0, x \rangle)$$

so that the lemma is equivalent to showing that

$$\frac{|e^{\langle \theta - \theta_0, x \rangle} - 1|}{\|\theta - \theta_0\|} \le K \max_{j \le m} \exp(\langle \theta_j - \theta_0, x \rangle).$$

[1] For complex functions, Osgood's lemma shows that if a function is continuous and holomorphic in each variable individually, it is holomorphic. For a treatment of such ideas in an engineering context, see, e.g. [209, Ch. 1].

We can assume without loss of generality that $\theta_0 = \mathbf{0}$ by a translation. By convexity $e^{-a} \geq 1 - a$ for all $a \in \mathbb{R}$, so $1 - e^a \leq |a|$ when $a \leq 0$. Conversely, if $a > 0$, then $ae^a \geq e^a - 1$ (note that $\frac{\partial}{\partial a}(ae^a) = ae^a + e^a \geq e^a$), so dividing by $\|x\|$, we see that

$$\frac{|e^{\langle \theta, x\rangle} - 1|}{\|\theta\| \, \|x\|} \leq \frac{|e^{\langle \theta, x\rangle} - 1|}{|\langle \theta, x\rangle|} \leq \frac{\max\{\langle \theta, x\rangle e^{\langle \theta, x\rangle}, |\langle \theta, x\rangle|\}}{|\langle \theta, x\rangle|} \leq e^{\langle \theta, x\rangle} + 1.$$

As $\theta \in C$, Lemma 3.6.1 then implies that

$$\frac{|e^{\langle \theta, x\rangle} - 1|}{\|\theta\|} \leq \|x\| \left(e^{\langle \theta, x\rangle} + 1\right) \leq K \max_j e^{\langle \theta_j, x\rangle},$$

as desired. □

With the lemmas in hand, we can demonstrate a dominating function for the derivatives. Fix $\theta_0 \in \operatorname{int} \Theta$ and for $\theta \in \Theta$, define

$$g(\theta, x) = \frac{\exp(\langle \theta, x\rangle) - \exp(\langle \theta_0, x\rangle) - \exp(\langle \theta_0, x\rangle)\langle x, \theta - \theta_0\rangle}{\|\theta - \theta_0\|} = \frac{e^{\langle \theta, x\rangle} - e^{\langle \theta_0, x\rangle} - \langle \nabla_\theta e^{\langle \theta_0, x\rangle}, \theta - \theta_0\rangle}{\|\theta - \theta_0\|}.$$

Then $\lim_{\theta \to \theta_0} g(\theta, x) = 0$ by the differentiability of $t \mapsto e^t$. Lemmas 3.6.1 and 3.6.2 show that if we take any collection $\{\theta_j\}_{j=1}^m \subset \Theta$ for which $\theta \in \operatorname{int} \operatorname{Conv}\{\theta_j\}$, then for closed $C \subset \operatorname{int} \operatorname{Conv}\{\theta_j\}$, there exists a constant K such that

$$|g(\theta, x)| \leq \frac{|\exp(\langle \theta, x\rangle) - \exp(\langle \theta_0, x\rangle)|}{\|\theta - \theta_0\|} + \|x\| \exp(\langle \theta_0, x\rangle) \leq K \max_j \exp(\langle \theta_j, x\rangle)$$

for all $\theta \in C$. As $\int \max_j e^{\langle \theta_j, x\rangle} d\mu(x) \leq \sum_{j=1}^m \int e^{\langle \theta_j, x\rangle} d\mu(x) < \infty$, the dominated convergence theorem thus implies that

$$\lim_{\theta \to \theta_0} \int g(\theta, x) d\mu(x) = 0,$$

and so $M(\theta) = \exp(A(\theta))$ is differentiable in θ, as

$$M(\theta) = M(\theta_0) + \left\langle \int x e^{\langle \theta_0, x\rangle} d\mu(x), \theta - \theta_0 \right\rangle + o(\|\theta - \theta_0\|).$$

It is evident that we have the derivative

$$\nabla M(\theta) = \int \nabla \exp(\langle \theta, x\rangle) d\mu(x).$$

Analyticity Over the subset $\Theta_{\mathbb{C}} := \{\theta + iz \mid \theta \in \operatorname{int} \Theta, z \in \mathbb{R}^d\}$ (where $i = \sqrt{-1}$ is the imaginary unit), we can extend the preceding results to demonstrate that M is analytic on $\Theta_{\mathbb{C}}$. We first simply note that for $a, b \in \mathbb{R}$, $\exp(a + ib) = \exp(a)\exp(ib)$ and $|\exp(a + ib)| = \exp(a)$, i.e. $|e^z| = e^{\operatorname{Re} z}$ for $z \in \mathbb{C}$, and so Lemmas 3.6.1 and 3.6.2 follow *mutatis-mutandis* as in the real case. These are enough for the application of the dominated convergence theorem above, and we use that $\exp(\cdot)$ is analytic to conclude that $\theta \mapsto M(\theta)$ is analytic on $\Theta_{\mathbb{C}}$.

3.7 Bibliography and further reading

Exponential family models have been a bedrock of statistical learning for several decades, and accordingly, there are many excellent textbooks and other treatments of the material whose surface we only scratch here. Brown [90] gives a comprehensive and deep treatment of exponential family models and their analytic properties, and our theoretical results typically follow from his. Efron [178] gives an elegant treatment that combines the theory and practice of exponential families, showing how their theory impacts practical applications in current statistical methodology, including in survival analysis, controlling false discoveries, and missing data. Wainwright and Jordan [423] develop many of the connections between exponential families, graphical models—probabilistic models of processes on networks and graphs—and convex analysis. Their treatment also identifies the centrality of many notions of entropy and information, via convex duality, to the theory of exponential families; we will revisit this in Part III of this book.

Our treatment of estimation and learning in exponential families in Section 3.2.2 and for generalized linear models in Section 3.4.1 was brief and, essentially, heuristic. Estimation in these models admits supremely clean calculations and asymptotics, and therefore often forms the introductory material in any class on asymptotics and statistical estimation. A particular favorite is van der Vaart [407], who, in addition to providing a bevy of advanced material on asymptotic convergence of estimators, gives a particularly nice analysis of moment-matching estimators [407, Ch. 4].

3.8 Exercises

Exercise 3.1: In Example 3.1.4, give the sufficient statistic ϕ and an explicit formula for the log partition function $A(\theta, K)$ so that we can write $p_{\theta,K}(x) = \exp(\langle\theta, \phi_1(x)\rangle + \langle K, \phi_2(x)\rangle - A(\theta, K))$.

Exercise 3.2: Consider the binary logistic regression model in Example 3.4.2, and let $\ell(\theta; x, y) = -\log p_\theta(y \mid x)$ be the associated log loss.

(i) Give the Hessian $\nabla^2_\theta \ell(\theta; x, y)$.

(ii) Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ be a sample. Give a sufficient condition for the minimizer of the empirical log loss

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i, y_i)$$

to be unique that depends only on the vectors $\{x_i\}$. *Hint.* A convex function h is strictly convex if and only if its Hessian $\nabla^2 h$ is positive definite.

Exercise 3.3 (Bounded range regression as an exponential family model)**:** Suppose that we know $Y \in [-b, b]$ and wish to model it via an exponential family model with density

$$p_\theta(y \mid x) = \exp(yx^\top \theta - A(\theta \mid x)) \mathbf{1}\left\{y \in [-b, b]\right\},$$

which is non-zero only for $-b \le y \le b$.

(a) Give the log partition $A(\theta \mid x)$.

(b) Compute $\mathbb{E}[Y \mid x^\top \theta = s]$.

Exercise 3.4: Give the Fisher information (3.4.2) for the following generalized linear models:

(a) Linear regression (Example 3.4.1).

(b) Poisson regression (Example 3.4.4).

Exercise 3.5: Let $\{P_\theta\}_{\theta \in \mathbb{R}^d}$ be an exponential family model with log partition $A(\theta)$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. For $h > 0$, consider the sequence of hypothesis tests

$$H_0 : f(\theta) \le 0 \quad \text{versus} \quad H_{1,n} : f(\theta) \ge \frac{1}{\sqrt{n}} \cdot h. \tag{3.8.1}$$

(a) Let θ_0 be a parameter satisfying $f(\theta_0) = 0$ and $\nabla^2 A(\theta_0) \succ 0$. Use the Bretagnolle-Huber inequality to show that there exists a sequence of the parameters of the form $\theta_n = \theta_0 + v/\sqrt{n}$ such that

$$\liminf_n \inf_\Psi \mathbb{P}_{H_0}(\Psi(X_1^n) \neq 0) + \mathbb{P}_{H_1}(\Psi(X_1^n) \neq 1) \ge 1 - \sqrt{1 - \exp(-c^2/2)},$$

where

$$\sqrt{n} f(\theta_n) \ge c \sqrt{\nabla f(\theta_0)^\top \nabla^2 A(\theta_0)^{-1} \nabla f(\theta_0)} + o(1).$$

That is, the scaling $h^2 = \nabla f(\theta_0)^\top \nabla^2 A(\theta_0)^{-1} \nabla f(\theta_0)$ provides one threshold where testing (3.8.1) is asymptotically (nearly) impossible.

(b) Interpret the result of part (a).

Exercise 3.6* (Achievability in a Gaussian test)**:** This exercise complements Exercise 3.5 to demonstrate achievability, but it presumes familiarity with asymptotics in statistics. Assume the conditions of Exercise 3.5 that $\{P_\theta\}$ is an exponential family and f is continuously differentiable.

(a) Fix $\theta^\star$ and let $X_i \stackrel{\text{iid}}{\sim} P_{\theta^\star}$ be an i.i.d. sample of size n, and assume that the maximum likelihood estimator $\widehat{\theta}_n$ satisfies the approximation (3.2.6) uniformly near $\theta^\star$, that is,

$$\sqrt{n}\left(\widehat{\theta}_n - (\theta^\star + v/\sqrt{n})\right) \underset{\theta^\star + v/\sqrt{n}}{\overset{d}{\rightsquigarrow}} \mathsf{N}\left(0, \nabla^2 A(\theta^\star)^{-1}\right),$$

where $\overset{d}{\rightsquigarrow}_{\theta_n}$ denotes convergence in distribution under the distributions P_{θ_n}. Define the test

$$\Psi_n := \begin{cases} \text{Reject } H_0 & \text{if } f(\widehat{\theta}_n) \ge c/\sqrt{n} \\ \text{Accept } H_0 & \text{if } f(\widehat{\theta}_n) \le c/\sqrt{n} \end{cases}$$

and $\sigma_f^2(\theta) := \nabla f(\theta)^\top \nabla^2 A(\theta)^{-1} \nabla f(\theta)$. Show that if $f(\theta^\star) \le 0$, then

$$\limsup_n P_{\theta^\star}(\Psi_n \text{ rejects}) \le \Phi\left(-c/\sigma_f(\theta^\star)\right),$$

while assuming $f(\theta^\star) = 0$, then under the sequence of alternatives $P_{\theta^\star + v/\sqrt{n}}$,

$$\liminf_n P_{\theta^\star + v/\sqrt{n}}(\Psi_n \text{ accepts}) = \Phi\left(\frac{c - \nabla f(\theta^\star)^\top v}{\sigma_f(\theta^\star)}\right),$$

where Φ is the standard Gaussian CDF.

(b) Explain why the preceding result implies that the scaling in part (a) is the "right" behavior, i.e., that the fundamental complexity of testing H_0 against $H_{1,n}$ reposes on the scale of $f(\theta)$ relative to

$$\frac{1}{n} \nabla f(\theta)^\top \nabla^2 A(\theta)^{-1} \nabla f(\theta).$$

Part I

Concentration, information, stability, and generalization

Précis

The ways in which distributions become close to one another forms the thread connecting the different parts of this book. In the context of Part I, we leverage this perspective first to show ways in which samples from a population behave similarly to the entire population. Letting $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i}$ denote the empirical distribution of a sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P$ from a population P, where $\mathbf{1}_{X_i}$ denotes a point mass at X_i, we would like to know how close P_n is to P. As a simple example, we typically expect the sample mean $\mathbb{E}_{P_n}[X] = \frac{1}{n}\sum_{i=1}^n X_i$ to be close to the population mean $\mathbb{E}_P[X]$, as the laws of large numbers and central limit theorem show. Concentration inequalities can guarantee this in more quantitative ways, and information-theoretic techniques will allow us to extend this closeness to more sophisticated phenomena, demonstrating and leveraging different types of stability of estimators. For example, we will show how machine-learned models, based on a sample P_n, generalize to the full population, and show how random matrices concentrate around their expectations. We culminate the section by showing how to leverage (information-theoretic) closeness of distributions P_n and P_n' to guarantee different desiderata for estimators, including limiting false discoveries in data analysis and privacy. "Stability" provides the principle that unifies each of the concepts, and throughout this part will mean that changing the input sample to a procedure induces at most slight changes in the relevant output, where the metrics for these changes varies with the problem.

Chapter 4

Concentration inequalities

Understanding how a random variable X behaves—by bounding the probability it deviates far from its expected behavior—provides a fundamental way to demonstrate the performance of many data-based procedures. By showing that a random variable is typically near its expectation, for example, this will allow us to prove that estimation and learning procedures have performance guarantees or that different decoding and encoding schemes work with high probability. In this chapter, we develop tools for controlling these deviations. We begin with classical tail inequalities, then introduce sub-Gaussian and sub-exponential random variables and their basic equivalences. We next develop martingale inequalities and conclude with matrix concentration results and applications.

4.1 Basic tail inequalities

We begin with a simple goal: given a random variable X, how does X concentrate around its mean? That is, assuming without loss of generality that $\mathbb{E}[X] = 0$, how well can we bound

$$\mathbb{P}(X \geq t) \quad \text{and} \quad \mathbb{P}(X \leq -t)$$

for $t \geq 0$? We begin with the three most classical inequalities for this purpose: the Markov, Chebyshev, and Chernoff bounds, which are all instances of the same technique.

The basic inequality on which all others build is Markov's inequality.

Proposition 4.1.1 (Markov's inequality)**.** *Let X be a nonnegative random variable, meaning that $X \geq 0$ with probability 1. Then*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof For any nonnegative random variable,

$$\mathbb{P}(X \geq t) = \mathbb{E}[\mathbf{1}\left\{X \geq t\right\}] \leq \mathbb{E}[(X/t)\mathbf{1}\left\{X \geq t\right\}] \leq \mathbb{E}[X]/t,$$

as $X/t \geq 1$ whenever $X \geq t$. □

It is possible to tighten this (evocatively, to Markov's equality) via a randomization argument, which Exercise 4.1 details.

When we know more about a random variable than that its expectation is finite, we can give more powerful bounds on the probability that the random variable deviates from its typical values. As we shall see, each of these results is but an application of Proposition 4.1.1. The first step in this direction, Chebyshev's inequality, requires two moments:

Proposition 4.1.2 (Chebyshev's inequality)**.** *Let X be a random variable with $\mathrm{Var}(X) < \infty$. Then*

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \frac{\mathrm{Var}(X)}{t^2} \quad \text{and} \quad \mathbb{P}(X - \mathbb{E}[X] \leq -t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

for all $t \geq 0$.

Proof We prove only the upper tail result, as the lower tail is identical. We first note that $X - \mathbb{E}[X] \geq t$ implies that $(X - \mathbb{E}[X])^2 \geq t^2$. The random variable $Z = (X - \mathbb{E}[X])^2$ is non-negative, so Markov's inequality gives $\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \mathbb{P}(Z \geq t^2) \leq \mathbb{E}[Z]/t^2$, and $\mathbb{E}[Z] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$. □

Chebyshev's inequality, combined with our first tensorization argument for sums, shows that we expect the typical fluctuations of a sum of identically distributed random variables with variance σ^2 to be roughly not much larger than $\sigma/\sqrt{n}$. Indeed, letting $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the mean of independent random variables X_i with $\mathrm{Var}(X_i) = \sigma^2$,

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \quad \text{so} \quad \mathbb{P}(\overline{X}_n \geq \mathbb{E}[\overline{X}_n] + t) \leq \frac{\sigma^2}{nt^2}.$$

That is, typical fluctuations should rarely be much more than $t \approx \frac{\sigma}{\sqrt{n}}$.

If a random variable has a moment generating function—exponential moments—we can give bounds that elegantly extend to sums of random variables. First, we recall that

$$\varphi_X(\lambda) := \mathbb{E}[e^{\lambda X}]$$

is the moment generating function of the random variable X. Then we have the Chernoff bound.

Proposition 4.1.3. *For any random variable X and $t \geq 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} = \varphi_X(\lambda) e^{-\lambda t}$$

for all $\lambda \geq 0$.

Proof This is another application of Markov's inequality: for $\lambda > 0$, we have $e^{\lambda X} \geq e^{\lambda t}$ if and only if $X \geq t$, so that $\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda X}]/e^{\lambda t}$. □

The freedom in the bound—that it holds simultaneously for all $\lambda \geq 0$—makes optimizing over the right-hand side natural. Thus, taking the infimum over all $\lambda \geq 0$ in Proposition 4.1.3 gives the more standard Chernoff (large deviation) bound

$$\mathbb{P}(X \geq t) \leq \exp\left(\inf_{\lambda \geq 0} \log \varphi_X(\lambda) - \lambda t\right).$$

Example 4.1.4 (Gaussian random variables)**:** When X is a mean-zero Gaussian variable with variance σ^2, we have

$$\varphi_X(\lambda) = \mathbb{E}[\exp(\lambda X)] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{4.1.1}$$

To see this, we compute the integral; we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\lambda x - \frac{1}{2\sigma^2}x^2\right) dx \\ &= e^{\frac{\lambda^2\sigma^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\lambda\sigma^2)^2\right) dx}_{=1},\end{aligned}$$

because this is simply the integral of the Gaussian density.

As a consequence of the equality (4.1.1) and the Chernoff bound technique (Proposition 4.1.3), we see that for X Gaussian with variance σ^2, we have

$$\mathbb{P}(X \geq \mathbb{E}[X] + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right) \quad \text{and} \quad \mathbb{P}(X \leq \mathbb{E}[X] - t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

for all $t \geq 0$. Indeed, we have $\log \varphi_{X-\mathbb{E}[X]}(\lambda) = \frac{\lambda^2\sigma^2}{2}$, and $\inf_\lambda\{\frac{\lambda^2\sigma^2}{2} - \lambda t\} = -\frac{t^2}{2\sigma^2}$, which is attained by $\lambda = \frac{t}{\sigma^2}$. $\diamond$

4.1.1 Sub-Gaussian random variables

The convenience of the quadratic moment generating function of the Gaussian suggests a class of random variables whose deviations we might control with similar ease, which leads us to develop *sub-Gaussian* random variables.

Definition 4.1. *A random variable X is* sub-Gaussian with parameter σ^2 *if*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$. We also say such a random variable is σ^2-sub-Gaussian.

Gaussian random variables satisfy Definition 4.1 with equality (and, because σ^2 is a bound, the inequality holds for all parameters $\tau^2 \geq \sigma^2$ when $X \sim \mathsf{N}(0, \sigma^2)$). Happily, many other random variables are also sub-Gaussian, and we detail several examples.

Example 4.1.5 (Random signs, i.e., Rademacher variables)**:** The random variable X uniform on $\{-1, 1\}$ is 1-sub-Gaussian. Indeed,

$$\mathbb{E}[\exp(\lambda X)] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \frac{1}{2}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{(\lambda^2)^k}{2^k k!} = \exp\left(\frac{\lambda^2}{2}\right),$$

as claimed. $\diamond$

Bounded random variables are also sub-Gaussian:

Example 4.1.6 (Bounded random variables and Hoeffding's lemma)**:** Suppose that for some $a \leq b$, X satisfies $X \in [a, b]$. Then Hoeffding's lemma states that

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right), \tag{4.1.2}$$

so that X is $(b-a)^2/4$-sub-Gaussian. $\diamond$

We give two arguments to prove the inequality in Example 4.1.6, both of which preview important techniques. The first gives a somewhat weaker statement via a *symmetrization* argument, where we replace X with a symmetrized version of itself, a common technique throughout mathematics. Let $\varepsilon \in \{-1, 1\}$ be a Rademacher variable, so that $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$, and let X' be an independent copy of X, so that $\mathbb{E}[X'] = \mathbb{E}[X]$. We have

$$\begin{aligned}\varphi_{X-\mathbb{E}[X]}(\lambda) &= \mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X']))\right] \le \mathbb{E}\left[\exp(\lambda(X - X'))\right] \\ &= \mathbb{E}\left[\exp(\lambda\varepsilon(X - X'))\right],\end{aligned}$$

where the inequality follows from Jensen's inequality and the last equality is a consequence of the symmetry of $X - X'$ about 0. Using the result of Example 4.1.5,

$$\mathbb{E}\left[\exp(\lambda\varepsilon(X - X'))\right] \le \mathbb{E}\left[\exp\left(\frac{\lambda^2(X - X')^2}{2}\right)\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{2}\right),$$

where the final inequality is immediate from the fact that $|X - X'| \le b - a$.

A sharper technique explicitly bounds the logarithm of the moment generating function of X via *exponential tilts* of its density. In particular, letting X be mean zero for simplicity, if we let

$$\psi(\lambda) = \log \varphi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}],$$

then

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \quad \text{and} \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[Xe^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2},$$

where we can interchange the order of taking expectations and derivatives whenever $\psi(\lambda)$ is finite in a neighborhood of 0 (this follows via Proposition 3.2.2, which also shows that $\varphi_X(\lambda)$ is analytic and $\psi(\lambda)$ is convex). Notably, if X has density p then the random variable Y_λ with density

$$p_\lambda(y) := \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda X}]}p(y)$$

satisfies

$$\psi'(\lambda) = \mathbb{E}[Y_\lambda] \quad \text{and} \quad \psi''(\lambda) = \mathbb{E}[Y_\lambda^2] - \mathbb{E}[Y_\lambda]^2 = \mathrm{Var}(Y_\lambda).$$

One can exploit this in many ways, which the exercises and coming chapters do.

As a first example, we demonstrate equality (4.1.2). When $X \in [a, b]$, the tilted random variable Y_λ still satisfies $Y_\lambda \in [a, b]$. The variance of *any* bounded random variable $B \in [a, b]$ satisfies

$$\mathrm{Var}(B) = \mathbb{E}[(B - \mathbb{E}[B])^2] = \inf_t \mathbb{E}[(B - t)^2] \le \mathbb{E}\left[\left(B - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4},$$

so $\mathrm{Var}(Y_\lambda) = \psi''(\lambda) \le \frac{(b-a)^2}{4}$. Integrating, we have

$$\psi'(0) = 0 \quad \text{and} \quad \psi'(\lambda) \le \frac{(b-a)^2}{4}\lambda,$$

so

$$\log \varphi_X(\lambda) = \psi(\lambda) = \int_0^\lambda \psi'(u)du \le \int_0^\lambda \frac{(b-a)^2}{4}udu = \frac{\lambda^2(b-a)^2}{8}.$$

We can give sharper sub-Gaussian constants for Bernoulli random variables.

Example 4.1.7 (Bernoulli random variables)**:** Let $X \sim \mathsf{Bernoulli}(p)$, so that $X = 1$ with probability p and $X = 0$ otherwise. Then a strengthening of Hoeffding's lemma (also, essentially, due to Hoeffding) is that

$$\log \mathbb{E}[e^{\lambda(X-p)}] \le \frac{\sigma^2(p)}{2}\lambda^2 \quad \text{for} \quad \sigma^2(p) := \frac{1-2p}{2\log\frac{1-p}{p}}.$$

Here we take the limits as $p \to \{0, \frac{1}{2}, 1\}$ and have $\sigma^2(0) = 0$, $\sigma^2(1) = 0$, and $\sigma^2(\frac{1}{2}) = \frac{1}{4}$. Because $p \mapsto \sigma^2(p)$ is concave and symmetric about $p = \frac{1}{2}$ (see Exercise 4.12), this inequality is always sharper than that of Example 4.1.6. We defer derivation of this guarantee to Chapter 7, where Exercises 7.9 and 7.10 use a refined version of Pinsker's inequality to prove the result. $\diamond$

Chernoff bounds for sub-Gaussian random variables are immediate, as they have (more or less) the same concentration properties as Gaussian random variables, a consequence of the nice analytical properties of their moment generating functions (that their logarithms are at most quadratic). Thus, using the technique of Example 4.1.4, we obtain the following proposition.

Proposition 4.1.8. *Let X be a σ^2-sub-Gaussian random variable. Then for all $t \ge 0$,*

$$\mathbb{P}(X - \mathbb{E}[X] \ge t) \vee \mathbb{P}(X - \mathbb{E}[X] \le -t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Tensorization and sums of random variables. Chernoff bounds extend naturally to sums of independent random variables, because moment generating functions of sums of independent random variables become products of moment generating functions.

Proposition 4.1.9. *Let $X_1, X_2, \ldots, X_n$ be independent σ_i^2-sub-Gaussian random variables. Then*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)\right] \le \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right) \quad \textit{for all } \lambda \in \mathbb{R},$$

that is, $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian.

Proof We assume w.l.o.g. that the X_i are mean zero. We have by independence and sub-Gaussianity that

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\right]\mathbb{E}[\exp(\lambda X_n)] \le \exp\left(\frac{\lambda^2\sigma_n^2}{2}\right)\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\right].$$

Applying this technique inductively to $X_{n-1}, \ldots, X_1$, we obtain the desired result. $\square$

Propositions 4.1.8 and 4.1.9 show that sums of sub-Gaussian random variables concentrate around their expectations, as in the following concentration inequality.

Corollary 4.1.10. *Let X_i be independent σ_i^2-sub-Gaussian random variables. Then for all $t \ge 0$*

$$\max\left\{\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right), \mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \le -t\right)\right\} \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

The classical Hoeffding bound follows when we couple Example 4.1.6 with Corollary 4.1.10: if $X_i \in [a_i, b_i]$, then $\sum_{i=1}^n X_i$ is $\frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2$-sub-Gaussian, so

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

To give another interpretation of these inequalities, assume that X_i are independent and σ^2-sub-Gaussian, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Then

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2\sigma^2}\right) \quad \text{so} \quad \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \le \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}}\right) \ge 1 - \delta$$

for $\delta \in (0, 1)$ by taking $t = \sqrt{(2\sigma^2/n)\log(2/\delta)}$.

The next theorem demonstrates several equivalent formulations of sub-Gaussianity.

Theorem 4.1.11. *Let X be a random variable and $\sigma^2 \ge 0$. The following are all equivalent, meaning that there are numerical constant factors K_j such that if statement (i) holds with parameter K_i, then statement (j) holds with parameter $K_j \le CK_i$, where C is a numerical constant.*

(1) Sub-Gaussian tails: $\mathbb{P}(|X| \ge t) \le 2\exp(-\frac{t^2}{K_1\sigma^2})$ for all $t \ge 0$.

(2) Sub-Gaussian moments: $\mathbb{E}[|X|^k]^{1/k} \le K_2\sigma\sqrt{k}$ for all k.

(3) Super-exponential moment: $\mathbb{E}[\exp(X^2/(K_3\sigma^2))] \le e$.

If in addition X is mean zero, each of these is equivalent to

(4) Sub-Gaussian moment generating function: $\mathbb{E}[\exp(\lambda X)] \le \exp(K_4\lambda^2\sigma^2)$ for all $\lambda \in \mathbb{R}$.

This result is standard in the literature on concentration and random variables; we defer its proof to the technical Section 4.4.1.

For completeness, we can give a tighter result than part (3) of the preceding theorem, giving a concrete upper bound on squares of sub-Gaussian random variables. The technique used in the example, to introduce an independent random variable for auxiliary randomization, is a common and useful technique in probabilistic arguments (similar to our use of symmetrization in Example 4.1.6).

Example 4.1.12 (Sub-Gaussian squares)**:** Let X be a mean-zero σ^2-sub-Gaussian random variable. Then for $\lambda \ge 0$,

$$\mathbb{E}[\exp(\lambda X^2)] \le \frac{1}{[1 - 2\sigma^2\lambda]_+^{\frac{1}{2}}}, \tag{4.1.3}$$

and expression (4.1.3) holds with equality for $X \sim \mathsf{N}(0, \sigma^2)$.

To see this result, we focus on the Gaussian case first and assume (for this case) without loss of generality (by scaling) that $\sigma^2 = 1$. Assuming that $\lambda < \frac{1}{2}$, we have

$$\mathbb{E}[\exp(\lambda Z^2)] = \int \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2} - \lambda)z^2} dz = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1-2\lambda}{2}z^2} dz = \frac{\sqrt{2\pi}}{\sqrt{1-2\lambda}} \frac{1}{\sqrt{2\pi}},$$

the final equality a consequence of the fact that $\int e^{-\frac{1}{2\sigma^2}z^2} dz = \sqrt{2\pi\sigma^2}$. When $\lambda \ge \frac{1}{2}$, the above integrals are all infinite, giving the equality in expression (4.1.3).

For the more general inequality, we recall that if Z is a $\mathsf{N}(0,1)$ random variable, then $\mathbb{E}[\exp(tZ)] = \exp(\frac{t^2}{2})$. So if $Z \sim \mathsf{N}(0,1)$ is independent of X,

$$\mathbb{E}[\exp(\lambda X^2)] = \mathbb{E}[\exp(\sqrt{2\lambda}XZ)] = \mathbb{E}\left[\mathbb{E}[\exp(\sqrt{2\lambda}XZ) \mid Z]\right] \stackrel{(i)}{\leq} \mathbb{E}\left[\exp(\lambda\sigma^2 Z^2)\right] \stackrel{(ii)}{=} \frac{1}{[1-2\sigma^2\lambda]_+^{\frac{1}{2}}},$$

where inequality (i) follows because X is sub-Gaussian, and (ii) because $Z \sim \mathsf{N}(0,1)$. $\diamond$

4.1.2 Sub-exponential random variables

In many cases, random variables may have finite moment generating function only in a neighborhood of 0, or we may wish to obtain sharper control of the moment generating near 0 than is globally possible. This leads to us to the slightly weaker condition that a random variable is *sub-exponential*, which—for a mean-zero random variable—means that its moment generating function exists in a neighborhood of zero.

Definition 4.2. *A random variable X is* sub-exponential *with parameters (τ^2, b) if for $|\lambda| \leq 1/b$,*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\frac{\lambda^2\tau^2}{2}\right).$$

It is clear from Definition 4.2 that a σ^2-sub-Gaussian random variable is $(\sigma^2, 0)$-sub-exponential.

A variety of random variables are sub-exponential. As a first example, χ^2-random variables are sub-exponential with constant values for τ and b:

Example 4.1.13: Let $X = Z^2$, where $Z \sim \mathsf{N}(0,1)$. We claim that

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(2\lambda^2) \text{ for } \lambda \leq \frac{1}{4}. \tag{4.1.4}$$

Indeed, for $\lambda < \frac{1}{2}$ we have (recall Example 4.1.12) that

$$\mathbb{E}[\exp(\lambda(Z^2 - \mathbb{E}[Z^2]))] = \exp\left(-\frac{1}{2}\log(1-2\lambda) - \lambda\right) \stackrel{(\star)}{\leq} \exp\left(\lambda + 2\lambda^2 - \lambda\right)$$

where inequality $(\star)$ holds because $-\log(1-2\lambda) \leq 2\lambda + 4\lambda^2$ for $\lambda \leq \frac{1}{4}$. $\diamond$

As a second example, we can show that bounded random variables are sub-exponential. As they are also sub-Gaussian, this is somewhat trivial; however, this shows some of the promise that we may obtain sharper control over the moment generating function in a neighborhood of 0, and we may in turn leverage that for sharper concentration inequalities than are possible using only sub-Gaussian techniques.

Example 4.1.14 (Bennett's inequality)**:** When $X \leq b$ and $\mathbb{E}[X] = 0$, we can provide one-sided bounds on the moment generating function for $\lambda \geq 0$ without assuming any lower tail. In this case, we claim the bound

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right).$$

To see this, we control the derivatives of the moment generating function $\varphi(\lambda) = \mathbb{E}[e^{\lambda X}]$, which is finite for all $\lambda \geq 0$. Then

$$\varphi'(\lambda) = \mathbb{E}[Xe^{\lambda X}] \;\text{ and }\; \varphi''(\lambda) = \mathbb{E}[X^2 e^{\lambda X}] \leq e^{\lambda b}\mathbb{E}[X^2] = e^{\lambda b}\sigma^2.$$

Computing a Taylor expansion with remainder, we have

$$\begin{aligned}\varphi(\lambda) &= \varphi(0) + \lambda\varphi'(0) + \int_0^\lambda \int_0^u \varphi''(s)\, ds\, du \\ &\leq 1 + 0 + \int_0^\lambda \int_0^u \sigma^2 e^{sb}\, ds\, du = 1 + \frac{\sigma^2}{b}\int_0^\lambda (e^{ub} - 1)du = 1 + \frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right),\end{aligned}$$

as desired. Then using that $1 + x \leq e^x$, we obtain *Bennett's moment generating inequality*,

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right)\right) \quad \text{for } \lambda \geq 0. \tag{4.1.5}$$

Inequality (4.1.5) always holds, and for λb near 0, we have $e^{\lambda b} - 1 - \lambda b \approx \frac{\lambda^2 b^2}{2}$, yielding an (almost) sub-Gaussian bound. $\diamond$

Example 4.1.15 (Bounded random variables are sub-exponential)**:** Let X be a mean zero random variable taking values in $[-b, b]$ with variance $\sigma^2 = \mathbb{E}[X^2]$ (we are guaranteed that $\sigma^2 \leq b^2$ in this case). Applying Bennett's inequality (4.1.5) for $-X \leq b$, we obtain that

$$\max\left\{\mathbb{E}[\exp(\lambda X)], \mathbb{E}[\exp(-\lambda X)]\right\} \leq \exp\left(\frac{\sigma^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right)\right) \quad \text{for } \lambda \geq 0.$$

For $0 \leq t \leq \frac{1}{2}$, we have $e^t - 1 - t \leq \frac{3}{5}t^2$, so that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{3\lambda^2\sigma^2}{5}\right) \quad \text{for } |\lambda| \leq \frac{1}{2b}. \tag{4.1.6}$$

by canceling the b^2 factor. $\diamond$

In both of the inequalities (4.1.5) and (4.1.6), if the variance $\sigma^2 \ll b^2$, the absolute bound on X, we obtain much tighter control on the moment generating function of X than typical sub-Gaussian bounds using only the boundedness $X \in [-b, b]$ allow.

More broadly, we can show a result similar to Theorem 4.1.11.

Theorem 4.1.16. *Let X be a random variable and $\sigma \geq 0$. Then—in the sense of Theorem 4.1.11—the following statements are all equivalent for suitable numerical constants $K_1, \ldots, K_4$.*

(1) Sub-exponential tails: $\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t}{K_1\sigma})$ for all $t \geq 0$

(2) Sub-exponential moments: $\mathbb{E}[|X|^k]^{1/k} \leq K_2\sigma k$ for all $k \geq 1$.

(3) Existence of an exponential moment: $\mathbb{E}[\exp(|X|/(K_3\sigma))] \leq e$.

If in addition X is mean zero, each of these is equivalent to

(4) Sub-exponential moment generating function: $\mathbb{E}[\exp(\lambda X)] \leq \exp(K_4\lambda^2\sigma^2)$ for $|\lambda| \leq K_4'/\sigma$.

See Section 4.4.2 for the proof, which is similar to that for Theorem 4.1.11.

While the concentration properties of sub-exponential random variables are not quite so nice as those for sub-Gaussian random variables (recall Hoeffding's inequality, Corollary 4.1.10), we can give sharp tail bounds for sub-exponential random variables. We first give a simple "Bernstein-type" bound—so called because it involves both the variance and uniform bound—on deviation probabilities.

Proposition 4.1.17. *Let X be a mean-zero (τ^2, b)-sub-exponential random variable. Then for all $t \geq 0$,*

$$\mathbb{P}(X \geq t) \vee \mathbb{P}(X \leq -t) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\tau^2}, \frac{t}{b}\right\}\right).$$

Proof The proof is an application of the Chernoff bound technique; we prove only the upper tail as the lower tail is similar. We have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} \stackrel{(i)}{\leq} \exp\left(\frac{\lambda^2\tau^2}{2} - \lambda t\right),$$

inequality (i) holding for $|\lambda| \leq 1/b$. To minimize the last term in λ, we take $\lambda = \min\{\frac{t}{\tau^2}, 1/b\}$, which gives the result. □

Comparing with sub-Gaussian random variables, which have $b = 0$, we see that Proposition 4.1.17 gives a similar result for small t—essentially the same concentration sub-Gaussian random variables—while for large t, the tails decrease only exponentially in t.

Tensorization and sums of random variables. We can also give a tensorization identity similar to Proposition 4.1.9.

Proposition 4.1.18. *Let $X_1, \ldots, X_n$ be independent mean-zero sub-exponential random variables, where X_i is (σ_i^2, b_i)-sub-exponential. Then for any vector $a \in \mathbb{R}^n$, we have*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n a_i X_i\right)\right] \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n a_i^2\sigma_i^2}{2}\right) \quad \text{for} \quad |\lambda| \leq \frac{1}{b_*},$$

where $b_* = \max_i b_i|a_i|$. That is, $\langle a, X\rangle$ is $(\sum_{i=1}^n a_i^2\sigma_i^2, \max_i b_i|a_i|)$-sub-exponential.

Proof We apply an inductive technique similar to that used in the proof of Proposition 4.1.9. First, for any fixed i, we know that if $|\lambda| \leq \frac{1}{b_i|a_i|}$, then $|a_i\lambda| \leq \frac{1}{b_i}$ and so

$$\mathbb{E}[\exp(\lambda a_i X_i)] \leq \exp\left(\frac{\lambda^2 a_i^2\sigma_i^2}{2}\right).$$

We inductively apply the preceding inequality, which applies if $|\lambda| \leq \frac{1}{b_i|a_i|}$ for all i. We have

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n a_i X_i\right)\right] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda a_i X_i)] \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 a_i^2\sigma_i^2}{2}\right),$$

which is our desired result. □

As in the case of sub-Gaussian random variables, a combination of the tensorization property—that the moment generating functions of sums of sub-exponential random variables are well-behaved—of Proposition 4.1.18 and the concentration inequality (4.1.17) immediately yields the following Bernstein-type inequality. (See also Vershynin [412].)

Corollary 4.1.19. *Let $X_1, \ldots, X_n$ be independent mean-zero (σ_i^2, b_i)-sub-exponential random variables (Definition 4.2). Define $b_* := \max_i b_i$. Then for all $t \geq 0$ and all vectors $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \vee \mathbb{P}\left(\sum_{i=1}^n a_i X_i \leq -t\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\sum_{i=1}^n a_i^2\sigma_i^2}, \frac{t}{b_* \left\|a\right\|_\infty}\right\}\right).$$

It is instructive to study the structure of the bound of Corollary 4.1.19. Notably, the bound is similar to the Hoeffding-type bound of Corollary 4.1.10 (holding for σ^2-sub-Gaussian random variables) that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\left\|a\right\|_2^2 \sigma^2}\right),$$

so that for small t, Corollary 4.1.19 gives sub-Gaussian tail behavior. For large t, the bound is weaker. However, in many cases, Corollary 4.1.19 can give finer control than naive sub-Gaussian bounds. Indeed, suppose that the random variables X_i are i.i.d., mean zero, and satisfy $X_i \in [-b, b]$ with probability 1, but have variance $\sigma^2 = \mathbb{E}[X_i^2] \leq b^2$ as in Example 4.1.15. Then Corollary 4.1.19 implies that

$$\mathbb{P}\left(\sum_{i=1}^n a_i X_i \geq t\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{5}{6}\frac{t^2}{\sigma^2\left\|a\right\|_2^2}, \frac{t}{2b\left\|a\right\|_\infty}\right\}\right). \tag{4.1.7}$$

When applied to a standard mean (and with a minor simplification that $5/12 < 1/3$) with $a_i = \frac{1}{n}$, we obtain the bound that $\frac{1}{n}\sum_{i=1}^n X_i \leq t$ with probability at least $1 - \exp(-n\min\{\frac{t^2}{3\sigma^2}, \frac{t}{4b}\})$. Written differently, we take $t = \max\{\sigma\sqrt{\frac{3\log\frac{1}{\delta}}{n}}, \frac{4b\log\frac{1}{\delta}}{n}\}$ to obtain

$$\frac{1}{n}\sum_{i=1}^n X_i \leq \max\left\{\sigma\frac{\sqrt{3\log\frac{1}{\delta}}}{\sqrt{n}}, \frac{4b\log\frac{1}{\delta}}{n}\right\} \quad \text{with probability } 1 - \delta.$$

The sharpest such bound possible via more naive Hoeffding-type bounds is $b\sqrt{2\log\frac{1}{\delta}}/\sqrt{n}$, which has substantially worse scaling.

The exercises ask you to work out further variants of these results, including the sub-exponential behavior of quadratic forms of Gaussian random vectors (see Exercises 4.10 and 4.11).

Further conditions and examples

There are a number of examples and conditions sufficient for random variables to be sub-exponential. One common condition, the so-called *Bernstein* condition, controls the higher moments of a random variable X by its variance. In this case, we say that X satisfies the b-Bernstein condition if

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{k!}{2}\sigma^2 b^{k-2} \quad \text{for } k = 3, 4, \ldots, \tag{4.1.8}$$

where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2$. In this case, the following lemma controls the moment generating function of X. This result is essentially present in Theorem 4.1.16, but it provides somewhat tighter control with precise constants.

Lemma 4.1.20. *Let X be a random variable satisfying the Bernstein condition (4.1.8). Then*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right) \quad \text{for } |\lambda| \le \frac{1}{b}.$$

Thus a random variable satisfying Condition (4.1.8) is $(2\sigma^2, 2b)$-sub-exponential.

Proof Without loss of generality we assume $\mu = 0$. We expand the moment generating function by noting that

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}[X^k]}{k!} &\stackrel{(i)}{\le} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}|\lambda b|^{k-2} \\ &= 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{[1-b|\lambda|]_+}\end{aligned}$$

where inequality (i) uses the Bernstein condition (4.1.8), because $\lambda^k\mathbb{E}[X^k] \le |\lambda|^k|\mathbb{E}[X^k]|$. Noting that $1 + x \le e^x$ gives the result. □

As one final example, we return to Bennett's inequality (4.1.5) from Example 4.1.15. Note that this bound is only one-sided, as it requires only that the $X_i \le b$.

Proposition 4.1.21 (Bennett's inequality)**.** *Let X_i be independent mean-zero random variables with $\mathrm{Var}(X_i) = \sigma_i^2$ and $X_i \le b$. Then for $h(t) := (1+t)\log(1+t) - t$ and $\sigma^2 := \sum_{i=1}^n \sigma_i^2$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)\right).$$

Proof Using the standard Chernoff bound argument coupled with inequality (4.1.5), we see that

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(\sum_{i=1}^n \frac{\sigma_i^2}{b^2}\left(e^{\lambda b} - 1 - \lambda b\right) - \lambda t\right).$$

Letting $h(t) = (1+t)\log(1+t) - t$ as in the statement of the proposition and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$, we minimize over $\lambda \ge 0$, setting $\lambda = \frac{1}{b}\log(1 + \frac{bt}{\sigma^2})$. Substituting into our Chernoff bound application gives the proposition. □

Using averages instead of sums, for $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$, Bennett's inequality becomes

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{n\sigma^2}{b^2}h\left(\frac{bt}{\sigma^2}\right)\right). \tag{4.1.9}$$

Exercise 4.13 walks through an argument that $\frac{n\sigma^2}{b^2}h(\frac{bt}{\sigma^2}) \ge \frac{nt^2}{2\sigma^2 + \frac{2}{3}bt}$, yielding the classical Bernstein inequality that under the same conditions as Proposition 4.1.21,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}bt}\right). \tag{4.1.10}$$

4.1.3* Orlicz norms

Sub-Gaussian and sub-exponential random variables are examples of a broader class of random variables belonging to what are known as *Orlicz spaces*. Take any convex function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$ and $\psi(t) \to \infty$ as $t \uparrow \infty$, called an Orlicz function. (As $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ are convex, they are non-decreasing.) Then the Orlicz norm of a random variable X is

$$\|X\|_\psi := \inf \left\{t > 0 \mid \mathbb{E}[\psi(|X|/t)] \le 1\right\}. \tag{4.1.11}$$

That this is a norm is not completely trivial, though a few properties are immediate: clearly $\|aX\|_\psi = |a| \|X\|_\psi$, and we have $\|X\|_\psi = 0$ if and only if $X = 0$ with probability 1. The key result is that in fact, $\|\cdot\|_\psi$ is actually convex, which then guarantees that it is a norm.

Proposition 4.1.22. *The function $\|\cdot\|_\psi$ is convex on the space of random variables.*

Proof Because ψ is convex and non-decreasing, $x \mapsto \psi(|x|)$ is convex. Thus, its *perspective transform* $\psi_{\rm per}(|x|, t) := t\psi(|x|/t)$ is jointly convex in both $t \ge 0$ and x (see Appendix B.2.5, as well as the development of f-divergences in Section 2.2.3). This joint convexity of $\psi_{\rm per}$ implies that for any random variables X_0 and X_1 and t_0, t_1,

$$\mathbb{E}[\psi_{\rm per}(|\lambda X_0 + (1-\lambda)X_1|, \lambda t_0 + (1-\lambda)t_1)] \le \lambda \mathbb{E}[\psi_{\rm per}(|X_0|, t_0)] + (1-\lambda)\mathbb{E}[\psi_{\rm per}(|X_1|, t_1)].$$

Now note that $\mathbb{E}[\psi(|X|/t)] \le 1$ if and only if $t\mathbb{E}[\psi(|X|/t)] \le t$, so if $\|X_0\|_\psi < t_0$ and $\|X_1\|_\psi < t_1$, then $\|\lambda X_0 + (1-\lambda)X_1\| < \lambda t_0 + (1-\lambda)t_1$ and take $t_i \downarrow \|X_i\|_\psi$. □

Because $\|\cdot\|_\psi$ is convex and positively homogeneous, we certainly have

$$\|X + Y\|_\psi = 2\|(X+Y)/2\|_\psi \le \|X\|_\psi + \|Y\|_\psi,$$

that is, the triangle inequality holds. This implies that centering a variable can never increase its norm by much:

$$\|X - \mathbb{E}[X]\|_\psi \le \|X\|_\psi + \|\mathbb{E}[X]\|_\psi \le \|X\|_\psi + \|X\|_\psi$$

by Jensen's inequality, so that $\|X - \mathbb{E}[X]\|_\psi \le 2\|X\|_\psi$.

We can recover several standard norms on random variables, including some we have already implicitly used. The first are the classical L^p norms, where we take $\psi(t) = t^p$, where we see that

$$\inf\{t > 0 \mid \mathbb{E}[|X|^p/t^p] \le 1\} = \mathbb{E}[|X|^p]^{1/p}.$$

We also have the *sub-Gaussian* and *sub-exponential* norms, which we denote by considering the functions

$$\psi_p(x) := \exp\left(|x|^p\right) - 1.$$

These induce the *Orlicz ψ_p-norms*, as for $p \ge 1$, these are convex (as they are the composition of the increasing convex function $\exp(\cdot)$ applied to the nonnegative convex function $|\cdot|^p$). Theorem 4.1.11 shows that we have a *sub-Gaussian* norm

$$\|X\|_{\psi_2} := \inf\left\{t > 0 \mid \mathbb{E}[\exp(X^2/t^2)] \le 2\right\}, \tag{4.1.12}$$

while Theorem 4.1.16 shows a *sub-exponential norm* (or Orlicz ψ_1-norm)

$$\|X\|_{\psi_1} := \inf\left\{t > 0 \mid \mathbb{E}[\exp(|X|/t)] \le 2\right\}. \tag{4.1.13}$$

Many relationships follow immediately from the definitions (4.1.12) and (4.1.13). For example, the definition of the ψ_p-norms immediately implies that a sub-Gaussian random variable (whether or not it is mean zero) has a sub-exponential square:

Lemma 4.1.23. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential, as*

$$\|X\|_{\psi_2}^2 = \left\|X^2\right\|_{\psi_1}.$$

Additionally,

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Proof The first statement is trivial, so we prove only the second. Because $xy \le \frac{x^2}{2\eta} + \frac{\eta y^2}{2}$ for any x, y, and any $\eta > 0$, for any $t > 0$ we have

$$\mathbb{E}[\exp(|XY|/t)] \le \mathbb{E}\left[\exp\left(\frac{X^2}{2\eta t} + \frac{\eta Y^2}{2t}\right)\right] \le \mathbb{E}[\exp(X^2/\eta t)]^{1/2}\mathbb{E}[\exp(\eta Y^2/t)]^{1/2}$$

by the Cauchy-Schwarz inequality. In particular, if we take $t = \|X\|_{\psi_2} \|Y\|_{\psi_2}$, then the choice $\eta = \|X\|_{\psi_2}^2 / \|Y\|_{\psi_2}^2$ gives $\mathbb{E}[\exp(X^2/\eta t)] \le 2$ and $\mathbb{E}[\exp(\eta Y^2/t)] \le 2$, so that $\mathbb{E}[\exp(|XY|/t)] \le 2$. □

By tracing through the arguments in the proofs of Theorems 4.1.11 and 4.1.16, we can also see that we have the equivalences

$$\|X\|_{\psi_2} \asymp \sup_{k \in \mathbb{N}} \frac{1}{\sqrt{k}} \mathbb{E}[|X|^k]^{1/k} \quad \text{and} \quad \|X\|_{\psi_1} \asymp \sup_{k \in \mathbb{N}} \frac{1}{k} \mathbb{E}[|X|^k]^{1/k},$$

where $\asymp$ denotes upper and lower bounds by numerical constants. We leave this as an exercise.

The arguments we use to prove Theorems 4.1.11 and 4.1.16 also show the following result, which gives explicit constants connecting sub-exponential behavior with the ψ_1-norm.

Corollary 4.1.24. *Let X be any random variable with $\|X\|_{\psi_1} < \infty$. Then for all $t \ge 0$,*

$$\mathbb{P}(|X| \ge t) \le 2\exp\left(-t/\|X\|_{\psi_1}\right)$$

and if $\mathbb{E}[X] = 0$, then X is $(8\|X\|_{\psi_1}^2, 2\|X\|_{\psi_1})$-sub-exponential.

Proof The first statement is nearly trivial: by the Chernoff bounding method

$$\mathbb{P}(|X| \ge t) \le \mathbb{E}\left[\exp\left(|X|/\|X\|_{\psi_1}\right)\right]\exp(-t/\|X\|_{\psi_1}) \le 2\exp(-t/\|X\|_{\psi_1})$$

by definition of the ψ_1-norm. For the second, we mimic the proof of Theorem 4.1.16: because $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge t)dt$ for $Z \ge 0$, we have

$$\frac{\mathbb{E}[|X|^k]}{\|X\|_{\psi_1}^k} \le \int_0^\infty \mathbb{P}(|X|/\|X\|_{\psi_1} \ge t^{1/k})dt = k\int_0^\infty \mathbb{P}(|X|/\|X\|_{\psi_1} \ge u)u^{k-1}du \le 2k\int_0^\infty u^{k-1}e^{-u}du$$

using the substitution $u^k = t$. Rearranging yields $\mathbb{E}[|X|^k] \le 2\|X\|_{\psi_1}^k \Gamma(k+1) = 2\|X\|_{\psi_1}^k k!$. Then computing the moment generating function, we obtain

$$\mathbb{E}\left[\exp(\lambda X/\|X\|_{\psi_1})\right] \le 1 + \sum_{k=2}^\infty \frac{\lambda^k \mathbb{E}[|X|^k]}{\|X\|_{\psi_1}^k k!} \le 1 + 2\sum_{k=2}^\infty \lambda^k = 1 + \frac{2\lambda^2}{1-|\lambda|}$$

for $|\lambda| < 1$. For $|\lambda| \le \frac{1}{2}$, we use $1 + x \le e^x$ to obtain $\mathbb{E}[\exp(\lambda X/\|X\|_{\psi_1})] \le \exp(4\lambda^2)$, which is the desired result. □

4.1.4 An application of concentration: random projections

Consider nearest-neighbor (or k-nearest-neighbor) classification schemes. We have a sequence of data points as pairs (u_i, y_i), where the vectors $u_i \in \mathbb{R}^d$ have labels $y_i \in \{1, \ldots, L\}$, where L is the number of possible labels. Given a new point $u \in \mathbb{R}^d$ that we wish to label, we find the k-nearest neighbors to u in the sample $\{(u_i, y_i)\}_{i=1}^n$, then assign u the majority label of these k-nearest neighbors (ties are broken randomly). Unfortunately, it can be prohibitively expensive to store high-dimensional vectors and search over large datasets to find near vectors, and as such we use randomized projections to store lower dimensional approximations. By a careful use of concentration, we can show this yields strong approximation guarantees.

We consider the following variant: given n points $u_1, \ldots, u_n$, we wish to construct a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$, where $m \ll d$, such that

$$\left\|\Phi u_i - \Phi u_j\right\|^2 \in (1 \pm \epsilon) \left\|u_i - u_j\right\|^2.$$

Depending on the norm chosen, this task may be impossible; for the Euclidean (ℓ_2) norm, however, such an embedding is easy to construct using the *Johnson-Lindenstrauss* embedding, which uses Gaussian random variables with $m = O(\frac{1}{\epsilon^2} \log n)$. Key to the construction is that the size m is *independent* of the dimension d, only depending on the number of points n.

Example 4.1.25 (Johnson-Lindenstrauss)**:** Let the matrix $\Phi \in \mathbb{R}^{m \times d}$ have entries

$$\Phi_{ij} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, 1/m),$$

and let $\Phi_i \in \mathbb{R}^d$ denote the ith row of this matrix. Let $0 \le \delta \le 1$. We claim that

$$m \ge \frac{8}{\epsilon^2}\left[2 \log n + \log \frac{1}{\delta}\right] \quad \text{implies} \quad \left\|\Phi u_i - \Phi u_j\right\|_2^2 \in (1 \pm \epsilon) \left\|u_i - u_j\right\|_2^2$$

for all pairs u_i, u_j with probability at least $1 - \delta$. In particular, $m \gtrsim \frac{\log n}{\epsilon^2}$ is sufficient to achieve accurate dimension reduction with high probability.

To see this, note that for any fixed vector u,

$$\frac{\langle \Phi_i, u \rangle}{\left\|u\right\|_2} \sim \mathsf{N}(0, 1/m), \quad \text{and} \quad \frac{\left\|\Phi u\right\|_2^2}{\left\|u\right\|_2^2} = \sum_{i=1}^m \langle \Phi_i, u / \left\|u\right\|_2 \rangle^2$$

is a sum of independent scaled χ^2-random variables. In particular, we have $\mathbb{E}[\left\|\Phi u / \left\|u\right\|_2\right\|_2^2] = 1$, and using the χ^2-concentration result of Example 4.1.13 yields

$$\begin{aligned}\mathbb{P}\left(\left|\left\|\Phi u\right\|_2^2 / \left\|u\right\|_2^2 - 1\right| \ge \epsilon\right) &= \mathbb{P}\left(m \left|\left\|\Phi u\right\|_2^2 / \left\|u\right\|_2^2 - 1\right| \ge m\epsilon\right) \\ &\le 2 \inf_{|\lambda| \le \frac{1}{4}} \exp\left(2m\lambda^2 - \lambda m \epsilon\right) = 2 \exp\left(-\frac{m\epsilon^2}{8}\right),\end{aligned}$$

the last inequality holding for $\epsilon \in [0, 1]$. Now, using the union bound applied to each of the pairs (u_i, u_j) in the sample, we have

$$\mathbb{P}\left(\text{there exist } i \ne j \text{ s.t. } \left|\left\|\Phi(u_i - u_j)\right\|_2^2 - \left\|u_i - u_j\right\|_2^2\right| \ge \epsilon \left\|u_i - u_j\right\|_2^2\right) \le 2\binom{n}{2} \exp\left(-\frac{m\epsilon^2}{8}\right).$$

Taking $m \ge \frac{8}{\epsilon^2} \log \frac{n^2}{\delta} = \frac{16}{\epsilon^2} \log n + \frac{8}{\epsilon^2} \log \frac{1}{\delta}$ yields that with probability at least $1 - \delta$, we have $\left\|\Phi u_i - \Phi u_j\right\|_2^2 \in (1 \pm \epsilon) \left\|u_i - u_j\right\|_2^2$. $\diamond$

Exercise 4.14 uses a related strategy of identifying nearest points to implement a simple communication scheme with a randomized codebook.

4.2 Martingale methods

The next set of tools we consider constitute our first look at arguments based on *stability*, that is, how quantities that do not change very much when a single observation changes should concentrate. In this case, we would like to understand more general quantities than sample means, developing a few basic tools to understand when functions $f(X_1, \ldots, X_n)$ of independent random variables X_i concentrate around their expectations. Roughly, we expect that if changing the value of one x_i does not significantly change $f(x_1^n)$—it is stable—then $f(X_1^n)$ should concentrate.

To develop the tools to do this, we go through an approach based on martingales, a deep subject in probability theory. We give a high-level treatment of martingales, taking an approach that does not require measure-theoretic considerations.

Definition 4.3. *Let $M_1, M_2, \ldots$ be an $\mathbb{R}$-valued sequence of random variables. They are a* martingale *if there exist another sequence of random variables $\{Z_1, Z_2, \ldots\} \subset \mathcal{Z}$ and sequence of functions $f_n : \mathcal{Z}^n \to \mathbb{R}$ such that*

$$\mathbb{E}[M_n \mid Z_1^{n-1}] = M_{n-1} \quad \text{and} \quad M_n = f_n(Z_1^n)$$

for all $n \geq 2$. We say that the sequence M_n is adapted *to $\{Z_n\}$.*

In general, one replaces the sequence $Z_1, Z_2, \ldots$ with an increasing sequence of σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots$, and M_n is $\mathcal{F}_n$-measurable, but Definition 4.3 is sufficient for our purposes. We also will find it convenient to study *differences* of martingales, so that we make the following

Definition 4.4. *A sequence $D_1, D_2, \ldots$ of random variables form a* martingale difference sequence *if $M_n := \sum_{i=1}^n D_i$ is a martingale.*

Equivalently, there is a sequence of random variables Z_n and functions $g_n : \mathcal{Z}^n \to \mathbb{R}$ such that

$$\mathbb{E}[D_n \mid Z_1^{n-1}] = 0 \quad \text{and} \quad D_n = g_n(Z_1^n)$$

for all $n \geq 1$, where $\mathbb{E}[D_1] = 0$.

There are numerous examples of martingale sequences. The classical ones involve random walks.

Example 4.2.1 (Symmetric random walk)**:** Let $D_n \in \{\pm 1\}$ be uniform and independent. Then D_n form a martingale difference sequence adapted to themselves (that is, we may take $Z_n = D_n$), and $M_n = \sum_{i=1}^n D_i$ is a martingale. $\diamond$

Example 4.2.2: Let D_n be independent random variables with $\mathbb{E}[D_n] = 0$ and $\mathbb{E}[|D_n|] < \infty$. Then D_n form a martingale difference sequence adapted to themselves. $\diamond$

A more sophisticated example, to which we will frequently return and that suggests the potential usefulness of martingale constructions, is the *Doob martingale* associated with a function f.

Example 4.2.3 (Doob martingales)**:** Let $f : \mathcal{X}^n \to \mathbb{R}$ be an otherwise arbitrary function, and let $X_1, \ldots, X_n$ be arbitrary random variables. The difference sequence

$$D_i := \mathbb{E}[f(X_1^n) \mid X_1^i] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$$

defines the Doob martingale. By inspection, the D_i are functions of X_1^i, and we have

$$\begin{aligned}\mathbb{E}[D_i \mid X_1^{i-1}] &= \mathbb{E}[\mathbb{E}[f(X_1^n) \mid X_1^i] \mid X_1^{i-1}] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] \\ &= \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}] = 0\end{aligned}$$

by the tower property of expectations. Thus, the D_i satisfy Definition 4.4 of a martingale difference sequence, and

$$M_n := \sum_{i=1}^n D_i = f(X_1^n) - \mathbb{E}[f(X_1^n)],$$

so the Doob martingale captures the difference between f and its expectation. $\diamond$

4.2.1 Sub-Gaussian martingales and Azuma-Hoeffding inequalities

With these motivating ideas introduced, we turn to definitions that allow generalizations of our concentration inequalities for sub-Gaussian sums to sub-Gaussian martingales.

Definition 4.5. *Let $\{D_n\}$ be a martingale difference sequence adapted to $\{Z_n\}$. Then D_n is a* σ_n^2-sub-Gaussian *martingale difference if for all n and $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda D_n) \mid Z_1^{n-1}] \le \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right)$$

Immediately we have the Azuma-Hoeffding inequalities, which generalize the earlier tensorization identities for sub-Gaussian random variables.

Theorem 4.2.4 (Azuma-Hoeffding)**.** *Let $\{D_n\}$ be a σ_n^2-sub-Gaussian martingale difference sequence. Then $M_n = \sum_{i=1}^n D_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian, and moreover,*

$$\max\left\{\mathbb{P}(M_n \ge t), \mathbb{P}(M_n \le -t)\right\} \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \quad \text{for all } t \ge 0.$$

Proof Letting Z_n be the sequence to which the D_n are adapted, we write

$$\begin{aligned}\mathbb{E}[\exp(\lambda M_n)] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda D_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^n e^{\lambda D_i} \mid Z_1^{n-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda D_i} \mid Z_1^{n-1}\right] \mathbb{E}[e^{\lambda D_n} \mid Z_1^{n-1}]\right]\end{aligned}$$

because $D_1, \ldots, D_{n-1}$ are functions of Z_1^{n-1}. Then we use Definition 4.5, which implies that $\mathbb{E}[e^{\lambda D_n} \mid Z_1^{n-1}] \le e^{\lambda^2 \sigma_n^2/2}$, to obtain

$$\mathbb{E}[\exp(\lambda M_n)] \le \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda D_i}\right] \exp\left(\frac{\lambda^2 \sigma_n^2}{2}\right).$$

Repeating the same argument for $n-1, n-2, \ldots, 1$ gives that

$$\log \mathbb{E}[\exp(\lambda M_n)] \le \frac{\lambda^2}{2} \sum_{i=1}^n \sigma_i^2.$$

The second claims apply Chernoff bounds via Proposition 4.1.8 and that $\mathbb{E}[M_n] = 0$. $\square$

As an immediate corollary, we recover Proposition 4.1.9: the sum $M_n = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$ forms a martingale. A second corollary gives what is typically termed the Azuma inequality:

Corollary 4.2.5. *Let D_i be a bounded difference martingale difference sequence, meaning that $|D_i| \le c$. Then $M_n = \sum_{i=1}^n D_i$ satisfies*

$$\mathbb{P}(n^{-1/2} M_n \ge t) \vee \mathbb{P}(n^{-1/2} M_n \le -t) \le \exp\left(-\frac{t^2}{2c^2}\right) \quad \text{for } t \ge 0.$$

Thus, bounded increment random walks typically remain within $\pm\sqrt{n}$ of their expectations after n steps. The bibliographic section provides references to Bennett- and Bernstein-type extensions of these inequalities.

4.2.2 Examples and bounded differences

We develop several applications of the Azuma-Hoeffding inequalities (Theorem 4.2.4) to functions satisfying certain stability conditions. We first define the collections of functions we consider.

Definition 4.6 (Bounded differences)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$ for some space $\mathcal{X}$. Then f satisfies* bounded differences *with constants c_i if for each $i \in \{1, \ldots, n\}$, all $x_1^n \in \mathcal{X}^n$, and $x_i' \in \mathcal{X}$ we have*

$$|f(x_1^{i-1}, x_i, x_{i+1}^n) - f(x_1^{i-1}, x_i', x_{i+1}^n)| \le c_i.$$

The bounded differences (or McDiarmid's) inequality classically relates Definition 4.6 to concentration:

Proposition 4.2.6 (Bounded differences inequality)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$ satisfy bounded differences with constants c_i, and let X_i be independent random variables. $f(X_1^n) - \mathbb{E}[f(X_1^n)]$ is $\frac{1}{4}\sum_{i=1}^n c_i^2$-sub-Gaussian, and*

$$\mathbb{P}\left(f(X_1^n) - \mathbb{E}[f(X_1^n)] \ge t\right) \vee \mathbb{P}\left(f(X_1^n) - \mathbb{E}[f(X_1^n)] \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof The idea is to show that f has $c_i^2/4$-sub-Gaussian Doob martingale (Example 4.2.3) and apply the Azuma-Hoeffding inequality. To that end, define $D_i = \mathbb{E}[f(X_1^n) \mid X_1^i] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$, and note that $\sum_{i=1}^n D_i = f(X_1^n) - \mathbb{E}[f(X_1^n)]$. The random variables

$$L_i := \inf_x \mathbb{E}[f(X_1^n) \mid X_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$$
$$U_i := \sup_x \mathbb{E}[f(X_1^n) \mid X_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1}]$$

evidently satisfy $L_i \le D_i \le U_i$, and moreover, we have

$$\begin{aligned} U_i - L_i &\le \sup_{x_1^{i-1}} \sup_{x,x'} \left\{\mathbb{E}[f(X_1^n) \mid X_1^{i-1} = x_1^{i-1}, X_i = x] - \mathbb{E}[f(X_1^n) \mid X_1^{i-1} = x_1^{i-1}, X_i = x']\right\} \\ &= \sup_{x_1^{i-1}} \sup_{x,x'} \int \left(f(x_1^{i-1}, x, x_{i+1}^n) - f(x_1^{i-1}, x', x_{i+1}^n)\right) dP(x_{i+1}^n) \le c_i, \end{aligned}$$

where we have used the independence of the X_i and Definition 4.6 of bounded differences. Consequently, Hoeffding's Lemma (Example 4.1.6) implies that $\mathbb{E}[e^{\lambda D_i} \mid X_1^{i-1}] \le \exp(\lambda^2 c_i^2/8)$, so Theorem 4.2.4 gives the proposition. □

We give two examples satisfying the conditions of Proposition 4.2.6; we will revisit each.

Example 4.2.7 (Bounded random vectors)**:** Let $\mathbb{B}$ be a Banach space—a complete normed vector space—with norm $\|\cdot\|$. Let X_i be independent bounded random vectors in $\mathbb{B}$ satisfying $\mathbb{E}[X_i] = 0$ and $\|X_i\| \le c$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Then by the triangle inequality, the quantity $f(X_1^n) = \|\overline{X}_n\|$ satisfies bounded differences:

$$|f(x_1^{i-1}, x, x_{i+1}^n) - f(x_1^{i-1}, x', x_{i+1}^n)| \le \frac{1}{n}\left\|x - x'\right\| \le \frac{2c}{n}.$$

Consequently, if X_i are independent, for all $t \ge 0$ the sample mean satisfies

$$\max\left\{\mathbb{P}\left(\left\|\overline{X}_n\right\| - \mathbb{E}\left[\left\|\overline{X}_n\right\|\right] \ge t\right), \mathbb{P}\left(\left\|\overline{X}_n\right\| - \mathbb{E}\left[\left\|\overline{X}_n\right\|\right] \le -t\right)\right\} \le \exp\left(-\frac{nt^2}{2c^2}\right). \tag{4.2.1}$$

So the norm of averages of bounded random vectors concentrates about its expectation.

The challenge becomes to control the *expectation* term in the concentration bound (4.2.1). When the norm $\|\cdot\|$ is Euclidean, this is easy: assuming the X_i belong to a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|x\|^2 = \langle x, x\rangle$, the Cauchy-Schwarz inequality implies that

$$\mathbb{E}\left[\left\|n\overline{X}_n\right\|\right]^2 \le \mathbb{E}\left[\left\|n\overline{X}_n\right\|^2\right] = \sum_{i=1}^n \mathbb{E}[\|X_i\|^2].$$

Assuming the X_i are independent with $\mathbb{E}[\|X_i\|^2] \le \sigma^2$, inequality (4.2.1) implies

$$\mathbb{P}\left(\left\|\overline{X}_n\right\| \ge \frac{\sigma}{\sqrt{n}} + t\right) \le \exp\left(-\frac{nt^2}{2c^2}\right),$$

for all $t \ge 0$. $\diamond$

We can specialize Example 4.2.7 to a situation that is very important for treatments of concentration, sums of random vectors, and generalization bounds in machine learning.

Example 4.2.8 (Rademacher complexities)**:** The frequent uses of this special case of Example 4.2.7 justify more specialized consideration. Let $\mathcal{X}$ be a space and $\mathcal{F}$ a collection of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\varepsilon_i \in \{-1, 1\}$ be a collection of independent random signs. Then the *empirical Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F} \mid x_1^n) := \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right],$$

where the expectation is over the random signs ε_i. The *Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}[R_n(\mathcal{F} \mid X_1^n)],$$

the expectation of the empirical Rademacher complexities. If $b_0 \le f \le b_1$ for $f \in \mathcal{F}$, the Rademacher complexity satisfies bounded differences, because for any two sequences x_1^n and z_1^n differing in only element j, we have

$$n|R_n(\mathcal{F} \mid x_1^n) - R_n(\mathcal{F} \mid z_1^n)| \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}|\varepsilon_j(f(x_j) - f(z_j))|\right] \le b_1 - b_0.$$

The Rademacher complexity $R_n(\mathcal{F} \mid X_1^n)$ is thus $\frac{(b_1-b_0)^2}{4n}$-sub-Gaussian by Theorem 4.2.4. $\diamond$

Example 4.2.9 (Rademacher complexity as a random vector)**:** We may think of Rademacher complexities as norms on certain spaces. Indeed, if we consider a vector space $\mathcal{L}$ of linear functions on $\mathcal{F}$, we can define the $\mathcal{F}$-seminorm $\|L\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |L(f)|$ on $\mathcal{L}$. In this case, we may consider the symmetrized empirical distributions

$$P_n^0 := \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{1}_{X_i} \quad f \mapsto P_n^0 f := \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)$$

as elements of this vector space $\mathcal{L}$. (Here we have used $\mathbf{1}_{X_i}$ to denote the point mass at X_i.) Then the Rademacher complexity is the expected norm of P_n^0, a random vector, as in Example 4.2.7. This perspective shows that any general results we prove about random vectors will carry over immediately to versions of the Rademacher complexity. $\diamond$

4.3* Matrix concentration

In this section we develop matrix analogues of the concentration inequalities for scalar sums in Section 4.1, including matrix Hoeffding and Bernstein inequalities. In addition to intellectual interest, their importance in probability theory more broadly, and the ways that they admit interesting information-theoretic arguments, these will also be useful when we develop convergence guarantees for statistical estimators in the next chapter. Nonetheless, this chapter captures more advanced material, and understanding the techniques here will not be essential for the remainder of the book.

Our main goal will be to bound maximal eigenvalues (or operator norms) of symmetric or Hermitian matrices, that is, for sums $S_n = \sum_{i=1}^n X_i$ of independent matrices, of deviation probabilities

$$\mathbb{P}(\lambda_{\max}(S_n) \geq t) \quad \text{or} \quad \mathbb{P}(\lambda_{\min}(S_n) \leq -t),$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote maximal and minimal eigenvalues, respectively. Our approach will be to generalize the approach using moment generating functions, though this becomes non-trivial because there is no immediately obvious analogue of the tensorization identities we have for scalars. While in the scalar case, for a sum $S_n = \sum_{i=1}^n X_i$ of independent random variables, we have

$$e^{\lambda S_n} = \prod_{i=1}^n e^{\lambda X_i},$$

such an identity fails for matrices, because their exponentials (typically) fail to commute.

To develop the basic matrix concentration equalities we provide, we require a brief review of matrix calculus and operator functions. We shall typically work with Hermitian matrices $A \in \mathbb{C}^{d \times d}$, meaning that $A = A^*$, where A^* denotes the Hermitian (conjugate) transpose of A, whose entries are $(A^*)_{ij} = \overline{A}_{ji}$, the conjugate of A_{ji}. We work in this generality for two reasons: first, because such matrices admit the spectral decompositions we require to develop the operators we use, and second, because we often will encounter random matrices with symmetric distributions, meaning that $X \stackrel{\text{dist}}{=} -X$, which can lead to ambiguity were we to describe A as symmetric.

With this, we give a brief review of some properties of Hermitian matrices and some associated matrix operators. Let $\mathcal{H}_d := \{A \in \mathbb{C}^{d \times d} \mid A^* = A\}$ be the Hermitian matrices. The spectral theorem states that any $A \in \mathcal{H}_d$ admits the spectral decomposition $A = U \Lambda U^*$, where Λ is the diagonal matrix of the (necessarily) real eigenvalues of A and $U \in \mathbb{C}^{n \times n}$ is unitary, so that $U^* U = U U^* = I$. For a function $f : \mathbb{R} \to \mathbb{R}$, we can then define its *operator extension* to $\mathcal{H}_d$ by

$$f(A) := U \operatorname{diag}\left(f(\lambda_1(A)), \ldots, f(\lambda_n(A))\right) U^*,$$

where A has spectral decomposition $A = U\Lambda U^*$ and $\lambda_i(A)$ denotes the ith eigenvalue of A. Because we wish to mimic the approach based on moment generating functions that yields our original sub-Gaussian and sub-exponential concentration inequalities in Section 4.1, the most important function for us will be the exponential

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k,$$

where we recall the convention that $A^0 = I$ whenever A is Hermitian.

A Hermitian matrix A is *positive definite*, denoted $A \succ 0$, if $x^*Ax > 0$ for all $x \neq 0$, and is positive semidefinite (PSD), which we denote by $A \succeq 0$, if $x^*Ax \geq 0$ for all vectors x. Positive definiteness is then equivalent to the condition that $\lambda_i(A) > 0$ for all eigenvalues of A, while semidefiniteness that $\lambda_i(A) \geq 0$. We also use the standard semidefinite ordering, so that $A \succeq B$ means that $A - B \succeq 0$. For $A \in \mathcal{H}_d$, we evidently have $\exp(A) \succ 0$. The familiar trace $\mathrm{tr}(A) = \sum_{j=1}^d A_{jj}$ of a square matrix allows us to define inner products, where for general complex matrices $A, B \in \mathbb{C}^{m \times n}$ we define $\langle A, B \rangle = \mathrm{tr}(A^*B)$, while the space of Hermitian matrices admits the *real* inner product $\langle A, B \rangle := \mathrm{tr}(A^*B) = \mathrm{tr}(AB)$. (See Exercise 4.15.) The spectral theorem also shows the standard identity that $\mathrm{tr}(A) = \sum_{j=1}^d \lambda_j(A)$ for $A \in \mathcal{H}_d$, and that the operator norm

$$\|A\|_{\mathrm{op}} := \sup_{u,v \in \mathbb{S}^{d-1}} |u^*Av| = \max_{j \leq d} |\lambda_j(A)|.$$

To analogize our approach with real-valued random variables, we begin with the Chernoff bound, Proposition 4.1.3. Here, we have the following observation:

Proposition 4.3.1. *For any random Hermitian matrix X,*

$$\mathbb{P}(\lambda_{\max}(X) \geq t) \leq \mathrm{tr}(\mathbb{E}[e^{\lambda X}])e^{-\lambda t}$$

for all $\lambda \geq 0$ and $t \geq 0$.

Proof Apply the standard Chernoff bound to the random variable $\lambda_{\max}(X)$, which gives that for any $\lambda > 0$ that

$$\mathbb{P}(\lambda_{\max}(X) \geq t) \leq \mathbb{E}[e^{\lambda \cdot \lambda_{\max}(X)}]e^{-\lambda t}.$$

Then observe that by definition of the matrix exponential, we have $e^{\lambda \cdot \lambda_{\max}(X)} \leq \mathrm{tr}(e^{\lambda X})$, because $e^{\lambda X}$ has all positive eigenvalues. □

We would like now to provide some type of general tensorization identity for matrices, in analogy with Propositions 4.1.9 or 4.1.18. Unfortunately, this breaks down: for Hermitian A, B, we have

$$e^{A+B} = e^A e^B$$

if and only if A and B commute, so that they are simultaneously diagonalizable, and even more [342, Section 3], $\mathrm{tr}(e^{A+B}) = \mathrm{tr}(e^A e^B)$ if and only if $AB = BA$. Nonetheless, we have the following inequality, the key to extending the standard one-dimensional approach to concentration:

Proposition 4.3.2 (The Golden-Thompson inequality)**.** *Let A, B be Hermitian matrices. Then*

$$\mathrm{tr}(e^{A+B}) \leq \mathrm{tr}(e^A e^B).$$

While the proof is essentially elementary, it is not central to our development, so we defer it to Section 4.4.3. With Proposition 4.3.2 in hand, however, we can develop matrix analogues of the Hoeffding and Bernstein-type concentration bounds in Section 4.1.

We begin with Azuma-Hoeffding-type bounds, which analogize Theorem 4.2.4. To perform the iterative "peeling" off of individual terms via symmetrization in a sum of random matrices, we use the following result:

Lemma 4.3.3 (A matrix symmetrization inequality)**.** *Let H be an arbitrary (fixed) Hermitian matrix, X be a mean-zero Hermitian matrix, and $\varepsilon \sim \mathsf{Uniform}\{\pm 1\}$. Then*

$$\mathrm{tr}(\mathbb{E}[e^{H+X}]) \le \mathrm{tr}(\mathbb{E}[e^{H+2\varepsilon X}]).$$

Proof Let X' be an independent copy of X. Then because the trace exponential $\mathrm{tr}(e^X)$ is convex on the Hermitian matrices (see Exercise 4.17), we have

$$\mathrm{tr}(\mathbb{E}[e^{H+X}]) = \mathrm{tr}(\mathbb{E}[e^{H+(X-\mathbb{E}[X'])}]) \le \mathrm{tr}(\mathbb{E}[e^{H+X-X'}])$$

by Jensen's inequality. Introducing the random sign $\varepsilon \in \{\pm 1\}$, we have by symmetry that $X - X' \stackrel{\mathrm{dist}}{=} \varepsilon(X - X')$, and so

$$\begin{aligned}\mathrm{tr}(\mathbb{E}[e^{H+X}]) \le \mathbb{E}[\mathrm{tr}(e^{H+\varepsilon X-\varepsilon X'})] &= \mathbb{E}[\mathrm{tr}(e^{H/2+\varepsilon X+H/2-\varepsilon X'})] \\ &\le \mathbb{E}[\mathrm{tr}(e^{H/2+\varepsilon X}e^{H/2-\varepsilon X'})],\end{aligned}$$

where the second inequality follows from Proposition 4.3.2. Now we use that for Hermitian matrices A, B, the Cauchy-Schwartz inequality implies $\mathrm{tr}(AB) \le \|A\|_{\mathrm{Fr}} \|B\|_{\mathrm{Fr}} = \mathrm{tr}(A^2)^{1/2} \mathrm{tr}(B^2)^{1/2}$, so that

$$\mathbb{E}[\mathrm{tr}(e^{H/2+\varepsilon X}e^{H/2-\varepsilon X'})] \le \mathbb{E}[\mathrm{tr}(e^{H+2\varepsilon X})^{1/2} \mathrm{tr}(e^{H-2\varepsilon X'})^{1/2}] \le \mathbb{E}[\mathrm{tr}(e^{H+2\varepsilon X})]^{1/2}\mathbb{E}[\mathrm{tr}(e^{H-2\varepsilon X'})]^{1/2}$$

again by Cauchy-Schwarz. Because $\varepsilon X \stackrel{\mathrm{dist}}{=} -\varepsilon X'$, the lemma follows. □

This allows us to perform the type of "peeling-off" argument (or tensorization), addressing one term in the sum at a time, that gives tight enough moment generating function bounds.

Theorem 4.3.4. *Let $X_1, \ldots, X_n \in \mathcal{H}_d$ be independent, mean-zero, and satisfy $\|X_i\|_{\mathrm{op}} \le b_i$. Define $S_n = \sum_{i=1}^n X_i$. Then for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\mathrm{tr}(e^{\lambda S_n})] \le d \exp\left(2\lambda^2 \sum_{i=1}^n b_i^2\right).$$

Proof By iterated expectation and Lemma 4.3.3, we have

$$\mathrm{tr}(\mathbb{E}[e^{\lambda S_n}]) \le \mathrm{tr}(\mathbb{E}[e^{\lambda S_{n-1}+2\lambda\varepsilon X_n}]) \le \mathrm{tr}(\mathbb{E}[e^{\lambda S_{n-1}}]) \left\|\mathbb{E}[e^{2\lambda\varepsilon X_n}]\right\|_{\mathrm{op}},$$

where $\varepsilon \in \{\pm 1\}$ is an independent random sign and we have used independence. Now, we use the following calculation [399]: if X is Hermitian and ε a random sign, then

$$\mathbb{E}[e^{\varepsilon X}] \preceq \mathbb{E}[e^{X^2/2}]. \tag{4.3.1}$$

Deferring the argument for the operator inequality (4.3.1), note that it immediately implies $\mathbb{E}[e^{2\lambda\varepsilon X}] \preceq \mathbb{E}[e^{2\lambda^2 X^2}]$. The convexity of the operator norm and that $\|X_n\|_{\rm op} \le b_n$ then imply

$$\left\|\mathbb{E}[e^{2\lambda\varepsilon X_n}]\right\|_{\rm op} \le \left\|\mathbb{E}[e^{2\lambda^2 X_n^2}]\right\|_{\rm op} \le \mathbb{E}\left[e^{2\lambda^2\|X_n^2\|_{\rm op}}\right] \le e^{2\lambda^2 b_n^2}$$

Repeating the argument by iteratively peeling off the term X_{n-i} for S_{n-2} through S_1 then yields

$$\mathbb{E}[\mathrm{tr}(\exp(\lambda S_n))] \le \mathrm{tr}(I)\prod_{i=1}^n \exp(2\lambda^2 b_i^2) = d\exp\left(2\lambda^2\sum_{i=1}^n b_i^2\right).$$

To see inequality (4.3.1), note that for any positive semidefinite A, we have $A \preceq tA$ for $t \ge 1$. Then because $X^{2k} \succeq 0$ for all $k \in \mathbb{N}$ and $(2k)! \ge 2^k k!$, we have

$$\mathbb{E}[e^{\varepsilon X}] = I + \sum_{k=1}^\infty \frac{\mathbb{E}[X^{2k}]}{(2k)!} \preceq I + \sum_{k=1}^\infty \frac{\mathbb{E}[(X^2)^k]}{2^k k!} = \mathbb{E}[e^{X^2/2}],$$

where we used symmetry to eliminate terms with odd powers. □

Theorem 4.3.4 immediately implies the following corollary, whose argument parallels those in Section 4.1 (e.g., Corollary 4.1.10).

Corollary 4.3.5 (Matrix Hoeffding inequality)**.** *Let $X_i \in \mathcal{H}_d$ be independent mean-zero Hermitian matrices with $\|X_i\|_{\rm op} \le b_i$. Then $S_n := \sum_{i=1}^n X_i$ satisfies*

$$\mathbb{P}\left(\|S_n\|_{\rm op} \ge t\right) \le 2d\exp\left(-\frac{t^2}{8\sum_{i=1}^n b_i^2}\right).$$

If we have more direct bounds on $\mathbb{E}[e^{\lambda X_i}]$, then we can also employ those via a similar "peeling off" the last term argument. By carefully controlling matrix moment generating functions in a way similar to that we did in Example 4.1.15 to obtain sub-exponential behavior for bounded random variables, we can give a *matrix Bernstein-type* inequality.

Theorem 4.3.6. *Let X_i be independent mean-zero Hermitian matrices with $\|X_i\|_{\rm op} \le b$ and $\left\|\mathbb{E}[X_i^2]\right\|_{\rm op} \le \sigma_i^2$. Then $S_n = \sum_{i=1}^n X_i$ satisfies*

$$\mathbb{P}\left(\|S_n\|_{\rm op} \ge t\right) \le 2d\exp\left(-\min\left\{\frac{t^2}{4\sum_{i=1}^n \sigma_i^2}, \frac{3t}{4b}\right\}\right).$$

The proof is similar to that of Theorem 4.3.4, so we leave it as an extended exercise (Exercise 4.18).

We unpack the theorem a bit to give some intuition. Given a variance bound σ^2 such that $\mathbb{E}[X_i^2] \preceq \sigma^2 I$, the theorem states that

$$\mathbb{P}\left(\left\|n^{-1}S_n\right\|_{\rm op} \ge t\right) \le 2d\exp\left(-\min\left\{\frac{nt^2}{4\sigma^2}, \frac{3nt}{4b}\right\}\right).$$

Letting $\delta \in (0,1)$ be arbitrary and setting $t = \max\{\frac{2\sigma}{\sqrt{n}}\sqrt{\log\frac{2d}{\delta}}, \frac{4b}{3n}\log\frac{2d}{\delta}\}$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_{\rm op} \le \max\left\{\frac{2\sigma}{\sqrt{n}}\sqrt{\log\frac{2d}{\delta}}, \frac{4b}{3n}\log\frac{2d}{\delta}\right\}$$

with probability at least $1-\delta$. So we see the familiar sub-Gaussian and sub-exponential scaling of the random sum.

4.4 Technical proofs

4.4.1 Proof of Theorem 4.1.11

(1) **implies** (2) Let $K_1 = 1$. Using the change of variables identity that for a nonnegative random variable Z and any $k \geq 1$ we have $\mathbb{E}[Z^k] = k \int_0^\infty t^{k-1} \mathbb{P}(Z \geq t) dt$, we find

$$\mathbb{E}[|X|^k] = k \int_0^\infty t^{k-1} \mathbb{P}(|X| \geq t) dt \leq 2k \int_0^\infty t^{k-1} \exp\left(-\frac{t^2}{\sigma^2}\right) dt = k\sigma^k \int_0^\infty u^{k/2-1} e^{-u} du,$$

where for the last inequality we made the substitution $u = t^2/\sigma^2$. Noting that this final integral is $\Gamma(k/2)$, we have $\mathbb{E}[|X|^k] \leq k\sigma^k \Gamma(k/2)$. Because $\Gamma(s) \leq s^s$ for $s \geq 1$, we obtain

$$\mathbb{E}[|X|^k]^{1/k} \leq k^{1/k} \sigma \sqrt{k/2} \leq e^{1/e} \sigma \sqrt{k}.$$

Thus (2) holds with $K_2 = e^{1/e}$.

(2) **implies** (3) Let $\sigma = \sup_{k \geq 1} k^{-\frac{1}{2}} \mathbb{E}[|X|^k]^{1/k}$, so that $K_2 = 1$ and $\mathbb{E}[|X|^k] \leq k^{\frac{k}{2}} \sigma^k$ for all k. For $K_3 \in \mathbb{R}_+$, we thus have

$$\mathbb{E}[\exp(X^2/(K_3\sigma^2))] = \sum_{k=0}^\infty \frac{\mathbb{E}[X^{2k}]}{k! K_3^k \sigma^{2k}} \leq \sum_{k=0}^\infty \frac{\sigma^{2k}(2k)^k}{k! K_3^k \sigma^{2k}} \stackrel{(i)}{\leq} \sum_{k=0}^\infty \left(\frac{2e}{K_3}\right)^k$$

where inequality (i) follows because $k! \geq (k/e)^k$, or $1/k! \leq (e/k)^k$. Noting that $\sum_{k=0}^\infty \alpha^k = \frac{1}{1-\alpha}$, we obtain (3) by taking $K_3 = 2e(1 - \frac{1}{e}) \approx 3.437$.

(3) **implies** (4) Let us take $K_3 = 1$ and recall the assumption that $\mathbb{E}[X] = 0$. We claim that (4) holds with $K_4 = \frac{3}{4}$. We prove this result for both small and large λ. First, note the (non-standard) inequality that $e^x \leq x + e^{\frac{9x^2}{16}}$ for all x. Then we have

$$\mathbb{E}[\exp(\lambda X)] \leq \underbrace{\mathbb{E}[\lambda X]}_{=0} + \mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right]$$

Now note that for $|\lambda| \leq \frac{4}{3\sigma}$, we have $9\lambda^2\sigma^2/16 \leq 1$, and so by Jensen's inequality,

$$\mathbb{E}\left[\exp\left(\frac{9\lambda^2 X^2}{16}\right)\right] = \mathbb{E}\left[\exp(X^2/\sigma^2)^{\frac{9\lambda^2\sigma^2}{16}}\right] \leq e^{\frac{9\lambda^2\sigma^2}{16}}.$$

For large λ, we use the simpler Fenchel-Young inequality, that is, that $\lambda x \leq \frac{\lambda^2}{2c} + \frac{cx^2}{2}$, valid for all $c \geq 0$. Then we have for any $0 \leq c \leq 2$ that

$$\mathbb{E}[\exp(\lambda X)] \leq e^{\frac{\lambda^2\sigma^2}{2c}} \mathbb{E}\left[\exp\left(\frac{cX^2}{2\sigma^2}\right)\right] \leq e^{\frac{\lambda^2\sigma^2}{2c}} e^{\frac{c}{2}}$$

by Jensen's inequality. If $|\lambda| \geq \frac{4}{3\sigma}$, then $\frac{1}{2} \leq \frac{9}{32}\lambda^2\sigma^2$, and

$$\mathbb{E}[\exp(\lambda X)] \leq \inf_{c \in [0,2]} e^{[\frac{1}{2c} + \frac{9c}{32}]\lambda^2\sigma^2} = \exp\left(\frac{3\lambda^2\sigma^2}{4}\right).$$

(3) **implies** (1) Assume (3) holds with $K_3 = 1$. Then for $t \geq 0$ we have

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(X^2/\sigma^2 \geq t^2/\sigma^2) \leq \mathbb{E}[\exp(\lambda X^2/\sigma^2)] \exp\left(-\frac{\lambda t^2}{\sigma^2}\right)$$

for all $\lambda \geq 0$. For $\lambda \leq 1$, Jensen's inequality implies $\mathbb{E}[\exp(\lambda X^2/\sigma^2)] \leq \mathbb{E}[\exp(X^2/\sigma^2)]^\lambda \leq e^\lambda$ by assumption (3). Set $\lambda = \log 2 \approx .693$.

(4) **implies** (1) This is Proposition 4.1.8, with $K_4 = \frac{1}{2}$ and $K_1 = 2$.

4.4.2 Proof of Theorem 4.1.16

(1) **implies** (2) As in the proof of Theorem 4.1.11, we use that for a nonnegative random variable Z we have $\mathbb{E}[Z^k] = k \int_0^\infty t^{k-1} \mathbb{P}(Z \geq t) dt$. Let $K_1 = 1$. Then

$$\mathbb{E}[|X|^k] = k \int_0^\infty t^{k-1} \mathbb{P}(|X| \geq t) dt \leq 2k \int_0^\infty t^{k-1} \exp(-t/\sigma) dt = 2k\sigma^k \int_0^\infty u^{k-1} \exp(-u) du,$$

where we used the substitution $u = t/\sigma$. Thus we have $\mathbb{E}[|X|^k] \leq 2\Gamma(k+1)\sigma^k$, and using $\Gamma(k+1) \leq k^k$ yields $\mathbb{E}[|X|^k]^{1/k} \leq 2^{1/k} k\sigma$, so that (2) holds with $K_2 \leq 2$.

(2) **implies** (3) Let $K_2 = 1$, and note that

$$\mathbb{E}[\exp(|X|/(K_3\sigma))] = \sum_{k=0}^\infty \frac{\mathbb{E}[|X|^k]}{K_3^k \sigma^k k!} \leq \sum_{k=0}^\infty \frac{k^k}{k!} \cdot \frac{1}{K_3^k} \stackrel{(i)}{\leq} \sum_{k=0}^\infty \left(\frac{e}{K_3}\right)^k,$$

where inequality (i) used that $k! \geq (k/e)^k$. Taking $K_3 = e^2/(e-1) < 5$ gives the result.

(3) **implies** (1) If $\mathbb{E}[\exp(|X|/\sigma)] \leq e$, then for $t \geq 0$

$$\mathbb{P}(|X| \geq t) \leq \mathbb{E}[\exp(|X|/\sigma)] e^{-t/\sigma} \leq e^{1-t/\sigma}.$$

Combining this with the trivial bound of one gives

$$\mathbb{P}(|X| \geq t) \leq e^{1-t/\sigma} \wedge 1 \leq 2e^{-\frac{2t}{5\sigma}},$$

so that (1) holds with $K_1 = \frac{5}{2}$.

(2) **if and only if** (4) Assume that (2) holds with $K_2 = 1$, and let $\sigma = \sup_{k \geq 1} \frac{1}{k} \mathbb{E}[|X|^k]^{1/k}$. Then because $\mathbb{E}[X] = 0$,

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \sum_{k=2}^\infty \frac{\lambda^k \mathbb{E}[X^k]}{k!} \leq 1 + \sum_{k=2}^\infty \frac{(k|\lambda|\sigma)^k}{k!} \leq 1 + \sum_{k=2}^\infty (e|\lambda|\sigma)^k,$$

where we have used that $k! \geq (k/e)^k$. When $|\lambda| < \frac{1}{\sigma e}$, evaluating the geometric series yields

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \frac{(e\lambda\sigma)^2}{1 - e|\lambda|\sigma}.$$

For $|\lambda| \leq \frac{1}{2e\sigma}$, we obtain $\mathbb{E}[e^{\lambda X}] \leq 1 + 2e^2\sigma^2\lambda^2$, and as $1 + x \leq e^x$ this implies (4).

For the opposite direction, assume (4) holds with $K_4 = K_4' = 1$. Then $\mathbb{E}[\exp(\lambda X/\sigma)] \leq \exp(1)$ for $\lambda \in [-1, 1]$, and (3) holds. The preceding parts imply the remainder of the equivalence.

4.4.3 Proof of Proposition 4.3.2

The key insight is to rewrite the matrix exponential e^{A+B} as a limit of sums of matrices, then work more directly with traces of powers. To that end, we use the Lie product formula

$$\lim_{n\to\infty} \left(\exp(A/n)\exp(B/n)\right)^n = \exp(A+B). \tag{4.4.1}$$

We leave the proof of the equality (4.4.1) as Exercise 4.16. Using it, however, it is evidently sufficient to prove that there exists some sequence of integers $n \to \infty$ where along this sequence,

$$\operatorname{tr}\left(\left(e^{A/n}e^{B/n}\right)^n\right) \le \operatorname{tr}(e^A e^B). \tag{4.4.2}$$

Letting $\gamma(A)$ be the vector of A's singular values, recall that the Schatten p-norm of a matrix $\|A\|_p := \operatorname{tr}((AA^*)^{p/2})^{1/p} = \|\gamma(A)\|_p$, the ℓ_p-norm of its singular values, where $p = 2$ gives the Euclidean or Frobenius norm $\|A\|_2 = (\sum_{i,j} |A_{ij}|^2)^{1/2}$. This norm gives a generalized Hölder-type inequality for powers of 2, that is, $n \in \{2^k\}_{k\in\mathbb{N}}$, which we can in turn use to prove the Golden-Thompson inequality. In particular, we demonstrate that for n a power of 2,

$$|\operatorname{tr}(A_1 \cdots A_n)| \le \|A_1\|_n \cdots \|A_n\|_n . \tag{4.4.3}$$

To see this inequality, we proceed inductively. Because the trace defines the inner product $\langle A, B\rangle = \operatorname{tr}(A^*B)$, for $n = 2$, the Cauchy-Schwarz inequality implies

$$|\operatorname{tr}(A_1 A_2)| = |\langle A_1^*, A_2\rangle| \le \|A_1\|_2 \|A_2\|_2 .$$

We now perform an induction, where we have demonstrated the base case $n = 2$. Then for $n \ge 4$ a power of 2, we have by the inductive hypothesis that inequality (4.4.3) holds for $n/2$ that

$$|\operatorname{tr}(A_1 \cdots A_n)| \le \|A_1A_2\|_{n/2} \cdots \|A_{n-1}A_n\|_{n/2} .$$

Now consider an arbitrary pair of matrices A, B. We will demonstrate that $\|AB\|_{n/2} \le \|A\|_n \|B\|_n$, which will then evidently imply inequality (4.4.3). For these, we have

$$\|AB\|_{n/2}^{n/2} = \operatorname{tr}(\underbrace{ABB^*A^* \cdots ABB^*A^*}_{n/4 \text{ times}}) = \operatorname{tr}\left((A^*ABB^*)^{n/4}\right)$$

by the cyclic property of the trace. Using the inductive hypothesis again with $n/4$ copies of each of the matrices A^*A and BB^*, we thus have

$$\begin{aligned}\|AB\|_{n/2}^{n/2} &\le \operatorname{tr}\left((A^*ABB^*)^{n/4}\right) \\ &\le \|A^*A\|_{n/2}^{n/4} \|BB^*\|_{n/2}^{n/4} = \operatorname{tr}\left((A^*A)^{n/2}\right)^{1/2} \operatorname{tr}\left((BB^*)^{n/2}\right)^{1/2} = \|A\|_n^{n/2} \|B\|_n^{n/2} .\end{aligned}$$

That is, we have $\|AB\|_{n/2} \le \|A\|_n \|B\|_n$ for any A, B as desired, giving inequality (4.4.3).

We apply inequality (4.4.3) to powers of products of Hermitian matrices A, B. We have

$$\operatorname{tr}\left((AB)^n\right) \le \|AB\|_n^n = \operatorname{tr}\left((ABB^*A^*)^{n/2}\right) = \operatorname{tr}\left((A^*ABB^*)^{n/2}\right) = \operatorname{tr}\left((A^2B^2)^{n/2}\right)$$

because $A = A^*$ and $B = B^*$. Recognizing that A^2 and B^2 are Hermitian, we repeat this argument to obtain

$$\operatorname{tr}\left((A^2B^2)^{n/2}\right) \le \operatorname{tr}\left((A^4B^4)^{n/4}\right) \le \cdots \le \operatorname{tr}(A^nB^n)$$

for $n \in \{2^k\}_{k\in\mathbb{N}}$. Replacing A and B by e^A and e^B, which are Hermitian positive definite, we obtain

$$\operatorname{tr}\left((e^Ae^B)^n\right) \le \operatorname{tr}(e^{nA}e^{nB}) \quad \text{for } n \in \{2^k\}_{k\in\mathbb{N}}.$$

This is inequality (4.4.2) once we replace A and B by A/n and B/n.

4.5 Bibliography and further reading

There is a plethora of material, both pedagogical and research-focused, on concentration inequalities, random matrices, and their connections with entropy. Notable references include Vershynin's extraordinarily readable lecture notes [412], upon which our proof of Theorem 4.1.11 is based, and his book [413]. The comprehensive book of Boucheron, Lugosi, and Massart [81] provides coverage of many of the most important techniques in concentration, including the "entropy method" (which we do not cover here) and several information-theoretic methods we will develop later in this book. Buldygin and Kozachenko [92] provide a deep treatment of sub-Gaussian-like random variables and metric structures one may build on top of them, especially focusing on spaces of random processes.

There exist many extensions of the martingale inequalities we present in Section 4.2, particularly to cases where we control the variance of the martingales. Freedman's paper [199] provides an early version of these ideas, which have had many applications, including in optimal rates of convergence in online learning [261, 8]. Howard, Ramdas, McAuliffe, and Sekhon [242] give sharp concentration inequalities for many stochastic processes, generalizing the classical Bennett, Bernstein, Hoeffding, and Freedman inequalities, extending them to settings including Banach-space valued martingales.

The motivation we provide of the Johnson-Lindenstrauss lemma arises from work in computer science on fast methods for nearest neighbors. This line of research begins with Indyk and Motwani [249], and continues through a variety of works, including Indyk [248] and work on locality-sensitive hashing by Andoni et al. [18]. Johnson and Lindenstrauss [258] developed their original approach in the study of Banach spaces [258]. Computing low-dimensional embeddings of high-dimensional data also admits several practical considerations impacted by theoretical contributions, with work on achieving sharper constants [128] and how to use structured matrices to compute the embeddings Φu faster (see, for example, Achlioptas [6], and Ailon and Chazelle [15]).

Ahlswede and Winter [13] pioneered the Chernoff-bound and trace inequality approach to matrix concentration inequalities; Petz [342] provides an excellent survey of trace inequalities that forms the mathematical basis for much of this work. Our versions of the matrix-Hoeffding and matrix-Bernstein inequalities in Section 4.3* are somewhat loose, and it is possible to do much better. We have borrowed the proof of Lemma 4.3.3 from Tropp [399, Lemma 7.6]. By using a sharper convexity inequality than the Golden-Thompson bound, specifically, Lieb's concavity inequality that

$$f(A) := \operatorname{tr}\left(\exp(H + \log A)\right) \tag{4.5.1}$$

is concave in $A \succ 0$, one can provide sharper concentration. Tropp [399] develops this approach, Exercises 4.19 and 4.20 outline it. Summarizing, if $X_i \in \mathcal{H}_d$ are independent mean-zero Hermitian matrices with $X_i^2 \preceq A_i^2$ for each i, then the sum $S_n := \sum_{i=1}^n X_i$ satisfies

$$\mathbb{P}(\|S_n\|_{\rm op} \geq t) \leq 2d \exp\left(-\frac{t^2}{8\sigma^2}\right) \quad \text{where} \quad \sigma^2 = \left\|\sum_{i=1}^n A_i^2\right\|_{\rm op}, \tag{4.5.2}$$

which improves the matrix Hoeffding inequality (Corollary 4.3.5). Similarly, if $X_i \in \mathcal{H}_d$ satisfy that $\|X_i\|_{\rm op} \leq b$ and the semidefinite variance bound $\mathbb{E}[X_i^2] \preceq A_i^2$, then

$$\mathbb{P}\left(\|S_n\|_{\rm op} \geq t\right) \leq 2d \exp\left(-c \min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right), \quad \text{where} \quad \sigma^2 = \left\|\sum_{i=1}^n A_i^2\right\|_{\rm op}, \tag{4.5.3}$$

improving Theorem 4.3.6, the matrix Bernstein inequality. Mackey et al. [298] use the method of exchangeable pairs to develop strong concentration inequalities for random matrices. The proof of the Golden-Thompson inequality (Proposition 4.3.2) comes from Terry Tao's blog [392].

4.6 Exercises

Exercise 4.1 (Sharper concentration via randomization)**:** Throughout this question, let $U \sim \mathsf{Uniform}[0, 1]$ be an independent uniform random variable.

(a) Show that if $X \geq 0$, then $\mathbb{P}(X \geq Ut) \leq \mathbb{E}[X]/t$ for all $t \geq 0$. Why is this stronger than Markov's inequality (Proposition 4.1.1)?

(b) Give a sufficient condition for part (a) to be sharp.

(c) Let X be a mean-zero σ^2-sub-Gaussian random variable. Show that

$$\mathbb{P}\left(X \geq \sigma t + \frac{\sigma \log U}{t}\right) \leq \exp\left(-\frac{t^2}{2}\right).$$

Why is this stronger than Proposition 4.1.8?

Exercise 4.2 (Variance lower bounds on sub-Gaussian parameters)**:**

(a) Let X be σ^2-sub-Gaussian, that is, $\mathbb{E}[\exp(\lambda X)] \leq \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbb{R}$. Show that $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] \leq \sigma^2$.

(b) Let X be a random variable with $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = \sigma^2 > 0$. Show that

$$\liminf_{|\lambda| \to \infty} \frac{1}{|\lambda|} \log \mathbb{E}[\exp(\lambda X)] > 0.$$

Exercise 4.3 (Mills ratio)**:** Let $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ be the density of a standard Gaussian, $Z \sim \mathsf{N}(0, 1)$, and $\Phi(t) = \int_{-\infty}^{t} \phi(u)du$ its cumulative distribution function.

(a) Show that $\mathbb{P}(Z \geq t) \leq \frac{1}{t}\phi(t)$ for all $t \geq 0$. *Hint.* Define $g(t) = 1 - \Phi(t) - \frac{1}{t}\phi(t)$ and argue that it is negative for $t \geq 0$.

(b) Define

$$g(t) := 1 - \Phi(t) - \frac{t}{t^2 + 1}\phi(t).$$

Show that $g(0) = \frac{1}{2}$, $g'(t) < 0$ for all $t \geq 0$, and that $\lim_{t\to\infty} g(t) = 0$.

(c) Conclude that for all $t \geq 0$,

$$\frac{t}{t^2 + 1}\phi(t) \leq \mathbb{P}(Z \geq t) \leq \frac{1}{t}\phi(t).$$

Exercise 4.4 (Likelihood ratio bounds and concentration)**:** Consider a data release problem, where given a sample x, we release a sequence of data $Z_1, Z_2, \ldots, Z_n$ belonging to a discrete set $\mathcal{Z}$, where Z_i may depend on Z_1^{i-1} and x. We assume that the data has limited information about x in the sense that for any two samples x, x', we have the likelihood ratio bound

$$\frac{p(z_i \mid x, z_1^{i-1})}{p(z_i \mid x', z_1^{i-1})} \leq e^{\varepsilon}.$$

Let us control the amount of "information" (in the form of an updated log-likelihood ratio) released by this sequential mechanism. Fix x, x', and define

$$L(z_1, \ldots, z_n) := \log \frac{p(z_1, \ldots, z_n \mid x)}{p(z_1, \ldots, z_n \mid x')}.$$

(a) Show that, assuming the data Z_i are drawn conditional on x,

$$\mathbb{P}\left(L(Z_1, \ldots, Z_n) \geq n\varepsilon(e^\varepsilon - 1) + t\right) \leq \exp\left(-\frac{t^2}{2n\varepsilon^2}\right).$$

Equivalently, show that

$$\mathbb{P}\left(L(Z_1, \ldots, Z_n) \geq n\varepsilon(e^\varepsilon - 1) + \varepsilon\sqrt{2n\log(1/\delta)}\right) \leq \delta.$$

(b) Let $\gamma \in (0, 1)$. Give the largest value of ε you can that is sufficient to guarantee that for any test $\Psi : \mathcal{Z}^n \to \{x, x'\}$, we have

$$P_x(\Psi(Z_1^n) \neq x) + P_{x'}(\Psi(Z_1^n) \neq x') \geq 1 - \gamma,$$

where P_x and $P_{x'}$ denote the sampling distribution of Z_1^n under x and x', respectively.

Exercise 4.5 (Marcinkiewicz-Zygmund inequality)**:** Let X_i be independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^p] < \infty$, where $1 \leq p < \infty$. Prove that

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right] \leq C_p \mathbb{E}\left[\left(\sum_{i=1}^n |X_i|^2\right)^{p/2}\right]$$

where C_p depends only on p. As a corollary, derive that if $\mathbb{E}[|X_i|^p] \leq \sigma^p$ and $p \geq 2$, then

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right|^p\right] \leq C_p \frac{\sigma^p}{n^{p/2}}.$$

Hint. For any fixed $x \in \mathbb{R}^n$, if ε_i are i.i.d. uniform signs $\varepsilon_i \in \{\pm 1\}$, then $\varepsilon^T x$ is sub-Gaussian.

Exercise 4.6 (A vector Marcinkiewicz-Zygmund inequality)**:** Let $X_i \in \mathbb{R}^d$ be independent vectors with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[\|X_i\|_2^p] < \infty$, where $1 \leq p < \infty$. Prove that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|_2^p\right] \leq C_p \mathbb{E}\left[\left(\sum_{i=1}^n \|X_i\|_2^2\right)^{p/2}\right]$$

where C_p is a constant depending only on p.

Exercise 4.7 (Small balls and anti-concentration)**:** Let X be a nonnegative random variable satisfying $\mathbb{P}(X \leq \epsilon) \leq c\epsilon$ for some $c < \infty$ and all $\epsilon > 0$. Argue that if X_i are i.i.d. copies of X, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \geq 1 - \exp(-2n\left[1/2 - 2ct\right]_+^2) \text{ for all } t.$$

Exercise 4.8 (Lipschitz functions remain sub-Gaussian)**:** Let X be σ^2-sub-Gaussian and $f : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz, meaning that $|f(x) - f(y)| \le L|x - y|$ for all x, y. Prove that there exists a numerical constant $C < \infty$ such that $f(X)$ is $CL^2\sigma^2$-sub-Gaussian.

Exercise 4.9 (Sub-Gaussian maxima)**:** Let $X_1, \ldots, X_n$ be σ^2-sub-Gaussian mean-zero (not necessarily independent) random variables. Show that

(a) $\mathbb{E}[\max_i X_i] \le \sqrt{2\sigma^2 \log n}$.

(b) There exists a numerical constant $C < \infty$ such that $\mathbb{E}[\max_i |X_i|^p] \le (Cp\sigma^2 \log n)^{p/2}$.

Exercise 4.10: Let $Z \sim \mathsf{N}(0, 1)$.

(a) Use the Cauchy-Schwarz inequality to show that

$$\mathbb{E}\left[\exp(\lambda((\mu + Z)^2 - \mu^2))\right] \le \exp\left(4\lambda^2\mu^2 + \lambda + \frac{2\lambda^2}{[1 - 4|\lambda|]_+}\right).$$

(b) Use a direct integration argument, as in Examples 4.1.12 and 4.1.13, to show that

$$\mathbb{E}\left[\exp(\lambda(\mu + Z)^2)\right] \le \exp\left(\lambda\mu^2 + \frac{2\lambda^2}{1 - 2\lambda} - \frac{1}{2}\log(1 - 2\lambda)\right)$$

for $\lambda < \frac{1}{2}$. Use this to prove that

$$\mathbb{E}\left[\exp\left(\lambda\left((\mu + Z)^2 - (\mu^2 + 1)\right)\right)\right] \le \exp\left(\frac{2\lambda^2\mu^2}{1 - 2\lambda} + \frac{\lambda^2}{[1 - 2|\lambda|]_+}\right).$$

Hint. It may be useful to use that $-\log(1 - x) \le -x + \frac{x^2}{2[1-|x|]_+}$ for all $x \in \mathbb{R}$.

Exercise 4.11: Let $Z \sim \mathsf{N}(0, I_d)$, and let $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ be otherwise arbitrary. Using the result of Exercise 4.10, show that $\|AZ - b\|_2^2$ is $(4(\|A\|_{\mathrm{Fr}}^2 + 2\|b\|_2^2), 4\|A\|_{\mathrm{op}}^2)$-sub-exponential. *Hint.* Use the singular value decomposition $A = U\Gamma V^\top$ of A, and note that $V^\top Z \sim \mathsf{N}(0, I_d)$.

Exercise 4.12: Let $s(p) = \frac{1-2p}{\log\frac{1-p}{p}}$. Show that s is concave on $[0, 1]$.

Exercise 4.13: This exercise demonstrates the classical Bernstein inequality (4.1.10) by providing concrete upper bounds on Bennett's inequality. In particular, defining $h(t) = (1 + t)\log(1 + t) - t$, we will show that

$$h(t) \ge \frac{t^2}{2(1 + t/3)}.$$

(a) Fix $c > 0$ and define $g(t) = \frac{t^2}{1+ct}$. Show that $g(0) = 0$, $g'(0) = 0$, and $g''(t) = \frac{2}{(1+ct)^3}$.

(b) Show that $\frac{1}{2}g''(t) \le \frac{1}{1+t}$ for all $t \ge 0$ if and only if $c \ge \frac{1}{3}$.

(c) By comparing derivatives, conclude that $g(t) \le 2h(t)$ for all t when $c \ge \frac{1}{3}$. Show that this implies inequality (4.1.10).

Exercise 4.14 (A naive randomized codebook)**:** Consider a *memoryless binary symmetric channel* Q, given an input bit $b \in \{0,1\}$, outputs $z \in \{0,1\}$ with $Q(Z = b \mid b) = 1-\epsilon$ and $Q(Z = 1-b \mid b) = \epsilon$, and given $b_1^n \in \{0,1\}^n$, outputs $Q(Z_1^n = z_1^n \mid b_1^n) = \prod_{i=1}^n Q(Z_i = z_i \mid b_i)$. We assume we have a codebook $\{x_i\}_{i=1}^m$ of codewords $x_i \in \{0,1\}^d$. To send a message i, we send the bit string $x_i \in \{0,1\}^d$ through the channel Q, which generates $Z \in \{0,1\}^d$ with conditionally independent coordinates $Q(Z_j = x_{ij} \mid x_{ij}) = 1 - \epsilon$.

(a) Assume that for some $c > 0$, we have$\|x_i - x_j\|_1 \geq cd$ for each pair $i \neq j$. Show that if $Q(\cdot \mid x_i)$ denotes the distribution of Z conditional on transmitting x_i, then

$$Q\left(\|Z - x_i\|_1 \geq \min_{j \neq i} \|Z - x_j\|_1 \mid x_i\right) \leq m \exp\left(-\frac{1}{2}cd(1-2\epsilon)^2\right).$$

(b) Show that for x_i, $i = 1, \ldots, m$, uniform as above,

$$\mathbb{P}\left(\text{any pair } i \neq j \text{ satisfies } \|x_i - x_j\|_1 < \frac{d}{2} - dt\right) \leq \binom{m}{2} \exp\left(-2dt^2\right).$$

(c) Define the *maximum likelihood decoder* $\widehat{i} = \operatorname{argmin}_{i \in [m]} \|Z - x_i\|_1$. Show that if the codebook $\{x_i\}_{i=1}^m$ has elements uniform on $\{0,1\}^d$ and

$$d \geq \max\left\{32 \log m, \frac{8 \log \frac{m}{\delta}}{(1-2\epsilon)^2}\right\},$$

then with probability at least $1 - 1/m$ over the draw of the codebook, the maximum probability of error in transmitting a message is at most δ, that is, $\max_{i \leq m} Q(\widehat{i} \neq i \mid x_i) \leq \delta$.

Exercise 4.15 (Inner products on complex matrices)**:** Recall that $\langle \cdot, \cdot \rangle$ is a *complex* inner product on a vector space V if it satisfies the following for all $x, y, z \in V$:

(i) $\langle x, x \rangle \geq 0$, with $\langle x, x \rangle = 0$ if and only if $x = 0$.

(ii) It is conjugate symmetric, so that $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

(iii) It is conjugate linear in its first argument, so that $\langle \alpha x + y, z \rangle = \overline{\alpha} \langle x, z \rangle + \langle y, z \rangle$ for all $\alpha \in \mathbb{C}$.

The vector space V is real with real inner product if property (ii) is replaced with the symmetry $\langle x, y \rangle = \langle y, x \rangle$ and linearity (iii) holds for $\alpha \in \mathbb{R}$.

(a) Show that the space of complex $m \times n$ matrices $\mathbb{C}^{m \times n}$ has complex inner product $\langle A, B \rangle := \operatorname{tr}(A^* B)$.

(b) Show that the space $\mathcal{H}_n$ of $n \times n$ Hermitian matrices is a real vector space with inner product $\langle A, B \rangle := \operatorname{tr}(A^* B)$, and that consequently $\langle A, B \rangle \in \mathbb{R}$.

Exercise 4.16 (The Lie product formula)**:** Let A and B be symmetric (or Hermitian) matrices.

(a) Prove the Lie product formula (4.4.1), that is,

$$\lim_{n \to \infty} \left(\exp(A/n) \exp(B/n)\right)^n = \exp(A + B).$$

Hint. One argument proceeds as follows. Let $O(\epsilon)$ denote a matrix E such that $\|E\|_{\rm op} \lesssim \epsilon$. First, demonstrate that

$$e^{A/n} = I + \frac{1}{n}A + O(n^{-2}).$$

Then show that for any matrix A, we have $(I + n^{-1}A + o(n^{-1}))^n \to \exp(A)$. Combine these.

(b) Give an example of matrices A and B that do not commute and for which $\exp(A + B) \neq \exp(A)\exp(B)$.

Exercise 4.17: Define the trace exponential function $f(X) := \tr(e^X)$ on the Hermitian matrices.

(a) Prove that f is monotone for the semidefinite order, that is, if $X \preceq Y$, then $f(X) \leq f(Y)$. *Hint.* It is enough to show that for any $A \succeq 0$, the one-dimensional function $h(t) := f(X + tA)$ is monotone in t, or even that $h'(0) \geq 0$.

(b) Prove that the trace exponential is convex on the Hermitian matrices, that is, $f(X) := \tr(e^X)$ is convex. *Hint.* It is enough to show that for any X, V Hermitian that $h(t) := f(X + tV)$ is convex in t, for which it in turn suffices to show that $h''(0) \geq 0$.

Exercise 4.18 (The matrix-Bernstein inequality)**:** In this question, we prove Theorem 4.3.6.

(a) Let $X_i \in \mathcal{H}_d$ be independent Hermitian matrices and $S_n = \sum_{i=1}^n X_i$. Use the Golden-Thompson inequality (Proposition 4.3.2) to show that for all $\lambda \in \mathbb{R}$,

$$\tr(\mathbb{E}[e^{\lambda S_n}]) \leq d \prod_{i=1}^n \left\|\mathbb{E}[e^{\lambda X_i}]\right\|_{\rm op}.$$

(b) Extend Example 4.1.15 to the matrix-valued case. Demonstrate that if X is a mean-zero Hermitian random matrix with $\|X\|_{\rm op} \leq b$, then for all $|\lambda| < \frac{3}{b}$,

$$\mathbb{E}[\exp(\lambda X)] \preceq I + \frac{1}{1 - b|\lambda|/3}\frac{\lambda^2 \mathbb{E}[X^2]}{2}.$$

(c) Use parts (a) and (b) to show that for $|\lambda| \leq \frac{3}{2b}$,

$$\tr(\mathbb{E}[e^{\lambda S_n}]) \leq d \exp\left(\lambda^2 \sum_{i=1}^n \sigma_i^2\right).$$

(d) Prove Theorem 4.3.6.

Exercise 4.19: We use Lieb's concavity inequality (4.5.1) to obtain a stronger matrix Hoeffding inequality. Let $X_1, \ldots, X_n \in \mathcal{H}_d$ be independent mean-zero Hermitian matrices and $S_n = \sum_{i=1}^n X_i$.

(a) Let X be a random Hermitian matrix and $X^2 \preceq A^2$. Show that

$$\log \mathbb{E}[e^{\lambda \varepsilon X} \mid X] \preceq \frac{\lambda^2}{2}A^2.$$

Hint. Use that the matrix logarithm is operator monotone, that is, if $A \preceq B$, then $\log A \preceq \log B$.

(b) Assume that $X_n^2 \preceq A_n^2$. Show that

$$\mathbb{E}[\mathrm{tr}(\exp(\lambda S_n))] \le \mathbb{E}\left[\mathrm{tr}\exp\left(\lambda S_{n-1} + 2\lambda^2 A_n^2\right)\right].$$

(c) Show that if $X_i^2 \preceq A_i^2$ for each i, then

$$\mathbb{E}[\mathrm{tr}(\exp(\lambda S_n))] \le \exp\left(2\lambda^2 \sum_{i=1}^n A_i^2\right).$$

(d) Show that if $\sigma^2 \ge \|\sum_{i=1}^n A_i^2\|_{\mathrm{op}}$, then for $t \ge 0$,

$$\mathbb{P}(\lambda_{\max}(S_n) \ge t) \le d\exp\left(-\frac{t^2}{8\sigma^2}\right).$$

(e) Give an example of random Hermitian matrices where the preceding bound is much sharper than Corollary 4.3.5.

Exercise 4.20: In this question, we use Lieb's concavity inequality (4.5.1) to demonstrate a sharper matrix Bernstein inequality than Theorem 4.3.6.

(a) Define the matrix cumulant generating function $\phi_{X_i}(\lambda) := \log \mathbb{E}[\exp(\lambda X_i)]$. Show that

$$\mathbb{E}[\mathrm{tr}(\exp(\lambda S_n))] \le \mathrm{tr}\exp\left(\sum_{i=1}^n \phi_{X_i}(\lambda)\right).$$

(b) Using Exercise 4.18 part (b), show that if $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] \preceq \Sigma_i$, and $\|X_i\|_{\mathrm{op}} \le b$ for each i, then for $|\lambda| < \frac{3}{b}$, we have

$$\mathbb{E}[\mathrm{tr}(e^{\lambda S_n})] \le d\exp\left(\frac{1}{1 - b|\lambda|/3}\frac{\lambda^2}{2}\sigma^2\right) \quad \text{where} \quad \sigma^2 := \left\|\sum_{i=1}^n \Sigma_i\right\|_{\mathrm{op}}.$$

(c) Show that there exists a numerical constant $c > 0$ such that for all $t \ge 0$,

$$\mathbb{P}\left(\|S_n\|_{\mathrm{op}} \ge t\right) \le 2d\exp\left(-c\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right).$$

Why is this sharper than Theorem 4.3.6?

Exercise 4.21 (Bounds on the incomplete Gamma function)**:** For many concentration inequalities, it is useful to bound the incomplete Gamma function, which for $p > 0$ and $a \ge 0$ is

$$\Gamma(p, a) := \int_a^\infty t^{p-1}e^{-t}dt.$$

(a) Use integration by parts to show that $(1 - \frac{p-1}{a})\Gamma(p, a) \le a^{p-1}e^{-a}$ for $p \ge 1$ and $a > 0$.

(b) Let $C > 1$ and $p \ge 1$. Show that if $a \ge \frac{C}{C-1}(p-1)$, then

$$\Gamma(p, a) \le Ca^{p-1}e^{-a}.$$

Chapter 5

Estimation and generalization

We put concentration inequalities to use in demonstrating that a variety of estimation, learning, and related procedures converge. We will investigate two related desiderata, one arising more from machine learning and the other from statistical estimation. Prediction problems best motivate the first. Imagine a binary classification problem with data in pairs (X, Y), $Y \in \{-1, 1\}$, where we seek a function $f : \mathcal{X} \to \mathbb{R}$ that typically predicts the correct sign of Y, minimizing

$$P(Yf(X) \le 0)$$

over f in a class $\mathcal{F}$. Because we only have a sample $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i,Y_i}$ of the population P, we can measure performance only on P_n, and so we develop guarantees that P_n and P are close in that

$$|P_n(Yf(X) \le 0) - P(Yf(X) \le 0)|$$

is small uniformly across $f \in \mathcal{F}$.

Finding f to minimize $P_n(Yf(X) \le 0)$ may be computationally challenging, so often we relax the problem to convex minimization, where for a convex decreasing $\ell : \mathbb{R} \to \mathbb{R}_+$, we wish to minimize $L(f) := \mathbb{E}_P[\ell(Yf(X))]$. Figure 5.1 illustrates this technique with the binary logistic loss $\ell(t) = \log(1 + e^{-t})$ superimposed on the zero-one loss $\ell_{0\text{-}1}(t) = \mathbf{1}\{t \le 0\}$. By demonstrating either

i. a *uniform law* for the empirical loss $\widehat{L}_n(f) := \mathbb{E}_{P_n}[\ell(Yf(X))]$, that is, that

$$\sup_{f\in\mathcal{F}} |\widehat{L}_n(f) - L(f)| \to 0, \quad \text{or}$$

ii. that we can construct an estimate $\widehat{f}$ for which $L(\widehat{f}) = \inf_{f\in\mathcal{F}} L(f) + o(1)$,

we obtain predictive guarantees. Sections 5.1 and 5.2 provide tools for such uniform laws.

In Section 5.3 of this chapter, we turn from purely predictive guarantees to estimation. There, we wish to recover the parameter minimizing an expected loss $L(\theta) = \mathbb{E}_P[\ell(\theta, Z)]$, such as the squared error $\ell(\theta, (x, y)) = (\langle\theta, x\rangle - y)^2$, because θ itself encodes a phenomenon of interest.

5.1 Uniformity and metric entropy

The first set of results we consider are *uniform laws of large numbers*, where the goal is to bound means uniformly over different classes of functions. Frequently, such results are called *Glivenko-Cantelli* laws, after the original Glivenko-Cantelli theorem, which shows that empirical distributions

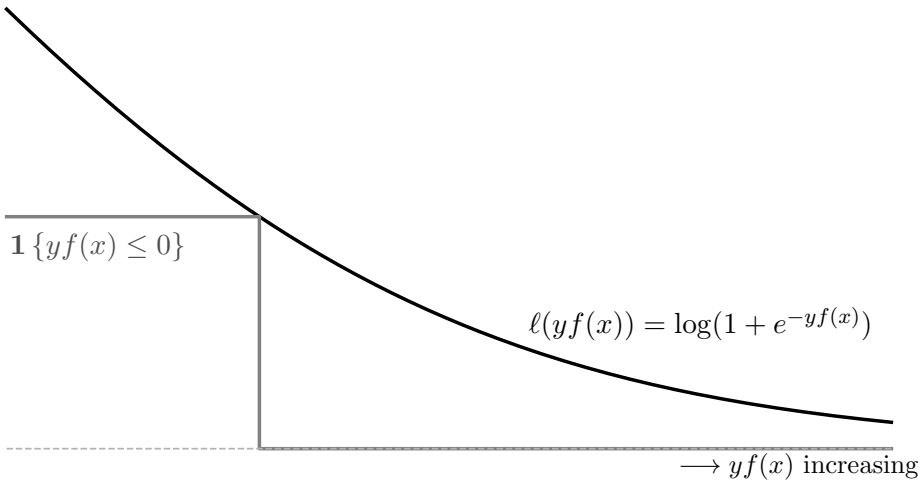


Figure 5.1. Illustration of the margin-based loss $\ell(t) = \log(1 + e^{-t})$ acting as a surrogate for the zero-one loss $\ell_{0\text{-}1}(t) = \mathbf{1}\{t \le 0\}$.

uniformly converge. We will first develop the ideas using bounded differences inequalities, which we follow by introducing metric entropy (volume-like considerations). We will leverage these ideas in the next two sections to develop generalization bounds and guarantees on estimation.

5.1.1 Symmetrization and uniform laws

The starting point is to define uniform laws of large numbers. We adopt notation (as in Example 4.2.9) we will use throughout the remainder of the book, reminding readers as we go. For a sample $X_1, \ldots, X_n$ on a space $\mathcal{X}$, denote the empirical distribution on $\{X_i\}_{i=1}^n$ by

$$P_n := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i},$$

where $\mathbf{1}_{X_i}$ is a point mass at X_i. Then for functions f on $\mathcal{X}$, we let

$$P_n f := \mathbb{E}_{P_n}[f(X)] = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

denote the empirical expectation of f evaluated on the sample, and

$$Pf := \mathbb{E}_P[f(X)] = \int f(x) dP(x)$$

denotes expectation under a probability P. With this notation, we study *uniform laws of large numbers* of the form

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \to 0, \tag{5.1.1}$$

where convergence is in one of several modes of probabilistic convergence. Viewing P_n and P as linear functionals on the space of maps $\mathcal{F}$, we may define the (semi)norm $\|\cdot\|_{\mathcal{F}}$ for $L : \mathcal{F} \to \mathbb{R}$ by

$$\|L\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |L(f)|,$$

in which case the convergence (5.1.1) is equivalent to $\|P_n - P\|_{\mathcal{F}} \to 0$. Thus, roughly, we only ask questions about when random vectors converge to their expectations.[1]

We begin with bounded $f : \mathcal{X} \to [a, b]$ for finite $a \le b$. The bounded differences inequality (Proposition 4.2.6) immediately implies that $\|P_n - P\|_{\mathcal{F}}$ concentrates around its expectation:

Proposition 5.1.1. *Let $\mathcal{F}$ be as above. Then*

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} \ge \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + t\right) \le \exp\left(-\frac{2nt^2}{(b-a)^2}\right) \quad \text{for } t \ge 0.$$

Proof Let P_n and P_n' be empirical distributions differing only in observations X_i and X_i'. Then

$$\begin{aligned}\sup_{f\in\mathcal{F}} |P_n f - Pf| - \sup_{f\in\mathcal{F}} |P_n' f - Pf| &\le \sup_{f\in\mathcal{F}} \left\{|P_n f - Pf| - |P_n' f - Pf|\right\} \\ &\le \frac{1}{n} \sup_{f\in\mathcal{F}} |f(X_i) - f(X_i')| \le \frac{b-a}{n}\end{aligned}$$

by the triangle inequality. The lower bound $\frac{a-b}{n}$ is similar, and Proposition 4.2.6 gives the result. □

Proposition 5.1.1 shows that, to provide high-probability concentration of $\|P_n - P\|_{\mathcal{F}}$, it is (at least when $\mathcal{F}$ is bounded) sufficient to control the expectation $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$. We take this approach through the remainder of this section, developing tools to simplify bounding this quantity.

Our starting point consists of inequalities relating expectations to *symmetrized* quantities, which are frequently easier to control than their non-symmetrized parts. This symmetrization technique is common in probability theory, theoretical statistics, and machine learning. For centered random variables, symmetrized and non-symmetrized quantities have identical expectations to within constants, making it equivalent to analyze the symmetrized quantity [284, 134].

Proposition 5.1.2. *Let X_i be independent random vectors on a space with norm $\|\cdot\|$ and let $\varepsilon_i \in \{-1, 1\}$ be independent random signs. Then for any $p \ge 1$,*

$$2^{-p}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - \mathbb{E}[X_i])\right\|^p\right] \le \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \le 2^p\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|^p\right]$$

In the proof of the upper bound, we could also show the bound

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \le 2^p\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - \mathbb{E}[X_i])\right\|^p\right],$$

so we may analyze whichever is more convenient.

Proof We prove the upper bound, leaving the lower bound as Exercise 5.1. Introduce independent copies X_i' of the X_i, and use Jensen's inequality and the convexity of $\|\cdot\|^p$ to observe that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i'])\right\|^p\right] \le \mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - X_i')\right\|^p\right].$$

Because the distribution of $X_i - X_i'$ is symmetric, $X_i - X_i' \stackrel{\text{dist}}{=} \varepsilon_i(X_i - X_i')$, and thus

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \le \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i(X_i - X_i')\right\|^p\right].$$

[1] Some readers may worry about measurability here. Our applications will be in separable spaces, we may take suprema without concern for measurability, and consequently we omit such details.

Multiplying and dividing by 2^p, Jensen's inequality then gives

$$\mathbb{E}\left[\left\|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right\|^p\right] \le 2^p \mathbb{E}\left[\left\|\frac{1}{2}\sum_{i=1}^n \varepsilon_i (X_i - X_i')\right\|^p\right] \le 2^{p-1}\left[\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|^p\right] + \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i'\right\|^p\right]\right].$$

Recognize that X_i and X_i' have identical distribution. □

5.1.2 Rademacher complexity bounds

To apply the symmetrization ideas to obtain generalization, we (re)introduce *Rademacher complexities*, a general symmetrized object over empirical measures around which researchers have built a substantial set of tools. We use the symmetrized empirical measure

$$P_n^0 := \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{X_i}, \quad \text{so} \quad P_n^0 f = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i).$$

The expectation of $\|P_n^0\|_{\mathcal{F}}$ is thus the Rademacher complexity (Examples 4.2.8 and 4.2.9), so Proposition 5.1.2 immediately implies a Rademacher complexity bound:

Corollary 5.1.3. *Let $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$ and $X_i \stackrel{\text{iid}}{\sim} P$. Then $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \le 2\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$.*

From Corollary 5.1.3, it is evident that by controlling the expectation of the symmetrized process $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$ we can derive concentration inequalities and uniform laws of large numbers. For example, we immediately obtain that for $t \ge 0$, when $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [a, b]$, then

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{F}} \ge 2\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + t\right) \le \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Many examples of uniform laws of large numbers thus reduce to developing bounds on the expectation $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$, which is possible using techniques we develop in Section 5.1.3 and in later chapters. A frequent application of these symmetrization ideas is to risk minimization problems, and to better match the development of these ideas, we return to the notation of Rademacher complexities, so that $R_n(\mathcal{F}) := \mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$. We begin with a now classical result [304].

Proposition 5.1.4 (Massart's finite class bound)**.** *Let $\mathcal{F}$ be any collection of functions with $f : \mathcal{X} \to \mathbb{R}$, and assume that $\sigma_n^2 := n^{-1}\mathbb{E}[\max_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)^2] < \infty$. Then*

$$R_n(\mathcal{F}) \le \frac{\sqrt{2\sigma_n^2 \log(2|\mathcal{F}|)}}{\sqrt{n}}.$$

Proof For fixed x_1^n, the random variable $\sum_{i=1}^n \varepsilon_i f(x_i)$ is $\sum_{i=1}^n f(x_i)^2$-sub-Gaussian. Define $\sigma^2(x_1^n) := n^{-1}\max_{f\in\mathcal{F}} \sum_{i=1}^n f(x_i)^2$. Because $\mathbb{E}[\max_{j\le n} Z_j] \le \sqrt{2\sigma^2 \log n}$ if the Z_j are σ^2-sub-Gaussian (Exercise 4.9),

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{\sqrt{2\sigma^2(x_1^n)\log(2|\mathcal{F}|)}}{\sqrt{n}}.$$

Jensen's inequality ensures that $\mathbb{E}[\sqrt{\cdot}] \le \sqrt{\mathbb{E}[\cdot]}$, giving the result. $\square$

A refinement of Massart's finite class bound applies when $\mathcal{F}$ may be infinite but, on a collection $X_1, \ldots, X_n$, the functions $f \in \mathcal{F}$ take fewer values. We define the *empirical shatter coefficient* of a collection of points $x_1, \ldots, x_n$ by $\mathsf{S}_{\mathcal{F}}(x_1^n) := \mathrm{card}\{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$, the number of distinct vectors $(f(x_1), \ldots, f(x_n))$ over $f \in \mathcal{F}$. The *shatter coefficient* $\mathsf{S}_{\mathcal{F}}(n) := \sup_{x_1^n} \mathsf{S}_{\mathcal{F}}(x_1^n)$ is the maximum of the empirical shatter coefficients over $x_1^n \in \mathcal{X}^n$. It is clear that $\mathsf{S}_{\mathcal{F}}(n) \le |\mathcal{F}|$ always, but by only counting distinct values, we have the following corollary.

Corollary 5.1.5. *Let $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ and $\sigma_n^2 := n^{-1}\mathbb{E}[\max_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)^2] < \infty$. Then*

$$R_n(\mathcal{F}) \le \frac{\sqrt{2\sigma_n^2 \log(2\mathsf{S}_{\mathcal{F}}(n))}}{\sqrt{n}}.$$

Typical classes with small shatter coefficients include Vapnik-Chervonenkis classes of functions; we do not discuss these here, instead referring to any of several broad treatments [411, 407, 375, 422].

Among the most useful of the calculus rules are the *comparison inequalities* for Rademacher sums, which allow us to compose function classes and maintain small complexity measurers. We state it here; the proof is complex, so we provide references in the bibliographic discussion.

Theorem 5.1.6 (Ledoux-Talagrand Contraction)**.** *Let $T \subset \mathbb{R}^n$ be an arbitrary set and let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and satisfy $\phi_i(0) = 0$. Then for any nondecreasing convex function $\Phi : \mathbb{R} \to \mathbb{R}_+$,*

$$\mathbb{E}\left[\Phi\left(\frac{1}{2}\sup_{t \in T}\left|\sum_{i=1}^n \phi_i(t_i)\varepsilon_i\right|\right)\right] \le \mathbb{E}\left[\Phi\left(\sup_{t \in T}|\langle t, \varepsilon\rangle|\right)\right].$$

Corollaries suggests the theorem's power and applicability. Let $\phi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz, and for a function class $\mathcal{F}$ define $\phi \circ \mathcal{F} = \{\phi \circ f \mid f \in \mathcal{F}\}$. Then the theorem provides control of Rademacher complexities of compositions.

Corollary 5.1.7. *Let $\mathcal{F}$ be an arbitrary function class and ϕ be L-Lipschitz. Then*

$$R_n(\phi \circ \mathcal{F}) \le 2LR_n(\mathcal{F}) + |\phi(0)|/\sqrt{n}.$$

Proof The result follows from Theorem 5.1.6; we simply recenter our functions. Indeed, we have

$$\begin{aligned}
R_n(\phi \circ \mathcal{F} \mid x_1^n) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(f(x_i)) - \phi(0)) + \frac{1}{n}\sum_{i=1}^n \varepsilon_i\phi(0)\right|\right] \\
&\le \mathbb{E}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(\phi(f(x_i)) - \phi(0))\right|\right] + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\phi(0)\right|\right] \\
&\le 2LR_n(\mathcal{F}) + \frac{|\phi(0)|}{\sqrt{n}},
\end{aligned}$$

by Theorem 5.1.6, as $g(t) = \phi(t) - \phi(0)$ satisfies $g(0) = 0$, and because $\mathbb{E}[|\sum_{i=1}^n \varepsilon_i|] \le \sqrt{n}$. $\square$

5.1.3 Metric entropy, coverings, and packings

When the class of functions $\mathcal{F}$ is finite, the union bound guarantees that $P_n f$ is close to Pf for all $f \in \mathcal{F}$. When $\mathcal{F}$ is infinite, we require a different set of tools to address uniform laws. In many cases, the bounded differences inequality in Proposition 5.1.1 means all we really need to do is to control the expectation $\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]$, though the techniques we develop here will have broader use.

The basic object we wish to control is a measure of the size of the space on which we work. We consider vectors $\theta \in \Theta$, where Θ is a non-empty set with an associated (semi)metric ρ. For estimation and in our optimality results in the later chapters, a natural way to measure the size of the set is via the number of balls of a fixed radius $\delta > 0$ required to cover it.

Definition 5.1 (Covering number)**.** *Let Θ be a set with (semi)metric ρ. A* δ-cover *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_N\}$ such that for any point $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ with $\rho(\theta, \theta_v) \le \delta$. The* δ-covering number *of Θ is*

$$N(\delta, \Theta, \rho) := \inf \left\{ N \in \mathbb{N} \mid \text{there exists a } \delta\text{-cover } \theta_1, \ldots, \theta_N \text{ of } \Theta \right\}.$$

For example, any interval $[a, b]$ on the real line has $N(\delta, [a, b], |\cdot|) \asymp (b - a)/\delta$.

The *metric entropy* of the set Θ is the logarithm of its covering number $\log N(\delta, \Theta, \rho)$. We can define a related measure (relevant for constructing lower bounds) of size that relates to the number of disjoint balls of radius $\delta > 0$ that can be placed into the set Θ.

Definition 5.2 (Packing number)**.** *A* δ-packing *of the set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_M\}$ such that for all distinct $v, v' \in \{1, \ldots, M\}$, we have $\rho(\theta_v, \theta_v') \ge \delta$. The* δ-packing number *of Θ is*

$$M(\delta, \Theta, \rho) := \sup \left\{ M \in \mathbb{N} \ : \ \text{there exists a } \delta\text{-packing } \theta_1, \ldots, \theta_M \text{ of } \Theta \right\}.$$

Figures 5.2 and 5.3 give examples of (respectively) a covering and a packing of the same set.

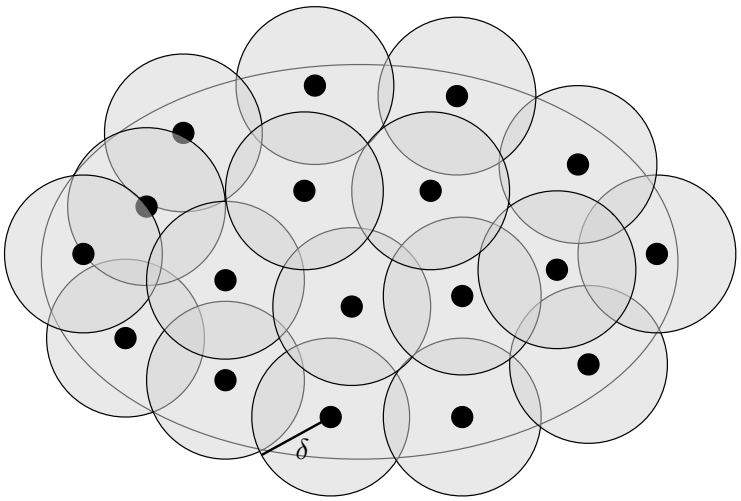


Figure 5.2. A δ-covering of the elliptical set by balls of radius δ.

A proof by contradiction (see Exercise 5.3) relates packing and covering numbers directly:

Lemma 5.1.8. *For $\delta \ge 0$, the packing and covering numbers satisfy the inequalities*

$$M(2\delta, \Theta, \rho) \le N(\delta, \Theta, \rho) \le M(\delta, \Theta, \rho).$$

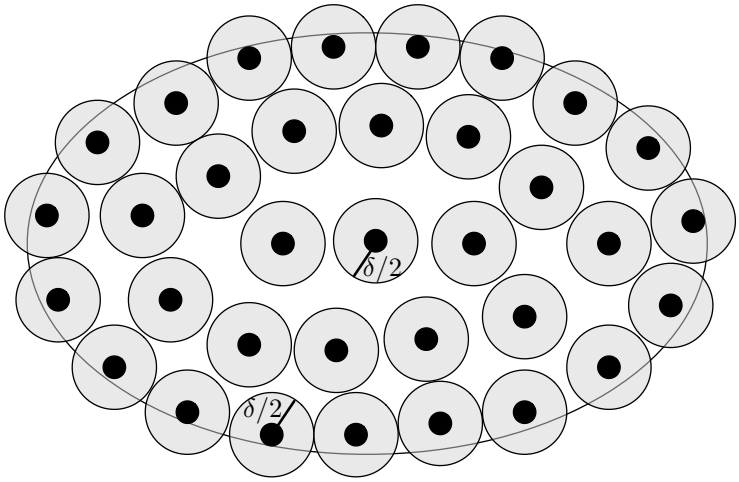


Figure 5.3. A δ-packing of the elliptical set, where balls have radius $\delta/2$. No balls overlap, and each center of the packing satisfies $\|\theta_v - \theta_{v'}\| \geq \delta$.

This lemma shows that packing and covering numbers have the same scaling in the radius δ.

As one example of the metric entropy, consider a set of functions $\mathcal{F}$ with reasonable covering numbers (metric entropy) in the supremum norm.

Example 5.1.9 (The "standard" covering number guarantee)**:** Let $\mathcal{F}$ consist of functions $f : \mathcal{X} \to [-b, b]$ and let the metric $\rho(f, g) = \|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq t\right) \leq \exp\left(-\frac{nt^2}{18b^2} + \log N(t/3, \mathcal{F}, \|\cdot\|_\infty)\right). \tag{5.1.2}$$

When the covering numbers $N(t, \mathcal{F}, \|\cdot\|_\infty)$ grow sub-exponentially in t—so $\log N(t) \ll nt^2$—we have the (essentially) sub-Gaussian tail bound (5.1.2). Example 5.2.12 gives one typical case. Indeed, fix a minimal $t/3$-cover $\{f_1, \ldots, f_N\}$ of $\mathcal{F}$ in $\|\cdot\|_\infty$ of size $N := N(t/3, \mathcal{F}, \|\cdot\|_\infty)$. Then for any $f \in \mathcal{F}$ and the function f_i satisfying $\|f - f_i\|_\infty \leq t/3$, we have

$$|P_n f - Pf| \leq |P_n f - P_n f_i| + |P_n f_i - Pf_i| + |Pf_i - Pf| \leq |P_n f_i - Pf_i| + \frac{2t}{3}.$$

The Azuma-Hoeffding inequality (Theorem 4.2.4) and a union bound guarantee that

$$\mathbb{P}\left(\max_{i \leq N} |P_n f_i - Pf_i| \geq t\right) \leq \exp\left(-\frac{nt^2}{2b^2} + \log N\right).$$

Replace t with $t/3$ to obtain inequality (5.1.2). $\diamond$

The following volume bound on covering with norm balls makes explicit the relationships between packing, covering, and the volume of sets $\Theta \subset \mathbb{R}^d$.

Lemma 5.1.10. *Let $\mathbb{B}$ denote the unit $\|\cdot\|$-ball in $\mathbb{R}^d$. Then*

$$\left(\frac{1}{\delta}\right)^d \leq N(\delta, \mathbb{B}, \|\cdot\|) \leq \left(1 + \frac{2}{\delta}\right)^d.$$

Proof For the lower bound, if the points $v_1, \ldots, v_N$ are a δ-cover of $\mathbb{B}$, then

$$\mathrm{Vol}(\mathbb{B}) \leq \sum_{i=1}^N \mathrm{Vol}(\delta\mathbb{B} + v_i) = N \,\mathrm{Vol}(\delta\mathbb{B}) = N \,\mathrm{Vol}(\mathbb{B})\delta^d.$$

In particular, $N \geq \delta^{-d}$. To upper bound $N(\delta, \mathbb{B}, \|\cdot\|)$, let $\mathcal{V}$ be a δ-packing of $\mathbb{B}$ with maximal cardinality $|\mathcal{V}| = M(\delta, \mathbb{B}, \|\cdot\|) \geq N(\delta, \mathbb{B}, \|\cdot\|)$ (recall Lemma 5.1.8). The collection $\{\delta\mathbb{B} + v_i\}_{i=1}^M$ covers $\mathbb{B}$ (as otherwise, we could put an additional element in the packing $\mathcal{V}$), and the balls $\{\frac{\delta}{2}\mathbb{B} + v_i\}$ are disjoint. Consequently, we find that

$$M\left(\frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}) = M \,\mathrm{Vol}\left(\frac{\delta}{2}\mathbb{B}\right) \leq \mathrm{Vol}\left(\mathbb{B} + \frac{\delta}{2}\mathbb{B}\right) = \left(1 + \frac{\delta}{2}\right)^d \mathrm{Vol}(\mathbb{B}).$$

Dividing both sides by $(\delta/2)^d \,\mathrm{Vol}(\mathbb{B})$ gives the result. □

5.1.4 Application: concentration of covariance matrices

Let us apply Lemma 5.1.10 to concentration of covariance matrices; we explore more in the exercises as well. We consider *sub-Gaussian random vectors*, where $X \in \mathbb{R}^d$ is σ^2-sub-Gaussian if

$$\mathbb{E}[\exp(\langle u, X - \mathbb{E}[X]\rangle)] \leq \exp\left(\frac{\sigma^2}{2}\|u\|_2^2\right) \tag{5.1.3}$$

for all $u \in \mathbb{R}^d$. For example, $X \sim \mathsf{N}(0, I_d)$ is immediately 1-sub-Gaussian, and $X \in [-b, b]^d$ with independent entries is b^2-sub-Gaussian. Let X_i be independent σ^2-sub-Gaussian isotropic random vectors, so $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i X_i^\top] = I_d$. Then we can give concentration guarantees for the sample covariance in the operator norm $\|A\|_{\mathrm{op}} := \sup\{\langle u, Av\rangle \mid \|u\|_2 = \|v\|_2 = 1\}$.

Proposition 5.1.11. *Let X_i be independent isotropic and σ^2-sub-Gaussian vectors. Then there is a numerical constant C such that the sample covariance $\Sigma_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ satisfies*

$$\|\Sigma_n - I_d\|_{\mathrm{op}} \leq C\sigma^2\left[\frac{d + \log\frac{1}{\delta}}{n} + \sqrt{\frac{d + \log\frac{1}{\delta}}{n}}\right]$$

with probability at least $1 - \delta$.

Proof We begin with an intermediate lemma.

Lemma 5.1.12. *Let A be symmetric and $\{u_i\}_{i=1}^N$ be an ϵ-cover of the unit ℓ_2 ball $\mathbb{B}_2^d$. Then*

$$(1 - 2\epsilon)\|A\|_{\mathrm{op}} \leq \max_{i,j \leq N}\langle u_i, Au_j\rangle \leq \|A\|_{\mathrm{op}}.$$

Proof The second inequality is trivial. For the first, fix any $u, v \in \mathbb{B}_2^d$. Then for the i and j such that $\|u - u_i\|_2 \leq \epsilon$ and $\|v - u_j\|_2 \leq \epsilon$, we have

$$\begin{aligned}\langle v, Au\rangle &= \langle v - u_j, Au\rangle + \langle u_j, Au\rangle \\ &= \langle v - u_j, Au\rangle + \langle u_j, A(u - u_i)\rangle + \langle u_j, Au_i\rangle \leq 2\epsilon\|A\|_{\mathrm{op}} + \langle u_j, Au_i\rangle\end{aligned}$$

by definition of the operator norm. Taking a supremum over u and v gives the final result. □

Let $E_i = X_i X_i^\top - I$ and define the average error $\overline{E}_n = \frac{1}{n}\sum_{i=1}^n E_i$. Lemma 5.1.12 shows that

$$(1 - 2\epsilon)\left\|\overline{E}_n\right\|_{\mathrm{op}} \leq \max_{u,v \in \mathcal{N}}\langle u, \overline{E}_n v\rangle$$

for any ϵ-cover $\mathcal{N}$ of the ℓ_2-ball. The error $\langle u, E_i v\rangle = \langle u, X_i\rangle\langle v, X_i\rangle - \langle u, v\rangle$ is sub-exponential, as it is a mean 0 product of sub-Gaussians; in particular, Theorem 4.1.16 and Lemma 4.1.23 imply

$$\mathbb{E}[\exp(\lambda\langle u, E_i u\rangle)] \le \exp\left(C\lambda^2\sigma^4\right) \quad \text{for} \quad |\lambda| \le \frac{1}{C\sigma^2}.$$

for a numerical constant $C < \infty$. Taking $\epsilon = \frac{1}{4}$ in the covering $\mathcal{N}$, a union bound yields

$$\mathbb{P}(\left\|\overline{E}_n\right\|_{\rm op} \ge t) \le \mathbb{P}\left(\max_{u,v\in\mathcal{N}} \langle u, \overline{E}_n v\rangle \ge t/2\right) \le |\mathcal{N}|^2 \cdot \max_{u,v\in\mathcal{N}} \mathbb{P}\left(\langle u, n\overline{E}_n v\rangle \ge nt/2\right).$$

Sums of sub-exponential random variables remain sub-exponential, so Corollary 4.1.19 implies

$$\mathbb{P}\left(\left\|\overline{E}_n\right\|_{\rm op} \ge t\right) \le |\mathcal{N}|^2 \exp\left(-c\min\left\{\frac{nt^2}{\sigma^4}, \frac{nt}{\sigma^2}\right\}\right),$$

where $c > 0$ is a numerical constant. We apply Lemma 5.1.10, which guarantees that $|\mathcal{N}| \le 9^d$. □

5.2 Generalization bounds

We build on our ideas on uniform laws of large numbers and Rademacher complexities to demonstrate their applications in statistical machine learning problems, focusing on *empirical risk minimization* procedures. We consider a setting as follows: we have a sample $Z_1, \ldots, Z_n \in \mathcal{Z}$ drawn i.i.d. according to some (unknown) distribution P, and we have a collection of functions $\mathcal{F}$ from which we wish to select an f that "fits" the data well, according to some loss measure $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. That is, we wish to find a function $f \in \mathcal{F}$ minimizing the *risk*, or *population loss*

$$L(f) := \mathbb{E}_P[\ell(f, Z)]. \tag{5.2.1}$$

In general, however, we only have access to the risk via the empirical distribution of the Z_i, and we often choose f by minimizing the empirical risk

$$\widehat{L}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i). \tag{5.2.2}$$

We then say we have *generalization* if the performance of f on $\widehat{L}_n$ is similar to its performance on the population $L(f)$. Examples make this abstract formulation more concrete.

Example 5.2.1 (Binary classification problems)**:** In the *binary classification problem*, data Z_i come in pairs (X, Y), where $X \in \mathcal{X}$ consists of covariates (independent variables) and $Y \in \{-1, 1\}$ labels example X. The class $\mathcal{F}$ consists of functions $f : \mathcal{X} \to \mathbb{R}$, and the goal is to minimize

$$\mathbb{P}(\text{sign}(f(X)) \ne Y),$$

that is, minimize the risk $\mathbb{E}[\ell(f, Z)]$ for the 0-1 loss $\ell(f, (x, y)) = \mathbf{1}\{f(x)y \le 0\}$. ◇

Example 5.2.2 (Multiclass classification)**:** Instead of $Y \in \{-1, 1\}$ in binary classification, we assume that $Y \in [k] = \{1, \ldots, k\}$ for some $k \ge 2$, and $\mathcal{F}$ consists of functions $f : \mathcal{X} \to \mathbb{R}^k$. The

goal is to find a function f such that, if $Y = y$ correctly labels the point x, then $f_y(x) > f_l(x)$ for all $l \neq y$. That is, we seek $f \in \mathcal{F}$ minimizing

$$\mathbb{P}\left(\text{there exists } l \neq Y \text{ such that } f_l(X) \geq f_Y(X)\right).$$

The loss function is the zero-one loss $\ell(f, (x, y)) = \mathbf{1}\left\{\max_{l \neq y} f_l(x) \geq f_y(x)\right\}$. $\diamond$

Example 5.2.3 (Binary classification with linear functions)**:** In the standard statistical learning setting, the data x belong to $\mathbb{R}^d$, and we assume that a set $\Theta \subset \mathbb{R}^d$ indexes the function class $\mathcal{F}$, so $\mathcal{F} = \{f_\theta : f_\theta(x) = \theta^\top x, \theta \in \Theta\}$. We may use the zero-one loss $\ell_{\rm zo}(f_\theta, (x, y)) := \mathbf{1}\left\{y\theta^\top x \leq 0\right\}$ or the hinge or logistic losses

$$\ell_{\rm hinge}(f_\theta, (x, y)) = \left[1 - yx^\top\theta\right]_+ \quad \text{and} \quad \ell_{\rm logit}(f_\theta, (x, y)) = \log(1 + \exp(-yx^\top\theta)).$$

The (convex) hinge and logistic losses yield computationally tractable estimation problems. $\diamond$

We ask the following main question: given a sample $Z_1, \ldots, Z_n$, if we choose some $\widehat{f}_n \in \mathcal{F}$ based on this sample, can we guarantee that it generalizes to unseen data? In particular, can we guarantee that (with high probability) we have the empirical risk bound

$$\widehat{L}_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \ell(\widehat{f}_n, Z_i) \leq L(\widehat{f}_n) + \epsilon \tag{5.2.3}$$

for some small ϵ? If we allow $\widehat{f}_n$ to be arbitrary, this becomes clearly impossible: consider the classification example 5.2.1, and take the "hash" function that sets $\widehat{f}_n(x) = y$ if the pair (x, y) is in the sample and $\widehat{f}_n(x) = -1$ otherwise. Then $\widehat{L}_n(\widehat{f}_n) = 0$, while there is no useful bound on $L(\widehat{f}_n)$.

5.2.1 Finite and countable classes of functions

To get bounds of the form (5.2.3), we require a few assumptions that are not too onerous. Throughout this section, we assume that for any fixed function f, the loss $\ell(f, Z)$ is σ^2-sub-Gaussian, so

$$\mathbb{E}_P\left[\exp\left(\lambda(\ell(f, Z) - L(f))\right)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \tag{5.2.4}$$

for all $f \in \mathcal{F}$. (Recall the risk functional $L(f) = \mathbb{E}_P[\ell(f, Z)]$.) For example, for the zero-one loss, inequality (5.2.4) holds with $\sigma^2 = \frac{1}{4}$ by Hoeffding's lemma. To guarantee a bound of the form (5.2.4) for a function $\widehat{f}$ dependent on the data, in this section we would like to bound

$$\mathbb{P}\left(\text{there exists } f \in \mathcal{F} \text{ s.t. } L(f) > \widehat{L}_n(f) + t\right) \quad \text{or} \quad \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\widehat{L}_n(f) - L(f)\right| > t\right).$$

Such uniform bounds are sufficient to guarantee that the empirical risk is a good proxy for the true risk L, even when $\widehat{f}_n$ is chosen based on the data.

Recalling that our set of predictors $\mathcal{F}$ is finite or countable, suppose that for $f \in \mathcal{F}$, we have a complexity measure $c(f)$ such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} \leq 1. \tag{5.2.5}$$

This inequality should look familiar to the Kraft inequality (Theorem 2.4.2) from source coding. Essentially, the bound (5.2.5) implies that we can encode each function with code of length $c(f)$ nats, and so the next theorem captures the intuition that if we can encode f, we are guaranteed generalization. Indeed, as soon as we have such a penalty function, we have the following result.

Theorem 5.2.4. *Let the loss ℓ, distribution P on $\mathcal{Z}$, and function class $\mathcal{F}$ be such that $\ell(f, Z)$ is σ^2-sub-Gaussian for each $f \in \mathcal{F}$, and assume that the complexity inequality (5.2.5) holds. Then with probability at least $1 - \delta$ over the sample $Z_{1:n}$,*

$$L(f) \le \widehat{L}_n(f) + \sqrt{2\sigma^2 \frac{\log \frac{1}{\delta} + c(f)}{n}} \quad \text{for all } f \in \mathcal{F}.$$

Proof By sub-Gaussian concentration (Corollary 4.1.10) we have for any $t \ge 0$ and $f \in \mathcal{F}$ that

$$\mathbb{P}\left(L(f) \ge \widehat{L}_n(f) + t\right) \le \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

We replace t by $\sqrt{t^2 + 2\sigma^2 c(f)/n}$ to obtain

$$\mathbb{P}\left(L(f) \ge \widehat{L}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) \le \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right).$$

Then using a union bound, we have

$$\mathbb{P}\left(\exists\, f \in \mathcal{F} \text{ s.t. } L(f) \ge \widehat{L}_n(f) + \sqrt{t^2 + 2\sigma^2 c(f)/n}\right) \le \sum_{f \in \mathcal{F}} \exp\left(-\frac{nt^2}{2\sigma^2} - c(f)\right) \le \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

because $\sum_f e^{-c(f)} \le 1$. Setting $t^2 = 2\sigma^2 \log \frac{1}{\delta}/n$ gives the result. □

As one classical example of this setting, suppose we have a finite class of functions $\mathcal{F}$. Then $c(f) = \log |\mathcal{F}|$ clearly satisfies the summation guarantee (5.2.5), and we obtain

$$L(f) \le \widehat{L}_n(f) + \sqrt{2\sigma^2 \frac{\log \frac{1}{\delta} + \log |\mathcal{F}|}{n}} \quad \text{uniformly for } f \in \mathcal{F}$$

with probability at least $1 - \delta$. To make this even more concrete, consider the following example.

Example 5.2.5 (Floating point classifiers)**:** We implement a linear binary classifier using double-precision floating point values, that is, $f_\theta(x) = \theta^\top x$ for $\theta \in \mathbb{R}^d$ represented using d double-precision floating point numbers. Each coordinate of θ can represent at most 2^{64} values; in total, we have $|\mathcal{F}| \le 2^{64d}$. For the zero-one loss $\ell_{\rm zo}(f_\theta, (x, y)) = \mathbf{1}\left\{\theta^\top xy \le 0\right\}$,

$$L(f_\theta) \le \widehat{L}_n(f_\theta) + \sqrt{\frac{\log \frac{1}{\delta} + 45d}{2n}}$$

for all representable classifiers simultaneously with probability at least $1 - \delta$, as the zero-one loss is $\frac{1}{4}$-sub-Gaussian. (We use that $64 \log 2 < 45$.) ◇

By replacing δ with $\delta/2$ in the bounds of Theorem 5.2.4, a union bound yields the following two-sided corollary.

Corollary 5.2.6. *Under the conditions of Theorem 5.2.4, we have*

$$\left|\widehat{L}_n(f) - L(f)\right| \le \sqrt{2\sigma^2 \frac{\log \frac{2}{\delta} + c(f)}{n}} \quad \text{for all } f \in \mathcal{F}$$

with probability at least $1 - \delta$.

Bernstein-type inequalities, which leverage the variance of $\ell(f, Z)$, provide sharper convergence guarantees. Intuitively, if $f^\star = \operatorname{argmin}_{f\in\mathcal{F}} L(f)$, we might hope that because

$$\operatorname{Var}(\ell(f, Z)) = \mathbb{E}[\ell(f, Z)^2] - L(f)^2 \le L(f) - L(f)^2 = L(f)(1 - L(f))$$

when $0 \le \ell(f, z) \le 1$, we could apply a type of uniform convergence guarantee to obtain

$$L(\widehat{f}) \le L(f^\star) + O(1)\sqrt{\frac{L(f^\star)(1 - L(f^\star))}{n}}$$

with high probability. We cannot quite achieve this, but assuming we have a complexity measure $c(f)$ (i.e., encoding of $f \in \mathcal{F}$) satisfying the Kraft inequality (5.2.5), we have the following result.

Theorem 5.2.7. *Let $\ell \in [0, 1]$ be a bounded loss, the complexity penalty $c : \mathcal{F} \to \mathbb{R}_+$ satisfy inequality (5.2.5), and $\widehat{f} = \operatorname{argmin}_{f\in\mathcal{F}} \widehat{L}_n(f)$ be an empirical minimizer of $\widehat{L}_n$. Then for any $t \ge 0$, with probability at least $1 - 2e^{-t}$,*

$$L(\widehat{f}) \le \inf_{f\in\mathcal{F}} \left\{ L(f) + \sqrt{\frac{8L(f)}{3} \cdot \frac{c(\widehat{f}) + t}{n}} + \frac{16}{3} \cdot \frac{c(\widehat{f}) + t}{n} \right\}.$$

Exercise 5.2 outlines one proof of this result.

The bound in Theorem 5.2.7 is (typically) sharper than the guarantees that Theorem 5.2.4 and its corollaries provide, in that the deviations depend on the best value of $L(f)$. Indeed, the global sub-Gaussian constant σ^2 in Theorem 5.2.4 is always larger than the variance $\operatorname{Var}(\ell(f, Z))$, and if all we know is that $\ell(f, Z) \in [0, 1]$ it necessarily scales as $\sigma^2 \gtrsim 1$. (Recall Exercise 4.2.) As a particular example, if there exists some function $f^\star$ achieving perfect loss, that is, $\ell(f^\star, Z) = 0$ with probability 1 over Z, then Theorem 5.2.7 shows that $L(\widehat{f}) \le L(f^\star) + O(1)\frac{c(\widehat{f})+t}{n}$ with probability at least $1 - e^{-t}$. Theorem 5.2.4 cannot imply this $O(1/n)$ convergence rate.

5.2.2 Large classes

When the collection of functions is (uncountably) infinite, obtaining strong generalization bounds is more challenging, though many tools are available. The most basic, of which we give examples, leverage covering number bounds (as in Example 5.1.9). In the next chapter we provide alternative approaches based on an information-theoretic characterization of divergence measures, which provide guarantees with similar structure to those we present here.

Let us begin by considering a few examples, which combine with the Rademacher contraction inequality (Theorem 5.1.6) to derive explicit generalization bounds.

Example 5.2.8 (Rademacher complexity of the ℓ_2-ball)**:** Let $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$, and consider the class of linear functionals $\mathcal{F} := \{f_\theta(x) = \theta^\top x, \theta \in \Theta\}$. Then

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2},$$

because

$$R_n(\mathcal{F} \mid x_1^n) = \frac{r}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2\right] \le \frac{r}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2\right]} = \frac{r}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2},$$

as desired. $\diamond$

In high-dimensional situations, it is sometimes useful to consider more restrictive function classes, for example, those indexed by vectors in an ℓ_1-ball.

Example 5.2.9 (Rademacher complexity of the ℓ_1-ball)**:** Let $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r\}$, and consider the linear class $\mathcal{F} := \{f_\theta(x) = \theta^\top x, \theta \in \Theta\}$. Then

$$R_n(\mathcal{F} \mid x_1^n) = \frac{r}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_\infty\right].$$

Now, each coordinate j of $\sum_{i=1}^n \varepsilon_i x_i$ is $\sum_{i=1}^n x_{ij}^2$-sub-Gaussian, and thus using that $\mathbb{E}[\max_{j\le d} Z_j] \le \sqrt{2\sigma^2 \log d}$ for arbitrary σ^2-sub-Gaussian Z_j (see Exercise 4.9), we have

$$R_n(\mathcal{F} \mid x_1^n) \le \frac{r}{n}\sqrt{2\log(2d)\max_j \sum_{i=1}^n x_{ij}^2}.$$

To facilitate comparison with Example 5.2.8, suppose that the vectors x_i all satisfy $\|x_i\|_\infty \le b$. In this case, the preceding inequality implies that $R_n(\mathcal{F} \mid x_1^n) \le rb\sqrt{2\log(2d)}/\sqrt{n}$. In contrast, the ℓ_2-norm of such x_i may satisfy $\|x_i\|_2 = b\sqrt{d}$, so that the bounds of Example 5.2.8 scale instead as $rb\sqrt{d}/\sqrt{n}$, which can be exponentially larger. $\diamond$

These examples are sufficient to derive risk bounds for cases where we have a loss function applied to a class with finite Rademacher complexity, in which case it is possible to recenter the loss class to achieve generalization bounds. The coming proposition does precisely this in the case of margin-based binary classification. Consider points $(x, y) \in \mathcal{X} \times \{\pm 1\}$, and let $\mathcal{F}$ be an arbitrary class of functions $f : \mathcal{X} \to \mathbb{R}$ and $\mathcal{L} = \{(x, y) \mapsto \ell(yf(x))\}_{f\in\mathcal{F}}$ the induced collection of losses, as we briefly touch on in the introduction to this chapter. As typical examples, we might have $\ell(t) = [1 - t]_+$, $\ell(t) = e^{-t}$, or $\ell(t) = \log(1 + e^{-t})$. We have the following proposition.

Proposition 5.2.10. *Let $\mathcal{F}$ and $\mathcal{X}$ be such that $\sup_{x\in\mathcal{X}} |f(x)| \le M$ for $f \in \mathcal{F}$ and assume that ℓ is L-Lipschitz. Define the empirical and population risks $\widehat{L}_n(f) := P_n\ell(Yf(X))$ and $L(f) := P\ell(Yf(X))$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\widehat{L}_n(f) - L(f)| \ge 4LR_n(\mathcal{F}) + t\right) \le 2\exp\left(-\frac{nt^2}{2L^2M^2}\right) \quad \text{for } t \ge 0.$$

Proof We may recenter the class $\mathcal{L}$, that is, replace $\ell(\cdot)$ with $\ell(\cdot) - \ell(0)$, without changing $\widehat{L}_n(f) - L(f)$. Call this class $\mathcal{L}_0$, so that $\|P_n - P\|_{\mathcal{L}} = \|P_n - P\|_{\mathcal{L}_0}$. This recentered class satisfies bounded differences with constant $2ML$, as $|\ell(yf(x)) - \ell(y'f(x'))| \le L|yf(x) - y'f(x')| \le 2LM$, as in the proof of Proposition 5.1.1. Applying Proposition 5.1.1 and then Corollary 5.1.3 gives that $\mathbb{P}(\sup_{f\in\mathcal{F}} |\widehat{L}_n(f) - L(f)| \ge 2R_n(\mathcal{L}_0) + t) \le \exp(-\frac{nt^2}{2M^2L^2})$ for $t \ge 0$. Then applying the contraction inequality (Theorem 5.1.6) yields $R_n(\mathcal{L}_0) \le 2LR_n(\mathcal{F})$, giving the result. $\square$

An example application can make this clearer.

Example 5.2.11 (Support vector machines and hinge losses)**:** In the support vector machine problem, we receive data $(X_i, Y_i) \in \mathbb{R}^d \times \{\pm 1\}$, and we seek to minimize the average of the

losses $\ell(\theta; (x, y)) = \left[1 - y\theta^\top x\right]_+$. We assume that the space $\mathcal{X}$ has $\|x\|_2 \le b$ for $x \in \mathcal{X}$ and that $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$. Applying Proposition 5.2.10 gives

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |P_n\ell(\theta; (X, Y)) - P\ell(\theta; (X, Y))| \ge 4R_n(\mathcal{F}_\Theta) + t\right) \le 2\exp\left(-\frac{nt^2}{2r^2b^2}\right),$$

where $\mathcal{F}_\Theta = \{f_\theta(x) = \theta^\top x\}_{\theta\in\Theta}$. Example 5.2.8 gives

$$R_n(\mathcal{F}_\Theta) \le \frac{rb}{\sqrt{n}}.$$

That is, we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |P_n\ell(\theta; (X, Y)) - P\ell(\theta; (X, Y))| \ge \frac{4rb}{\sqrt{n}} + t\right) \le 2\exp\left(-\frac{nt^2}{2(rb)^2}\right),$$

so that P_n and P become close at rate roughly $rb/\sqrt{n}$ in this case. $\diamond$

Example 5.2.11 exhibits a "dimension free" convergence result—there is no explicit dependence on the dimension d of the problem except implicitly through the radii r and b. One consequence of this is that if x and θ belong to a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|x\|^2 = \langle x, x\rangle$, but for which we are guaranteed that $\|\theta\| \le r$ and $\|x\| \le b$, then the result still applies.

When we do not have the simplifying structure of $\ell(yf(x))$ identified in the preceding examples, we can still provide guarantees of generalization using the covering number guarantees Section 5.1.3 introduces. The most common and important case is when we have a Lipschitzian loss function in an underlying parameter θ.

Example 5.2.12 (Lipschitz functions over a norm-bounded parameter space)**:** Consider the parametric loss minimization problem

$$\underset{\theta\in\Theta}{\text{minimize}} \;\; L(\theta) := \mathbb{E}[\ell(\theta; Z)]$$

for a loss function ℓ that is M-Lipschitz (with respect to the norm $\|\cdot\|$) in its argument, where for normalization we assume $\inf_{\theta\in\Theta} \ell(\theta, z) = 0$ for each z. Then the metric entropy of Θ bounds the metric entropy of the loss class $\mathcal{F} := \{z \mapsto \ell(\theta, z)\}_{\theta\in\Theta}$ for the supremum norm $\|\cdot\|_\infty$. Indeed, for any pair θ, θ', we have

$$\sup_z |\ell(\theta, z) - \ell(\theta', z)| \le M\left\|\theta - \theta'\right\|,$$

and so an ϵ-cover of Θ is an $M\epsilon$-cover of $\mathcal{F}$ in supremum norm. In particular,

$$N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \le N(\epsilon/M, \Theta, \|\cdot\|).$$

Assume that $\Theta \subset \{\theta \mid \|\theta\| \le b\}$ for some finite r. Then Lemma 5.1.10 guarantees that $\log N(\epsilon, \Theta, \|\cdot\|) \le d\log(1 + 2/\epsilon) \lesssim d\log\frac{1}{\epsilon}$, and so the classical covering number argument in Example 5.1.9 gives

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |P_n\ell(\theta, Z) - P\ell(\theta, Z)| \ge t\right) \le \exp\left(-c\frac{nt^2}{r^2M^2} + Cd\log\frac{M}{t}\right),$$

where c, C are numerical constants. In particular, taking $t^2 \asymp \frac{M^2r^2d}{n}\log\frac{n}{\delta}$ gives that

$$|P_n\ell(\theta, Z) - P\ell(\theta, Z)| \lesssim \frac{Mr\sqrt{d\log\frac{n}{\delta}}}{\sqrt{n}}$$

with probability at least $1 - \delta$. $\diamond$

5.3 M-estimators and estimation

In many problems in statistics and machine learning, we seek to actually recover parameters of interest. For example, in a regression problem where we model

$$y = \langle x, \theta \rangle + \varepsilon,$$

we often care about the actual values of θ. Less prosaically, the lower bounds in the latter part are most informative when algorithms and upper bounds demonstrate their tightness.

To that end, we develop representative finite-sample results on the convergence of different estimators. We focus on *M-estimators*, which arise from minimization of a loss function $\ell(\theta, z)$ convex in θ, where z is problem data. Exponential families (Chapter 3) provide natural examples.

Example 5.3.1 (Exponential families and log loss)**:** Let $p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$ be the density of an exponential family. Then $\ell(\theta, x) := -\log p_\theta(x) = -\langle \theta, \phi(x) \rangle + A(\theta)$ is convex and $\mathcal{C}^\infty$ over $\Theta := \{\theta \mid A(\theta) < \infty\}$. $\diamond$

Example 5.3.2 (Logistic regression)**:** In binary logistic regression (Example 3.4.2) with labels $y \in \{\pm 1\}$, we have data $z = (x, y) \in \mathbb{R}^d \times \{\pm 1\}$, and the loss

$$\ell(\theta, z) = -\log p_\theta(y \mid x) = \log \left(1 + \exp(-yx^\top \theta)\right),$$

which is $\mathcal{C}^\infty$, domain all of $\mathbb{R}^d$, and Lipschitz-continuous derivatives of all orders. (Though the particular Lipschitz constant depends on x). $\diamond$

We will consider rates of convergence and estimation error for minimizers of the empirical loss

$$\widehat{L}_n(\theta) := P_n \ell(\theta, Z) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, Z_i),$$

where ℓ is convex in its first argument, which approximates the population loss

$$L(\theta) := \mathbb{E}_P[\ell(\theta, Z)],$$

where $Z_i \stackrel{\text{iid}}{\sim} P$ and $P_n = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Z_i}$ is the usual empirical distribution. For a closed convex set $\Theta \subset \mathbb{R}^d$, we study the *M-estimator*

$$\widehat{\theta}_n := \operatorname*{argmin}_{\theta \in \Theta} \widehat{L}_n(\theta), \tag{5.3.1}$$

providing prototypical arguments for its convergence. (Often we take $\Theta = \mathbb{R}^d$.)

Based on our results in the preceding sections, one idea is to use uniform convergence: if

$$\sup_{\theta \in \Theta} |\widehat{L}_n(\theta) - L(\theta)| \to 0,$$

then so long as the minimizer $\theta^\star$ of L is unique and $L(\theta) - L(\theta^\star)$ grows with the distance $\|\theta - \theta^\star\|$, we necessarily have $\widehat{\theta}_n \to \theta^\star$. This naive approach typically fails to achieve the correct convergence rates, and we therefore take another approach.

Recall that a twice differentiable function L is convex if and only if $\nabla^2 L(\theta) \succeq 0$ for all $\theta \in \operatorname{dom} L$. Typically such L have *quadratic growth* around the minimizer $\theta^\star$, meaning that near $\theta^\star$,

$$L(\theta) \geq L(\theta^\star) + \frac{\lambda}{2} \left\|\theta - \theta^\star\right\|_2^2.$$

Because the sampled $\widehat{L}_n$ is also convex and approximates L, we expect that for parameters θ far enough from $\theta^\star$ that the growth of $L(\theta)$ above $L(\theta^\star)$ dominates the noise inherent in the sampling, we necessarily have $\widehat{L}_n(\theta) > \widehat{L}_n(\theta^\star)$. Because the empirical minimizer $\widehat{\theta}_n$ satisfies $\widehat{L}_n(\widehat{\theta}_n) \leq \widehat{L}_n(\theta^\star)$, we would never choose such a distant parameter, implying a convergence rate. We are well-equipped to develop the convex analysis and sampling theory to make this argument rigorous.

5.3.1 Standard conditions and convex optimization

To provide cleaner results, we will consider a restricted collection of loss functions. Many families of losses satisfy the assumed conditions, and we can relax them using more sophisticated concentration inequalities, though as the sharpest convergence guarantees are not our main focus, we leave such refinements to exercises. We therefore make the following standing assumption.

Assumption A.5.1 (Standard conditions)**.** *For each $z \in \mathcal{Z}$, the losses $\ell(\theta, z)$ are convex in θ. There are constants $M_0, M_1, M_2 < \infty$ such that for each $z \in \mathcal{Z}$,*

(i) $\left\|\nabla \ell(\theta^\star, z)\right\|_2 \leq M_0$

(ii) $\left\|\nabla^2 \ell(\theta^\star, z)\right\|_{\rm op} \leq M_1$, *and*

(iii) the Hessian $\nabla^2 \ell(\theta, z)$ is M_2-Lipschitz continuous in a neighborhood of radius $r > 0$ around $\theta^\star$, meaning $\left\|\nabla^2 \ell(\theta_0, z) - \nabla^2 \ell(\theta_1, z)\right\|_{\rm op} \leq M_2 \left\|\theta_0 - \theta_1\right\|_2$ whenever $\left\|\theta_i - \theta^\star\right\|_2 \leq r$.

Additionally, the minimizer $\theta^\star = \operatorname{argmin}_\theta L(\theta)$ exists and for a $\lambda > 0$ satisfies

$$\nabla^2 L(\theta^\star) \succeq \lambda I.$$

These "standard conditions" frequently hold. As we see when we specialize our coming results to exponential family models in Section 5.3.4, Assumption A.5.1 holds essentially as soon as the family is minimal and $\phi(x)$ is bounded. The existence of minimizers can be somewhat more subtle to guarantee than the smoothness conditions (i)–(iii), though these are typically straightforward. (For more on the existence of minimizers, see Exercise 5.11.)

To quickly highlight the conditions, we revisit binary logistic and robust regression.

Example (Example 5.3.2 continued)**:** For logistic regression with labels $y \in \{\pm 1\}$, we have

$$\nabla \ell(\theta, (x, y)) = -\frac{1}{1 + e^{yx^\top \theta}} yx \quad \text{and} \quad \nabla^2 \ell(\theta, (x, y)) = p_\theta(y \mid x)(1 - p_\theta(y \mid x)) xx^\top.$$

Then conditions (i)–(iii) hold so long as $\sup_{x \in \mathcal{X}} \left\|x\right\|_2 < \infty$, with $M_0 = \sup_{x \in \mathcal{X}} \left\|x\right\|_2$ and $M_1 = \frac{1}{4} M_0^2$. Because $0 < p_\theta(y \mid x)(1 - p_\theta(y \mid x)) \leq \frac{1}{4}$ for any θ, x, y, *if* a minimizer exists then it is necessarily unique as soon as $\mathbb{E}[XX^\top] \succ 0$. $\diamond$

Example 5.3.3 (Robust regression)**:** In robust regression, we wish to best approximate responses $y \in \mathbb{R}$ via linear functions $x^\top \theta$, but because of outliers, do not use the squared loss. Thus, for a smooth symmetric convex function h with bounded derivatives, we take

$$\ell(\theta, (x, y)) = h(\langle x, \theta \rangle - y),$$

so that

$$\nabla \ell(\theta, (x, y)) = h'(\langle x, \theta \rangle - y)x \quad \text{and} \quad \nabla^2 \ell(\theta, (x, y)) = h''(\langle x, \theta \rangle - y)xx^\top.$$

The function $h(t) = \log(1 + e^t) + \log(1 + e^{-t})$, which satisfies $h'(t) = \frac{e^t - 1}{e^t + 1} \in [-1, 1]$ and $h''(t) = \frac{2e^t}{(e^t+1)^2} \in (0, 1/2]$, provides a prototypical example. So long as the covariates $x \in \mathcal{X}$ have finite radius $\mathrm{rad}_2(\mathcal{X}) := \sup_{x \in \mathcal{X}} \|x\|_2$, we obtain the Lipschitz constant bounds

$$M_0 \le \mathrm{rad}_2(\mathcal{X}), \quad M_1 \lesssim \mathrm{rad}_2^2(\mathcal{X}), \quad \text{and} \quad M_2 \lesssim \mathrm{rad}_2^3(\mathcal{X})$$

for Assumption A.5.1, parts (i)–(iii). In general, if h is symmetric with $h''(0) > 0$, then minimizers exist whenever Y is non-pathological and $\mathbb{E}[XX^\top] \succ 0$. Exercise 5.12 asks you to prove this last claim on existence of minimizers. $\diamond$

5.3.2 Some growth properties of convex functions

As we discuss above, we will roughly proceed in our analysis by showing that the growth of the loss function dominates the noise inherent in sampling. We therefore collect growth properties of convex functions, as they provide the fundamental building block for convergence analysis. First, we show that for any convex h, if there exists a shell $S = \{\theta \mid \|\theta - \theta_0\|_2 = r\}$ around a point θ_0 for which $h(\theta) > h(\theta_0)$ for all $\theta \in S$, then $\widehat{\theta} = \mathrm{argmin}_\theta h(\theta)$ satisfies $\|\widehat{\theta} - \theta_0\|_2 < r$. See Figure 5.4.

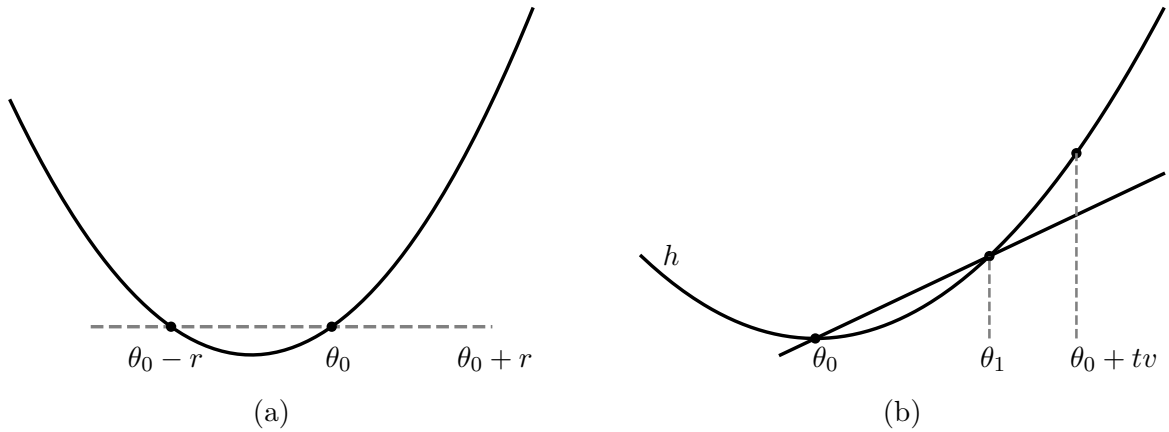


Figure 5.4. Minimizers of convex functions. (a) If $h(\theta) > h(\theta_0)$ for all θ satisfies $\|\theta - \theta_0\|_2 = r$, then minimizers of h satisfy $\|\theta - \theta_0\|_2 < r$. (b) Illustration of Lemmas 5.3.4 and 5.3.5. The convex function h grows at least linearly beyond θ_1.

Lemma 5.3.4. *Let h be convex, $\theta_0 \in \mathrm{dom}\, h$, v an arbitrary vector, and $\theta_1 = \theta_0 + v$. Then for all $t \ge 1$, $h(\theta_0 + tv) - h(\theta_0) \ge t(h(\theta_1) - h(\theta_0))$.*

Proof Let $\theta_t = \theta_0 + tv$ for $t \geq 1$. Then for $t \geq 1$, we can write $\theta_1 = \frac{1}{t}\theta_t + (1 - \frac{1}{t})\theta_0$, and so

$$h(\theta_1) = h\left(\left(1 - \frac{1}{t}\right)\theta_0 + \frac{1}{t}\theta_t\right) \leq \left(1 - \frac{1}{t}\right)h(\theta_0) + \frac{1}{t}h(\theta_t).$$

Rearranging yields

$$\left(1 - \frac{1}{t}\right)\left(h(\theta_1) - h(\theta_0)\right) \leq \frac{1}{t}\left(h(\theta_t) - h(\theta_1)\right),$$

and multiplying through by t and rearranging implies the desired result. □

Extending this to shells, we have the following result.

Lemma 5.3.5. *Let h be convex and $\theta_0 \in \operatorname{dom} h$ and assume that for some $\epsilon > 0$ and $\delta > 0$ that $h(\theta_0 + v) \geq h(\theta_0) + \delta$ for all $\|v\|_2 = \epsilon$. Then for all θ for which $\|\theta - \theta_0\|_2 \geq \epsilon$,*

$$h(\theta) - h(\theta_0) \geq \frac{\delta}{\epsilon}\|\theta - \theta_0\|_2.$$

Proof For any θ with $\|\theta - \theta_0\|_2 \geq \epsilon$, we can write $\theta = \theta_0 + tv$ for $v = \epsilon\frac{\theta - \theta_0}{\|\theta - \theta_0\|_2}$ and $t = \frac{\|\theta - \theta_0\|_2}{\epsilon}$. Apply Lemma 5.3.4 with the substitution $h(\theta_1) - h(\theta_0) \geq \delta$. □

We connect these results to the growth of smooth convex functions and existence of nearby minimizers by arguing that the minimizer of a convex function h is not too far from a benchmark point θ_0 at which h has upward curvature and small gradient. Figure 5.5 illustrates.

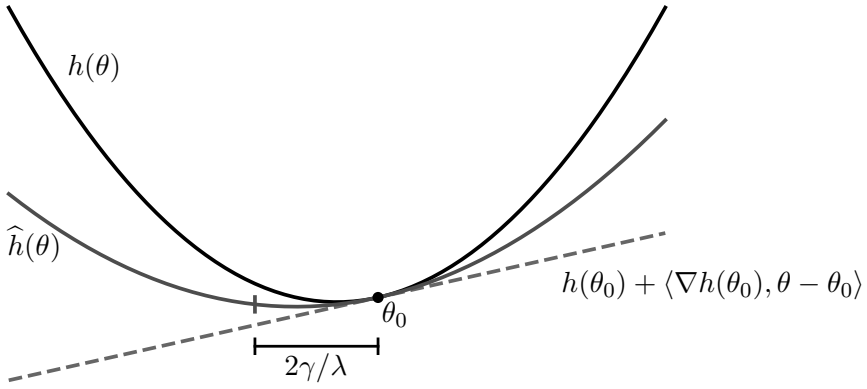


Figure 5.5. Illustration of Lemmas 5.3.6 and 5.3.7, where $\|\nabla h(\theta_0)\|_2 \leq \gamma$ and $\nabla^2 h(\theta) \succeq \lambda I$ for all θ. The quadratic approximation $\widehat{h}(\theta) = h(\theta_0) + \langle \nabla h(\theta_0), \theta - \theta_0 \rangle + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2$ lower bounds the true function h, and $\widehat{h}(\theta)$ (and hence h) must grow outside of the region $\|\theta - \theta_0\|_2 \leq \frac{2\gamma}{\lambda}$.

Lemma 5.3.6. *Let h be convex and $\lambda > 0$, $\gamma \geq 0$, and $\epsilon \geq \frac{2\gamma}{\lambda}$. Let θ_0 satisfy that $\|\nabla h(\theta_0)\|_2 \leq \gamma$ and $\nabla^2 h(\theta) \succeq \lambda I$ whenever $\|\theta - \theta_0\|_2 \leq \epsilon$. Then the minimizer $\widehat{\theta} = \operatorname{argmin}_\theta h(\theta)$ exists and satisfies*

$$\|\widehat{\theta} - \theta_0\|_2 \leq \frac{2\gamma}{\lambda}.$$

Proof By Taylor's theorem, for any θ there exists $\overline{\theta}$ on the line between θ and θ_0 with

$$h(\theta) = h(\theta_0) + \langle \nabla h(\theta_0), \theta - \theta_0 \rangle + \frac{1}{2}(\theta - \theta_0)^\top \nabla^2 h(\overline{\theta})(\theta - \theta_0).$$

Now, let us take θ such that $\|\theta - \theta_0\|_2 \le t$. Then by assumption $\nabla^2 h(\overline{\theta}) \succeq \lambda I$, and so we have

$$\begin{aligned} h(\theta) &\ge h(\theta_0) + \langle \nabla h(\theta_0), \theta - \theta_0 \rangle + \frac{\lambda}{2} \|\theta - \theta_0\|_2^2 \\ &\ge h(\theta_0) - \gamma \|\theta - \theta_0\|_2 + \frac{\lambda}{2} \|\theta - \theta_0\|_2^2 . \end{aligned}$$

by assumption that $\|\nabla h(\theta_0)\|_2 \le \gamma$ and the Cauchy-Schwarz inequality.

Fix $t \ge 0$. If we can show that $h(\theta) > h(\theta_0)$ for all θ satisfying $\|\theta - \theta_0\|_2 = t$, then Lemma 5.3.5 implies that $h(\theta) > h(\theta_0)$ whenever $\|\theta - \theta_0\|_2 \ge t$, so that necessarily $\|\widehat{\theta} - \theta_0\|_2 < t$. Returning to the previous display and letting $t = \|\theta - \theta_0\|_2$, note that

$$h(\theta) \ge h(\theta_0) - \gamma t + \frac{\lambda}{2} t^2,$$

and as $-\gamma t + \frac{\lambda}{2} t^2 = t(\frac{\lambda}{2} t - \gamma) > 0$ whenever $t > \frac{2\gamma}{\lambda}$. As by assumption $\nabla^2 h(\theta) \succeq \lambda I$ whenever $\|\theta - \theta_0\|_2 \le \epsilon$ for some $\epsilon \ge \frac{2\gamma}{\lambda}$, this implies the result. □

Lemma 5.3.7. *Let h be convex and assume that $\nabla^2 h$ is M_2-Lipschitz (part (iii) of Assumption A.5.1). Let $\lambda > 0$, $\gamma > 0$ satisfy $\gamma < \frac{\lambda^2}{8M_2}$. Then if both $\nabla^2 h(\theta_0) \succeq \lambda I$ and $\|\nabla h(\theta_0)\|_2 \le \gamma$, the minimizer $\widehat{\theta} = \operatorname{argmin}_\theta h(\theta)$ exists and satisfies*

$$\|\widehat{\theta} - \theta_0\|_2 \le \frac{4\gamma}{\lambda}.$$

Proof By Lemma 5.3.6, it is enough to show that $\nabla^2 h(\theta) \succeq \frac{\lambda}{2}$ for all θ with $\|\theta - \theta_0\|_2 \le \frac{4\gamma}{\lambda}$. For this, we use the M_2-Lipschitz continuity of $\nabla^2 h$ to obtain that for any θ with $\|\theta - \theta_0\|_2 = t$,

$$\nabla^2 h(\theta) \succeq \nabla^2 h(\theta_0) - M_2 \|\theta - \theta_0\|_2 I \succeq (\lambda - M_2 t) I.$$

So if $t \le \frac{\lambda}{2M_2}$ we have $\nabla^2 h(\theta) \succeq \frac{\lambda}{2} I$. Because $\frac{4\gamma}{\lambda} \le \frac{\lambda}{2M_2}$ by assumption, we have $\nabla^2 h(\theta) \succeq \frac{\lambda}{2} I$ whenever $\|\theta - \theta_0\|_2 \le \frac{4\gamma}{\lambda}$, yielding the result. □

5.3.3 Convergence analysis for convex M-estimators

By Lemma 5.3.7, to show a convergence rate guarantee for the empirical minimizer $\widehat{\theta}_n$, it is sufficient to demonstrate two (related) conditions: that for some sequence $\gamma_n \to 0$, we have

$$\left\|\nabla \widehat{L}_n(\theta^\star)\right\|_2 \le \gamma_n \tag{5.3.2a}$$

with high probability, and that for some $\lambda > 0$, with high probability

$$\nabla^2 \widehat{L}_n(\theta^\star) \succeq \lambda I. \tag{5.3.2b}$$

The convergence guarantees we develop in Chapter 4 provide precisely the tools to do this.

Theorem 5.3.8. *Let Assumption A.5.1 hold for the M-estimation problem (5.3.1). Let $\delta \in (0, \frac{1}{2})$, and define*

$$\gamma_n(\delta) := \frac{M_0}{\sqrt{n}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right) \quad \text{and} \quad \epsilon_n(\delta) := \frac{2\sqrt{2} M_1}{\sqrt{n}} \sqrt{\log \frac{2d}{\delta}}.$$

Then with probability at least $1 - 2\delta$, $\|\nabla \widehat{L}_n(\theta^\star)\|_2 \le \gamma_n(\delta)$ and $\|\nabla^2 \widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)\|_{\rm op} \le \epsilon_n(\delta)$. So long as $\epsilon_n(\delta) \le \frac{\lambda}{4}$ and $\gamma_n(\delta) \le \frac{9\lambda^2}{128 M_2}$, then with the same probability,

$$\left\|\widehat{\theta}_n - \theta^\star\right\|_2 \le \frac{16}{3} \frac{\gamma_n(\delta)}{\lambda}.$$

We defer the proof of the theorem to Section 5.3.5, instead providing commentary and a few examples of its application. Ignoring the numerical constants, the theorem roughly states the following: once n is large enough that

$$n \gtrsim \frac{M_2^2 M_0^2}{\lambda^4} \log \frac{1}{\delta} \quad \text{and} \quad n \gtrsim \frac{M_1^2}{\lambda^2} \log \frac{d}{\delta}, \tag{5.3.3}$$

with probability at least $1 - \delta$ we have

$$\|\widehat{\theta}_n - \theta^\star\|_2 \lesssim \frac{M_0}{\lambda \sqrt{n}} \sqrt{\log \frac{1}{\delta}}. \tag{5.3.4}$$

These finite sample results are order optimal for large n, as we will develop in the coming chapters on fundamental limits. Nonetheless, the conditions (5.3.3) are stronger than necessary, typically requiring that n be quite large. In Exercises 5.6 and 5.7, we explore a class of *quasi-self-concordant* losses, where the second derivative controls the third derivative, allowing more direct application of Lemma 5.3.6, reducing this sample size requirement. Exercise 5.15 weakens the moment assumptions while providing a related finite-sample convergence guarantee.

Example 5.3.9 (Logistic regression, Example 5.3.2 continued)**:** Assume the domain $\mathcal{X} \subset \mathbb{R}^d$ consists of vectors x with $\|x\|_2 \le \sqrt{d}$, for example, $\mathcal{X} \subset [-1, 1]^d$. In this case, $M_0 = \sqrt{d}$, while $M_1 \le \frac{1}{4} d$ and $M_2 \lesssim d^{3/2}$. If the Hessian $\nabla^2 L(\theta^\star) = \mathbb{E}[p_{\theta^\star}(Y \mid X)(1 - p_{\theta^\star}(Y \mid X)) X X^\top]$ has minimal eigenvalue $\lambda_{\min}(\nabla^2 L(\theta^\star)) \gtrsim 1$, Theorem 5.3.8 applies once $n \gtrsim d^4 \log \frac{1}{\delta}$. $\diamond$

When n is large enough, the guarantee (5.3.4) allows us to also make the heuristic asymptotic expansions for the exponential family models in Section 3.2.2 hold in finite samples. Assume $\theta^\star \in \operatorname{int} \Theta$, so $\nabla L(\theta^\star) = 0$, and let the conclusions of Theorem 5.3.8 hold (that is, $\|\nabla^2 \widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)\|_{\rm op} \le \epsilon_n(\delta)$ and so on). Then once we know that $\widehat{\theta}_n$ exists, by a Taylor expansion we can write

$$\begin{aligned} 0 = \nabla \widehat{L}_n(\widehat{\theta}_n) &= \nabla \widehat{L}_n(\theta^\star) + (\nabla^2 \widehat{L}_n(\theta^\star) + E_n)(\widehat{\theta}_n - \theta^\star) \\ &= \nabla \widehat{L}_n(\theta^\star) + (\nabla^2 L(\theta^\star) + E_n')(\widehat{\theta}_n - \theta^\star), \end{aligned}$$

where E_n is an error matrix satisfying $\|E_n\|_{\rm op} \le M_2 \|\widehat{\theta}_n - \theta^\star\|_2$ and $E_n' = E_n + \nabla^2 \widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)$ satisfies $\|E_n'\|_{\rm op} \le \|E_n\|_{\rm op} + \epsilon_n(\delta)$ by the triangle inequality and Theorem 5.3.8. Using that if $A \succ 0$ then $(A + E)^{-1} = A^{-1} - A^{-1} E A^{-1} + O(\|E\|_{\rm op}^2 / \lambda_{\min}(A)^3)$ (see Exercise 5.5), we therefore have

$$\widehat{\theta}_n - \theta^\star = -(\nabla^2 L(\theta^\star) + E_n')^{-1} \nabla \widehat{L}_n(\theta^\star) = -\nabla^2 L(\theta^\star)^{-1} \nabla \widehat{L}_n(\theta^\star) + R_n,$$

where the remainder vector R_n satisfies

$$\begin{aligned}\|R_n\|_2 &\lesssim \left\|\nabla^2 L(\theta^\star)^{-1} E_n' \nabla^2 L(\theta^\star)^{-1}\right\|_{\rm op} \left\|\nabla \widehat{L}_n(\theta^\star)\right\|_2 \\ &\lesssim \frac{1}{\lambda^2}\left(\frac{M_2 M_0 \sqrt{\log \frac{1}{\delta}}}{\lambda\sqrt{n}} + \epsilon_n(\delta)\right)\left\|\nabla \widehat{L}_n(\theta^\star)\right\|_2 \lesssim \frac{\log \frac{d}{\delta}}{\lambda^2 n} \cdot \left(\frac{M_2 M_0^2}{\lambda} + M_1\right)\end{aligned}$$

with probability at least $1-\delta$. We summarize this in the following corollary.

Corollary 5.3.10. *Let the conditions of Theorem 5.3.8 hold and $\theta^\star \in \operatorname{int}\Theta$. Then there is a problem-dependent constant C such that for $\delta > 0$ and $n \ge C \log \frac{1}{\delta}$, with probability at least $1-\delta$*

$$\widehat{\theta}_n - \theta^\star = -\nabla^2 L(\theta^\star)^{-1} \nabla \widehat{L}_n(\theta^\star) + R_n,$$

where the remainder R_n satisfies $\|R_n\|_2 \le C \cdot \frac{1}{n} \log \frac{d}{\delta}$. The constant C may be taken to be continuous in the problem parameters of Assumption A.5.1.

Corollary 5.3.10 highlights the salient terms governing error in estimation problems: the curvature of the loss near the optimum from $\nabla^2 L(\theta^\star)$ and the variance in the gradients $\nabla \widehat{L}_n(\theta^\star)$. As a final remark, let us imagine that the remainder term R_n in the corollary, in addition to being small with high probability, satisfies $\mathbb{E}[\|R_n\|_2^2] \le \frac{C}{n^2}$, where C is a problem-dependent constant. Let $G_n = -\nabla^2 L(\theta^\star)^{-1} \nabla \widehat{L}_n(\theta^\star)$ be the leading term in the expansion, which satisfies

$$\mathbb{E}[\|G_n\|_2^2] = \frac{1}{n} \mathrm{Var}(\nabla^2 L(\theta^\star)^{-1} \nabla \ell(\theta^\star, Z)) = \frac{\mathrm{tr}(\nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1})}{n}.$$

Then because $\|G_n + R_n\|_2^2 \le \|G_n\|_2^2 + \|R_n\|_2^2 + 2\|G_n\|_2 \|R_n\|_2$, we have the heuristic

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2\right] &= \mathbb{E}\left[\|G_n + R_n\|_2^2\right] \\ &= \mathbb{E}[\|G_n\|_2^2] + \mathbb{E}[\|R_n\|_2^2] \pm 2\mathbb{E}[\|G_n\|_2 \|R_n\|_2] \\ &\stackrel{(\star)}{=} \frac{\mathrm{tr}(\nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1})}{n} \pm \frac{C}{n^{3/2}}, \end{aligned} \tag{5.3.5}$$

where C is a problem-dependent constant and the step $(\star)$ is heuristic. See Exercise 5.8 for one approach to make this step rigorous.

5.3.4 Consequences for exponential families and generalized linear models

Working through applications of Corollary 5.3.10 with the exponential family and generalized linear models of Chapter 3 can make the results clearer. Recall that for an exponential family model with loss $\ell(\theta; x) = -\log p_\theta(x) = -\langle \theta, \phi(x)\rangle + A(\theta)$, we heuristically derived in expression (3.2.6) that if the data were i.i.d. from the exponential family model $P_{\theta^\star}$, then

$$\widehat{\theta}_n - \theta^\star \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1} \cdot \nabla^2 A(\theta^\star)^{-1}\right).$$

Corollary 5.3.10 presents one formalization. Assuming the sufficient statistics ϕ are bounded, we have $\nabla^2 L(\theta^\star) = \nabla^2 A(\theta^\star)$, and $\mathrm{Cov}_{\theta^\star}(\nabla \widehat{L}_n(\theta^\star)) = \frac{1}{n}\mathrm{Cov}_{\theta^\star}(\phi(X)) = \frac{1}{n}\nabla^2 A(\theta^\star)$. So

$$\widehat{\theta}_n - \theta^\star = -\nabla^2 A(\theta^\star)^{-1} \frac{1}{n} \sum_{i=1}^n (\phi(X_i) - \nabla A(\theta^\star)) + R_n,$$

where the remainder satisfies $\|R_n\|_2 \le C\frac{1}{n}\log\frac{d}{\delta}$ with probability at least $1-\delta$ when $n \ge C\log\frac{1}{\delta}$.

Small probability events (e.g., that the sampled Hessian matrix $\nabla^2\widehat{L}_n(\theta^\star)$ fails to be invertible) make obtaining finite sample expected bounds tedious. One simple device is to consider the estimator $\widehat{\theta}_n$ only on a "good" event $\mathcal{E}_n$ that occurs with high probability. The next corollary provides a prototypical result under the assumption that $X_i \stackrel{\text{iid}}{\sim} P_{\theta^\star}$ for an exponential family model with bounded data $\sup_{x\in\mathcal{X}}\|\phi(x)\|_2 < \infty$ and positive definite Hessian $\nabla^2 A(\theta^\star) \succ 0$.

Corollary 5.3.11. *Under the preceding conditions on the exponential family model $P_{\theta^\star}$, there exists a problem dependent constant $C < \infty$ such that the following holds: for any $k \ge 1$, there are events $\mathcal{E}_n$ with probability $\mathbb{P}(\mathcal{E}_n) \ge 1 - \frac{1}{n^k}$ and*

$$\mathbb{E}_{\theta^\star}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2 \cdot \mathbf{1}\left\{\mathcal{E}_n\right\}\right] \le \frac{1}{n}\operatorname{tr}\left(\nabla^2 A(\theta^\star)^{-1}\right) + \frac{Ck\log n}{n^{3/2}}.$$

The constant C may be taken continuous in $\theta^\star$.

Recalling the equality (3.3.2), we see that the Fisher information $\nabla^2 A(\theta)$ appears in a fundamental way for the exponential families. Proposition 3.5.1 shows that it necessarily lower bounds testing error; here it provides an upper bound on the convergence of the maximum likelihood estimator. Exercise 5.9 extends Corollary 5.3.11 to an equality to within lower order terms.

Proof Let $\delta = \delta_n > 0$ to be chosen and define the event $\mathcal{E}_n$ to be that $\|R_n\|_2 \le C\frac{1}{n}\log\frac{d}{\delta}$, which occurs with probability at least $1-\delta$. Let

$$G_n = -\nabla^2 L(\theta^\star)^{-1}\nabla\widehat{L}_n(\theta^\star) = -\nabla^2 A(\theta^\star)^{-1}P_n(\phi(X) - \nabla A(\theta^\star))$$

be the mean-zero gradient term. Then $\widehat{\theta}_n - \theta^\star = G_n + R_n$, and $\|G_n + R_n\|_2^2 \le \|G_n\|_2^2 + \|R_n\|_2^2 + 2\|G_n\|_2\|R_n\|_2$. On the event $\mathcal{E}_n$ we have $\|R_n\|_2 \le \frac{C}{n}\log\frac{d}{\delta}$, and so

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2\mathbf{1}\left\{\mathcal{E}_n\right\}\right] \le \mathbb{E}[\|G_n\|_2^2] + 2\mathbb{E}[\|R_n\|_2^2\,\mathbf{1}\left\{\mathcal{E}_n\right\}]^{1/2}\mathbb{E}[\|G_n\|_2^2]^{1/2} + \frac{C^2}{n^2}\log^2\frac{d}{\delta}.$$

Now note that $\mathbb{E}[\|G_n\|_2^2] = \frac{1}{n}\operatorname{tr}(\nabla^2 A(\theta^\star)^{-1})$, and set $\delta = \frac{1}{n^k}$. □

These ideas extend to the generalized linear models of Section 3.4. For the abstract generalized linear model of predicting a target $y \in \mathcal{Y}$ from covariates $x \in \mathcal{X}$, we have

$$\ell(\theta, (x,y)) = -\log p_\theta(y \mid x) = -\phi(x,y)^\top\theta + A(\theta \mid x).$$

Because the log partition A is $\mathcal{C}^\infty$, the smoothness conditions in Assumption A.5.1 then reduce to the boundedness $\sup_{x\in\mathcal{X},y\in\mathcal{Y}}\|\phi(x,y)\|_2 < \infty$. Assuming a minimizer $\theta^\star = \operatorname{argmin}_\theta L(\theta)$ exists, the (local) strong convexity condition that $\nabla^2 L(\theta^\star) \succ 0$ becomes $\mathbb{E}[\operatorname{Cov}_{\theta^\star}(\phi(X,Y) \mid X)] \succ 0$. Exercise 5.11 gives sufficient conditions for the existence of minimizers in GLMs. For logistic regression (Example 3.4.2), these conditions correspond to a bound on the covariate data x, that $\mathbb{E}[XX^\top] \succ 0$, and that for each X, the label Y is non-deterministic.

An argument completely parallel to that for Corollary 5.3.10 provides convergence rates for generalized linear model estimators. We avoid the assumption of model fidelity, instead assuming that $\theta^\star = \operatorname{argmin}_\theta L(\theta)$ exists and $\nabla^2 L(\theta^\star) = \mathbb{E}[\nabla^2 A(\theta^\star \mid X)] \succ 0$, so that $\theta^\star$ is unique.

Corollary 5.3.12. *Let the preceding conditions hold and p_θ be a generalized linear model. Then there exists a problem-dependent constant $C < \infty$ such the following holds: for any $\delta \in (0,1)$ and for all $n \ge C \log \frac{1}{\delta}$, with probability at least $1 - \delta$*

$$\widehat{\theta}_n - \theta^\star = -\mathbb{E}[\nabla^2 A(\theta^\star \mid X)]^{-1} P_n \left(\phi(X,Y) - \nabla A(\theta^\star \mid X)\right) + R_n,$$

where the remainder $\|R_n\|_2 \le \frac{C}{n} \log \frac{d}{\delta}$.

When the model $Y \mid X = x \sim P_\theta(\cdot \mid x)$ is correct, then $\mathrm{Cov}(\phi(X,Y) \mid X) = \nabla^2 A(\theta^\star \mid X)$, and so (for a sequence of events $\mathcal{E}_n$ with probability at least $1 - 1/n^k$), we have

$$\mathbb{E}_{\theta^\star}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2 \mathbf{1}\left\{\mathcal{E}_n\right\}\right] \le \frac{\mathrm{tr}(\mathbb{E}[\nabla^2 A(\theta^\star \mid X)]^{-1})}{n} + \frac{Ck \log n}{n^{3/2}}.$$

The trace of the inverse Fisher information (3.3.2) in the generalized linear model characterizes this bound: the "larger" the information, the better estimation accuracy we can guarantee.

5.3.5 Proof of Theorem 5.3.8

The two key steps in the proof of the theorem are lemmas providing the guarantees (5.3.2).

Lemma 5.3.13. *Let Assumption A.5.1 hold. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\left\|\nabla \widehat{L}_n(\theta^\star)\right\|_2 \le \frac{M_0}{\sqrt{n}}\left(1 + \sqrt{2 \log \frac{1}{\delta}}\right).$$

Proof The function $z_1^n \mapsto \|\nabla \widehat{L}_n(\theta)\|_2$ satisfies bounded differences: for any two empirical samples P_n, P_n' differing in only observation i,

$$\begin{aligned}\left|\left\|P_n \nabla \ell(\theta, Z)\right\|_2 - \left\|P_n' \nabla \ell(\theta, Z)\right\|_2\right| &\le \left\|P_n \nabla \ell(\theta, Z) - P_n' \nabla \ell(\theta, Z)\right\|_2 \\ &\le \frac{1}{n}\left\|\nabla \ell(\theta, Z_i) - \nabla \ell(\theta, Z_i')\right\|_2 \le \frac{2M_0}{n}\end{aligned}$$

by Assumption A.5.1.(i). Because $\mathbb{E}[\|\nabla \widehat{L}_n(\theta^\star)\|_2] \le M_0/\sqrt{n}$, Proposition 4.2.6 gives that

$$\mathbb{P}\left(\left\|\nabla \widehat{L}_n(\theta^\star)\right\|_2 \ge M_0/\sqrt{n} + t\right) \le \exp\left(-\frac{nt^2}{2M_0^2}\right)$$

for all $t \ge 0$. Solving for t in $\exp(-\frac{nt^2}{2M_0^2}) = \delta$ yields the lemma. □

Lemma 5.3.14. *Let Assumption A.5.1 hold. Then for $\delta \in (0,1)$, with probability at least $1 - \delta$*

$$\left\|\nabla^2 \widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)\right\|_{\mathrm{op}} \le \frac{2\sqrt{2} M_1}{\sqrt{n}} \sqrt{\log \frac{2d}{\delta}}.$$

Proof Because $\nabla^2 \widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(\theta, Z_i)$ and $\|\nabla^2 \ell(\theta, z)\|_{\mathrm{op}} \le M_1$ by Assumption A.5.1.ii, Corollary 4.3.5 implies that

$$\mathbb{P}\left(\left\|\nabla^2 \widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)\right\|_{\mathrm{op}} \ge t\right) \le 2d \exp\left(-\frac{nt^2}{8M_1^2}\right).$$

Set $t = \sqrt{\frac{8M_1^2}{n}}\sqrt{\log\frac{2d}{\delta}}$. □

Let $\epsilon_n(\delta)$ be the bound on the right side of Lemma 5.3.14. Then with probability at least $1-\delta$,

$$\nabla^2 \widehat{L}_n(\theta^\star) \succeq \nabla^2 L(\theta^\star) - \epsilon_n(\delta) I_d \geq \frac{3\lambda}{4} I_d$$

by the assumption that n is large enough and $\nabla^2 L(\theta^\star) \succeq \lambda I$; we therefore have the first condition of Lemma 5.3.7 where $3\lambda/4$ replaces λ. Lemma 5.3.13 gives that with probability at least $1-\delta$, $\|\nabla \widehat{L}_n(\theta^\star)\|_2 \leq \gamma_n(\delta)$. Now use the assumption that $\gamma_n(\delta) \leq \frac{9\lambda^2}{128 M_2}$, so that Lemma 5.3.7 implies

$$\|\widehat{\theta}_n - \theta^\star\|_2 \leq \frac{16}{3}\frac{\gamma_n(\delta)}{\lambda}.$$

5.4 Bibliography and further reading

The treatment of estimation and convergence guarantees here follows mostly classical lines, with a few modifications to reflect our more modern understanding of the connections between convexity, concentration of random matrices, and estimation. Many classical results on uniform laws of large numbers rely on symmetrization techniques, essentially providing control on Rademacher complexities (as in Section 5.2.2). The foundational books on probability and asymptotics of Ledoux and Talagrand [284], van der Vaart and Wellner [408], and de la Peña and Giné [134] provide excellent, if fairly advanced, coverage; Boucheron et al. [80] survey of concentration inequalities in the context of classification problems in machine learning. Wainwright [422] provides comprehensive coverage of finite-sample guarantees in statistical estimation and machine learning, especially for high-dimensional problems.

Applications of Rademacher complexity in statistical machine learning exploded with the publication of Bartlett and Mendelson's paper [47]. Bartlett and Mendelson leveraged the Ledoux-Talagrand Rademacher contraction inequality (Theorem 5.1.6; see, for example, Ledoux and Talagrand [284, Theorem 4.12] or Koltchinskii [268, Theorem 2.2] for a proof) to provide generalization guarantees for many margin-based statistical learning problems. They, along with Bousquet and Koltchinskii [48, 267], leveraged Talagrand's Bernstein-type extensions [390] of the uniform convergence guarantees in Proposition 5.1.1 to give sharper convergence guarantees based on "localized Rademacher complexities." At some level, these approaches are similar to the one we employ in Section 5.3, relying on a two-stage approach: first, one argues that an estimator $\widehat{\theta}_n$ must be near a population minimizer $\theta^\star$, and then one argues that fluctuations of the empirical process $\widehat{L}_n(\theta) - L(\theta)$ must necessarily be small in a neighborhood of $\theta^\star$. These types of techniques are classical in asymptotic statistics, providing a way to obtain general rates of convergence in estimation problems; see, for example, van der Vaart and Wellner [408, Chapter 3.4] or van der Vaart [407, Chapter 5.8].

5.5 Exercises

Exercise 5.1: This exercise completes the proof of Proposition 5.1.2. In the notation of the proposition, let $Y_i = X_i - \mathbb{E}[X_i]$ and $\mathcal{I}_\pm = \{i \in [n] \mid \varepsilon_i = \pm 1\}$.

(a) Use that $\sum_{i=1}^n \varepsilon_i Y_i = \sum_{i \in \mathcal{I}_+} (Y_i - \mathbb{E}[Y_i]) - \sum_{i \in \mathcal{I}_-} (Y_i - \mathbb{E}[Y_i])$ to argue that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i Y_i\right\|^p\right] \le 2^{p-1}\mathbb{E}\left[\left\|\sum_{i \in \mathcal{I}_+} Y_i + \sum_{i \in \mathcal{I}_-} \mathbb{E}[Y_i]\right\|^p\right] + 2^{p-1}\mathbb{E}\left[\left\|\sum_{i \in \mathcal{I}_-} Y_i + \sum_{i \in \mathcal{I}_+} \mathbb{E}[Y_i]\right\|^p\right].$$

(b) Conclude the lower bound in Proposition 5.1.2.

Exercise 5.2: This exercise outlines a proof of Theorem 5.2.7. Recall that $\mathcal{F}$ is a function class with penalty mapping $c : \mathcal{F} \to \mathbb{R}_+$ satisfying the Kraft inequality (5.2.5), the losses $\ell(f, z) \in [0, 1]$, and we let $L(f) = \mathbb{E}[\ell(f, Z)]$ and $\widehat{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i)$.

(a) Show that there exists a numerical constant $C > 0$ such that for any $f \in \mathcal{F}$ and $t \ge 0$,

$$\mathbb{P}\left(\widehat{L}_n(f) \ge L(f) + t\right) \vee \mathbb{P}\left(\widehat{L}_n(f) \le L(f) - t\right) \le \exp\left(-C \min\left\{\frac{nt^2}{L(f)(1 - L(f))}, nt\right\}\right).$$

For the remainder of the problem, the constant $C > 0$ remains the same as that in part (a).

(b) Show that for any $\epsilon > 0$ and any fixed $f^\star \in \mathcal{F}$, if $L(f) \ge L(f^\star) + 2\epsilon$, then

$$\begin{aligned}\mathbb{P}\left(\widehat{L}_n(f) \le L(f^\star) + \epsilon\right) &\le \exp\left(-C \min\left\{\frac{n\epsilon^2}{L(f^\star)(1 - L(f^\star)) + 2\epsilon(1 - 2\epsilon)}, n\epsilon\right\}\right) \\ &\le \exp\left(-C \min\left\{\frac{n\epsilon^2}{L(f^\star) + 2\epsilon}, n\epsilon\right\}\right).\end{aligned}$$

Hint. Define $\epsilon_f := L(f) - L(f^\star) \ge 2\epsilon$ and consider the monotonicity properties of $\epsilon \mapsto \frac{\epsilon^2}{b + \epsilon(1-\epsilon)}$.

(c) For $t > 0$, define

$$\epsilon(f, t) := \frac{2(c(f) + t)}{Cn} + \sqrt{L(f^\star)}\sqrt{\frac{c(f) + t}{Cn}}.$$

Show that if $L(f) \ge L(f^\star) + 2\epsilon(f, t)$ then

$$\mathbb{P}\left(\widehat{L}_n(f) \le L(f^\star) + \epsilon(f, t)\right) \le \exp(-c(f) - t)$$

(d) Define $\mathcal{F}_t := \{f \in \mathcal{F} \mid L(f) \ge L(f^\star) + 2\epsilon(f, t)\}$. Show that if $\widehat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \widehat{L}_n(f)$, then

$$\mathbb{P}(\widehat{f} \in \mathcal{F}_t) \le e^{-t}.$$

(e) Conclude Theorem 5.2.7.

Exercise 5.3: Prove Lemma 5.1.8.

Exercise 5.4: Consider a classification problem with logistic loss $\ell(\theta; (x, y)) = \log(1 + \exp(-y\theta^\top x))$, where $\theta \in \Theta := \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le r\}$, $y \in \{\pm 1\}$, and $\mathcal{X} \subset \{x \in \mathbb{R}^d \mid \|x\|_\infty \le b\}$. Let $\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \widehat{L}_n(\theta)$. Show that with probability at least $1 - \delta$ over $(X_i, Y_i) \stackrel{\text{iid}}{\sim} P$,

$$L(\widehat{\theta}_n) \le \inf_{\theta \in \Theta} L(\theta) + C\frac{rb\sqrt{\log \frac{d}{\delta}}}{\sqrt{n}}$$

where $C < \infty$ is a numerical constant.

Exercise 5.5: Let $A \succ 0$ be a positive definite matrix, and let E be Hermitian and satisfy $\|E\|_{\rm op} < \lambda_{\min}(A)$. Define $S_k := A^{-1} + \sum_{i=1}^k (-1)^i (A^{-1}E)^i A^{-1}$.

(a) Show that for any $k \in \mathbb{N}$,

$$(A + E)S_k = I + (-1)^k \left(EA^{-1}\right)^{k+1}.$$

(b) Argue that $S_\infty := \lim_k S_k$ exists and $(A + E)^{-1} = S_\infty = A^{-1} + \sum_{i=1}^\infty (-1)^i (A^{-1}E)^i A^{-1}$.

(c) Let $\gamma = \|E\|_{\rm op} / \lambda_{\min}(A) < 1$. Show that

$$\left\|(A + E)^{-1} - S_k\right\|_{\rm op} \le \gamma^{k+1} \frac{1}{\lambda_{\min}(A)(1 - \gamma)}.$$

Exercise 5.6: Let $f : \mathbb{R} \to \mathbb{R}$ be a three-times differentiable convex function. We say f is *C-quasi-self-concordant (q.s.c.)* if $|f'''(t)| \le C f''(t)$ for all $t \in \mathbb{R}$ and $f'' > 0$.

(a) Define $g(t) = \log f''(t)$. Show that if f is C-q.s.c., then $|g'(t)| \le C$ and so for any $s \in \mathbb{R}$,

$$e^{-C|s|} f''(t) \le f''(t + s) \le e^{C|s|} f''(t).$$

(b) Show that the function $f(t) = \log(1 + e^t) + \log(1 + e^{-t})$ is q.s.c., and give its self-concordance parameter.

(c) Show that the function $f(t) = \log(1 + e^t)$ is q.s.c., and give its self-concordance parameter.

(d) Let f be C-q.s.c. and for a fixed $x \in \mathbb{R}^d$, define $h(\theta) := f(\langle \theta, x \rangle)$. Show that for any θ, θ_0 with $\Delta = \theta - \theta_0$, h satisfies

$$e^{-C|\langle \Delta, x \rangle|} \nabla^2 h(\theta) \preceq \nabla^2 h(\theta_0) \preceq e^{C|\langle \Delta, x \rangle|} \nabla^2 h(\theta).$$

Exercise 5.7 (Quasi self-concordant M-estimators [338])**:** Consider a prediction problem of predicting targets y from vectors $x \in \mathbb{R}^d$. A loss ℓ is a *C-quasi self-concordant* loss if we can write

$$\ell(\theta, (x, y)) = h(\langle \theta, x \rangle, y),$$

where for each y, $h(\cdot, y)$ is a C-q.s.c. function (recall Exercise 5.6).

(a) Show that logistic regression with loss $\ell(\theta, (x, y)) = \log(1 + e^{-y\langle x, \theta \rangle})$ and robust linear regression with loss $\ell(\theta, (x, y)) = \log(1 + e^{y - \langle x, \theta \rangle}) + \log(1 + e^{\langle x, \theta \rangle - y})$ are both 1-q.s.c. losses.

For the remainder of the problem, assume that the data $\mathcal{X} \subset \mathbb{R}^d$ satisfy $\|x\|_2 \le \sqrt{d}$ for all $x \in \mathcal{X}$. Let $L(\theta) = \mathbb{E}[\ell(\theta, (X, Y))]$ and $\theta^\star = \operatorname{argmin}_\theta L(\theta)$. Assume that $\nabla^2 L(\theta^\star) \succeq \lambda I$, where $\lambda > 0$ is fixed, and let $\widehat{L}_n(\theta) = P_n \ell(\theta, (X, Y))$ as usual.

(b) Show that if $\|\theta - \theta^\star\|_2 \le 1/\sqrt{d}$, then $\nabla^2 \widehat{L}_n(\theta) \succeq e^{-C} \nabla^2 \widehat{L}_n(\theta^\star)$.

(c) Argue that if $t = \|\theta - \theta^\star\|_2 \le \frac{1}{\sqrt{d}}$ and $\|\nabla \widehat{L}_n(\theta^\star)\|_2 \le \gamma$, then

$$\widehat{L}_n(\theta) \ge \widehat{L}_n(\theta^\star) - \gamma t + \frac{e^{-C}\lambda_{\min}(\nabla^2 \widehat{L}_n(\theta^\star))}{2} t^2.$$

(d) Let ℓ be a 1-q.s.c. loss with $|h'(t,y)| \le 1$ and $|h''(t,y)| \le 1$ for $t \in \mathbb{R}$. Give a result similar to Theorem 5.3.8, but show that the conclusions hold with probability at least $1 - \delta$ as soon as

$$n \gtrsim \frac{d^2}{\lambda^2} \log \frac{d}{\delta}.$$

Exercise 5.8 (Truncation to obtain a moment bound)**:** Let $B < \infty$. Show that under the conditions of Corollary 5.3.10,

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2 \wedge B\right] \le \frac{\mathrm{tr}(\nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1})}{n} + \frac{C \log n}{n^{3/2}},$$

where C is a problem-dependent constant.

Exercise 5.9: Let $\widehat{\theta}_n = \mathrm{argmin}_\theta \widehat{L}_n(\theta)$ for an M-estimation problem satisfying the conditions of Corollary 5.3.10. Show that for any $k \ge 1$, there are events $\mathcal{E}_n$ with $\mathbb{P}(\mathcal{E}_n) \ge 1 - n^{-k}$ and for which

$$\left|\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|_2^2 \mathbf{1}\left\{\mathcal{E}_n\right\}\right] - \frac{1}{n} \mathrm{tr}\left(\nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1}\right)\right| \le \frac{Ck \log n}{n^{3/2}},$$

where C is a problem-dependent constant.

Exercise 5.10: This problem provides sufficient conditions for log-loss minimizers to exist for exponential family models.

(a) Let P_θ be a minimal exponential family (Definition 3.2) with density $p_\theta(x) = \exp(\theta^\top x - A(\theta))$ with respect to a base measure μ. Show that for any $\theta^\star \in \mathrm{dom}\, A$, the well-specified population loss $L(\theta) = -\mathbb{E}_{\theta^\star}[\log p_\theta(X)]$ has unique minimizer $\theta^\star$.

For the remainder of the problem, we no longer assume the model is well-specified. For a measure μ on a set $\mathcal{X}$, recall that the essential supremum of a function $f : \mathcal{X} \to \mathbb{R}$ is

$$\operatorname*{ess\,sup}_\mu(f) := \inf\left\{t \in \mathbb{R} \mid \mu\left(\{f(x) \ge t\}\right) = 0\right\}.$$

We say that a measure μ on $\mathbb{R}^d$ *essentially covers* a vector $v \in \mathbb{R}^d$ if

$$v^\top u < \operatorname*{ess\,sup}_\mu\{x^\top u\} \quad \text{for all } u \ne 0. \tag{5.5.1}$$

(b) Let $\{P_\theta\}$ be the exponential family with density $p_\theta(x) = \exp(\theta^\top x - A(\theta))$ with respect to the measure μ. Let X be a random variable for which μ essentially covers (5.5.1) the mean $\mathbb{E}[X]$. Show that $L(\theta) = -\mathbb{E}[\log p_\theta(X)]$ has a minimizer.

Hint. A continuous convex function h has a minimizer if it is coercive, meaning that $h(\theta) \to \infty$ whenever $\|\theta\|_2 \to \infty$. Corollary C.3.6, part (i) may be useful.

Exercise 5.11: We extend Exercise 5.10 to generalized linear models, where $L(\theta) = \mathbb{E}[\ell(\theta, (X, Y))]$.

(a) Consider Poisson regression, $\ell(\theta, (x, y)) = -yx^\top\theta + e^{\theta^\top x}$. Show that L has a unique minimizer if $\mathbb{E}[X] = 0$ and $\mathrm{Cov}(X) = \mathbb{E}[XX^\top] \succ 0$.

(b) Consider logistic regression, $\ell(\theta, (x, y)) = -yx^\top\theta + \log(1 + e^{\theta^\top x})$ for $y \in \{0, 1\}$. Show that L has a unique minimizer if $\mathbb{E}[XX^\top] \succ 0$ and $0 < \mathbb{P}(Y = 1 \mid X) < 1$ with probability 1 over X.

(c) Consider a generalized linear model with densities $p_\theta(y \mid x) = \exp(\phi(x, y)^\top\theta - A(\theta \mid x))$ w.r.t. a base measure $\mu(\cdot \mid x)$ on $y \in \mathcal{Y}$, and assume for simplicity that $\mu(\mathcal{Y} \mid x) = 1$ for all x. Assume that for each vector $v \in \mathbb{S}^{d-1}$ and $x \in \mathcal{X}$,

$$\operatorname*{ess\,sup}_{\mu(\cdot|x)}\{v^\top\phi(x, y)\} \geq \mathbb{E}_P[v^\top\phi(x, Y) \mid X = x],$$

and the set of x for which strict inequality holds has positive P-probability (recall the essential cover condition (5.5.1)). Show that for $\ell(\theta, (x, y)) = -\log p_\theta(y \mid x)$, a minimizer of L exists. You may assume $\mathbb{E}[|A(\theta \mid X)|] < \infty$ for $\|\theta\|_2 \leq 1$.

Hint. The techniques to solve Exercise 5.10 may be useful. In addition, see Exercise 4.2.

Exercise 5.12: Consider the robust regression setting of Example 5.3.3, and let $h \geq 0$ be symmetric, convex, and $\mathcal{C}^2$ near 0. Assume that for any (measurable) subset $\mathcal{X}_0 \subset \mathcal{X}$, $\mathbb{E}[Y \mid X \in \mathcal{X}_0]$ exists and is finite, and $\mathbb{E}[XX^\top] \succ 0$. Show that a minimizer of $L(\theta) := \mathbb{E}[h(\langle\theta, X\rangle - Y)]$ exists. *Hint.* Show that L is coercive. Corollary C.3.6, part (i) may be useful.

Exercise 5.13 (The delta method for approximate sums)**:** Let $T : \mathbb{R}^d \to \mathbb{R}^p$ be a differentiable function with derivative matrix $\dot{T}(\theta) \in \mathbb{R}^{p\times d}$, so that $T(\theta+\Delta) = T(\theta)+\dot{T}(\theta)\Delta+o(\|\Delta\|)$ as $\Delta \to 0$. For an i.i.d. sequence Z_i and remainder term R_n, let $\widehat{\theta}_n \in \mathbb{R}^d$ be a sequence of random vectors with

$$\widehat{\theta}_n - \theta = P_nZ + R_n.$$

(a) Assume $\mathbb{E}[\|Z_i\|_2^2] < \infty$ and that for each $\epsilon > 0$, $\mathbb{P}(\|R_n\|_2 \geq \epsilon/\sqrt{n}) \to 0$ as $n \to \infty$. Show that

$$\widehat{\theta}_n - \theta = \dot{T}(\theta)P_nZ + R_n',$$

where the remainder R_n' also satisfies $\mathbb{P}(\|R_n'\|_2 \geq \epsilon/\sqrt{n}) \to 0$ for all $\epsilon > 0$.

(b) Assume that T is locally smooth enough that for some $K < \infty$ and $\delta > 0$,

$$\left\|T(\theta + \Delta) - T(\theta) - \dot{T}(\theta)\Delta\right\|_2 \leq K\|\Delta\|_2^2 \quad \text{when} \quad \|\Delta\|_2 \leq \delta.$$

Assume there exist $C_0, C_1 < \infty$ such that $\|R_n\|_2 \leq \frac{C_0t}{n}$ with probability at least $1 - e^{-t}$ and that $\mathbb{P}(\|P_nZ\|_2 \geq t) \leq C_1\exp(-nt^2/\sigma^2)$ for $t \geq 0$. Give a quantitative version of part (a).

Exercise 5.14 (Moment bounds on random matrices [109])**:** Let $X_i \in \mathcal{H}_d$ be mean-zero random Hermitian matrices, $p \geq 2$, and assume $\mathbb{E}[\|X_i\|_{\mathrm{op}}^p] < \infty$ for each i. Define the sum $S_n := \sum_{i=1}^n X_i$.

(a) Define the random variables

$$\sigma_n^2 := \left\|\sum_{i=1}^n X_i^2\right\|_{\mathrm{op}} \quad \text{and} \quad b_n := \max_{i\leq n}\|X_i\|_{\mathrm{op}}.$$

Use symmetrization and the matrix Bernstein inequality (4.5.3) to show that for a numerical constant $c < \infty$,

$$\mathbb{E}\left[\|S_n\|_{\rm op}^p\right] \le (cp\log d)^{p/2}\mathbb{E}[\sigma_n^p] + (cp\log d)^p\mathbb{E}[b_n^p],$$

Hint. It may be useful to recall that if $Z \ge 0$, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge t)dt$, and the bounds of Exercise 4.21 on the incomplete Gamma function may be helpful.

(b) Let X_i be i.i.d. and assume that $\mathbb{E}[\|X_i\|_{\rm op}^p] < \infty$ for some $p \ge 2$. Show that

$$\mathbb{E}\left[\left\|n^{-1}S_n\right\|_{\rm op}^p\right]^{1/p} \le O(1)\left(\sqrt{\frac{\log d}{n}} + \frac{\log^{1/p} d}{n^{1-1/p}}\right)\mathbb{E}[\|X_1\|_{\rm op}^p]^{1/p}.$$

Exercise 5.15* (A moment-based estimation guarantee)**:** This question relaxes Assumption A.5.1 to obtain convergence guarantees under local smoothness conditions for M-estimators assuming only moment-based bounds on smoothness constants, providing a stronger version of Theorem 5.3.8.

Assumption A.5.2. *Let $\theta^\star$ minimize $L(\theta)$ and $k \ge 4$. There exist $M_0, M_1, M_2 : \mathcal{Z} \to \mathbb{R}_+$ satisfying*

(i) $\|\nabla\ell(\theta^\star, z)\|_2 \le M_0(z)$, *where* $\mathbb{E}[M_0(Z)^k]^{1/k} \le \overline{M_0}$.

(ii) $\left\|\nabla^2\ell(\theta^\star, z)\right\|_2 \le M_1(z)$, *where* $\mathbb{E}[M_1(Z)^k]^{1/k} \le \overline{M_1}$.

(iii) $\nabla^2\ell(\theta, z)$ *is* $M_2(z)$*-Lipschitz in a radius* $r > 0$ *ball around* $\theta^\star$, *where* $\mathbb{E}[M_2(Z)^k]^{1/k} \le \overline{M_2}$.

(iv) There exists $\lambda > 0$ *such that* $\nabla^2 L(\theta^\star) \succeq \lambda I$.

Under Assumption A.5.2, prove the following for all $t \ge 0$:

(a) Once $n \ge \log^2 d$, with probability at least $1 - t$,

$$\left\|\nabla\widehat{L}_n(\theta^\star)\right\|_2 \le \gamma_n(t) \quad \text{and} \quad \left\|\nabla^2\widehat{L}_n(\theta^\star) - \nabla^2 L(\theta^\star)\right\|_{\rm op} \le \epsilon_n(t)$$

where $\gamma_n(t) := t^{-1/k} \cdot \frac{C_k}{\sqrt{n}}\overline{M_0}$ and $\epsilon_n(t) := t^{-1/k}\frac{C_k\overline{M_1}\sqrt{\log d}}{\sqrt{n}}$ for a C_k depending only on k.

(b) With probability at least $1 - t$,

$$\frac{1}{n}\sum_{i=1}^n M_2(Z_i) \le \mathbb{E}[M_2(Z)] + t^{-1/k}\frac{C_k}{\sqrt{n}}\overline{M_2}.$$

(c) There exist numerical constants $0 < c_0, c_1 < \infty$ such that with probability at least $1 - t$, so long as $n \ge \log^2 d$, $\epsilon_n(t) \le \frac{\lambda}{4}$, and $\gamma_n(t) \le c_0\frac{\lambda^2}{\mathbb{E}[M_2(Z)]}$, and $t^{-1/k}\frac{C_k}{\sqrt{n}}\overline{M_2} \le \mathbb{E}[M_2(Z)]$, then

$$\|\widehat{\theta}_n - \theta^\star\|_2 \le c_1\frac{\gamma_n(t)}{\lambda}.$$

Hint. Use the Marcinkiewicz-Zygmund inequality in Exercises 4.5 and 4.6. It extends to $d \times d$ Hermitian matrices [109, Thm. A.1]: if $X_i \in \mathcal{H}_d$ are mean-zero, their sum $S_n := \sum_{i=1}^n X_i$ satisfies

$$\mathbb{E}\left[\|S_n\|_{\rm op}^k\right]^{1/k} \le 2\sqrt{e\max\{k, 2\log d\}}\left\|\left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{1/2}\right\|_{\rm op} + 4e\max\{k, 2\log d\}\mathbb{E}\left[\max_{i\le n}\|X_i\|_{\rm op}^k\right]^{1/k}.$$

Chapter 6

Generalization and stability

Concentration inequalities provide powerful techniques for demonstrating when random objects that are functions of collections of independent random variables—whether sample means, functions with bounded variation, or collections of random vectors—behave similarly to their expectations. This chapter continues the exploration of these ideas by incorporating the central thesis of this book: that information theory's connections to statistics center around measuring when and how two probability distributions get close to one another. At a high level, we remain focused on the main objects of the preceding chapters, where we have a population probability distribution P on a space $\mathcal{X}$ and some collection of functions $f : \mathcal{X} \to \mathbb{R}$. We wish to understand when the empirical distribution

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i},$$

defined by the sample $X_i \stackrel{\text{iid}}{\sim} P$, will be close to the population P as measured by f. Following the notation we introduce in Section 5.1, for $Pf := \mathbb{E}_P[f(X)]$, we again ask that

$$P_n f - Pf = \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}_P[f(X)] \right)$$

be small simultaneously for all f.

In this chapter, however, we develop a family of tools based around *PAC (probably approximately correct) Bayesian* bounds, where we slightly perturb the functions f of interest to average them in some way; when these perturbations keep $P_n f$ stable, we expect that $P_n f \approx Pf$, that is, the sample generalizes to the population. These perturbations allow us to bring the tools of the divergence measures we have developed to bear on the problems of convergence and generalization. They also allow us to go beyond the "basic" concentration inequalities to situations with interaction, where a data analyst may evaluate some functions of P_n, then adaptively choose additional queries or analyses to do on the sample X_1^n. This breaks standard statistical analyses—which assume an *a priori* specified set of hypotheses or questions to be answered—but is possible to address once we can limit the information the analyses release in precise ways that information-theoretic tools allow. Even more, in the next chapter we show how they form the basis for *transportation inequalities*, powerful tools for concentration of measure. Modern work has also shown how to leverage these techniques, coupled with computation, to provide non-vacuous bounds on learning for complicated scenarios and models to which all classical bounds fail to apply, such as deep learning.

6.1 The variational representation of Kullback-Leibler divergence

The starting point of all of our generalization bounds is a surprisingly simple variational result, which relates expectations, moment generating functions, and the KL-divergence in one single equality. By relating means with moment generating functions and divergences, this inequality allows us to prove generalization bounds based on information-theoretic tools and stability. Sometimes, it will be helpful to use quantized versions of the divergence, so we recall that simple functions are those of the form $g(x) = \sum_{i=1}^k c_i \mathbf{1}\{x \in A_i\}$ for (measurable) sets A_i, scalars c_i, and k finite.

Theorem 6.1.1 (Donsker-Varadhan variational representation)**.** *Let P and Q be distributions on a common space $\mathcal{X}$. Then*

$$D_{\rm kl}\left(P\|Q\right) = \sup_g \left\{\mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]\right\},$$

where the supremum is taken over measurable functions $g : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_Q[e^{g(X)}] < \infty$. We can also take the supremum over bounded simple functions g.

The proof of this result has similarities to that of Hoeffding's lemma (Example 4.1.6) in that we *tilt* the distribution Q by an exponential e^g to verify the equality. For readers familiar with convex optimization, a discrete version of this result may be intuitive, because the function $\log(\sum_{j=1}^k e^{x_j})$ is the convex conjugate of the negative entropy. Exercises 6.1 and 6.2 explore this duality.

Proof We may assume that P is absolutely continuous with respect to Q, meaning that $Q(A) = 0$ implies that $P(A) = 0$, as otherwise both sides are infinite by inspection (even with simple functions in the supremum). Thus, it is no loss of generality to let P and Q have densities p and q.

Attainment in the equality is easy: we take $g(x) = \log \frac{p(x)}{q(x)}$, so that $\mathbb{E}_Q[e^{g(X)}] = 1$. To show that the right-hand side is never larger than $D_{\rm kl}\left(P\|Q\right)$ requires more work. Let g be any function such that $\mathbb{E}_Q[e^{g(X)}] < \infty$, and define the random variable $Z_g(x) = e^{g(x)}/\mathbb{E}_Q[e^{g(X)}]$, so $\mathbb{E}_Q[Z_g] = 1$. Then using that P and Q have density ratio $\frac{p(X)}{q(X)} = \frac{dP}{dQ}$, we have

$$\begin{aligned}
\mathbb{E}_P[\log Z_g] = \mathbb{E}_P\left[\log \frac{p(X)}{q(X)} + \log\left(Z_g(X)\frac{q(X)}{p(X)}\right)\right] &= D_{\rm kl}\left(P\|Q\right) + \mathbb{E}_P\left[\log\left(Z_g \frac{dQ}{dP}\right)\right] \\
&\le D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_P\left[\frac{dQ}{dP} Z_g\right] \\
&= D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_Q[Z_g].
\end{aligned}$$

As $\mathbb{E}_Q[Z_g] = 1$, using that $\mathbb{E}_P[\log Z_g] = \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]$ gives the result.

For the sufficiency of simple functions, we use the definition (2.2.1) of the KL-divergence as a supremum over partitions. For a finite partition $\mathcal{A}$, define the bounded simple function

$$g_{\mathcal{A}}(x) := \sum_{A \in \mathcal{A}: P(A)>0, Q(A)>0} \mathbf{1}\{x \in A\} \log \frac{P(A)}{Q(A)}.$$

Then $D_{\rm kl}\left(P\|Q \mid \mathcal{A}\right) = \mathbb{E}_P[g_{\mathcal{A}}(X)] - \log \mathbb{E}_Q[e^{g_{\mathcal{A}}(X)}]$ because sets with $Q(A) = 0$ have $P(A) = 0$ and contribute nothing (as do sets with $P(A) = 0$, using the convention $0 \log \frac{0}{Q(A)} = 0$). Take a supremum over partitions $\mathcal{A}$. □

The Donsker-Varadhan variational representation already gives a hint that we can use some information-theoretic techniques to control the difference between an empirical sample and its expectation, at least in an average sense. In particular, we see that for any function g, we have

$$\mathbb{E}_P[g(X)] \le D_{\rm kl}\left(P\|Q\right) + \log \mathbb{E}_Q[e^{g(X)}]$$

for any random variable X. Turning this around, suppose that we consider a collection of functions $\mathcal{F}$ and put two probability measures π and π_0 on $\mathcal{F}$, and consider $P_n f - Pf$, where we consider f a random variable $f \sim \pi$ or $f \sim \pi_0$. Then a consequence of the Donsker-Varadhan theorem is that

$$\int (P_n f - Pf) d\pi(f) \le D_{\rm kl}\left(\pi\|\pi_0\right) + \log \int \exp(P_n f - Pf) d\pi_0(f)$$

for any π, π_0. While this inequality is a bit naive—bounding a difference by an exponent seems wasteful—it has substantial applications when we can upper bound the KL-divergence $D_{\rm kl}\left(\pi\|\pi_0\right)$.

6.2 PAC-Bayes bounds

Probably-approximately-correct (PAC) Bayesian bounds proceed from a perspective similar to that of the covering numbers and covering entropies we develop in Section 5.1, where if for a collection of functions $\mathcal{F}$ there is a finite subset (a cover) $\{f_v\}$ such that each $f \in \mathcal{F}$ is "near" one of the f_v, then we need only control deviations of $P_n f$ from Pf for the elements of $\{f_v\}$. In PAC-Bayes bounds, we instead average functions f with other functions, and this averaging allows a similar family of guarantees and applications.

Let $\mathcal{F}$ be a collection of σ^2-sub-Gaussian functions $f : \mathcal{X} \to \mathbb{R}$, which we recall (Definition 4.1) means that $\mathbb{E}[e^{\lambda(f(X) - Pf)}] \le \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$, where $Pf = \mathbb{E}_P[f(X)] = \int f(x) dP(x)$ denotes the expectation of f under P. The main theorem of this section shows that averages of the squared error $(P_n f - Pf)^2$ of the empirical distribution P_n to P converge quickly to zero for *all* averaging distributions π on functions $f \in \mathcal{F}$ so long as each f is σ^2-sub-Gaussian, with the caveat that we pay a cost for different choices of π. The key is that we first fix a prior distribution π_0 on $\mathcal{F}$.

Theorem 6.2.1. *Let Π be the collection of all probability distributions on the set $\mathcal{F}$ and let π_0 be a fixed prior probability distribution on $f \in \mathcal{F}$. With probability at least $1 - \delta$,*

$$\int (P_n f - Pf)^2 d\pi(f) \le \frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi\|\pi_0\right) + \log \frac{2}{\delta}}{n} \quad \textit{simultaneously for all } \pi \in \Pi.$$

Proof The key is to combine Example 4.1.12 with the variational representation that Theorem 6.1.1 provides for KL-divergences. We state Example 4.1.12 as a lemma here.

Lemma 6.2.2. *Let Z be a σ^2-sub-Gaussian random variable. Then for $\lambda \ge 0$,*

$$\mathbb{E}[e^{\lambda Z^2}] \le \frac{1}{\sqrt{[1 - 2\sigma^2\lambda]_+}}.$$

Without loss of generality, we assume that $Pf = 0$ for all $f \in \mathcal{F}$, and recall that $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ is the empirical mean of f. Then $P_n f$ is σ^2/n-sub-Gaussian, and Lemma 6.2.2 implies that $\mathbb{E}[\exp(\lambda(P_n f)^2)] \le \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}$ for any f, and so for any prior π_0 on f we have

$$\mathbb{E}\left[\int \exp(\lambda(P_n f)^2) d\pi_0(f)\right] \le \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}.$$

Consequently, taking $\lambda = \lambda_n := \frac{3n}{8\sigma^2}$, we obtain

$$\mathbb{E}\left[\int \exp(\lambda_n (P_n f)^2) d\pi_0(f)\right] = \mathbb{E}\left[\int \exp\left(\frac{3n}{8\sigma^2}(P_n f)^2\right) d\pi_0(f)\right] \le 2.$$

Markov's inequality thus implies that

$$\mathbb{P}\left(\int \exp\left(\lambda_n (P_n f)^2\right) d\pi_0(f) \ge \frac{2}{\delta}\right) \le \delta, \tag{6.2.1}$$

where the probability is over $X_i \stackrel{\text{iid}}{\sim} P$.

Now, we use the Donsker-Varadhan equality (Theorem 6.1.1). Letting $\lambda > 0$, we define the function $g(f) = \lambda (P_n f)^2$, so that for any two distributions π and π_0 on $\mathcal{F}$, we have

$$\frac{1}{\lambda}\int g(f) d\pi(f) = \int (P_n f)^2 d\pi(f) \le \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \int \exp(\lambda (P_n f)^2) d\pi_0(f)}{\lambda}.$$

This holds without any probabilistic qualifications, so using the application (6.2.1) of Markov's inequality with $\lambda = \lambda_n$, we thus see that with probability at least $1 - \delta$ over $X_1, \ldots, X_n$, simultaneously for all distributions π,

$$\int (P_n f)^2 d\pi(f) \le \frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{2}{\delta}}{n}.$$

This is the desired result (as we have assumed that $Pf = 0$ w.l.o.g.). □

By Jensen's inequality (or Cauchy-Schwarz), it is immediate from Theorem 6.2.1 that we also have

$$\int |P_n f - Pf| d\pi(f) \le \sqrt{\frac{8\sigma^2}{3} \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{2}{\delta}}{n}} \text{ simultaneously for all } \pi \in \Pi \tag{6.2.2}$$

with probability at least $1 - \delta$, so $\mathbb{E}_\pi[|P_n f - Pf|]$ is of order $1/\sqrt{n}$ with high probability. The inequality (6.2.2) is the original form of McAllester's PAC-Bayes bound, with slightly sharper constants and improved logarithmic dependence. The key is that *stability*, in the form of a prior π_0 and posterior π closeness, allows us to control the deviations of random variables and functions with high probability. (We intend no Bayesian modeling assumptions; "posterior" simply means any data-dependent distribution on $\mathcal{F}$.)

Let us give an example that recovers the classic covering-based uniform law in Example 5.2.12 using these tools to illustrate the approach. The basic idea is that by appropriate choice of prior π_0 and "posterior" π, smooth classes of functions admit generalization guarantees.

Example 6.2.3 (A uniform law for Lipschitz functions)**:** Consider a case as in Section 5.2, where we let $L(\theta) = P\ell(\theta, Z)$ for some $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$. Let $\mathbb{B}_2^d = \{v \in \mathbb{R}^d \mid \|v\|_2 \le 1\}$ be the ℓ_2-ball in $\mathbb{R}^d$ and assume that $\Theta \subset r\mathbb{B}_2^d$ and additionally that $\theta \mapsto \ell(\theta, z)$ is M-Lipschitz for all $z \in \mathcal{Z}$. For simplicity, we assume that $\ell(\theta, z) \in [0, 2Mr]$ for all $\theta \in \Theta$ (we may simply relativize our bounds by replacing ℓ by $\ell(\cdot, z) - \inf_{\theta \in \Theta} \ell(\theta, z) \in [0, 2Mr]$).

If $\widehat{L}_n(\theta) = P_n \ell(\theta, Z)$, then Theorem 6.2.1 implies that

$$\int |\widehat{L}_n(\theta) - L(\theta)| d\pi(\theta) \le \sqrt{\frac{8M^2 r^2}{3n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{2}{\delta}\right]}$$

for all π with probability at least $1-\delta$. Now, let $\theta_0 \in \Theta$ be arbitrary, and for $\epsilon > 0$ (to be chosen later) take π_0 to be uniform on $(r+\epsilon)\mathbb{B}_2^d$ and π to be uniform on $\theta_0 + \epsilon\mathbb{B}_2^d$. Then we immediately see that $D_{\rm kl}\left(\pi\|\pi_0\right) = d\log(1+\frac{r}{\epsilon})$. Moreover, $\int \widehat{L}_n(\theta)d\pi(\theta) \in \widehat{L}_n(\theta_0) \pm M\epsilon$ and similarly for $L(\theta)$, by the M-Lipschitz continuity of ℓ. For any fixed $\epsilon > 0$, therefore,

$$|\widehat{L}_n(\theta_0) - L(\theta_0)| \le 2M\epsilon + \sqrt{\frac{8M^2r^2}{3n}\left[d\log\left(1+\frac{r}{\epsilon}\right) + \log\frac{2}{\delta}\right]}$$

simultaneously for all $\theta_0 \in \Theta$, with probability at least $1-\delta$. Choosing $\epsilon = \frac{rd}{n}$ recovers Example 5.2.12: with high probability, $|\widehat{L}_n(\theta) - L(\theta)| \le O(1)Mr\sqrt{\frac{d}{n}\log\frac{n}{d}}$ for all $\theta \in \Theta$. $\diamond$

On the one hand, the result in Example 6.2.3 is satisfying: it applies to any Lipschitz function and provides a uniform bound. On the other hand, when we compare to the results achievable for specially structured linear function classes via Rademacher complexity bounds—such as Proposition 5.2.10 and Example 5.2.11—we have weaker results, in that they depend on the dimension explicitly. We will give an example presently showing how to address some of these issues.

6.2.1 Relative bounds

To provide finer control on $P_nf - Pf$ than the results so far, we adapt Bennett- and and Bernstein-type inequalities (Proposition 4.1.21) to obtain concentration guarantees that depend not on gross bounds on the sub-Gaussian parameters associated with f but on variances of functions $f \in \mathcal{F}$. The approach is similar to our derivation in Theorem 6.2.1, where we show that the moment generating function of a quantity like $P_nf - Pf$ is small (Eq. (6.2.1)). We relate this—via the Donsker-Varadhan representation in Theorem 6.1.1—to the quantities we wish to control. In the next proposition, we provide relative bounds on the deviations of functions from their means. To make this precise, let $\sigma^2(f) := \mathrm{Var}(f(X))$ be the variance of a function $f : \mathcal{X} \to \mathbb{R}$. We assume the class $\mathcal{F}$ satisfies the Bernstein condition (4.1.8) with parameter b,

$$\left|\mathbb{E}\left[(f(X) - Pf)^k\right]\right| \le \frac{k!}{2}\sigma^2(f)b^{k-2} \quad \text{for } k = 3, 4, \ldots. \tag{6.2.3}$$

By relating variance to higher moments of $f \in \mathcal{F}$, we can demonstrate the following result.

Proposition 6.2.4. *Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying the Bernstein condition (6.2.3). Then for fixed $|\lambda| \le \frac{1}{2b}$, with probability at least $1-\delta$,*

$$\lambda\int Pfd\pi(f) - \lambda^2\int\sigma^2(f)d\pi(f) \le \lambda\int P_nfd\pi(f) + \frac{1}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right]$$

simultaneously for all $\pi \in \Pi$.

Proof Lemma 4.1.20 implies that if $|\mathbb{E}[(X-\mu)^k]| \le \frac{k!}{2}\sigma^2b^{k-2}$ for $k \ge 2$,

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le \exp(\lambda^2\sigma^2)$$

for $|\lambda| \le 1/(2b)$. As a consequence, for any f in our collection $\mathcal{F}$, we see that if we define

$$\Delta_n(f,\lambda) := \lambda\left[P_nf - Pf - \lambda\sigma^2(f)\right],$$

we have

$$\mathbb{E}[\exp(n\Delta_n(f,\lambda))] = \mathbb{E}[\exp(\lambda(f(X) - Pf) - \lambda^2\sigma^2(f))]^n \le 1$$

for all n, $f \in \mathcal{F}$, and $|\lambda| \le \frac{1}{2b}$. For any fixed measure π_0 on $\mathcal{F}$, Markov's inequality implies that

$$\mathbb{P}\left(\int \exp(n\Delta_n(f,\lambda))d\pi_0(f) \ge \frac{1}{\delta}\right) \le \delta. \tag{6.2.4}$$

As in the proof of Theorem 6.2.1, the Donsker-Varadhan Theorem 6.1.1 implies that

$$n\int \Delta_n(f,\lambda)d\pi(f) \le D_{\rm kl}\left(\pi\|\pi_0\right) + \log\int \exp(n\Delta_n(f,\lambda))d\pi_0(f)$$

for all distributions π. Using inequality (6.2.4), we obtain that with probability at least $1-\delta$,

$$\int \Delta_n(f,\lambda)d\pi(f) \le \frac{1}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right]$$

for all π. As this holds for any $|\lambda| \le 1/(2b)$, the desired result follows by rearranging and choosing the sign of λ as necessary. □

We would like to optimize over the bound in Proposition 6.2.4 by choosing the "best" λ. If we *could* choose the optimal λ, by rearranging Proposition 6.2.4 we would obtain the bound

$$\begin{aligned}\mathbb{E}_\pi[Pf] &\le \mathbb{E}_\pi[P_nf] + \inf_{\lambda>0}\left\{\lambda\mathbb{E}_\pi[\sigma^2(f)] + \frac{1}{n\lambda}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right]\right\} \\ &= \mathbb{E}_\pi[P_nf] + 2\sqrt{\frac{\mathbb{E}_\pi[\sigma^2(f)]}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right]}\end{aligned}$$

simultaneously for all π, with probability at least $1-\delta$. This approach has problems: first, we cannot arbitrarily choose λ in Proposition 6.2.4, and second, this putative bound depends on the unknown variance $\sigma^2(f)$. It is thus of interest to develop situations in which we can replace unknown population quantities on the right side of the bound with known quantities.

To that end, let us consider a relative error condition related to the Bernstein condition (6.2.3): for each $f \in \mathcal{F}$,

$$\sigma^2(f) \le bPf. \tag{6.2.5}$$

This condition is natural when f takes nonnegative values, for example, when $f(X) = \ell(\theta, X)$ for some loss ℓ and parameter θ of a model. If the functions f are nonnegative and have upper bound b, then certainly $\sigma^2(f) \le \mathbb{E}[f(X)^2] \le b\mathbb{E}[f(X)] = bPf$, so that condition (6.2.5) holds. Revisiting Proposition 6.2.4, we rearrange to obtain the following theorem.

Theorem 6.2.5. *Let $\mathcal{F}$ be a collection of functions satisfying the Bernstein condition (6.2.3) as in Proposition 6.2.4, and in addition, assume the variance-bounding condition (6.2.5). Then for any $0 \le \lambda \le \frac{1}{2b}$, with probability at least $1-\delta$,*

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_nf] + \frac{\lambda b}{1-\lambda b}\mathbb{E}_\pi[P_nf] + \frac{1}{\lambda(1-\lambda b)}\frac{1}{n}\left[D_{\rm kl}\left(\pi\|\pi_0\right) + \log\frac{1}{\delta}\right]$$

for all π.

Proof For $\lambda > 0$, use condition (6.2.5) to see that

$$\lambda \mathbb{E}_\pi[Pf] - \lambda^2 b \mathbb{E}_\pi[Pf] \le \lambda \mathbb{E}_\pi[Pf] - \lambda^2 \mathbb{E}_\pi[\sigma^2(f)],$$

apply Proposition 6.2.4, and divide both sides of the resulting inequality by $\lambda(1 - \lambda b)$. □

To make this uniform in λ, we choose multiple values of λ and apply a union bound. Let $1 + \eta = \frac{1}{1-\lambda b}$, or $\eta = \frac{\lambda b}{1-\lambda b}$ and $\frac{1}{\lambda b(1-\lambda b)} = \frac{(1+\eta)^2}{\eta}$, so the inequality in Theorem 6.2.5 is equivalent to

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{(1+\eta)^2}{\eta} \frac{b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right].$$

For $0 \le \eta \le 1$, this implies

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{1}{\eta} \frac{b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right] + \frac{3b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta} \right].$$

Take $\eta_1 = 1/n, \eta_2 = 2/n, \ldots, \eta_n = 1$. Then optimizing over $\eta \in \{\eta_1, \ldots, \eta_n\}$ (equivalent, to within a $1/n$ factor, to optimizing over $0 < \eta \le 1$) and applying a union bound, we obtain

Corollary 6.2.6. *Let the conditions of Theorem 6.2.5 hold. Then with probability at least* $1 - \delta$,

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + 2\sqrt{\frac{b \mathbb{E}_\pi[P_n f]}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta} \right]} + \frac{1}{n} \left(\mathbb{E}_\pi[P_n f] + 5b \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta} \right] \right),$$

simultaneously for all π *on* $\mathcal{F}$.

Proof By a union bound, for each $\eta \in \{\eta_k = k/n\}_{k=1}^n$,

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + \eta \mathbb{E}_\pi[P_n f] + \frac{1}{\eta} \frac{b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta} \right] + \frac{3b}{n} \left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta} \right].$$

We consider two cases. First, assume that $\mathbb{E}_\pi[P_n f] \le \frac{b}{n}(D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta})$. Then taking $\eta = 1$ gives the result. In the second, we have $\mathbb{E}_\pi[P_n f] > \frac{b}{n}(D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta})$, and we can set

$$\eta_\star = \sqrt{\frac{\frac{b}{n}(D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta})}{\mathbb{E}_\pi[P_n f]}} \in (0, 1).$$

Choose η to be the smallest value η_k with $\eta_k \ge \eta_\star$, so that $\eta_\star \le \eta \le \eta_\star + \frac{1}{n}$. □

We may sharpen the $\log \frac{n}{\delta}$ terms in the corollary to $\log \frac{\log n}{\delta}$; see Exercise 6.4.

6.2.2 A large-margin guarantee

Let us revisit the loss minimization central to Section 5.2 and Example 6.2.3 in the context of Corollary 6.2.6. We investigate an approach to achieve convergence guarantees that are (nearly) dimension-independent, focusing on 0-1 losses in binary classification. Consider data $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$, where we make predictions $\langle \theta, x \rangle$, and for a *margin penalty* $\gamma \ge 0$ define the loss

$$\ell_\gamma(\theta; (x, y)) = \mathbf{1}\left\{ \langle \theta, x \rangle y \le \gamma \right\}.$$

We call $\langle \theta, x \rangle y$ the *margin* of θ on the pair (x, y), so when the margin is large, $\langle \theta, x \rangle$ has the same sign as y and is "confident" (far from zero). Define the expected and empirical margin γ losses

$$L_\gamma(\theta) := P\ell_\gamma(\theta; (X, Y)) \text{ and } \widehat{L}_\gamma(\theta) := P_n\ell_\gamma(\theta; (X, Y)).$$

Consider the following scenario: the data x lie in a ball of radius b, so that $\|x\|_2 \le b$; note that the losses ℓ_γ and ℓ_0 satisfy the Bernstein (6.2.3) and self-bounding (6.2.5) conditions with constant 1 as they take values in $\{0, 1\}$. We then have the following proposition.

Proposition 6.2.7. *Let the above conditions on the data (x, y) hold, and let the margin $\gamma > 0$ and radius $r < \infty$. Then with probability at least $1 - \delta$,*

$$P\big(\langle \theta, X \rangle Y \le 0\big) \le \left(1 + \frac{1}{n}\right) P_n\big(\langle \theta, X \rangle Y \le \gamma\big) + C\sqrt{\frac{r^2 b^2 \log \frac{n}{\delta}}{\gamma^2 n} \cdot P_n\big(\langle \theta, X \rangle Y \le \gamma\big)} + C\frac{r^2 b^2 \log \frac{n}{\delta}}{\gamma^2 n}$$

simultaneously for all $\|\theta\|_2 \le r$, where C is a numerical constant.

Proposition 6.2.7 provides a "dimension-free" guarantee depending only on the ℓ_2-norms $\|\theta\|_2$ and $\|x\|_2$, so it can apply in infinite dimensional spaces. The key to the inequality is that if we can find a large margin predictor—for example, by minimizing a convex loss of the form

$$\underset{\|\theta\|_2 \le r}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n \varphi(\langle X_i, \theta \rangle Y_i)$$

for some decreasing convex surrogate $\varphi : \mathbb{R} \to \mathbb{R}_+$, e.g. $\varphi(t) = [1 - t]_+$ or $\varphi(t) = \log(1 + e^{-t})$, then we obtain generalization guarantees relative to the empirical margin γ. As a particular instantiation of this approach, suppose we obtain a perfect classifier with positive margin: a vector θ with $\|\theta\|_2 \le r$ such that $\langle \theta, X_i \rangle Y_i \ge \gamma$ for each $i = 1, \ldots, n$. Then Proposition 6.2.7 guarantees that

$$P(\langle \theta, X \rangle Y \le 0) \le C\frac{r^2 b^2 \log \frac{n}{\delta}}{\gamma^2 n} \text{ with probability at least } 1 - \delta.$$

Proof Let π_0 be $\mathsf{N}(0, \tau^2 I)$ for some $\tau > 0$ to be chosen, and let π be $\mathsf{N}(\widehat{\theta}, \tau^2 I)$ for some $\widehat{\theta} \in \mathbb{R}^d$ satisfying $\|\widehat{\theta}\|_2 \le r$. Then Corollary 6.2.6 implies that

$$\begin{aligned}
&\mathbb{E}_\pi[L_\gamma(\theta)] \\
&\le \mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + 2\sqrt{\frac{\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)]}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta}\right]} + \frac{1}{n}\left(\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + C\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{n}{\delta}\right]\right) \\
&\le \mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + 2\sqrt{\frac{\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)]}{n}\left[\frac{r^2}{2\tau^2} + \log \frac{n}{\delta}\right]} + \frac{1}{n}\left(\mathbb{E}_\pi[\widehat{L}_\gamma(\theta)] + C\left[\frac{r^2}{2\tau^2} + \log \frac{n}{\delta}\right]\right)
\end{aligned}$$

with probability at least $1 - \delta$, where we use that $D_{\rm kl}\left(\mathsf{N}(\theta, \tau^2 I) \| \mathsf{N}(0, \tau^2 I)\right) = \|\theta\|_2^2 / (2\tau^2)$.

We use the margin assumption. If $Z \sim \mathsf{N}(0, \tau^2 I)$, then for any fixed θ_0, x, y we have

$$\ell_0(\theta_0; (x, y)) - \mathbb{P}(Z^\top x \ge \gamma) \le \mathbb{E}[\ell_\gamma(\theta_0 + Z; (x, y))] \le \ell_{2\gamma}(\theta_0; (x, y)) + \mathbb{P}(Z^\top x \ge \gamma)$$

where the middle expectation is over $Z \sim \mathsf{N}(0, \tau^2 I)$. Using the $\tau^2 \|x\|_2^2$-sub-Gaussianity of $Z^\top x$, we obtain immediately that if $\|x\|_2 \le b$, we have

$$\ell_0(\theta_0; (x, y)) - \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) \le \mathbb{E}[\ell_\gamma(\theta_0 + Z; (x, y))] \le \ell_{2\gamma}(\theta_0; (x, y)) + \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right).$$

Returning to our earlier bound, if $\|x\|_2 \le b$ for all $x \in \mathcal{X}$, then with probability at least $1-\delta$, simultaneously for all $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \le r$,

$$L_0(\theta) \le \widehat{L}_{2\gamma}(\theta) + 2\exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) + 2\sqrt{\frac{\widehat{L}_{2\gamma}(\theta) + \exp(-\frac{\gamma^2}{2\tau^2 b^2})}{n}\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]} + \frac{1}{n}\left(\widehat{L}_{2\gamma}(\theta) + \exp\left(-\frac{\gamma^2}{2\tau^2 b^2}\right) + C\left[\frac{r^2}{2\tau^2} + \log\frac{n}{\delta}\right]\right).$$

Setting $\tau^2 = \frac{\gamma^2}{2b^2 \log n}$ and replacing 2γ with γ gives the proposition. □

6.2.3 A mutual information bound

Mutual information provides an alternative perspective on the PAC-Bayesian bounds that Theorem 6.2.1 gives and will also be central to the interactive data analysis setting in the next section. Assume the setting of Theorem 6.2.1, so that $\mathcal{F}$ consists of σ^2-sub-Gaussian functions. Let us adopt the following observational model: we observe $X_1^n \stackrel{\text{iid}}{\sim} P$, and then conditional on the sample X_1^n, draw a (random) function $F \in \mathcal{F}$ following the distribution $\pi(\cdot \mid X_1^n)$. Fixing the prior π_0, Theorem 6.2.1 guarantees that with probability at least $1-\delta$ over X_1^n,

$$\mathbb{E}[(P_n F - PF)^2 \mid X_1^n] \le \frac{8\sigma^2}{3n}\left[D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right) + \log\frac{2}{\delta}\right],$$

where the expectation is taken over $F \sim \pi(\cdot \mid X_1^n)$, leaving the sample fixed. Now, consider choosing π_0 to be the average over all samples X_1^n of π, that is, $\pi_0(\cdot) = \mathbb{E}_P[\pi(\cdot \mid X_1^n)]$, the expectation taken over $X_1^n \stackrel{\text{iid}}{\sim} P$. Then by definition of mutual information,

$$I(F; X_1^n) = \mathbb{E}_P\left[D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right)\right],$$

and by Markov's inequality we have

$$\mathbb{P}(D_{\rm kl}\left(\pi(\cdot \mid X_1^n) \| \pi_0\right) \ge K \cdot I(F; X_1^n)) \le \frac{1}{K}$$

for $K \ge 0$. Combining these, we obtain the following corollary.

Corollary 6.2.8. *Let F be chosen according to any distribution $\pi(\cdot \mid X_1^n)$ conditional on the sample X_1^n. Then with probability at least $1-\delta_0-\delta_1$ over the sample $X_1^n \stackrel{\text{iid}}{\sim} P$,*

$$\mathbb{E}[(P_n F - PF)^2 \mid X_1^n] \le \frac{8\sigma^2}{3n}\left[\frac{I(F; X_1^n)}{\delta_0} + \log\frac{2}{\delta_1}\right].$$

This corollary shows that if we have any procedure—say, a learning procedure or otherwise—that limits the information between a sample X_1^n and an output F, then we are guaranteed that F generalizes. Tighter analyses are possible, though not our focus here; the reader should already have an inkling that limiting information between input samples and outputs may be fruitful.

6.3 Interactive and adaptive data analysis

A major challenge in modern data analysis is that analyses often fail to follow the (sometimes ossified) prescriptions of classical statistics and the scientific method. In the scientific method, one proposes a hypothesis, the status quo or some other belief, and then designs an experiment to falsify that hypothesis. Upon performing the experiment, there are two options: either the experimental results contradict the hypothesis so that the hypothesis is false (we reject the null), or the hypothesis remains consistent with available data. In classical Fisherian statistics, we develop a null hypothesis H_0 before observing data, we draw a sample $X \in \mathcal{X}$, and then for a test statistic $T : \mathcal{X} \to \mathbb{R}$ with observed value $t_{\rm observed} = T(X)$, we compute the probability under the null of observing something as extreme as what we observed, that is, the p-value $p = P_{H_0}(T(X) \geq t_{\rm observed})$.

Yet modern data analyses differ from this pristine perspective. We often have a number of hypotheses to test, not a single one. For example, in biological applications, we may wish to investigate the associations between the expression of number of genes and a particular phenotype or disease; each gene j then corresponds to a null hypothesis $H_{0,j}$ that gene j is independent of the phenotype. There are numerous approaches to addressing such multiple testing problems, including familywise error control and false discovery rate control [239, 57, 58, 287].

Even these approaches to multiple testing and high-dimensional problems do not capture modern data analyses. Indeed, in many fields, researchers use one or a few main datasets, writing papers and performing multiple analyses on the same dataset. For example, in medicine, the UK Biobank dataset [389] has tens of thousands of citations (as of 2026), many of which build on one another. Even in situations without a shared dataset, analyses present researchers with many degrees of freedom. A researcher may study a summary statistic of their sampled data, performing some simple data exploration—which statisticians have advocated for 50 years [402, 146]—but this introduces *potential* comparisons a researcher might make. This "garden of forking paths," as Gelman and Loken [207] term it, causes challenges even for honest researchers: because we make decisions that may depend on the data observed, these approaches invalidate classical statistical analyses.

To that end, we now consider *interactive*, or *adaptive*, data analyses, where we perform data analyses sequentially, computing new functions on a fixed sample $X_1, \ldots, X_n$ after observing some initial information about the sample. If the function we decide to compute on the data X_1^n is chosen without much information about the data, then its value on the sample should be similar to its values on the full population. This insight dovetails with what we have seen: appropriate "stability" in information can guarantee good future performance.

6.3.1 The interactive setting

We consider a stylized approach to interactive data analysis, as it captures many challenges while being broad enough for applications. In particular, we focus on *statistical queries*, where a data analyst wishes to evaluate expectations

$$\mathbb{E}_P[\phi(X)] \tag{6.3.1}$$

of functionals $\phi : \mathcal{X} \to \mathbb{R}$ under the population P using a sample $X_1^n \stackrel{\rm iid}{\sim} P$. Numerous problems are solvable via statistical queries (6.3.1). Means use $\phi(x) = x$, while we can compute variances using the queries $\phi_1(x) = x$ and $\phi_2(x) = x^2$, as $\mathrm{Var}(X) = \mathbb{E}_P[\phi_2(X)] - \mathbb{E}_P[\phi_1(X)]^2$.

Classical algorithms for the statistical query problem return sample means $P_n\phi := \frac{1}{n}\sum_{i=1}^n \phi(X_i)$. When the queries to be answered are not chosen adaptively, we can answer a large number accu-

rately; indeed, if we have a finite collection Φ of σ^2-sub-Gaussian $\phi : \mathcal{X} \to \mathbb{R}$, then

$$\mathbb{P}\left(\max_{\phi \in \Phi} |P_n \phi - P\phi| \geq \sqrt{\frac{2\sigma^2}{n}(\log(2|\Phi|) + t)}\right) \leq e^{-t} \quad \text{for } t \geq 0$$

by Corollary 4.1.10 (sub-Gaussian concentration) and a union bound. Thus, so long as $|\Phi|$ is not exponential in the sample size n, we expect uniformly high accuracy.

Example 6.3.1 (Risk minimization via statistical queries)**:** Assume the loss-minimization setting (5.2.2), where the losses $\ell(\theta, X_i)$ are convex and differentiable in θ. Then gradient descent applied to $\widehat{L}_n(\theta) = P_n \ell(\theta, X)$ converges to a minimizing value of $\widehat{L}_n$. We can implement gradient descent by a sequence of statistical queries $\phi(x) = \nabla_\theta \ell(\theta, x)$ (handling them coordinatewise), iterating

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k P_n \phi^{(k)}, \tag{6.3.2}$$

where $\phi^{(k)} = \nabla_\theta \ell(\theta^{(k)}, x)$ and α_k is a stepsize. $\diamond$

In Example 6.3.1, we see we *interact* with the dataset, because each sequential query $\phi^{(k)}$ depends on the previous $k-1$ queries, so standard non-adaptive bounds fail to apply.

We thus consider an interactive version of the statistical query estimation problem. There are now two parties: an analyst (or statistician) who issues queries $\phi : \mathcal{X} \to \mathbb{R}$, and a mechanism that answers the queries. We index our functionals ϕ by $t \in \mathcal{T}$ for a (possibly infinite) set $\mathcal{T}$, so we have a collection $\{\phi_t\}_{t \in \mathcal{T}}$. We thus have the following scheme:

Input: Sample X_1^n drawn i.i.d. P, collection $\{\phi_t\}_{t \in \mathcal{T}}$ of possible queries
Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Mechanism responds with answer A_k approximating $P\phi = \mathbb{E}_P[\phi(X)]$ using X_1^n

Figure 6.1: The interactive statistical query setting

In the iteration 6.1, we *interactively* choose $T_1, T_2, \ldots, T_k$, where T_i may depend on the approximations of $\mathbb{E}_P[\phi_{T_j}(X)]$ for $j < i$, that is, on the results of our previous queries. Our goal is to accurately compute expectations $P\phi_T = \mathbb{E}_P[\phi_T(X)]$ when the index T may depend on X_1^n. The setting in Figure 6.1 clearly breaks with the classical statistical setting in which an analysis is pre-specified before collecting data, but more closely captures modern data exploration practices.

6.3.2 Second moment errors and mutual information

Our derivation begins with the following result, which follows from more or less identical arguments to those for our PAC-Bayesian bounds, and demonstrates that if the selection T reveals little information about X_1^n, then $P_n \phi_T$ tracks $P\phi_T$ in mean-squared error and bias.

Theorem 6.3.2. *Let $\{\phi_t\}_{t \in \mathcal{T}}$ be a collection of σ^2-sub-Gaussian functions $\phi_t : \mathcal{X} \to \mathbb{R}$. Then for any random variable T and any $\lambda > 0$,*

$$\mathbb{E}[(P_n \phi_T - P\phi_T)^2] \leq \frac{1}{\lambda}\left[I(X_1^n; T) - \frac{1}{2}\log\left[1 - 2\lambda\sigma^2/n\right]_+\right]$$

and

$$|\mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T]| \le \sqrt{\frac{2\sigma^2}{n} I(X_1^n; T)}$$

where the expectations are taken over T and the sample X_1^n.

Proof The proof is similar to that of our first basic PAC-Bayes result in Theorem 6.2.1. Assume w.l.o.g. that $P\phi_t = 0$ for all $t \in \mathcal{T}$, noting that then $P_n\phi_t$ is σ^2/n-sub-Gaussian. We prove the first result first. Lemma 6.2.2 implies that $\mathbb{E}[\exp(\lambda(P_n\phi_t)^2)] \le \left[1 - 2\lambda\sigma^2/n\right]_+^{-1/2}$ for each $t \in \mathcal{T}$. As a consequence, we obtain via the Donsker-Varadhan equality (Theorem 6.1.1) that

$$\begin{aligned}
\lambda \mathbb{E}\left[\int (P_n\phi_t)^2 d\pi(t)\right] &\stackrel{(i)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi \| \pi_0\right)] + \mathbb{E}\left[\log \int \exp(\lambda(P_n\phi_t)^2) d\pi_0(t)\right] \\
&\stackrel{(ii)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi \| \pi_0\right)] + \log \mathbb{E}\left[\int \exp(\lambda(P_n\phi_t)^2) d\pi_0(t)\right] \\
&\stackrel{(iii)}{\le} \mathbb{E}[D_{\rm kl}\left(\pi \| \pi_0\right)] - \frac{1}{2}\log\left[1 - 2\lambda\sigma^2/n\right]_+
\end{aligned}$$

for all distributions π on $\mathcal{T}$, which may depend on P_n, where the expectation $\mathbb{E}$ is taken over the sample $X_1^n \stackrel{\rm iid}{\sim} P$. (Here inequality (i) is Theorem 6.1.1, inequality (ii) is Jensen's inequality, and inequality (iii) is Lemma 6.2.2.) Let π_0 be the marginal distribution on T (marginally over all observations X_1^n), and let π denote the posterior of T conditional on the sample X_1^n. Then $\mathbb{E}[D_{\rm kl}\left(\pi \| \pi_0\right)] = I(X_1^n; T)$ by definition of the mutual information, giving the squared error bound.

For the second result, note that the Donsker-Varadhan equality implies

$$\lambda \mathbb{E}\left[\int P_n\phi_t d\pi(t)\right] \le \mathbb{E}[D_{\rm kl}\left(\pi \| \pi_0\right)] + \log \int \mathbb{E}[\exp(\lambda P_n\phi_t)] d\pi_0(t) \le I(X_1^n; T) + \frac{\lambda^2\sigma^2}{2n}.$$

Dividing both sides by λ gives $\mathbb{E}[P_n\phi_T] \le \sqrt{2\sigma^2 I(X_1^n; T)/n}$, and performing the same analysis with $-\phi_T$ gives the second result of the theorem. □

The key in the theorem is that if the mutual (Shannon) information $I(X; T)$ between the sample X and T is small, then the expected squared error can be small. To make this a bit clearer, we instantiate the theorem by taking $\lambda = \frac{n}{2e\sigma^2}$.

Corollary 6.3.3. *Let the conditions of Theorem 6.3.2 hold. Then*

$$\mathbb{E}[(P_n\phi_T - P\phi_T)^2] \le \frac{2e\sigma^2}{n} I(X_1^n; T) + \frac{5\sigma^2}{4n}.$$

Consequently, if we can limit the amount of information any particular query T (i.e., ϕ_T) contains about the sample X_1^n, we guarantee small second moment errors $(P_n\phi_T - P\phi_T)^2$.

6.3.3 Limiting interaction in interactive analyses

Let us return to the interactive data analysis setting of Figure 6.1, where we recall the application of estimating mean functionals $P\phi$ for $\phi \in \{\phi_t\}_{t\in\mathcal{T}}$. When we consider the full extent of our goal—to provide accurate approximations to $P\phi$ for queries ϕ chosen adaptively for any population distribution P and $\phi : \mathcal{X} \to [-1, 1]$—we can construct quite perverse situations, where if we compute sample expectations $P_n\phi$ exactly, one round of interaction identifies a query ϕ whose empirical and population means differ by nearly as much as possible.

Example 6.3.4 (Exact answering allows arbitrary corruption)**:** We draw a sample X_1^n of size n on a sample space $\mathcal{X} = [m]$ with $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}([m])$, where $m \geq 2n$. Let Φ be the collection of all functions $\phi : [m] \to [-1, 1]$, so that $\mathbb{P}(|P_n\phi - P\phi| \geq t) \leq \exp(-nt^2/2)$ for any fixed ϕ. Suppose that in the interactive scheme in Figure 6.1, we release answers $A = P_n\phi$. Consider the query

$$\phi(x) = (n+1)^{1-x} \quad \text{for } x = 1, 2, \ldots, m.$$

Then by inspection, we have base-$(n+1)$ expansion

$$\begin{aligned} P_n\phi &= \sum_{j=1}^{m} (n+1)^{-j} \operatorname{card}(\{X_i \mid X_i = j\}) \\ &= \frac{1}{n+1} \operatorname{card}(\{X_i \mid X_i = 1\}) + \ldots + \frac{1}{(n+1)^m} \operatorname{card}(\{X_i \mid X_i = m\}). \end{aligned}$$

Since each count lies in $\{0, 1, \ldots, n\}$, this expansion is unique, so $P_n\phi$ allows exactly reconstructing the sample counts. Then if we define a second query $\phi_2(x) = 1$ for $x \in X_1^n$ and $\phi_2(x) = -1$ for $x \notin X_1^n$, we see that $P\phi_2 \leq \frac{2n}{m} - 1$, while $P_n\phi_2 = 1$. The gap is thus

$$\mathbb{E}[P_n\phi_2 - P\phi_2] \geq 2 - \frac{2n}{m},$$

which tends to the worst possible value 2. $\diamond$

Even in less capricious examples, challenges arise with a single round of interaction.

Example 6.3.5 (Correlation analysis)**:** We observe vectors $X \in \{-1, 1\}^k$ and "labels" $Y \in \{-1, 1\}$. We assume the vectors X are uniform on $\{-1, 1\}^k$ and independent of Y, but we imagine an analyst who seeks a vector v with $\|v\|_2 = 1$ such that the correlation between $v^\top X$ and Y is high, meaning $v^\top X$ is associated with Y. For any *a priori* fixed $v_1, \ldots, v_k, v_{k+1} \in \mathbb{S}^{k-1}$, we have

$$\mathbb{E}[\max_j (v_j^\top (P_n YX))^2] \leq O(1) \frac{\log k}{n}.$$

If instead we allow a *single* interaction, the dependence on k worsens exponentially. By issuing queries $v = e_1, \ldots, e_k$, the k standard basis vectors, and setting $V_{k+1} = P_n YX / \|P_n YX\|_2$,

$$\mathbb{E}[(V_{k+1}^\top (P_n YX))^2] = \mathbb{E}[\|P_n YX\|_2^2] = \frac{k}{n}.$$

An analyst allowed to interact with the dataset may be able to "discover" correlations that are certainly false in the population, which in this case has $PXY = 0$. $\diamond$

(Exercise 6.7 asks for details in the calculations in Example 6.3.5.)

Classical multiplicity control in hypothesis testing would need to account for all possible analyses one might have done otherwise. Instead, we control information in the indices we select. We seek some type of *stability* in the information algorithms provide to a data analyst—intuitively, if small changes to a sample do not change the behavior of an analyst substantially, then we expect to obtain reasonable generalization bounds. If outputs of an analysis procedure carry little information about a particular sample (but instead provide information about a population), then Corollary 6.3.3 shows that estimates we obtain should be accurate.

To develop this stability theory, we require two conditions: first, that whatever quantity we develop for stability should *compose adaptively*, meaning that if we apply two stable (randomized) algorithms to a sample, they remain jointly stable even if we arbitrarily choose the second algorithm as a function of the output of the first. Second, our notion should bound the mutual information $I(X_1^n; T)$ between the sample X_1^n and queries T. This control on the mutual information guarantees that, by the data processing inequality, any downstream analysis we perform that depends only on T satisfies the same stability and information guarantees.

We consider mechanisms (randomized algorithms) $M : \mathcal{X}^n \to \mathcal{A}$, taking values in our index set $\mathcal{A}$, where $M(x_1^n) \in \mathcal{A}$ is a random variable that depends on the sample x_1^n. We recall our abuses of notation in defining KL-divergences (Sections 2.1.2–2.1.3), and for random variables $X \sim P$ and $Y \sim Q$ on the same space, we denote

$$D_{\rm kl}\left(X \| Y\right) := D_{\rm kl}\left(P \| Q\right).$$

We then ask for a type of average-case leave-one-out stability for the mechanisms M. In the definition, recall the notation $x_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Definition 6.1. *Let $\varepsilon \geq 0$. A randomized algorithm $M : \mathcal{X}^n \to \mathcal{A}$ is ε-*KL-stable *if for each $i \in \{1, \ldots, n\}$ there is a randomized witness $M_i : \mathcal{X}^{n-1} \to \mathcal{A}$ such that for every sample $x_1^n \in \mathcal{X}^n$,*

$$\frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(M(x_1^n) \| M_i(x_{\setminus i})\right) \leq \varepsilon.$$

Sample means provide easy examples satisfying Definition 6.1.

Example 6.3.6 (Gaussian noise addition for KL-stable means)**:** Let $x_i \in [-1, 1]$, and define $M(x_1^n) = \frac{1}{n}\sum_{i=1}^n x_i + Z$ and $M_i(x_{\setminus i}) = \frac{1}{n}\sum_{j \neq i} x_j + Z$, where $Z \sim \mathsf{N}(0, \tau^2)$. Then

$$\frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(M(x_1^n) \| M_i(x_{\setminus i})\right) = \frac{1}{2n\tau^2} \sum_{i=1}^n \frac{1}{n^2} x_i^2 \leq \varepsilon := \frac{1}{2\tau^2 n^2},$$

so the sample mean of bounded variables perturbed with Gaussian noise is KL-stable. $\diamond$

We can consider other types of noise addition:

Example 6.3.7 (Laplace noise addition for KL-stable means)**:** Let $L_{\mu,\tau}$ denote the Laplace distribution with shape τ and mean μ, with density $p(z) = \frac{1}{2\tau} \exp(-|z - \mu|/\tau)$. Let the conditions of Example 6.3.6 hold, but add scaled Laplace noise, that is, $M(x_1^n) = \frac{1}{n}\sum_{i=1}^n x_i + Z$ for $Z \sim L_{0,\tau}$. Then

$$\begin{aligned} D_{\rm kl}\left(L_{\mu_0,\tau} \| L_{\mu_1,\tau}\right) & = \frac{1}{\tau^2} \int_0^{|\mu_1 - \mu_0|} \exp(-z/\tau)(|\mu_1 - \mu_0| - z) dz \\ & = \exp\left(-\frac{|\mu_1 - \mu_0|}{\tau}\right) - 1 + \frac{|\mu_1 - \mu_0|}{\tau} \leq \frac{|\mu_1 - \mu_0|^2}{2\tau^2}, \end{aligned}$$

so the sample mean of variables in $[-1, 1]$ perturbed with Laplace noise is $\varepsilon = \frac{1}{2\tau^2 n^2}$-KL-stable, where τ is the shape parameter. $\diamond$

We must be precise about composition to demonstrate the adaptivity properties of KL-stable algorithms. Thus, for a mechanism $M : \mathcal{A}_0 \times \mathcal{X}^n \to \mathcal{A}_1$ that depends on an auxiliary input from $\mathcal{A}_0$, we tacitly extend Definition 6.1 to say that M is ε-stable if, for each $a \in \mathcal{A}_0$, the map

$$x_1^n \mapsto M(a, x_1^n)$$

is ε-stable (uniformly in a). With this specification, KL-stable algorithms satisfy our two desiderata: they compose adaptively and bound mutual information between observations X_1^n and answers $M(X_1^n)$ when the X_i are independent. The next two lemmas quantify this.

Lemma 6.3.8. *Let $M : \mathcal{X}^n \to \mathcal{A}_0$ and $M' : \mathcal{A}_0 \times \mathcal{X}^n \to \mathcal{A}_1$ be ε and ε'-KL-stable, respectively. Then $(M' \circ M(x_1^n), M(x_1^n))$ is $(\varepsilon + \varepsilon')$-KL-stable.*

Proof Let $M_i : \mathcal{X}^{n-1} \to \mathcal{A}_0$ be witnesses for the ε-KL-stability of M, and let $M_i'(a, \cdot)$ be witnesses for M' on input a. For each i, define the joint and leave-one-out results

$$J(x_1^n) := \left(M(x_1^n), M'(M(x_1^n), x_1^n)\right) \quad \text{and} \quad J_i(x_{\setminus i}) := \left(M_i(x_{\setminus i}), M_i'(M_i(x_{\setminus i}), x_{\setminus i})\right).$$

We show that the J_i witness $(\varepsilon + \varepsilon')$-KL-stability of J.

Let $P_M(\cdot \mid x_1^n)$ denote the distribution of $M(x_1^n)$. The KL chain rule (2.1.6) implies

$$D_{\rm kl}\left(J(x_1^n) \| J_i(x_{\setminus i})\right) = D_{\rm kl}\left(M(x_1^n) \| M_i(x_{\setminus i})\right) + \int D_{\rm kl}\left(M'(a, x_1^n) \| M_i'(a, x_{\setminus i})\right) dP_M(a \mid x_1^n)$$

for each i and $x_1^n \in \mathcal{X}^n$, and averaging over i yields

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(J(x_1^n) \| J_i(x_{\setminus i})\right) \\
&= \frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(M(x_1^n) \| M_i(x_{\setminus i})\right) + \int \frac{1}{n} \sum_{i=1}^n D_{\rm kl}\left(M'(a, x_1^n) \| M_i'(a, x_{\setminus i})\right) dP_M(a \mid x_1^n) \le \varepsilon + \varepsilon'
\end{aligned}$$

by definition of stability with auxiliary input a. □

The second key lemma bounds the mutual information of a random function $M(X_1^n)$ and a sample X_1^n by the average of KL-divergences.

Lemma 6.3.9. *Let X_i be independent. Then for any random variable M,*

$$I(M; X_1^n) \le \sum_{i=1}^n I(M; X_i \mid X_{\setminus i}) = \sum_{i=1}^n \int D_{\rm kl}\left(M(x_1^n) \| M_i(x_{\setminus i})\right) dP(x_1^n),$$

where $M_i(x_{\setminus i}) = M(x_1^{i-1}, X_i, x_{i+1}^n)$ is the random realization of M conditional on $X_{\setminus i} = x_{\setminus i}$.

Proof Without loss of generality we assume the X_i are discrete (recall the discussion of divergence, mutual information, and quantization in Section 2.2.2). In this case, we have

$$I(M; X_1^n) = \sum_{i=1}^n I(M; X_i \mid X_1^{i-1}) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) - H(X_i \mid M, X_1^{i-1}).$$

Because the X_i follow a product distribution, $H(X_i \mid X_1^{i-1}) = H(X_i)$, while $H(X_i \mid M, X_1^{i-1}) \ge H(X_i \mid M, X_{\setminus i})$ because conditioning reduces entropy. Consequently, we have

$$I(M; X_1^n) \le \sum_{i=1}^n H(X_i) - H(X_i \mid M, X_{\setminus i}) = \sum_{i=1}^n I(M; X_i \mid X_{\setminus i}).$$

To see the final equality, note that

$$\begin{aligned} I(M; X_i \mid X_{\setminus i}) &= \int_{\mathcal{X}^{n-1}} I(M; X_i \mid X_{\setminus i} = x_{\setminus i}) dP(x_{\setminus i}) \\ &= \int_{\mathcal{X}^{n-1}} \int_{\mathcal{X}} D_{\rm kl}\left(M(x_1^n) \| M(x_1^{i-1}, X_i, x_{i+1}^n)\right) dP(x_i) dP(x_{\setminus i}) \end{aligned}$$

by definition of mutual information as $I(X;Y) = \mathbb{E}_X[D_{\rm kl}\left(P_{Y|X} \| P_Y\right)]$. □

Combining Lemmas 6.3.8 and 6.3.9, we see nearly immediately that KL-stability implies a mutual information bound, and so interactive KL-stable algorithms bound information.

Theorem 6.3.10. *Let $M_1, \ldots, M_k$ be ε_i-KL-stable procedures, respectively, composed adaptively. Let X_i be independent. Then*

$$\frac{1}{n} I(M_1, \ldots, M_k; X_1^n) \le \sum_{i=1}^k \varepsilon_i.$$

Proof A trivial induction via Lemma 6.3.8 shows that the adaptive composition $(M_1, \ldots, M_k)$ is $\sum_{j=1}^k \varepsilon_j$-KL-stable. Let $M_{1\setminus i}^k$ denote the leave-one-out witnesses for the full composition. Then

$$I(M_1^k; X_1^n) \le \sum_{i=1}^n \int D_{\rm kl}\left(M_1^k(x_1^n) \| M_{1\setminus i}^k(x_{\setminus i})\right) dP(x_1^n) \le n \sum_{j=1}^k \varepsilon_j$$

by Lemma 6.3.9. Divide by n. □

6.3.4 Error bounds for a simple noise addition scheme

To build a generalizing interactive procedure, Theorem 6.3.10 shows it is sufficient to build a mechanism for the interaction that maintains KL-stability. Example 6.3.6 makes this easy for the class of bounded queries. Let $\Phi = \{\phi_t\}_{t \in \mathcal{T}}$ where $\phi_t : \mathcal{X} \to [-1, 1]$ be a collection of statistical queries. Then Theorem 6.3.10 and Example 6.3.6 show the following procedure is KL-stable.

Theorem 6.3.10 and Example 6.3.6 immediately imply the limiting information

$$\frac{1}{n} I(X_1^n; T_1, A_1, \ldots, T_k, A_k) \le \frac{k}{2\tau^2 n^2}. \tag{6.3.3}$$

We can use this bound to obtain the following theorem.

Theorem 6.3.11. *Let the indices T_i, $i = 1, \ldots, k$ be chosen using the mechanism in Figure 6.2 and $\tau^2 > 0$. Then*

$$\mathbb{E}\left[\max_{j \le k} (A_j - P\phi_{T_j})^2\right] \le \frac{2ek}{\tau^2 n^2} + \frac{5}{2n} + 4\tau^2(\log k + 1).$$

Input: Sample $X_1^n \in \mathcal{X}^n$ drawn i.i.d. P, collection $\{\phi_t\}_{t\in\mathcal{T}}$ of queries $\phi_t : \mathcal{X} \to [-1, 1]$
Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Mechanism draws independent $Z_k \sim \mathsf{N}(0, \tau^2)$ and responds with answer

$$A_k := P_n\phi + Z_k = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + Z_k.$$

Figure 6.2: Sequential Gaussian noise mechanism.

Temporarily deferring the proof, we provide a few remarks to situate the result.

By inspection, we can optimize τ^2 by setting $\tau^2 = \sqrt{k/(\log k + 1)}/n$ to yield upper bound

$$\mathbb{E}\left[\max_{j\le k}(A_j - P\phi_{T_j})^2\right] \le \frac{5}{2n} + 10\frac{\sqrt{k(1 + \log k)}}{n}.$$

Comparing to Example 6.3.5, we see a substantial improvement. While we do not achieve accuracy scaling with $\log k$, as we would if the queried functionals ϕ_t were completely independent of the sample, we see that we achieve mean-squared error of order

$$\frac{\sqrt{k \log k}}{n}$$

for k adaptively chosen queries. As an aside, and we will say more in the bibliographic details, this result is (to within logarithmic factors) rate optimal.

Proof To prove Theorem 6.3.11, we use a *monitor* technique. As a proof device, we may choose an index T_{k+1} in any way we desire as long as it is a function of the answers $A_1, \ldots, A_k$, external randomness, and objects independent of the data. Thus, we may choose

$$T_{k+1} := T_{k^\star} \quad \text{where} \quad k^\star = \operatorname*{argmax}_{j\le k}\{|A_j - P\phi_{T_j}|\},$$

which is a function of the k different $\varepsilon = \frac{1}{2\tau^2 n^2}$-KL-stable queries T_1^k and their answers A_1^k. Because the queries ϕ are 1-sub-Gaussian Corollary 6.3.3 shows that for $T = T_{k+1}$,

$$\begin{aligned}\mathbb{E}[(P_n\phi_T - P\phi_T)^2] &\le \frac{2e}{n} I(X_1^n; T_{k+1}) + \frac{5}{4n} \\ &\overset{(i)}{\le} \frac{2e}{n} I(X_1^n; T_1, A_1, \ldots, T_k, A_k) \overset{(6.3.3)}{\le} 2ek\varepsilon + \frac{5}{4n} = \frac{ek}{\tau^2 n^2} + \frac{5}{4n}, \end{aligned} \tag{6.3.4}$$

where inequality (i) follows from data processing and the dependence $X_1^n \to (A_i, T_i)_{i=1}^k \to T_{k+1}$. The next lemma, whose proof we leave as Exercise 6.10, controls the error from noise addition:

Lemma 6.3.12. *Let* $W_j \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$. *Then* $\mathbb{E}[\max_{j\le k} W_j^2] \le 2(\log k + 1)$.

Now note that $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, so the monitor inequality (6.3.4) implies

$$\mathbb{E}\left[\max_{j \le k}(A_j - P\phi_{T_j})^2\right] \le 2\mathbb{E}[(P_n\phi_T - P\phi_T)^2] + 2\mathbb{E}\left[\max_{j \le k}\{Z_j^2\}\right] \le \frac{2ek}{\tau^2 n^2} + \frac{5}{2n} + 4\tau^2(\log k + 1),$$

by Lemma 6.3.12. □

6.4 Bibliography and further reading

PAC-Bayes techniques originated with work of David McAllester [306, 307, 308], and we remark on his excellently readable tutorial [309]. The particular approaches we take to our proofs in Section 6.2 follow Catoni [103] and McAllester [308]. The PAC-Bayesian bounds we present, that simultaneously for *any* distribution π on $\mathcal{F}$, if $F \sim \pi$ then

$$\mathbb{E}[(P_nF - PF)^2 \mid X_1^n] \lesssim \frac{1}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log\frac{1}{\delta}\right]$$

with probability at least $1-\delta$ suggest that we can optimize them by choosing π carefully. For example, in the context of learning a statistical model parameterized by $\theta \in \Theta$ with losses $\ell(\theta; x, y)$, it is natural to attempt to find π minimizing

$$\mathbb{E}_\pi[P_n\ell(\theta; X, Y) \mid P_n] + C\sqrt{\frac{1}{n}D_{\rm kl}\left(\pi \| \pi_0\right)}$$

in π, where the expectation is taken over $\theta \sim \pi$. If this quantity has optimal value $\epsilon_n^\star$, then one is immediately guaranteed that for the population P, we have $\mathbb{E}_\pi[P\ell(\theta; X, Y)] \le \epsilon_n^\star + C\sqrt{\log\frac{1}{\delta}}/\sqrt{n}$. Langford and Caruana [274] take this approach, and Dziugaite and Roy [176] use it to give (the first) non-trivial bounds for deep learning models.

The questions of interactive data analysis begin at least several decades ago, perhaps most profoundly highlighted positively by Tukey's *Exploratory Data Analysis* [402]. Problems of scientific replicability have, conversely, highlighted many of the challenges of reusing data or peeking, even innocently, at samples before performing statistical analyses [252, 198, 207]. Our approach to formalizing these ideas and rigorously limiting information leakage draws from a more recent strain of work in the theoretical computer science literature, with major contributions from Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth and Bassily, Nissim, Smith, Steinke, Stemmer, and Ullman [174, 172, 173, 50, 51]. Our particular treatment most closely follows Feldman and Steinke [184], and Exercises 6.8, 6.9, and 6.11 show how to extend Theorem 6.3.11 to provide accuracy guarantees that depend on the variance of the queries rather than their gross range. The question of whether the result in Theorem 6.3.11 is sharp remains open; researchers have employed lower bound techniques leveraging "fingerprinting codes" to provide lower bounds in such data analysis problems [384, 330], which have deep connections to the optimality results based on "score correlations" that we develop in Chapter 11. Among algorithms that guarantee both the sample error $|A - P_n\phi_T|$ and population error $|A - P\phi_T|$ are small, Theorem 6.3.11 is (to within a logarithmic factor) unimprovable [330, 297]. The problems these techniques target also arise frequently in high-dimensional statistics, where one often wishes to estimate uncertainty and perform inference *after* selecting a model. While we do not touch on these problems, a few references in this direction include [62, 396, 245].

6.5 Exercises

Exercise 6.1 (A discrete Donsker-Varadhan representation via convex conjugates)**:** Let P and Q have finite supports $\{1, \ldots, k\}$, with p.m.f.s $p, q \in \Delta_k = \{v \in \mathbb{R}_+^k \mid \langle \mathbf{1}, v \rangle = 1\}$. Define $f_q(v) = \log(\sum_{j=1}^k q_j e^{v_j})$, which is convex in v (recall Proposition 3.2.1). Show that the conjugate

$$f_q^*(p) := \sup_{v \in \mathbb{R}^k} \{\langle p, v \rangle - f_q(v)\} = D_{\rm kl}\left(P \| Q\right).$$

Exercise 6.2 (Duality in Donsker-Varadhan)**:** Here, we give a converse result to Theorem 6.1.1, showing that for any function $h : \mathcal{X} \to \mathbb{R}$,

$$\log \mathbb{E}_Q[e^{h(X)}] = \sup_P \left\{\mathbb{E}_P[h(X)] - D_{\rm kl}\left(P \| Q\right)\right\}, \tag{6.5.1}$$

where the supremum is taken over probability measures. If Q has a density, the supremum may be taken over probability measures having a density.

(a) Show the equality (6.5.1) in the case that $\mathcal{X}$ is discrete by directly computing the supremum. (That is, let $|\mathcal{X}| = k$, and identify probability measures P and Q with vectors $p, q \in \mathbb{R}_+^k$.)

(b) Let Q have density q. Assume that $\mathbb{E}_Q[e^{h(X)}] < \infty$ and let

$$Z_h(x) = \exp(h(x))/\mathbb{E}_Q[\exp(h(X))],$$

so $\mathbb{E}_Q[Z_h(X)] = 1$. Let P have density $p(x) = Z_h(x)q(x)$. Show that

$$\log \mathbb{E}_Q[e^{h(X)}] = \mathbb{E}_P[h(X)] - D_{\rm kl}\left(P \| Q\right).$$

Why does this imply equality (6.5.1) in this case?

(c) If $\mathbb{E}_Q[e^{h(X)}] = +\infty$, then monotone convergence implies that $\lim_{B \uparrow \infty} \mathbb{E}_Q[e^{\min\{B, h(X)\}}] = +\infty$. Conclude (6.5.1).

Exercise 6.3 (An alternative PAC-Bayes bound)**:** Let $f : \Theta \times \mathcal{X} \to \mathbb{R}$, and let π_0 be a density on $\theta \in \Theta$. Use the dual form (6.5.1) of the variational representation of the KL-divergence to show that with probability at least $1 - \delta$ over the draw of $X_1^n \stackrel{\rm iid}{\sim} P$,

$$\int P_n f(\theta, X) \pi(\theta) d\theta \le \int \log \mathbb{E}_P\left[\exp(f(\theta, X))\right] \pi(\theta) d\theta + \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{1}{\delta}}{n}$$

simultaneously for all distributions π on Θ, where the expectation $\mathbb{E}_P$ is over $X \sim P$.

Exercise 6.4: In the proof of Corollary 6.2.6, instead choose $\eta \in H_n := \{1/n, 2/n, 4/n, \ldots, 2^k/n\}$, where $k \in \mathbb{N}$ is the smallest integer such that $2^k/n \ge 1$. Show that for a numerical constant $C < \infty$, under the conditions of the corollary, with probability at least $1 - \delta$,

$$\mathbb{E}_\pi[Pf] \le \mathbb{E}_\pi[P_n f] + C\sqrt{\frac{b\mathbb{E}_\pi[P_n f]}{n}\left[D_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{\log n}{\delta}\right]} + \frac{C}{n}\left(\mathbb{E}_\pi[P_n f] + bD_{\rm kl}\left(\pi \| \pi_0\right) + \log \frac{\log n}{\delta}\right).$$

Exercise 6.5 (A mean estimator for heavy-tailed distributions [104])**:** We use a PAC-Bayes bound to construct an estimator of the mean $\mathbb{E}[X]$ of a distribution with sub-Gaussian-like concentration that depends *only* on the second moments $\Sigma = \mathbb{E}[XX^\top]$ of the random vector X while only assuming that $\mathbb{E}[\|X\|^2] < \infty$. Let ψ be an odd function (i.e., $\psi(-t) = -\psi(t)$) satisfying

$$-\log(1 - t + t^2) \le \psi(t) \le \log(1 + t + t^2).$$

The function $\psi(t) = \min\{1, \max\{-1, t\}\}$ (the truncation of t to the range $[-1, 1]$) is such a function. Let π_θ be the normal distribution $\mathsf{N}(\theta, \sigma^2 I)$ and π_0 be $\mathsf{N}(0, \sigma^2 I)$.

(a) Let $\lambda > 0$. Use Exercise 6.3 to show that with probability at least $1 - \delta$, for all $\theta \in \mathbb{R}^d$

$$\frac{1}{\lambda} \int P_n \psi(\lambda \langle \theta', X \rangle) \pi_\theta(\theta') d\theta' \le \langle \theta, \mathbb{E}[X] \rangle + \lambda \left(\theta^\top \Sigma \theta + \sigma^2 \operatorname{tr}(\Sigma) \right) + \frac{\|\theta\|_2^2 / 2\sigma^2 + \log \frac{1}{\delta}}{n\lambda}.$$

(b) For $\lambda > 0$, define the "directional mean" estimator

$$E_n(\theta, \lambda) = \frac{1}{\lambda} \int P_n \psi(\lambda \langle \theta', X \rangle) \pi_\theta(\theta') d\theta'.$$

Give a choice of $\lambda > 0$ such that with probability $1 - \delta$,

$$\sup_{\theta \in \mathbb{S}^{d-1}} |E_n(\theta, \lambda) - \langle \theta, \mathbb{E}[X] \rangle| \le \frac{2}{\sqrt{n}} \sqrt{\left(\frac{1}{2\sigma^2} + \log \frac{1}{\delta} \right) \left(\|\Sigma\|_{\rm op} + \sigma^2 \operatorname{tr}(\Sigma) \right)},$$

where $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d \mid \|u\|_2 = 1\}$ is the unit sphere.

(c) Justify the following statement: choosing the vector $\widehat{\mu}_n$ minimizing

$$\sup_{\theta \in \mathbb{S}^{d-1}} |E_n(\theta, \lambda) - \langle \theta, \mu \rangle|$$

in μ guarantees that with probability at least $1 - \delta$,

$$\|\widehat{\mu}_n - \mathbb{E}[X]\|_2 \le \frac{4}{\sqrt{n}} \sqrt{\left(\frac{1}{2\sigma^2} + \log \frac{1}{\delta} \right) \left(\|\Sigma\|_{\rm op} + \sigma^2 \operatorname{tr}(\Sigma) \right)}.$$

(d) Give a choice of the prior/posterior variance σ^2 so that with probability at least $1 - \delta$,

$$\|\widehat{\mu}_n - \mathbb{E}[X]\|_2 \le \frac{4}{\sqrt{n}} \sqrt{\operatorname{tr}(\Sigma) + 2 \|\Sigma\|_{\rm op} \log \frac{1}{\delta}}.$$

Exercise 6.6 (Large-margin PAC-Bayes bounds)**:** Consider a multiclass prediction scenario in which data comes in pairs $(x, y) \in b\mathbb{B}_2^d \times [k]$, where $\mathbb{B}_2^d = \{v \in \mathbb{R}^d \mid \|v\|_2 \le 1\}$ denotes the ℓ_2-ball and $[k] = \{1, \ldots, k\}$. Using predictors $\theta_1, \ldots, \theta_k \in \mathbb{R}^d$, the prediction of y on an example x is

$$\widehat{y}(x) := \operatorname*{argmax}_{i \le k} \langle \theta_i, x \rangle.$$

We suffer an error whenever $\widehat{y}(x) \neq y$, and the *margin* of the classifier on pair (x, y) is

$$\langle \theta_y, x \rangle - \max_{i \neq y} \langle \theta_i, x \rangle = \min_{i \neq y} \langle \theta_y - \theta_i, x \rangle.$$

If $\langle \theta_y, x \rangle > \langle \theta_i, x \rangle$ for all $i \neq y$, the margin is then positive (and the prediction is correct).

(a) Develop an analogue of the bounds in Section 6.2.2 in this setting. To do so, (i) define the analogue of the margin-based loss ℓ_γ, (ii) show how Gaussian perturbations affect it, and (iii) prove an analogue of the bound in Proposition 6.2.7. Assume one of the conditions

$$\text{(C1)} \;\; \|\theta_i\|_2 \le r \text{ for all } i \qquad \text{(C2)} \;\; \sum_{i=1}^k \|\theta_i\|_2^2 \le kr^2$$

on the classification vectors θ_i. Specify the chosen condition.

(b) Describe a minimization procedure that uses convex optimization to find a large-margin multiclass classifier.

Exercise 6.7: Complete the calculations in Example 6.3.5.

Exercise 6.8 (A standard-deviation-based information bound)**:** Let $\Phi = \{\phi_t\}_{t\in\mathcal{T}}$ be a collection of functions $\phi_t : \mathcal{X} \to \mathbb{R}$, where each ϕ_t satisfies the Bernstein condition (4.1.8) with parameters $\sigma^2(\phi_t)$ and b, that is, $|\mathbb{E}[(\phi_t(X) - P\phi_t(X))^k]| \le \frac{k!}{2}\sigma^2(\phi_t)b^{k-2}$ for all $k \ge 3$ and $\mathrm{Var}(\phi_t(X)) = \sigma^2(\phi_t)$. Let $T \in \mathcal{T}$ be any random variable, which may depend on an observed sample X_1^n. Show that for all $C > 0$ and $|\lambda| \le \frac{C}{2b}$,

$$\left|\mathbb{E}\left[\frac{P_n\phi_T - P\phi_T}{\max\{C, \sigma(\phi_T)\}}\right]\right| \le \frac{1}{n|\lambda|} I(T; X_1^n) + |\lambda|.$$

Exercise 6.9 (A variance-based information bound)**:** Let $\Phi = \{\phi_t\}_{t\in\mathcal{T}}$ be a collection of functions $\phi_t : \mathcal{X} \to \mathbb{R}$, where each $\phi_t : \mathcal{X} \to [-1, 1]$. Let $\sigma^2(\phi_t) = \mathrm{Var}(\phi_t(X))$ and let $s_n^2(\phi) = P_n\phi^2 - (P_n\phi)^2$ be the sample variance of ϕ. Show that for all $C > 0$ and $0 \le \lambda \le C/4$,

$$\mathbb{E}\left[\frac{s_n^2(\phi_T)}{\max\{C, \sigma^2(\phi_T)\}}\right] \le \frac{1}{n\lambda} I(T; X_1^n) + 2.$$

Hint. If $0 \le x \le 1$, then $e^x \le 1 + 2x$, and if $X \in [0, 1]$, then $\mathbb{E}[e^X] \le 1 + 2\mathbb{E}[X] \le e^{2\mathbb{E}[X]}$. Use this to argue that $\mathbb{E}[e^{\lambda n P_n(\phi - P\phi)^2 / \max\{C,\sigma^2\}}] \le e^{2\lambda n}$ for any $\phi : \mathcal{X} \to [-1, 1]$ with $\mathrm{Var}(\phi) \le \sigma^2$, then apply the Donsker-Varadhan theorem.

Exercise 6.10: Using Mills ratio (Exercise 4.3), prove Lemma 6.3.12.

Exercise 6.11*: Consider the following scenario: let $\phi : \mathcal{X} \to [-1, 1]$ and let $\alpha > 0$, $\tau > 0$. Let $\mu = P_n\phi$ and $s^2 = P_n\phi^2 - \mu^2$. Define $\sigma^2 = \max\{\alpha s^2, \tau^2\}$, and assume that $\tau^2 \ge \frac{5\alpha}{n}$.

(a) Show that the mechanism with answer A_k defined by

$$A := P_n\phi + Z \quad \text{for} \quad Z \sim \mathsf{N}(0, \sigma^2)$$

is ε-KL-stable (Definition 6.1), where for a numerical constant $C < \infty$,

$$\varepsilon \le C \cdot \frac{s^2}{n^2\sigma^2} \cdot \left(1 + \frac{\alpha^2}{\sigma^2}\right).$$

(b) Show that if $\alpha^2 \le C'\tau^2$ for a numerical constant $C' < \infty$, then $\varepsilon \le O(1)\frac{1}{n^2\alpha}$.

Hint. Use Exercise 2.3, and consider the "alternative" mechanisms of sampling from

$$\mathsf{N}(\mu_{-i}, \sigma_{-i}^2) \;\; \text{where} \;\; \sigma_{-i}^2 = \max\{\alpha s_{-i}^2, \tau^2\}$$

for

$$\mu_{-i} = \frac{1}{n-1}\sum_{j \neq i} \phi(X_j) \;\; \text{and} \;\; s_{-i}^2 = \frac{1}{n-1}\sum_{j \neq i} \phi(X_j)^2 - \mu_{-i}^2.$$

Input: Sample $X_1^n \in \mathcal{X}^n$ drawn i.i.d. P, collection $\{\phi_t\}_{t \in \mathcal{T}}$ of queries $\phi_t : \mathcal{X} \to [-1, 1]$, parameters $\alpha > 0$ and $\tau > 0$
Repeat: for $k = 1, 2, \ldots$

i. Analyst chooses index $T_k \in \mathcal{T}$ and query $\phi := \phi_{T_k}$

ii. Set $s_k^2 := P_n\phi^2 - (P_n\phi)^2$ and $\sigma_k^2 := \max\{\alpha s_k^2, \tau^2\}$

iii. Mechanism draws independent $Z_k \sim \mathsf{N}(0, \sigma_k^2)$ and responds with answer

$$A_k := P_n\phi + Z_k = \frac{1}{n}\sum_{i=1}^n \phi(X_i) + Z_k.$$

Figure 6.3: Sequential Gaussian noise mechanism with variance sensitivity.

Exercise 6.12* (A general variance-dependent bound on interactive queries)**:** Consider the algorithm in Fig. 6.3. Let $\sigma^2(\phi_t) = \mathrm{Var}(\phi_t(X))$ be the variance of ϕ_t.

(a) Show that for $b > 0$ and for all $0 \le \lambda \le \frac{b}{2}$,

$$\mathbb{E}\left[\max_{j \le k} \frac{|A_j - P\phi_{T_j}|}{\max\{b, \sigma(\phi_{T_j})\}}\right] \le \frac{1}{n\lambda} I(X_1^n; T_1^k) + \lambda + \sqrt{2\log(ke)}\sqrt{\frac{4\alpha}{nb^2} I(X_1^n; T_1^k) + 2\alpha + \frac{\tau^2}{b^2}}.$$

(The constants are immaterial.)

(b) Using the result of Question 6.11, show that with appropriate choices of the parameters $\alpha, b, \tau^2, \lambda$ that for a numerical constant $C < \infty$

$$\mathbb{E}\left[\max_{j \le k} \frac{|A_j - P\phi_{T_j}|}{\max\{(k \log k)^{1/4}/\sqrt{n}, \sigma(\phi_{T_j})\}}\right] \le C\frac{(k \log k)^{1/4}}{\sqrt{n}}.$$

You may assume that k, n are large if necessary.

(c) How does this result improve over Theorem 6.3.11?

Chapter 7

Advanced concentration inequalities

While we have not made this explicit, the variation distance implicitly connects the leave-one-out stability notions in the bounded differences inequalities, as in Section 4.2.2, and distributional differences: if P_n and P_n' denote empirical distributions on x_1^n and $z_1^n \in \mathcal{X}^n$ and $f : \mathcal{X}^n \to \mathbb{R}$ is (permutation) symmetric and satisfies bounded differences with constant c, then

$$|f(x_1^n) - f(z_1^n)| \le cn \left\|P_n - P_n'\right\|_{\mathrm{TV}}.$$

There are functions f and vectors x_1^n, z_1^n achieving this with equality. This chapter shows how to use such bounds, which relate function differences to differences in various divergences and distance-like quantities, to concentration. By doing so, we can develop a perspective on concentration that is, nearly, completely information-theoretic, once we connect divergences to moment generating functions via duality, that is, the Donsker-Varadhan variational representation of the KL-divergence. By connecting divergences and moment generating functions, we will be able to develop some of the theory of transportation inequalities in Section 7.2*, and these tools allow us to show concentration of sophisticated random processes. As particular applications, we will relate them to isoperimetric inequalities (concentration on metric spaces) and to showing strong lower bound (converse) results in hypothesis testing.

7.1 From divergences to concentration and back

The Donsker-Varadhan representation of the KL-divergence, Theorem 6.1.1, has a dual form, which allows us to provide an important connection between moment generating functions and the KL-divergence. We state this as a corollary, then connect it to moment generating function bounds.

Corollary 7.1.1. *Let P and Q be distributions on a common space $\mathcal{X}$. Then*

$$D_{\mathrm{kl}}\left(P\|Q\right) = \sup_g \left\{\mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]\right\},$$

where the supremum is taken over measurable functions $g : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}_Q[e^{g(X)}] < \infty$. Conversely, for any measurable $g : \mathcal{X} \to \mathbb{R}$ and distribution Q on $\mathcal{X}$,

$$\log \mathbb{E}_Q[e^{g(X)}] = \sup_P \left\{\mathbb{E}_P[g(X)] - D_{\mathrm{kl}}\left(P\|Q\right)\right\},$$

where the supremum is taken over probability distributions P on $\mathcal{X}$ with $\mathbb{E}_P[g(X)] < \infty$.

Proof The first claim is Theorem 6.1.1. For the second, we assume that $\mathbb{E}_Q[e^{g(X)}] < \infty$. (See Exercise 7.1 for the case that $\mathbb{E}_Q[e^{g(X)}] = +\infty$.) Then via the first part of the corollary, for any distribution P for which $\mathbb{E}_P[g(X)] < \infty$ we have $D_{\rm kl}\left(P\|Q\right) \geq \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]$, that is,

$$\log \mathbb{E}_Q[e^{g(X)}] \geq \mathbb{E}_P[g(X)] - D_{\rm kl}\left(P\|Q\right).$$

To obtain equality, assume w.l.o.g. that Q has density q, and define P to have density $p(x) = \frac{e^{g(x)}}{\mathbb{E}_Q[e^{g(X)}]} q(x)$. This evidently integrates to 1, and

$$\mathbb{E}_P[g(X)] - D_{\rm kl}\left(P\|Q\right) = \frac{\mathbb{E}_Q[g(X)e^{g(X)}]}{\mathbb{E}_Q[e^{g(X)}]} - \mathbb{E}_Q\left[\frac{e^{g(X)}}{\mathbb{E}_Q[e^{g(X)}]} \log \frac{e^{g(X)}}{\mathbb{E}_Q[e^{g(X)}]}\right] = \log \mathbb{E}_Q[e^{g(X)}],$$

completing the proof. □

Corollary 7.1.1, yields the first building block for our development of *transportation inequalities*, which relate concentration and moment generating functions to KL-divergence measures. We have already used these ideas to develop the PAC-Bayesian bounds in Section 6.2. To set the stage, we use the variational representation to provide a fairly immediate alternative characterization of sub-Gaussian random variables.

Theorem 7.1.2. *Let X be a real-valued random variable. Then the following are equivalent:*

(i) For all $\lambda \geq 0$,

$$\log \mathbb{E}_P\left[e^{\lambda(X - \mathbb{E}_P[X])}\right] \leq \frac{\lambda^2\sigma^2}{2}.$$

(ii) For any probability distribution Q for which $D_{\rm kl}\left(Q\|P\right) < \infty$,

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \sqrt{2\sigma^2 D_{\rm kl}\left(Q\|P\right)}.$$

Proof To see that (i) implies (ii), note that swapping the roles of P and Q in Corollary 7.1.1 and taking $g(X) = \lambda(X - \mathbb{E}_P[X])$ yields

$$\lambda(\mathbb{E}_Q[X] - \mathbb{E}_P[X]) - D_{\rm kl}\left(Q\|P\right) \leq \frac{\lambda^2\sigma^2}{2}$$

for all distributions Q and all $\lambda \geq 0$. Rearranging, we obtain

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \frac{D_{\rm kl}\left(Q\|P\right)}{\lambda} + \frac{\lambda\sigma^2}{2},$$

and optimizing over $\lambda > 0$ yields (ii).

For the opposite direction, because $\inf_{\lambda > 0} \frac{a}{\lambda} + b\lambda = 2\sqrt{ab}$ for $a, b \geq 0$, we see that (ii) implies

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \frac{D_{\rm kl}\left(Q\|P\right)}{\lambda} + \frac{\lambda\sigma^2}{2}$$

for all $\lambda > 0$. Rewriting this by taking $g(X) = X - \mathbb{E}_P[X]$, we have $\lambda\mathbb{E}_Q[g(X)] - D_{\rm kl}\left(Q\|P\right) \leq \frac{\lambda^2\sigma^2}{2}$ for all $\lambda > 0$ and Q. Applying the second part of Corollary 7.1.1 and taking a supremum over Q yields (i). □

Theorem 7.1.2 allows another proof that bounded random variables are sub-Gaussian; the sheer slickness of the argument, which relies only on Pinsker's inequality, hints at the power of the approaches we will be able to develop. (Compare to Example 4.1.6.)

Corollary 7.1.3. *Let X be a random variable with $a \le X \le b$. Then X is $\frac{(b-a)^2}{4}$-sub-Gaussian.*

Proof Considering (measurable) functions g taking values in $[a, b]$, we have $\sup_{a \le g \le b} |Pg - Qg| = (b-a) \|P - Q\|_{\rm TV}$. As $a \le X \le b$, Pinsker's inequality (Proposition 2.2.9) implies

$$|\mathbb{E}_Q[X] - \mathbb{E}_P[X]| \le (b-a)\sqrt{\frac{1}{2} D_{\rm kl}\left(Q \| P\right)} = \sqrt{\frac{(b-a)^2}{2} D_{\rm kl}\left(Q \| P\right)}.$$

Theorem 7.1.2 therefore implies that X is $\frac{(b-a)^2}{4}$-sub-Gaussian. □

Pinsker's inequality that $\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(Q \| P\right)$ is the first example of a general family of *transportation inequalities*, which relate various distances on probability measures to the KL-divergence. We develop these inequalities and a few representative consequences in Section 7.2*. Before doing so, however, we present an application of the variational representation to estimating a covariance as well as presenting a generalization of Theorem 7.1.2 beyond sub-Gaussian random variables.

7.1.1 Concentration of covariance matrices via the variational representation

The variational representation of the KL-divergence in Corollary 7.1.1 forms the basis for many sophisticated concentration inequalities. Here, we show how it can provide a relatively immediate proof of concentration for random covariance matrices, where if $X_i \in \mathbb{R}^d$ are well-behaved random vectors, then

$$P_n X X^\top = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$$

concentrates well. This provides a useful comparison against the approaches to such concentration guarantees we develop in earlier chapters, including the matrix Chernoff bounds we outline in Section 4.3*, or covering numbers, as in Proposition 5.1.11.

By using PAC-Bayes bounds, we can avoid these techniques. We use the following consequence of the variational representation of the KL-divergence, which follows from Corollary 7.1.1.

Corollary 7.1.4. *Let $f : \Theta \times \mathcal{X} \to \mathbb{R}$ and π_0 be a density on $\theta \in \Theta$. Then for any $t \ge 0$, with probability at least $1 - e^{-t}$ over the draw $X_1^n \stackrel{\rm iid}{\sim} P$,*

$$\int P_n f(\theta, X) \pi(\theta) d\theta \le \int \log \mathbb{E}_P\left[\exp(f(\theta, X))\right] \pi(\theta) d\theta + \frac{D_{\rm kl}\left(\pi \| \pi_0\right) + t}{n}$$

simultaneously for all distributions π on Θ, where the expectation $\mathbb{E}_P$ is over $X \sim P$.

Exercise 6.3 asks you to prove this result, which follows by an application of Markov's inequality to the nonnegative random variable $\exp(n P_n f(\theta, X)) = \prod_{i=1}^n \exp(f(\theta, X_i))$.

We focus on d-dimensional sub-Gaussian vectors, with a slight modification of the definition (5.1.3). Because we investigate product processes of the form $v^\top X X^\top u$ as $v, u \in \mathbb{R}^d$ vary, we use Orlicz norms (Section 4.1.3*) because they provide bounds on products of random variables, where we recall that for a random variable Y, the Orlicz ψ_p-norm $\|Y\|_{\psi_p} = \inf\{t \ge 0 \mid \mathbb{E}[e^{|Y|^p / t^p}] \le 2\}$, and for any real-valued random variables Y_0, Y_1, $\|Y_0 Y_1\|_{\psi_1} \le \|Y_0\|_{\psi_2} \|Y_1\|_{\psi_2}$ (Lemma 4.1.23). We now say X is a σ^2-sub-Gaussian vector if

$$\|\langle X, u \rangle\|_{\psi_2} \le \sigma \|u\|_2 \quad \text{for } u \in \mathbb{R}^d.$$

We then have the following proposition, which provides the same guarantees as Proposition 5.1.11 for isotropic X_i, meaning $\mathbb{E}[X_i X_i^\top] = I_d$.

Proposition 7.1.5. *Let X_i be independent isotropic σ^2-sub-Gaussian vectors. Then with probability at least $1 - e^{-t}$,*

$$\left\|P_n XX^\top - I_d\right\|_{\rm op} \le \frac{8e^2\sigma^2}{(e-1)^2}\sqrt{\frac{2d+t}{n}}$$

so long as $n \ge 2d + t$. Otherwise, $\|P_n XX^\top - I_d\|_{\rm op} \le \frac{4e^2\sigma^2}{(e-1)^2}(1 + \frac{2d+t}{n})$.

Proof Fix $\epsilon > 0$ to be chosen, and let the prior π_0 be uniform on $(1+\epsilon)\mathbb{B}_2^d \times (1+\epsilon)\mathbb{B}_2^d$. For u, v with $\|u\|_2, \|v\|_2 \le 1$, let $\pi_{u,v}$ be uniform on the product $(u + \epsilon\mathbb{B}_2^d) \times (v + \epsilon\mathbb{B}_2^d)$, so for any vector x

$$\int \theta_1^\top x x^\top \theta_2 \pi_{u,v}(\theta_1, \theta_2) d\theta_1 d\theta_2 = u^\top x x^\top v.$$

Fix $\lambda \ge 0$ to be chosen, and define the P-mean zero function

$$f(\theta_1, \theta_2, x) := \theta_1^\top x x^\top \theta_2 - \langle \theta_1, \theta_2 \rangle.$$

As we assume $\|\langle u, X\rangle\|_{\psi_2} \le \sigma \|u\|_2$, the random variable $f(\theta_1, \theta_2, X)$ is sub-exponential:

$$\begin{aligned}\|f(\theta_1, \theta_2, X)\|_{\psi_1} = \left\|\theta_1^\top XX^\top \theta_2 - \langle \theta_1, \theta_2\rangle\right\|_{\psi_1} &\le 2\left\|\theta_1^\top XX^\top \theta_2\right\|_{\psi_1} \\ &\le 2\|\langle \theta_1, X\rangle\|_{\psi_2} \|\langle \theta_2, X\rangle\|_{\psi_2} \le 2(1+\epsilon)^2\sigma^2\end{aligned} \tag{7.1.1}$$

if $\|\theta_i\|_2 \le 1 + \epsilon$, by the triangle inequality and sub-multiplicativity (Lemma 4.1.23).

Let $\lambda \in \mathbb{R}$ and apply Corollary 7.1.4 to the function λf, which implies (via Jensen's inequality)

$$\lambda\left(u^\top P_n XX^\top v - \langle u, v\rangle\right) \le \frac{D_{\rm kl}\left(\pi_{u,v}\|\pi_0\right) + t}{n} + \log \int \mathbb{E}_P[\exp(\lambda f(\theta_1, \theta_2, X))]\pi_{u,v}(\theta_1, \theta_2) d\theta_1 d\theta_2$$

simultaneously for all $u, v \in \mathbb{B}_2^d$ with probability at least $1 - e^{-t}$. It remains to control the expectation $\mathbb{E}_P[\exp(\lambda f)]$ and the KL-divergence. For the former, we apply Corollary 4.1.24 and use inequality (7.1.1) to guarantee that f has finite ψ_1-norm, so for any $\theta_1, \theta_2 \in (1+\epsilon)\mathbb{B}_2^d$ we have

$$\mathbb{E}_P[\exp(\lambda f(\theta_1, \theta_2, X))] \le \exp\left(16\lambda^2(1+\epsilon)^4\sigma^4\right)$$

when $|\lambda| \le \frac{1}{4(1+\epsilon)^2\sigma^2}$. Substituting this above

$$\lambda\left(u^\top P_n XX^\top v - \langle u, v\rangle\right) \le \frac{D_{\rm kl}\left(\pi\|\pi_0\right) + t}{n} + 16\lambda^2(1+\epsilon)^4\sigma^4.$$

Finally, we evaluate the KL-divergence: for any $u, v \in \mathbb{B}_2^d$, we have

$$D_{\rm kl}\left(\pi_{u,v}\|\pi_0\right) = 2D_{\rm kl}\left(\mathsf{Uniform}(u + \epsilon\mathbb{B}_2^d)\|\mathsf{Uniform}((1+\epsilon)\mathbb{B}_2^d)\right) = 2d\log\frac{1+\epsilon}{\epsilon}$$

because $\pi_{u,v}$ and π_0 are product distributions. Take $\epsilon = \frac{1}{e-1}$ to give $D_{\rm kl}\left(\pi\|\pi_0\right) = 2d$ and $1 + \epsilon = \frac{e}{e-1}$. Dividing by λ (accounting for the sign as appropriate), we see that with probability at least $1 - e^{-t}$, simultaneously for all $\|u\|_2 = 1$ and $\|v\|_2 = 1$ we have

$$|u^\top P_n XX^\top v - \langle u, v\rangle| \le \frac{2d+t}{n\lambda} + 16\left(\frac{e}{e-1}\right)^4\sigma^4\lambda$$

whenever $0 \le \lambda \le \frac{(e-1)^2}{4e^2\sigma^2}$. Take $\lambda = \frac{(e-1)^2}{4e^2\sigma^2} \min\{\sqrt{\frac{2d+t}{n}}, 1\}$. $\square$

Proposition 7.1.5 admits elegant extensions to non-isotropic matrices, giving "dimension-free" concentration for sub-Gaussian random vectors. As one example, whose proof Exercise 7.6 develops, we can obtain a "dimension-free" concentration inequality for random covariance matrices. Define the *effective rank* of a matrix Σ

$$\mathrm{r}_{\mathrm{eff}}(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}. \tag{7.1.2}$$

Let M_i be independent positive semidefinite matrices with expectation $\mathbb{E}[M_i] = \Sigma$ and satisfying the non-isotropic sub-exponential inequality

$$\left\|u^\top M_i u\right\|_{\psi_1} \le \kappa^2 u^\top \Sigma u \quad \text{for all } u \in \mathbb{R}^d. \tag{7.1.3}$$

Notably, if $X_i \in \mathbb{R}^d$ are random vectors satisfying the non-isotropic sub-Gaussian condition that $\left\|u^\top X_i\right\|_{\psi_2}^2 \le \kappa^2 u^\top \Sigma u$, then $M_i = X_i X_i^\top$ satisfies condition (7.1.3).

Proposition 7.1.6 (Zhivotovskiy [439])**.** *Let $t \ge 0$ and let M_i be independent Hermitian positive semidefinite matrices satisfying $\mathbb{E}[M_i] = \Sigma$ and condition (7.1.3). Then so long as $n \ge 4\mathrm{r}_{\mathrm{eff}}(\Sigma) + t$, with probability at least $1 - e^{-t}$,*

$$\|P_n M - \Sigma\|_{\mathrm{op}} \le 20\kappa^2 \|\Sigma\|_{\mathrm{op}} \sqrt{\frac{4\mathrm{r}_{\mathrm{eff}}(\Sigma) + t}{n}}. \tag{7.1.4}$$

Because for all d we have $\mathrm{r}_{\mathrm{eff}}(\Sigma) \le d$, this always improves upon Proposition 7.1.5, but it can be much sharper. The (relatively) straightforward extension to the "dimension-free" concentration Proposition 7.1.6 exhibits highlights the power of the variational inequality approach.

7.1.2* The connection between moment generating functions and KL-divergence

The argument proving Theorem 7.1.2 extends beyond sub-Gaussian random variables to, essentially, any random variable whose moment generating function exists near 0, using a similar argument, because any log moment generating function necessarily looks locally quadratic. Recall that for random variable X with log moment generating function $\phi_X(\lambda) := \log \mathbb{E}[e^{\lambda X}]$ (the *cumulant generating function*), so long as $\phi_X(\lambda)$ is finite near 0, it is infinitely differentiable on the interior of its domain (Proposition 3.2.2), and we may take derivatives through expectations. We thus have

$$\phi_X'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \quad \text{and} \quad \phi_X''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2 = \mathrm{Var}_{P_\lambda}(X),$$

where P_λ denotes the distribution on X with density tilted by $\frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]}$. So long as X is mean zero and non-constant, we obtain $\phi_X'(0) = 0$ and $\phi_X''(\lambda) > 0$ for $\lambda \in \mathrm{int}\,\mathrm{dom}\,\phi_X$. (See also Exercise 7.3.)

To that end, we call a function ϕ *CGF (cumulant-generating-function) like* on the interval $[0, b)$ if ϕ is continuously differentiable on $[0, b)$, satisfies $\phi(0) = \phi'(0) = 0$, and is convex (here we take $\phi'(0)$ to be the right derivative). The *convex conjugate* of ϕ is

$$\phi^*(s) := \sup_{0 \le \lambda < b} \left\{\lambda s - \phi(\lambda)\right\},$$

which is strictly increasing in $s \geq 0$. (We say more about convex conjugates in the coming chapters; see also Appendix C, especially Section C.2.) For now, we only require the generalized inverse

$$(\phi^*)^{-1}(t) := \inf \{s \geq 0 \mid \phi^*(s) > t\}.$$

Because $\phi^*(s) \geq \lambda s - \phi(\lambda)$ for $0 < \lambda < b$, we have $\phi^*(s) \to \infty$ as $s \to \infty$, so $(\phi^*)^{-1}(t)$ exists.

Example 7.1.7 (Sub-Gaussian CGF-like functions)**:** If X is σ^2-sub-Gaussian, we have $\phi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$. For $\phi(\lambda) = \frac{\lambda^2\sigma^2}{2}$, we have $\phi^*(s) = \sup\{\lambda s - \frac{\lambda^2\sigma^2}{2}\} = \frac{s^2}{2\sigma^2}$, and

$$(\phi^*)^{-1}(t) = \sqrt{2\sigma^2 t} = \inf_{\lambda > 0} \frac{\lambda\sigma^2}{2} + \frac{t}{\lambda}$$

gives the inverse. (Recall the proof of Theorem 7.1.2, which relies on this transformation.) $\diamond$

Example 7.1.8 (Sub-exponential CGF-like functions)**:** When we have (τ^2, b)-sub-exponential random variables (Definition 4.2), we obtain CGF-like functions of the form $\phi(\lambda) = \frac{\tau^2\lambda^2}{2}$ for $0 \leq \lambda < \frac{1}{b}$. Taking suprema gives the convex conjugate

$$\phi^*(s) = \begin{cases} \frac{s^2}{2\tau^2} & \text{if } s \leq \frac{\tau^2}{b} \\ \frac{s}{b} - \frac{\tau^2}{2b^2} & \text{if } s \geq \frac{\tau^2}{b}. \end{cases}$$

This has the more complex inverse

$$(\phi^*)^{-1}(t) = \begin{cases} \sqrt{2\tau^2 t} & \text{if } 0 \leq t \leq \frac{\tau^2}{2b^2} \\ bt + \frac{\tau^2}{2b} & \text{otherwise} \end{cases}$$

in this case. $\diamond$

We can now generalize Theorem 7.1.2.

Theorem 7.1.9. *Let X be a real-valued random variable and ϕ be CGF-like on $[0, b)$. Then the following are equivalent:*

(i) For all $\lambda \in (0, b)$,

$$\log \mathbb{E}_P[e^{\lambda(X - \mathbb{E}_P[X])}] \leq \phi(\lambda).$$

(ii) For any probability distribution Q for which $D_{\rm kl}(Q\|P) < \infty$,

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq (\phi^*)^{-1}(D_{\rm kl}(Q\|P)).$$

Proof To see that (i) implies (ii), note that if $\phi(\lambda) \geq \log \mathbb{E}_P[e^{\lambda(X - \mathbb{E}_P[X])}]$, then as in the proof of Theorem 7.1.2, swapping the roles of P and Q in Corollary 7.1.1 and taking $g(X) = X - \mathbb{E}_P[X]$ yields that

$$\lambda(\mathbb{E}_Q[X] - \mathbb{E}_P[X]) - \phi(\lambda) \leq D_{\rm kl}(Q\|P)$$

for all distributions Q. Taking a supremum over $\lambda \in (0, b)$ on the left side then gives

$$\phi^*(\mathbb{E}_Q[X] - \mathbb{E}_P[X]) \leq D_{\rm kl}(Q\|P),$$

which implies part (ii).

For the converse direction, we require a technical lemma giving an inverse for convex conjugates of smooth increasing functions:

Lemma 7.1.10. *Let ϕ be CGF-like on $[0, b)$. Then*

$$(\phi^*)^{-1}(t) := \inf\{s \geq 0 \mid \phi^*(s) > t\} = \inf_{\lambda \in (0,b)} \frac{t + \phi(\lambda)}{\lambda}.$$

Additionally, $(\phi^*)^{-1}$ is concave and strictly increasing on $\mathbb{R}_+$, with $(\phi^*)^{-1}(t) \to \infty$ as $t \to \infty$.

Proof We prove the equality first. As in our discussion earlier, because $\phi^*(s) \geq \lambda s - \phi(\lambda)$ for any $\lambda \in (0, b)$, the set $\{s \geq 0 \mid \phi^*(s) > t\}$ is non-empty. Then $\phi^*(s) > t$ if and only if

$$\lambda s - \phi(\lambda) > t$$

for some $\lambda \in (0, b)$, that is, $s > \frac{t+\phi(\lambda)}{\lambda}$ for some $\lambda \in (0, b)$.

The concavity follows because $(\phi^*)^{-1}(t)$ is the infimum of linear functions of t. Because $\inf_{\lambda > 0} \phi(\lambda)/\lambda = \phi'(0) = 0$ by convexity (the slope $\phi'(\lambda)$ is non-decreasing in λ), we have $(\phi^*)^{-1}(t) \geq \inf_{\lambda \in (0,b)} t/\lambda = t/b$ for all $t \geq 0$. Coupled with concavity this implies $(\phi^*)^{-1}$ is strictly increasing, with $\lim_{t\to\infty}(\phi^*)^{-1}(t) = \infty$. □

Returning to the main thread of the argument, item (ii) then evidently implies

$$\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \frac{D_{\rm kl}\left(Q\|P\right) + \phi(\lambda)}{\lambda}$$

for all $\lambda \in (0, b)$, or, taking $g(X) = X - \mathbb{E}_P[X]$, that $\lambda \mathbb{E}_Q[g(X)] - D_{\rm kl}\left(Q\|P\right) \leq \phi(\lambda)$ for all Q. Applying Corollary 7.1.1 and taking a supremum over Q then yields $\log \mathbb{E}_P[e^{\lambda g(X)}] \leq \phi(\lambda)$. □

7.2* Transportation inequalities

In Corollary 7.1.3, we saw our first example of a *transportation inequality*: because

$$\|P - Q\|_{\rm TV} \leq \sqrt{\frac{1}{2} D_{\rm kl}\left(Q\|P\right)}$$

for any distributions P and Q, we saw that for any function f taking values in $[0, 1]$ we had

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \|P - Q\|_{\rm TV} \leq \sqrt{\frac{1}{2} D_{\rm kl}\left(Q\|P\right)},$$

and so *a fortiori* f is sub-Gaussian. We also have the characterization that $\sup_{f \in [0,1]} \mathbb{E}_Q[f] - \mathbb{E}_P[f] = \|P - Q\|_{\rm TV}$. To see this as a "transportation" inequality requires a bit more work.

7.2.1 Transportation between finitely supported distributions

For motivation, let us consider a finite setting, where X and Y take on values $1, \ldots, n$ (representing locations), where X has p.m.f. p and Y has p.m.f. q. We consider moving mass from $X \sim p$ to $Y \sim q$, and we can take one of two perspectives: the first is to look for a Markov kernel from $[n]$ to $[n]$, which specifies a transition from a location j to a location i and $P(X = i \mid Y = j)$ is the mass assigned, and we must have the consistency conditions

$$p_i = \sum_j P(X = i \mid Y = j) q_j \quad \text{and} \quad q_j = \sum_i P(Y = j \mid X = i) p_i,$$

where $p_i = P(X = i)$ and $q_j = P(Y = j)$. It is simpler to instead simply consider a joint distribution $\pi_{ij} = P(X = i, Y = j)$ satisfying the consistency conditions that

$$p_i = \sum_{j=1}^{n} \pi_{ij} \quad \text{and} \quad q_j = \sum_{i=1}^{n} \pi_{ij}, \tag{7.2.1}$$

which we see immediately is equivalent to the earlier display. We call a distribution satisfying the equality (7.2.1) a *coupling* of p and q.

Abstracting away from the index sets and letting $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, given a cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for moving mass, where $c(x, y)$ indicates the cost to transport one unit from x to y, the *transportation cost* between p and q is

$$W_c(p, q) := \inf_{\pi} \left\{ \mathbb{E}_\pi[c(X, Y)] \mid \pi \mathbf{1} = p, \; \pi^\top \mathbf{1} = q, \; \pi \geq 0 \right\}, \tag{7.2.2}$$

where the infimum (7.2.2) is over joint distributions π that we represent as a matrix for convenience. (We use the letter W because such quantities are frequently termed *Wasserstein distances* or *Kantorovich-Rubinstein distances*.) As it is a finite-dimensional linear program with compact and feasible constraint set (take $\pi = pq^\top$), the infimum (7.2.2) is attained and we can compute a dual with equal value [82]:

Proposition 7.2.1. *Let X and Y be random variables with finite supports $\mathcal{X}$ and $\mathcal{Y}$ and probability mass functions p and q, respectively. Then*

$$W_c(p, q) = \sup_{f:\mathcal{X}\to\mathbb{R}, g:\mathcal{Y}\to\mathbb{R}} \left\{ \mathbb{E}_p[f(X)] - \mathbb{E}_q[g(Y)] \mid f(x) - g(y) \leq c(x, y) \;\; \textit{for each } x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

Before proving the proposition, we use it to connect the transportation cost (7.2.2) to the variation distance. Let the cost $c(x, y) = \mathbf{1}\{x \neq y\}$, so that $c(x, y) = 0$ if $x = y$ and 1 otherwise. Then the transportation cost recovers total variation. The dual formulation makes this transparent, as for any g, the largest admissible f in Proposition 7.2.1 is

$$f_g(x) = \inf_y \{g(y) + \mathbf{1}\{x \neq y\}\}.$$

Similarly, the smallest admissible g is

$$g_f(y) = \sup_x \left\{ f(x) - \mathbf{1}\{x \neq y\} \right\},$$

Both transforms f_g and g_f are 1-Lipschitz with respect to the Hamming metric: observe that

$$f_g(x) - f_g(x') \leq \sup_y \left\{ \mathbf{1}\{x \neq y\} - \mathbf{1}\left\{x' \neq y\right\} \right\} = \mathbf{1}\left\{x \neq x'\right\},$$

and similarly for the lower bound and g_f. Consequently, we may restrict to 1-Lipschitz functions, and if f is 1-Lipschitz for the discrete metric, then $g_f(y) = f(y)$, so

$$\begin{aligned} W_c(p, q) &= \sup_f \left\{ \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)] \mid |f(x) - f(y)| \leq \mathbf{1}\{x \neq y\} \;\; \text{for all } x, y \right\} \\ &= \sup_{f:\mathcal{X}\to[0,1]} \left\{ \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)] \right\} = \|P - Q\|_{\mathrm{TV}}. \end{aligned}$$

Summarizing, for the 0-1 cost $c(x, y) = \mathbf{1}\{x \neq y\}$, the

$$\inf_{\pi} \mathbb{E}_\pi[\mathbf{1}\{X \neq Y\}] = \|P - Q\|_{\rm TV} = \sup_{f \in [0,1]} \{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]\}.$$

We return to the promised proof:

Proof of Proposition 7.2.1 Without loss of generality, we take $\mathcal{X} = \mathcal{Y} = [n]$, as everything is finite. Let $C = [c(i,j)]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ be the cost matrix, and treat $\pi \in \mathbb{R}_+^{n \times n}$ as a matrix, so $\langle C, \pi \rangle = \sum_{i,j} C_{ij}\pi_{ij}$ denotes their inner product. Then we have the optimization problem

$$\begin{array}{ll} \text{minimize} & \langle C, \pi \rangle \\ \text{subject to} & \pi \mathbf{1} = p, \; \pi^\top \mathbf{1} = q, \; \pi \geq 0. \end{array}$$

Introduce Lagrange multipliers $\alpha \in \mathbb{R}^n$ for the constraint $\pi\mathbf{1} = p$, $\beta \in \mathbb{R}^n$ for the constraint $\pi^\top \mathbf{1} = q$, and $\Lambda \in \mathbb{R}_+^{n \times n}$ for the nonnegativity constraint. This gives Lagrangian

$$L(\pi, \alpha, \beta, \Lambda) = \langle C, \pi \rangle + \alpha^\top(\pi\mathbf{1} - p) - \beta^\top(\pi^\top \mathbf{1} - q) - \langle \Lambda, \pi \rangle,$$

and taking an infimum to compute the dual, we see that

$$g(\alpha, \beta, \Lambda) := \inf_\pi L(\pi, \alpha, \beta, \Lambda) = \begin{cases} \beta^\top q - \alpha^\top p & \text{if } C + \alpha\mathbf{1}^\top - \mathbf{1}\beta^\top - \Lambda = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Because $\Lambda_{ij} \geq 0$, this is equivalent to the dual problem

$$\begin{array}{ll} \text{maximize} & \beta^\top q - \alpha^\top p \\ \text{subject to} & \beta_j - \alpha_i \leq C_{ij}, \quad i, j = 1, \ldots, n. \end{array}$$

Replacing α and β with function values $-f$ and $-g$ (respectively) gives the result. □

7.2.2 Transportation between general distributions

Proposition 7.2.1 holds in substantially more generality and yielding a plethora of powerful concentration inequalities. Given distributions P and Q on sets $\mathcal{X}$ and $\mathcal{Y}$, we let $\Pi(P, Q)$ denote the set of *couplings* between the distributions, meaning joint distributions on $\mathcal{X} \times \mathcal{Y}$ whose marginals are correct:

$$\Pi(P, Q) := \{\pi \text{ on } \mathcal{X} \times \mathcal{Y} \text{ s.t. } \pi(\cdot, \mathcal{Y}) = P(\cdot) \text{ and } \pi(\mathcal{X}, \cdot) = Q(\cdot)\},$$

analogizing the conditions (7.2.1). Then for a nonnegative cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the *transportation cost* between P and Q is

$$W_c(P, Q) := \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[c(X, Y)]. \tag{7.2.3}$$

As in the definition (7.2.2), we think of the coupling π as a "plan" for moving probability mass from P to Q via some sampling scheme $\pi(\cdot \mid X)$ or $\pi(\cdot \mid Y)$.

In analogy with the total variation distance, which uses the discrete distance $c(x, y) = \mathbf{1}\{x \neq y\}$, we can consider other costs on $\mathcal{X} \times \mathcal{Y}$. Notably, for any pair of functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, whenever $f(x) + g(y) \leq c(x, y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have the trivial inequality

$$\mathbb{E}_\pi[c(X, Y)] \geq \mathbb{E}_\pi[f(X) + g(Y)] = \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)].$$

Deep duality results extending Proposition 7.2.1 hold for these distances. The next theorem, a variant of results typically called the *Kantorovich duality theorems*, captures the main results.

Theorem 7.2.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{\infty\}$ be a lower semicontinuous cost function. Then for any distributions P and Q*

$$W_c(P, Q) = \sup \left\{ \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)] \mid f(x) + g(y) \le c(x, y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\},$$

and there is a coupling $\pi \in \Pi(P, Q)$ achieving $\mathbb{E}_\pi[c(X, Y)] = W_c(P, Q)$.

We provide a heuristic sketch of a proof of Theorem 7.2.2 in Section 7.2.4 (which more or less follows the proof of Proposition 7.2.1) to suggest why we might expect duality to hold, with pointers to more rigorous references. It is also typically the case that the supremum is attained by some functions f, g as well, but this is beyond our scope.

Most frequently, we consider distances as cost functions c, so that c satisfies the triangle inequality. In this case, we can restrict the supremum to be over only 1-Lipschitzian functions, where for a function f we define the Lipschitzian norm

$$\|f\|_{\text{Lip},c} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{c(x, y)}.$$

Indeed, we have the following corollary, whose proof we defer to Section 7.2.5.

Corollary 7.2.3. *In addition to the conditions of Theorem 7.2.2, assume that c is a distance on $\mathcal{X} \times \mathcal{X}$. Then*

$$W_c(P, Q) = \min_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[c(X, Y)] = \sup \left\{ \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \mid \|f\|_{\text{Lip},c} \le 1 \right\}$$

(Here we use min because the infimum defining W_c is attained.) For example, for the total variation distance, we have $c(x, y) = \mathbf{1}\{x \neq y\}$, and f is Lipschitz if $|f(x) - f(y)| \le 1$ for all x, y.

7.2.3 A tensorized transportation inequality

The power of transportation inequalities comes from their ability to tensorize—one can extend a one-dimensional transportation inequality to an n-dimensional one using properties of the KL-divergence and convexity. We give one such argument here, noting that more powerful approaches are possible (see the bibliographic section for references). Our starting point is a product distribution where each P_i satisfies a (modified) transportation cost inequality for an increasing convex function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\phi(W_c(Q, P_i)) \le D_{\rm kl}\left(Q \| P_i\right) \quad \text{for all } Q \tag{7.2.4}$$

for $i = 1, \ldots, n$. For example, Pinsker's inequality gives the bound (7.2.4) for any distribution P_i with cost $c(x, y) = \mathbf{1}\{x \neq y\}$ and $\phi(t) = 2t^2$. The following theorem shows how we can leverage the marginal transport inequality (7.2.4) into one involving n variables.

Theorem 7.2.4. *Let the distributions P_i on sets $\mathcal{X}_i$, $i = 1, \ldots, n$, satisfy the marginal transportation bound (7.2.4). Then the product $P = P_1 \times \cdots \times P_n$ satisfies*

$$\inf_{\pi \in \Pi(Q,P)} \sum_{i=1}^n \phi\left(\mathbb{E}_\pi[c(X_i, Y_i)]\right) \le D_{\rm kl}\left(Q \| P\right)$$

for all distributions Q on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Proof We induct on n, assuming that the claimed inequality holds for some value n (the base case being the assumed inequality (7.2.4)). By the chain rule (2.1.6) for the KL-divergence, where we use the notation there as well for $X_1^{n+1} \sim Q$ and $Y_1^{n+1} \sim P^{n+1}$, we have

$$D_{\rm kl}\left(Q\|P^{n+1}\right) = D_{\rm kl}\left(X_1^n\|Y_1^n \mid X_{n+1}\right) + D_{\rm kl}\left(X_{n+1}\|Y_{n+1}\right).$$

Let π_{n+1} be the optimal coupling between Q_{n+1} and P_{n+1}, and for values $x_{n+1} \in \mathcal{X}$, let π_x be the optimal coupling between $Q(X_1^n \in \cdot \mid X_{n+1} = x)$ and P^n (these exist by Theorem 7.2.2). Then by the induction hypothesis,

$$\sum_{i=1}^n \phi\left(\mathbb{E}_{\pi_x}[c(X_i, Y_i) \mid X_{n+1} = x]\right) \le D_{\rm kl}\left(X_1^n\|Y_1^n \mid X_{n+1} = x\right)$$

and

$$\phi\left(\mathbb{E}_{\pi_{n+1}}[c(X_{n+1}, Y_{n+1})]\right) \le D_{\rm kl}\left(X_{n+1}\|Y_{n+1}\right).$$

Integrating over $X_{n+1} \sim Q_{n+1}$, Jensen's inequality implies that

$$\sum_{i=1}^n \phi\left(\mathbb{E}_{\pi}[c(X_i, Y_i)]\right) + \phi\left(\mathbb{E}_{\pi_{n+1}}[c(X_{n+1}, Y_{n+1})]\right)$$
$$\le D_{\rm kl}\left(X_1^n\|Y_1^n \mid X_{n+1}\right) + D_{\rm kl}\left(X_{n+1}\|Y_{n+1}\right) = D_{\rm kl}\left(Q\|P^{n+1}\right).$$

The distribution $\pi = \mathbb{E}_{Q_{n+1}}[\pi_{X_{n+1}}]$ and the optimal coupling π_{n+1} have a consistent joint distribution by construction, giving the theorem. □

In Section 7.3*, we develop applications of Theorem 7.2.4 to isoperimetric inequalities, which involve showing that for certain distributions, slight enlargements of sets drastically increase their probability, and strong converse inequalities in hypothesis testing. Here, we present a few corollaries of the result.

Corollary 7.2.5 (Marton's transportation inequality)**.** *Let P_i be distributions on $\mathcal{X}$ and $P = P_1 \times \cdots \times P_n$. Then for any distribution Q on $\mathcal{X}^n$,*

$$\inf_{\pi \in \Pi(P,Q)} \sum_{i=1}^n \pi(X_i \neq Y_i)^2 \le \frac{1}{2} D_{\rm kl}\left(Q\|P\right).$$

Proof For any Q and $Y \sim P_i$, we have by Pinsker's inequality and Theorem 7.2.2 that

$$\inf_{\pi \in \Pi(P_i,Q)} \pi(X \neq Y) = \|Q - P_i\|_{\rm TV} \le \sqrt{\frac{1}{2} D_{\rm kl}\left(Q\|P_i\right)}.$$

This is the assumption of Theorem 7.2.4 with $\phi(t) = 2t^2$. □

While we have already proved the result by easier means, we can recover the bounded differences inequality (Proposition 4.2.6) as a corollary.

Corollary 7.2.6. *Let f satisfy bounded differences with constants $b = [b_i]_{i=1}^n$ (Definition 4.6) and let X_i be independent random variables. Then for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\lambda(f(X_1^n) - \mathbb{E}[f(X_1^n)])}\right] \le \exp\left(\frac{\lambda^2 \|b\|_2^2}{8}\right).$$

Proof By definition, f satisfies

$$|f(x_1^n) - f(y_1^n)| \le \sum_{i=1}^n b_i \mathbf{1}\left\{x_i \neq y_i\right\}.$$

We use Corollary 7.2.5 coupled with Theorem 7.1.2 to prove the inequality. Letting $Z(x_1^n) = f(x_1^n) - \mathbb{E}_P[f(X_1^n)]$, for any coupling $\pi \in \Pi(Q, P)$ on $Y_1^n \sim Q$ and $X_1^n \sim P$ we have

$$\mathbb{E}_Q[Z(Y_1^n)] - \mathbb{E}_P[Z(X_1^n)] \le \sum_{i=1}^n b_i \pi(Y_i \neq X_i) \le \left(\sum_{i=1}^n b_i^2\right)^{1/2} \left(\sum_{i=1}^n \pi(Y_i \neq X_i)^2\right)^{1/2}$$

by Cauchy-Schwarz. Applying Corollary 7.2.5 gives

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le \|b\|_2 \sqrt{D_{\rm kl}\left(Q\|P\right) / 2},$$

and applying the same argument to $-Z$ and using Theorem 7.1.2 implies the result. □

7.2.4 A heuristic proof of Theorem 7.2.2

For a formal statement of this result, and much more detail, we refer to Villani [414]; Theorem 4.1 of his book guarantees the existence of an optimal coupling so long as c is lower semicontinuous, while Theorem 5.10 provides the duality guarantees for $c \ge 0$.

Our proof is non-rigorous and mostly for intuition; to make the writing simpler, we assume P and Q have densities p and q for Lebesgue measure, and we will dispense with all measure-theoretic details. Letting π also be the density of (X, Y), we write the initial transport problem as

$$\begin{array}{ll} \text{minimize} & \mathbb{E}_\pi[c(X,Y)] \\ \text{subject to} & \int_{\mathcal{Y}} \pi(x,y)dy = p(x) \quad \text{and} \quad \int_{\mathcal{X}} \pi(x,y)dx = q(y) \end{array}$$

over $\pi \ge 0$. Introduce Lagrange multiplier functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ to obtain the Lagrangian

$$\begin{aligned} &L(\pi, f, g) \\ &= \int c(x,y)\pi(x,y)dxdy + \int_{\mathcal{X}} f(x)\left(p(x) - \int_{\mathcal{Y}} \pi(x,y)dy\right)dx + \int_{\mathcal{Y}} g(y)\left(q(y) - \int_{\mathcal{X}} \pi(x,y)dx\right)dy \\ &= \int \left(c(x,y) - f(x) - g(y)\right)\pi(x,y)dxdy + \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]. \end{aligned}$$

To obtain a dual problem, we minimize out the nonnegative function π. If there exists a pair (x, y) with $f(x) + g(y) > c(x, y)$ (again, eliding rigor), then we could send $\pi(x, y) \uparrow \infty$, making the first integral $-\infty$. Thus we obtain

$$\inf_{\pi \ge 0} L(\pi, f, g) = \begin{cases} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)] & \text{if } f(x) + g(y) \le c(x,y) \text{ all } x \in \mathcal{X}, y \in \mathcal{Y} \\ -\infty & \text{otherwise.} \end{cases}$$

Assuming strong duality obtains, this gives the associated dual problem and theorem.

7.2.5 Proof of Corollary 7.2.3

Recognize that given any function f, we can only improve the objective $\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]$ in Theorem 7.2.2 by taking g as large as possible. Accordingly, we define the c-conjugate of f by

$$f^c(y) := \inf_x \left\{c(x, y) - f(x)\right\},$$

which at each y is the largest possible value v satisfying $f(x) + v \le c(x, y)$ for all x. Setting $g = f^c$ therefore gives the equivalent problem

$$\sup \left\{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^c(Y)]\right\},$$

where we note that $f(x) + f^c(y) \le c(x, y)$ for all x, y. The conjugate f^c is Lipschitz with respect to the cost (distance) c: we have

$$f^c(x) - f^c(y) \le \sup_z \left\{c(z, x) - c(z, y)\right\} \le c(x, y)$$

by the triangle inequality, and the lower bound is similar, showing that f^c is Lipschitz.

We should set f as large as possible given f^c while satisfying $f(x) + f^c(y) \le c(x, y)$, meaning we choose

$$f^\star(x) = \inf_y \left\{c(x, y) - f^c(y)\right\}.$$

But for this we note that because f^c is Lipschitz,

$$c(x, y) - f^c(y) \ge c(x, y) - f^c(x) - c(x, y) = -f^c(x),$$

so $\inf_y\{c(x, y) - f^c(y)\} = -f^c(x)$. That is, $f^\star(x) = -f^c(x)$ and is therefore Lipschitz.

7.3* Two applications of transportation

We give two applications for the transportation inequalities we have developed. The first shows how transportation inequalities give concentration bounds that depend more directly on the underlying geometry of a set $\mathcal{X}$ and probability distribution P on $\mathcal{X}$. Such *isoperimetric inequalities* typically show that if a set A has probability $P(A) \ge \frac{1}{2}$, then slight enlargements of the set, e.g., $A + \epsilon\mathbb{B}$ for a ball $\mathbb{B}$, have $P(A + \epsilon\mathbb{B}) \approx 1$ even for small ϵ. We give one of the "basic" forms of such inequalities, then leverage the ideas to revisit optimality in hypothesis testing based on observations $X_i \stackrel{\text{iid}}{\sim} P$ and related *strong converses*. As we will see, product distributions P^n typically satisfy some type of isoperimetic inequality, and as tests using n observations involve accepting a region $A \subset \mathcal{X}^n$, increasing their probability of acceptance even a small amount (by enlarging A) can drastically decrease their probability of correctly rejecting a null distribution. This highlights one of the ways that concentration and convergence interact with information theory. In Section 12.4*, we will present further applications of these results to fundamental limits in statistics and communication.

7.3.1 Metric Gaussianity, transport inequalities, and expansion of sets

The connection between the transportation inequalities we have presented and concentration allows us to move beyond concentration of individual functions to more "geometric" concentration properties. The general formulation of such transportation inequalities is beyond our scope, but we provide the basic form to give the flavor of the results.

We begin by a general definition:

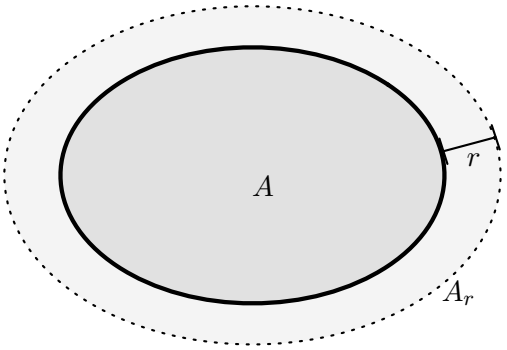


Figure 7.1. The r-blowup (or r-expansion) of the set A.

Definition 7.1. *Let $(\mathcal{X}, \rho)$ be a metric space with metric ρ, and let $p \in [1, 2]$. Define the cost function $c_p(x, y) := \rho^p(x, y)$. A probability distribution P on $\mathcal{X}$ satisfies an L^p* transportation cost inequality with constant σ^2, *or a $\mathrm{T}_p(\sigma^2)$-inequality, if for all distributions Q on $\mathcal{X}$*

$$W_{c_p}(Q, P)^{1/p} \le \sqrt{2\sigma^2 D_{\rm kl}\left(Q\|P\right)}.$$

We only consider $\mathrm{T}_1(\sigma^2)$-inequalities in this section, as they admit the most straightforward arguments. The bibliographic section provides pointers to other results.

Pinsker's inequality shows if we use the Hamming metric on $\mathcal{X}$, then *all* distributions P satisfy a $\mathrm{T}_1(\frac{1}{4})$-inequality. Marton's transportation inequality (Corollary 7.2.5) also shows that the product measure P^n satisfies a T_1-inequality, but with a constant that depends on n: let $d_{\rm ham}$ be the Hamming distance on $\mathcal{X}^n$. Then for any coupling π on $X_1^n \sim Q$ and $Y_1^n \sim P^n$, we have

$$\mathbb{E}_\pi[d_{\rm ham}(X_1^n, Y_1^n)] = \sum_{i=1}^n \pi(X_i \neq Y_i) \le \sqrt{n}\left(\sum_{i=1}^n \pi(X_i \neq Y_i)^2\right)^{1/2},$$

and so

$$W_{d_{\rm ham}}(Q, P^n) \le \sqrt{\frac{n}{2}}\sqrt{D_{\rm kl}\left(Q\|P^n\right)},$$

a $\mathrm{T}_1(\frac{n}{4})$-inequality.

The astounding consequence—perhaps not so astounding given the variational connection between sub-Gaussianity and the KL-divergence that the Donsker-Varadhan theorem provides—of such transport inequalities is that probability measure P satisfying them then inherits strong concentration of measure properties. To discuss this we require a bit of additional notation. For a metric space $(\mathcal{X}, \rho)$ and set $A \subset \mathcal{X}$, define the *r-blowup* or *r-expansion* of A by

$$A_r := \{x \in \mathcal{X} \mid \rho(x, A) \le r\},$$

(see Figure 7.1) where we recall the notation $\rho(x, A) = \inf_{y \in A} \rho(x, y)$. When a small blowup of A drastically increases its probability, then P exhibits strong concentration properties. A more quantitative version of this follows: we say P exhibits *Gaussian metric concentration on* $(\mathcal{X}, \rho)$ with constant $\kappa > 0$ if for some $K < \infty$,

$$P(A) \ge \frac{1}{2} \;\text{ implies }\; P(A_r) \ge 1 - Ke^{-\kappa r^2} \text{ for } r \ge 0. \tag{7.3.1}$$

Because we often think of r growing, the only important constant is $\kappa > 0$, and so sometimes we provide the weaker guarantee $P(A_r) \geq 1 - Ke^{-\kappa[r-r_0]_+^2}$ for some $r_0 \geq 0$. This latter guarantee implies the Gaussian metric concentration (7.3.1) for large r with a worse constant K and (asymptotically) equivalent κ.

By a clever conditioning argument, it is possible to show the *blowing up* lemma: whenever a distribution P satisfies a transportation inequality, it enjoys Gaussian metric concentration.

Theorem 7.3.1 (The blowing up lemma)**.** *Let P satisfy a $\mathrm{T}_1(\sigma^2)$-inequality. Then for any measurable set A with $P(A) > 0$ and $r \geq 0$,*

$$P(A) \cdot P(A_r^c) \leq \exp\left(-\frac{r^2}{4\sigma^2}\right)$$

and

$$P(A_r) \geq 1 - \exp\left(-\frac{1}{2\sigma^2}\left[r - \sqrt{2\sigma^2 \log \frac{1}{P(A)}}\right]_+^2\right).$$

Letting A be any set with $P(A) \geq \frac{1}{2}$, the first inequality in Theorem 7.3.1 implies that $P(A_r^c) \leq 2\exp(-\frac{r^2}{4\sigma^2})$, that is,

$$P(A_r) \geq 1 - 2\exp\left(-\frac{r^2}{4\sigma^2}\right).$$

The second inequality offers better constants in the exponent: taking $r_0 = \sqrt{2\sigma^2 \log 2}$ and $r = t + r_0$, we always have

$$P(A_{r_0+t}) \geq 1 - \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof For $A \subset \mathcal{X}$, define the conditional distribution

$$P_A(S) := P(A \cap S)/P(A) = P(S \mid A).$$

Without loss of generality, we may assume P_A and P have densities p_A and p with respect to some base measure μ, where $p_A(x) = p(x)\mathbf{1}\left\{x \in A\right\}/P(A)$. Then by inspection,

$$D_{\rm kl}\left(P_A \| P\right) = \int p_A(x) \log \frac{p_A(x)}{p(x)} d\mu(x) = \log \frac{1}{P(A)}.$$

For any sets A, B, we have

$$W_\rho(P_A, P_B) \leq W_\rho(P_A, P) + W_\rho(P_B, P)$$

by the triangle inequality, and so because P satisfies a $\mathrm{T}_1(\sigma^2)$-inequality,

$$W_\rho(P_A, P_B) \leq \sqrt{2\sigma^2 \log \frac{1}{P(A)}} + \sqrt{2\sigma^2 \log \frac{1}{P(B)}}. \tag{7.3.2}$$

Using the notation $\rho(A, B) = \inf_{x \in A} \inf_{y \in B} \rho(x, y)$ for the closest distance between A and B, for any coupling $\pi \in \Pi(P_A, P_B)$ on $X \in A$ and $Y \in B$, we can write

$$\mathbb{E}_\pi[\rho(X, Y)] \geq \inf_{x \in A} \inf_{y \in B} \rho(x, y) = \rho(A, B). \tag{7.3.3}$$

Substituting this into the bound (7.3.2), we have for any (measurable) sets A, B with $P(A) > 0$ and $P(B) > 0$ that

$$\rho(A, B) \le \sqrt{2\sigma^2 \log \frac{1}{P(A)}} + \sqrt{2\sigma^2 \log \frac{1}{P(B)}}.$$

From here, we can perform two manipulations. The first is to recognize that by the inequality $(\sqrt{a} + \sqrt{b})^2 = a + b + 2\sqrt{ab} \le 2a + 2b = (\sqrt{2(a+b)})^2$, we have $\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$, and so

$$\rho(A, B) \le \sqrt{4\sigma^2 \log \frac{1}{P(A)} \frac{1}{P(B)}}.$$

Now we take B be the set of points at least ρ distance r from A: recalling the r-blowup $A_r := \{y \in \mathcal{X} \mid d(y, A) \le r\}$ of A, we set $B = A_r^c$ and obtain $\rho(A, B) = \rho(A, A_r^c) \ge r$ for all $r \ge 0$, and $P(B) = P(A_r^c) = 1 - P(A_r)$. In particular,

$$r \le 2\sqrt{\sigma^2 \log \frac{1}{P(A)P(A_r^c)}}.$$

Rearranging gives $P(A)P(A_r^c) \le \exp(-\frac{r^2}{4\sigma^2})$, the first claim of the theorem. For the second, we can use inequality (7.3.2) directly to obtain

$$r \le W_\rho(P_A, P_{A_r^c}) \le \sqrt{2\sigma^2 \log \frac{1}{P(A)}} + \sqrt{2\sigma^2 \log \frac{1}{1 - P(A_r)}}.$$

Rearrange and solve for $P(A_r)$. □

Theorem 7.3.1 is particularly evocative for product measures on discrete spaces. In this case, we use the Hamming metric $d_{\rm ham}$, which satisfies a $\mathrm{T}_1(\frac{n}{4})$-transport cost inequality as in the discussion following Definition 7.1.

Corollary 7.3.2. *Let $A \subset \mathcal{X}^n$ be any measurable set and let $X_i \stackrel{\rm iid}{\sim} P$. Then for any $r \ge 0$,*

$$\mathbb{P}\left(d_{\rm ham}(X_1^n, A) \le r\right) \ge 1 - \exp\left(-\frac{2}{n}\left[r - \sqrt{\frac{n}{2}\log\frac{1}{P^n(A)}}\right]_+^2\right)$$

As a consequence, for any sequence of sets that have at least polynomial probability—the sets $A_n \subset \mathcal{X}^n$ satisfy $P^n(A_n) \ge n^{-p}$ for some $p < \infty$—neighborhoods expanded by only $\sqrt{n}$ in the Hamming metric necessarily cover nearly the entire probability mass. Indeed, in this case, $\log \frac{1}{P^n(A)} \le p \log n$, and so setting

$$r_n = c\sqrt{\frac{n}{2}} + \sqrt{\frac{pn}{2}\log n},$$

we have

$$\mathbb{P}(d_{\rm ham}(X_1^n, A_n) \le r_n) \ge 1 - \exp(-c^2) \text{ for all } n.$$

7.3.2 A weak and strong converse for hypothesis testing

Let us return to Section 2.3.1 to revisit the statistical question that underlies much of our development: a simple hypothesis test. We wish to test a null H_0 against an alternative H_1,

$$H_0 : X_1^n \stackrel{\text{iid}}{\sim} P_0 \text{ versus } H_1 : X_1^n \stackrel{\text{iid}}{\sim} P_1,$$

where P_0 and P_1 are different distributions. Section 2.3.1 connects bounds on the variation distance to lower bounds on the summed probability of error. Here, we give a different interpretation, fixing both P and Q and showing that if the type-I probability of error (a false positive, i.e., incorrectly rejecting H_0) is small, the type-II probability of error can only decrease at a certain rate.

To give intuition, let us first recall the optimal test, which gives an achievability result and shows the correct behavior of the asymptotic error. We assume that the variance of the log-likelihood ratio $v^2 := \mathrm{Var}_{P_0}(\log \frac{p_0(X)}{p_1(X)}) < \infty$. Define the log-likelihood ratio of a sequence x_1^n by

$$L_n(x_1^n) := \sum_{i=1}^n \log \frac{p_0(x_i)}{p_1(x_i)},$$

so that if $\Phi(t) = \mathbb{P}(Z \le t)$ is the standard normal CDF, the central limit theorem implies

$$P_0^n \left(L_n(X_1^n) \le nD_{\rm kl}\left(P_0 \| P_1\right) + v\sqrt{n}\Phi^{-1}(\epsilon) \right) \to \epsilon. \tag{7.3.4}$$

The Neyman-Pearson lemma states that at a given type-I error ϵ, the optimal test of P_0 against P_1 compares $L_n(X_1^n)$ against a threshold. That is, letting $\Psi = 0$ indicate accepting H_0 and $\Psi = 1$ indicate accepting H_1, subject to the constraint that $P_0^n(\Psi_n = 0) \ge 1 - \epsilon$ the optimal test is

$$\Psi_n(x_1^n) = \begin{cases} 0 : \text{accept } H_0 & \text{if } L_n(x_1^n) - nD_{\rm kl}\left(P_0 \| P_1\right) \ge v\sqrt{n}\Phi^{-1}(\epsilon) + a_n(\epsilon) \\ 1 : \text{accept } H_1 & \text{otherwise,} \end{cases}$$

where the sequence $a_n(\epsilon)$ is chosen so that $P_0^n(\Psi_n = 0) = 1 - \epsilon$. By the central limit calculation (7.3.4), we may take $a_n(\epsilon)$ so that $a_n(\epsilon)/\sqrt{n} \to 0$.

Now we consider the type-II error, that is, $P_1^n(\Psi_n = 0)$. We observe that

$$\begin{aligned} P_1^n(\Psi_n = 0) &= P_1^n \left(L_n(X_1^n) - nD_{\rm kl}\left(P_0 \| P_1\right) \ge v\sqrt{n}\Phi^{-1}(\epsilon) + a_n(\epsilon) \right) \\ &\stackrel{(i)}{\le} \mathbb{E}_{P_1^n}\left[e^{L_n(X_1^n)} \right] \exp\left(-nD_{\rm kl}\left(P_0 \| P_1\right) - v\sqrt{n}\Phi^{-1}(\epsilon) - a_n(\epsilon) \right) \\ &\stackrel{(ii)}{=} \exp\left(-nD_{\rm kl}\left(P_0 \| P_1\right) + v\sqrt{n}\Phi^{-1}(1-\epsilon) - a_n(\epsilon) \right), \end{aligned}$$

where inequality (i) is a Chernoff bound and equality (ii) holds because $\mathbb{E}_{P_1}[\exp(\log \frac{p_0(X)}{p_1(X)})] = 1$. Summarizing, we have the following result.

Proposition 7.3.3. *Let $\epsilon \in (0,1)$ and let Ψ_n be the level-ϵ likelihood ratio test of H_0 against H_1, and assume that $v^2 := \mathrm{Var}_{P_0}(\log \frac{p_0(X)}{p_1(X)}) < \infty$. Then*

$$\frac{1}{n} \log P_1^n(\Psi_n = 0) \le -D_{\rm kl}\left(P_0 \| P_1\right) + \frac{v}{\sqrt{n}}\Phi^{-1}(1-\epsilon) + o(1/\sqrt{n}).$$

Said differently, the probability of falsely rejecting P_1 tends to 0 exponentially quickly (in n) with rate $D_{\rm kl}\left(P_0\|P_1\right)$ and fluctuations of order $1/\sqrt{n}$ around this rate for any level $\epsilon > 0$. Classical techniques, relying on large-deviations bounds (see Exercises 7.4 and 7.5) provide one avenue for showing the optimality of these bounds.

Here, we derive converse results—lower bounds on $\log P_1^n(\Psi_n = 0)$—providing similar rates of convergence using the blowing-up lemma (Theorem 7.3.1) without relying so carefully on the particulars of the simple hypothesis test. We first present a so-called "weak converse", then extend the idea to a "strong converse." One should think of this distinction as follows: a weak converse states that if the probability of type-I error (false positive, or a false discovery) tends to zero, then the asymptotic type-II error (false negative) can only decrease to zero at a certain rate. This does not, however, eliminate the possibility of lower type-II errors by allowing a fixed $\epsilon > 0$ probability of error in testing the null H_0. The strong converse, on the other hand, eliminates this possibility: for any fixed $\epsilon > 0$, *no* improvement in (asymptotic) type-II error is possible. The style of argument we employ here extends beyond this simple setting, and Section 12.4* employs it to give similar fundamental limits in some communication and estimation problems.

Both converses we prove here follow from the data processing inequality for the KL-divergence.

Proposition 7.3.4 (Weak converse in hypothesis testing)**.** *Let P_0 and P_1 be arbitrary distributions, and let the test Ψ have level ϵ, that is, $P_0(\Psi = 1) \le \epsilon$. Then*

$$-\log \frac{1}{P_1(\Psi = 0)} \ge -\frac{D_{\rm kl}\left(P_0\|P_1\right)}{1-\epsilon} - \frac{\log 2}{1-\epsilon}.$$

Specializing to the product distribution case, we obtain that

$$-\log \frac{1}{P_1^n(\Psi_n = 0)} \ge -\frac{nD_{\rm kl}\left(P_0\|P_1\right)}{1-\epsilon} - \frac{\log 2}{1-\epsilon},$$

and so for any $\epsilon_n \to 0$, we have a limit matching Proposition 7.3.3, as

$$\liminf_{n\to\infty} -\frac{1}{n}\log \frac{1}{P_1^n(\Psi_n = 0)} \ge -D_{\rm kl}\left(P_0\|P_1\right).$$

Proof Let $A = \{x \in \mathcal{X} \mid \Psi(x) = 0\}$, and let $p_0 = P_0(A)$ and $p_1 = P_1(A)$. Recall our notation for the binary relative entropy $D_{\rm kl}\left(p\|q\right) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ and the binary entropy functional $h_2(p) = -p\log p - (1-p)\log(1-p) \le \log 2$. Then by the data-processing inequality

$$\begin{aligned} D_{\rm kl}\left(P_0\|P_1\right) \ge D_{\rm kl}\left(P_0(A)\|P_1(A)\right) &= p_0\log\frac{p_0}{p_1} + (1-p_0)\log\frac{1-p_0}{1-p_1} \\ &= p_0\log\frac{1}{p_1} + (1-p_0)\log\frac{1}{1-p_1} - h_2(p_0) \\ &\ge p_0\log\frac{1}{p_1} - h_2(p_0). \end{aligned}$$

Rearranging gives $-\log\frac{1}{p_1} \ge -\frac{1}{p_0}D_{\rm kl}\left(P_0\|P_1\right) - \frac{\log 2}{p_0}$, and substituting $p_0 \ge 1-\epsilon$ gives the result. □

Let us now develop the stronger result. A limitation of the method we employ here is that we must assume $\mathcal{X}$ is finite to enable the strongest applications of the blowing up lemma (Theorem 7.3.1). We focus on the product distributions P_0^n and P_1^n, and let $\mathcal{X}$ be the support of P_0.

We assume without loss of generality that $\inf_{x\in\mathcal{X}} P_1(\{x\}) > 0$, as otherwise, $D_{\rm kl}\left(P_0\|P_1\right) = \infty$ and any lower bound is trivial. The key insight, which Ahlswede et al. [14] develop, is to apply the data processing inequality to the enlargement $A_r = \{x_1^n \mid d_{\rm ham}(x_1^n, A) \le r\}$ of the acceptance set $A = \{x \in \mathcal{X}^n \mid \Psi_n(x) = 0\}$ rather than A itself; because of the blowing-up lemma, this set has nearly full measure, which in turn means that $P_1(\Psi_n = 0)$ cannot be too small.

Proposition 7.3.5 (Strong converse in hypothesis testing)**.** *Let P_0 and P_1 be distributions on the finite set $\mathcal{X}$ with $P_1 \gg P_0$, and let the test Ψ_n have level ϵ for testing P_0^n against P_1^n. Then*

$$-\frac{1}{n}\log\frac{1}{P_1^n(\Psi_n = 0)} \ge -D_{\rm kl}\left(P_0\|P_1\right) - O(1)\frac{\log n\sqrt{\log\frac{n}{1-\epsilon}}}{\sqrt{n}}$$

for all large enough n.

Proof Let $A = \{x_1^n \in \mathcal{X}^n \mid \Psi_n(x_1^n) = 0\}$ be the acceptance set of Ψ_n, and let $A_r = \{x \mid d_{\rm ham}(x, A) \le r\}$ be its r-blowup. For simplicity in notation, define $p_r = P_0^n(A_r)$ and $q_r = P_1^n(A_r)$. Then by the data-processing inequality as in the proof of Proposition 7.3.4, we have

$$nD_{\rm kl}\left(P_0\|P_1\right) = D_{\rm kl}\left(P_0^n\|P_1^n\right) \ge D_{\rm kl}\left(P_0^n(A_r)\|P_1^n(A_r)\right) = D_{\rm kl}\left(p_r\|q_r\right) \ge p_r\log\frac{1}{q_r} - h_2(p_r). \quad (7.3.5)$$

To provide bounds on $\log\frac{1}{q}$ for $q = P_1^n(A)$, not q_r, we use the following combinatorial lemma:

Lemma 7.3.6. *Define $q_\star = \min_{x\in\mathcal{X}} P_1(\{x\})$ and $K = \mathrm{card}(\mathcal{X})/q_\star$. For any $r \ge 0$,*

$$P_1^n(A) \le P_1^n(A_r) \le \sum_{k=0}^{\lfloor r\rfloor}\binom{n}{k}K^k P_1^n(A).$$

Proof Let $x, x' \in \mathcal{X}^n$. Then

$$P_1^n(\{x\}) = P_1^n(\{x'\})\prod_{i:x_i\neq x_i'}\frac{P_1(\{x_i\})}{P_1(\{x_i'\})} \ge P_1^n(\{x'\})q_\star^{d_{\rm ham}(x,x')}.$$

Thus if $d_{\rm ham}(x, x') = k$, then $P_1^n(\{x'\}) \le q_\star^{-k}P_1^n(\{x\})$. Summing over the Hamming ball of radius $\lfloor r\rfloor$ around x gives

$$P_1^n(\{x\}_r) \le P_1^n(\{x\})\sum_{k=0}^{\lfloor r\rfloor}\binom{n}{k}\left(\frac{\mathrm{card}(\mathcal{X})}{q_\star}\right)^k.$$

The union bound over $x \in A$ implies the lemma. □

We therefore obtain for $K = \frac{\mathrm{card}(\mathcal{X})}{q_\star}$ and $q := P_1^n(A) = P_1^n(\Psi_n = 0)$ that

$$P_1^n(A_r) = q_r \le \sum_{k=0}^{\lfloor r\rfloor}\binom{n}{k}K^k q \le (r+1)\left(\frac{Kne}{r}\right)^r q$$

for $r \ge 1$, by the standard binomial bound $\binom{n}{k} \le (\frac{ne}{k})^k$ and the harmless integer rounding of r. Substituting into the data processing bound (7.3.5), we obtain

$$nD_{\rm kl}\left(P_0\|P_1\right) \ge p_r\log\frac{1}{q} - r\log\frac{Kne}{r} - \log(r+1) - \log 2.$$

Finally, we use the blowing-up lemma (Theorem 7.3.1): fix $t \ge 0$ to be chosen and take

$$r = r(t) := t\sqrt{\frac{n}{2}} + \sqrt{\frac{n}{2}\log\frac{1}{1-\epsilon}}.$$

Then $p_r \ge 1 - e^{-t^2}$, and so

$$nD_{\rm kl}\left(P_0\|P_1\right) \ge (1 - e^{-t^2})\log\frac{1}{q} - r\log\frac{Kne}{r} - \log(r+1) - \log 2.$$

If we take $t = \sqrt{2\log n}$ then $r = O(1)\sqrt{n\log\frac{n}{1-\epsilon}}$, and so

$$-(1 - n^{-2})\log\frac{1}{q} \ge -nD_{\rm kl}\left(P_0\|P_1\right) - O(1)\sqrt{n\log^2 n \cdot \log\frac{n}{1-\epsilon} + n\log n \cdot \log K}.$$

Dividing by $n(1 - 1/n^2)$ gives the result. □

7.4 Discussion and bibliographic remarks

Many disciplines study and leverage concentration of measure phenomena, and our approach here—centered on transportation cost inequalities—presents only a narrow slice of this deep area. The books by Ledoux [283] and Boucheron, Lugosi, and Massart [81] provide more comprehensive treatments, and Wainwright [422] presents many of the ideas and connects them with high-dimensional estimation in statistical problems. Marton [302, 303] first developed the connections between transportation cost and concentration inequalities, using similar information-theoretic motivations as we do here; other substantial contributions in this direction include Bobkov and Götze [78] and Dembo and Zeitouni [140, 139]. Transportation of measure more broadly connects with a number of mathematical fields, and Villani [414] gives general results capturing the wealth of duality results and connections.

The particular approaches here, especially the connections between sub-Gaussianity and divergence measures, owe particular debt to Boucheron et al. [81]. Raginsky and Sason [351] provide an excellent survey of the connections between concentration of measure, information theory, and coding. These two references also extend beyond the "simple" T_1-inequalities to T_p-inequalities, which then allows applications to (dimension-free) concentration of more sophisticated processes.

Obtaining strong converses motivates a body of work within information theory, and we will return to more discussion in later chapters. Within the particular setting of (simple) hypothesis testing of $H_0 : X_i \stackrel{\rm iid}{\sim} P_0$ against $H_1 : X_i \stackrel{\rm iid}{\sim} P_1$, the question of optimal error exponents asks, among tests $\Psi_n : \mathcal{X}^n \to \{0, 1\}$ satisfying $\mathbb{P}(\Psi_n = 1 \mid H_0) \le \epsilon$, for the optimal decay in $\frac{1}{n}\log\mathbb{P}(\Psi_n = 0 \mid H_1)$, which typically satisfies $\liminf_n \frac{1}{n}\log\mathbb{P}(\Psi_n = 0 \mid H_1) = -D_{\rm kl}\left(P_0\|P_1\right)$, as in Proposition 7.3.3. Obtaining these optimal asymptotics is classical [74, 118]. More recent work asks for optimal second-order terms in the asymptotic errors, that is, for expansions of the form $\mathbb{P}(\Psi_n = 0 \mid H_1) = \exp(-nD_{\rm kl}\left(P_0\|P_1\right) + \sqrt{n}(1 + o(1))V(P_0, P_1))$, where $V(P_0, P_1)$ captures the fluctuations, again as in Proposition 7.3.3. The use of the "blowing-up" approach [14, 124, 351] can almost provide these results, but, as in Proposition 7.3.5, is usually loose by logarithmic factors and requires finite $\mathcal{X}$. The paper [292] provides a way around this as well as a clean and elegant development for showing such optimality results.

7.5 Exercises

Exercise 7.1: Complete the proof of Corollary 7.1.1 in the case that $\mathbb{E}_Q[e^{g(X)}] = +\infty$.

Exercise 7.2 (A discrete isoperimetric inequality)**:** Let $A \subset \mathbb{Z}^d$ be a finite subset of the d-dimensional integers. Let the projection mapping $\pi_j : \mathbb{Z}^d \to \mathbb{Z}^{d-1}$ be defined by

$$\pi_j(z_1, \ldots, z_d) = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_d)$$

so that we "project out" the jth coordinate, and define the projected sets.

$$\begin{aligned} A_j = \pi_j(A) &= \{\pi_j(z) : z \in A\} \\ &= \left\{z \in \mathbb{Z}^{d-1} : \text{ there exists } z_\star \in \mathbb{Z} \text{ such that } (z_1, z_2, \ldots, z_{j-1}, z_\star, z_j, \ldots, z_{d-1}) \in A\right\}. \end{aligned}$$

Prove the Loomis-Whitney inequality, that is, that

$$\mathrm{card}(A) \le \left(\prod_{j=1}^d \mathrm{card}(A_j)\right)^{\frac{1}{d-1}}.$$

Exercise 7.3: Let X be a non-constant and mean-zero random variable with moment generating function φ_X defined on a neighborhood of 0, and let

$$b_1 = \sup\left\{\lambda \ge 0 \mid \varphi_X(\lambda) < \infty\right\} \quad \text{and} \quad b_0 = \inf\left\{\lambda \le 0 \mid \varphi_X(\lambda) < \infty\right\}.$$

You may assume that φ_X is $\mathcal{C}^\infty$ on (b_0, b_1) (Proposition 3.2.2 guarantees this).

(a) Show that φ_X is strictly increasing on $(0, b_1)$ and strictly decreasing on $(b_0, 0)$.

(b) Show that φ_X' is strictly increasing on (b_0, b_1).

(c) Show that $\varphi_X(\lambda) \to \varphi_X(b_1)$ as $\lambda \uparrow b_1$ and $\varphi_X(\lambda) \to \varphi_X(b_0)$ as $\lambda \downarrow b_0$.

(d) Show that $\varphi_X'(\lambda) \to \mathbb{E}[Xe^{b_1 X}] > 0$ as $\lambda \uparrow b_1$.

(e) Show that the domain of φ_X, even when non-trivial, may be open or closed by giving random variables X and Y for which $\operatorname{dom}\varphi_X = [-1, 1]$ and $\operatorname{dom}\varphi_Y = (-1, 1)$.

(f) Let $\phi(\lambda) = \log\varphi(\lambda)$ be the cumulant generating function and recall the essential supremum $\operatorname{ess\,sup}_P X = \inf\{t \mid P(X \ge t) = 0\}$ of the variable X. Show that $\lim_{\lambda\uparrow b_1} \phi'(\lambda) = \operatorname{ess\,sup}_P X$.

Exercise 7.4 (Large deviations by tilting)**:** Let $X_i \stackrel{\text{dist}}{=} X$ be i.i.d. mean-zero, non-constant random variables with moment generating function $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ defined in a neighborhood of 0; let $\phi(\lambda) = \log\varphi_X(\lambda)$ be its logarithm, which is convex and $\mathcal{C}^\infty$ on the interior of its domain (Proposition 3.2.2). Define the convex conjugate (i.e., optimizer in the Chernoff bound)

$$\phi^*(t) := \sup_\lambda \left\{t\lambda - \phi(\lambda)\right\},$$

recognizing that for $t \ge 0$ it is no loss of generality to take $\lambda \ge 0$ above, as $\phi'(0) = \mathbb{E}[X] = 0$.

(a) Show the large deviation upper bound that for any $t \ge 0$,

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\overline{X}_n \ge t) \le -\phi^*(t).$$

(b) Let X have density p_0 and fix $\lambda \in \operatorname{int} \operatorname{dom} \phi$, $\lambda \ge 0$. Define the density

$$p_\lambda(x) := p_0(x) \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]}.$$

For $t := \mathbb{E}_{P_\lambda}[X]$, show that the set $A_\delta := \{x \in \mathbb{R}^n \mid t < \overline{x}_n \le t + \delta\}$ satisfies

$$\mathbb{P}(\overline{X}_n > t) \ge \exp\left(n\left(\phi(\lambda) - \lambda t\right) - n\lambda\delta\right) P_\lambda^n(A_\delta).$$

(c) Under the same conditions as part (b), show that for any $\delta > 0$, $\lim_n P_\lambda^n(A_\delta) = \frac{1}{2}$ and so if $\phi'(\lambda) = t$, then for all $\delta > 0$,

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\overline{X}_n > t) \ge -\phi^*(t) - \lambda\delta.$$

(d) Prove the large deviation bound that for $t \ge 0$,

$$-\inf_{u > t} \phi^*(u) \le \liminf_n \frac{1}{n} \log \mathbb{P}(\overline{X}_n > t) \le \limsup_n \frac{1}{n} \log \mathbb{P}(\overline{X}_n \ge t) \le -\phi^*(t). \tag{7.5.1}$$

Hint. Part (f) of Exercise 7.3 may be useful.

As an aside, if $t \ge 0$ belongs to the interior of the support of X, then each of the inequalities (7.5.1) becomes an equality.

Exercise 7.5 (The Chernoff information and optimal testing)**:** The *Chernoff information* between distributions P_0 and P_1, which without loss of generality have densities p_0 and p_1, is

$$C(P_0, P_1) := \sup_{\lambda \in [0,1]} -\log \int p_0^{1-\lambda}(x) p_1^\lambda(x) d\mu(x).$$

Consider a Bayesian hypothesis test of $H_0 : X_i \stackrel{\text{iid}}{\sim} P_0$ against $H_1 : X_i \stackrel{\text{iid}}{\sim} P_1$, where $D_{\rm kl}\left(P_0 \| P_1\right) > 0$, and there is a prior $\pi_v = \pi(H_v)$ for $v \in \{0, 1\}$. For a test $\Psi_n : \mathcal{X}^n \to \{0, 1\}$, let $\alpha_n = \mathbb{P}(\Psi_n = 1 \mid H_0)$ and $\beta_n = \mathbb{P}(\Psi_n = 0 \mid H_1)$, so that we wish to minimize

$$P_n^{\sf err} := \alpha_n \pi_0 + \beta_n \pi_1.$$

This question proves that

$$\inf_{\Psi_n} \lim_n \frac{1}{n} \log P_n^{\sf err} = -C(P_0, P_1). \tag{7.5.2}$$

(a) Using that the optimal test between H_0 and H_1 is a likelihood ratio test, define the random variable $L_n := \sum_{i=1}^n \log \frac{p_0(X_i)}{p_1(X_i)}$. Show that if $\alpha_n^* = \mathbb{P}(L_n < 0 \mid H_0)$ and $\beta_n^* = \mathbb{P}(L_n > 0 \mid H_1)$, then

$$\frac{1}{n} \log P_n^{\sf err} \ge \frac{1}{n} \min\{\log \pi_0, \log \pi_1\} + \frac{1}{n} \min\left\{\log \alpha_n^*, \log \beta_n^*\right\}.$$

(b) Use inequality (7.5.1) in Exercise 7.4 to show the limit (7.5.2).

Exercise 7.6 ("Dimension free" covariance estimation [439])**:** This exercise outlines a proof of Proposition 7.1.6. Let the conditions of the proposition hold.

(a) Fix $\beta > 0$ and $r > 0$, and define π_0 on $\Theta = \mathbb{R}^d \times \mathbb{R}^d$ by $\mathsf{N}(0, \beta^{-1}\Sigma) \times \mathsf{N}(0, \beta^{-1}\Sigma)$, that is, the product of two mean-zero normals with covariance $\beta^{-1}\Sigma$. Let P_u be the normal distribution $\mathsf{N}(u, \beta^{-1}\Sigma)$ restricted to the ball $u + r\mathbb{B}_2^d = \{x \mid \|x - u\|_2 \le r\}$, and define the posterior $\pi_{u,v} = P_u \times P_v$. Show that for any $u, v \in \Sigma^{1/2}\mathbb{S}^{d-1} = \{u \mid u^\top \Sigma^{-1} u = 1\}$, we have

$$D_{\rm kl}\left(\pi_{u,v} \| \pi_0\right) = 2 \log \frac{1}{C(r)} + \beta,$$

where $C(r) := \mathbb{P}(\|Z\|_2 \le r)$ for $Z \sim \mathsf{N}(0, \beta^{-1}\Sigma)$ is the normalization for P_u.

(b) Show that if $r = \sqrt{2\beta^{-1} \operatorname{tr}(\Sigma)}$ then $C(r) \ge \frac{1}{2}$, and conclude that $D_{\rm kl}\left(\pi_{u,v} \| \pi_0\right) \le 2 \log 2 + \beta$.

(c) Define

$$f(\theta_1, \theta_2, M) := \theta_1^\top \Sigma^{-1/2} (M - \Sigma) \Sigma^{-1/2} \theta_2.$$

Show that for fixed θ_1, θ_2, f is sub-exponential, specifically,

$$\|f(\theta_1, \theta_2, M)\|_{\psi_1} \le \kappa^2 (\|\theta_1\|_2^2 + \|\theta_2\|_2^2).$$

(d) Let $u, v \in \Sigma^{1/2}\mathbb{S}^{d-1}$. Show that for $(\theta_1, \theta_2) \sim \pi_{u,v}$ with the choice $r = \sqrt{2 \operatorname{tr}(\Sigma)/\beta}$, we have

$$\max\left\{\|\theta_1\|_2, \|\theta_2\|_2\right\} \le \sqrt{\|\Sigma\|_{\rm op}} + \sqrt{2\beta^{-1} \operatorname{tr}(\Sigma)}$$

(e) Use Corollary 7.1.4 to show that for any fixed $|\lambda| \le \frac{1}{16\kappa^2 \|\Sigma\|_{\rm op}}$, with probability at least $1 - e^{-t}$,

$$\lambda u^\top \Sigma^{-1/2} \left(P_n M - \Sigma\right) \Sigma^{-1/2} v \le 64 \lambda^2 \kappa^4 \|\Sigma\|_{\rm op} + \frac{2 \log 2 + 2 \mathrm{r}_{\rm eff}(\Sigma) + t}{n}$$

for all vectors $u, v \in \Sigma^{1/2}\mathbb{S}^{d-1}$ by appropriate choice of β.

(f) Demonstrate the conclusion (7.1.4) of Proposition 7.1.6. (The constants are unimportant.)

Exercise 7.7: In this question, we compare the convergence guarantee in Proposition 7.1.5 to that in Proposition 7.1.6. Let $X_i \stackrel{\rm iid}{\sim} \mathsf{N}(0, \Sigma)$ for some covariance $\Sigma \succeq 0$.

(a) Give the smallest value σ^2 for which for all $u \in \mathbb{R}^d$,

$$\|\langle X, u \rangle\|_{\psi_2} \le \sigma \|u\|_2.$$

(b) Give the smallest value κ^2 for which for all $u \in \mathbb{R}^d$,

$$\|\langle X, u \rangle\|_{\psi_2} \le \kappa \sqrt{u^\top \Sigma u}.$$

(c) Compare the guarantees Propositions 7.1.5 and 7.1.6 provide on the deviations $P_n X X^\top - \Sigma$. How far apart can they be?

Exercise 7.8 (An error lower bound)**:** Let P_0 and P_1 be distributions on $\mathcal{X}$, where $\mathcal{X}$ is finite. Let $\Psi_n : \mathcal{X}^n \to \{0, 1\}$ be a sequence of tests with small enough type-II error that

$$\gamma_n := \frac{1}{n} \log \frac{1}{P_1^n(\Psi_n = 0)} - D_{\rm kl}\left(P_0 \| P_1\right) \geq 0$$

for all large enough n. Show that there is a numerical constant $c > 0$ such that for all large enough n, the type-I error $\epsilon := P_0^n(\Psi_n = 1)$ satisfies

$$\epsilon \geq 1 - n \exp\left(-c \frac{n\gamma_n^2}{\log^2 n}\right).$$

Exercise 7.9: Define the *optimal upper sub-Gaussian constant* of a distribution P by

$$\sigma_+^2(P) := \inf\left\{\sigma^2 \mid \mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq \sqrt{2\sigma^2 D_{\rm kl}\left(Q\|P\right)} \text{ for all } Q \text{ s.t. } D_{\rm kl}\left(Q\|P\right) < \infty\right\}.$$

This question demonstrates the Kearns-Saul concentration guarantee [266, 333, 59] for Bernoulli random variables that if $X \sim \mathsf{Bernoulli}(p)$, then

$$\sigma_+^2(p) = \begin{cases} \frac{1-2p}{2\log\frac{1-p}{p}} & \text{if } p < \frac{1}{2} \\ p(1-p) & \text{if } p \geq \frac{1}{2}. \end{cases} \tag{7.5.3}$$

Define the binary divergence $d(q) = D_{\rm kl}\left(q\|p\right) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}$, and recognize that

$$r(q) = r_p(q) := \frac{(q-p)^2}{2d(q)} \quad \text{satisfies} \quad \sigma_+^2(p) = \sup_{q > p} r(q).$$

(a) Use Taylor's theorem to show that

$$d(q) = \int_p^q \frac{q-t}{t(1-t)} dt.$$

(b) Show that for $q > p$, that

$$r'(q) \text{ has the same sign as } s(q) := 2d(q) - (q-p)d'(q) = \int_p^q \frac{p+q-2t}{t(1-t)} dt.$$

(c) Show that if $p < \frac{1}{2}$, then for $q > p$, $s(q) < 0$ if and only if $q < 1 - p$, and $s(q) > 0$ if and only if $q > 1 - p$, and use this to conclude that $q^\star = 1 - p$ maximizes the ratio $r(q)$.

(d) Prove the claim (7.5.3).

Exercise 7.10 ([351], Remark 3.4.6)**:** Use inequality (7.5.3) to show that if $X \sim \mathsf{Bernoulli}(p)$, then X is $\sigma_\star^2(p) = \frac{1-2p}{2\log\frac{1-p}{p}}$-sub-Gaussian (where $\sigma_\star^2(0) = \sigma_\star^2(1) = 0$, and $\sigma_\star^2(\frac{1}{2}) = \frac{1}{4}$).

Chapter 8

Privacy and disclosure limitation

The previous chapters used stability and divergence bounds to control generalization and concentration, and we continue our development of stability ideas to show how they allow us to provide a new type of protection: privacy for study participants. Until the mid-2000s, the major challenge in this direction had been a satisfactory definition of privacy, because collection of side information often results in unforeseen compromises of private information. The introduction of *differential privacy*, which imposes a stability guarantee in the form of worst-case bounds on likelihood ratios to make neighboring datasets nearly indistinguishable, alleviated these challenges, providing a firm foundation on which to build private estimators. These ideas extend classical survey sampling mechanisms tracing to the 1960s, including Warner's *randomized response* [427], to provide a calculus of privacy. Consequently, we focus on privacy notions based on differential privacy and its cousins, developing the information-theoretic stability ideas helpful to understand these techniques.

8.1 Disclosure limitation, privacy, and definitions

We begin with a few cautionary tales and examples. A natural belief might be that, given only summary statistics of a large dataset, individuals in the data remain private. This is false. As an example, in 2008, Homer and colleagues [240] showed that even releasing aggregated genetic frequency statistics (e.g., frequency of single nucleotide polymorphisms (SNP) in microarrays) can allow resolution of individuals within a database. Consequently, the US National Institutes of Health (NIH), the Wellcome Trust, and the Broad Institute removed genetic summaries from public access (along with imposing stricter requirements for private access) [381, 117]. Linkage attacks, where a third party combines information from two or more databases to de-identify individuals, raise other issues; for example, a database containing item purchase times could be combined with a database of times individuals are online to identify the items individuals buy.

Dalenius [127] provides perhaps the strongest definition of data privacy, writing "nothing about an individual should be learnable from the database that cannot be learned without access to the database." Unfortunately, it is essentially impossible to reconcile this idea with scientific progress. Consider a study on smoking, which discovers that smoking causes cancer. In publishing the result, we "compromise" the privacy of everyone who smokes: we know they are more likely to get cancer.

In each of these cases, the biggest challenge is one of side information: how can we be sure that, when releasing a particular statistic, no adversary will be able to infer sensitive data about participants in our study? We articulate three conditions that researchers believe suffice for privacy, developing some vocabulary along the way. We term a (randomized) algorithm releasing data either

a privacy *mechanism*, consistent with much of the literature in privacy, or a *channel*, in keeping with our information-theoretic focus. We wish our privacy mechanism, which takes as input a sample $X_1^n \in \mathcal{X}^n$ and releases some Z, to satisfy the following desiderata:

D.1. If an individual participates in multiple private studies, there should be graceful degradation in the privacy protections rather than a catastrophic failure.

D.2. Further processing of the output Z of a private mechanism $X_1^n \to Z$, in the form of the Markov chain $X_1^n \to Z \to Y$, should not further compromise privacy.

D.3. The mechanism $X_1^n \to Z$ should be resilient to side information: even if the adversary knows something about an individual x, the adversary should learn little about whether x belongs to X_1^n. This should remain true even if the adversary *later* obtains more information about the individual. Operationally, given the output Z, an adversary knowing everyone in the study (excepting one individual) should be unable to test whether the individual belongs to the study.

In desideratum D.1, one should see echoes of Section 6.3's adaptive data analysis. In desideratum D.2, we see echoes of data processing inequalities. A Bayesian perspective, which we discuss in Section 8.1.2, will allow an elegant interpretation of the third desideratum D.3.

Before addressing these challenges, we establish notation. We consider randomized procedures acting on samples x_1^n, which we treat as equivalent to the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$. For two empirical distributions P_n and P_n' supported on $\{x_1, \ldots, x_n\}$ and $\{x_1', \ldots, x_n'\}$,

$$n \left\|P_n - P_n'\right\|_{\rm TV} = d_{\rm ham}(\{x_1, \ldots, x_n\}, \{x_1', \ldots, x_n'\}),$$

where the Hamming distance treats samples as multisets and ignores ordering. With this notational convenience in place, we take

$$\mathcal{P}_n = \left\{P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i} \mid x_i \in \mathcal{X}\right\}$$

as the set of n-point empirical distributions in $\mathcal{X}$, and we abuse notation to define $d_{\rm ham}(P_n, P_n') := n\left\|P_n - P_n'\right\|_{\rm TV}$ as the number of differing observations. A mechanism M is then a randomized mapping $M : \mathcal{P}_n \to \mathcal{Z}$, which we can identify with its induced Markov channel Q from $\mathcal{X}^n \to \mathcal{Z}$; we use the equivalent views as convenient.

The challenges of side information motivate Dwork et al.'s definition of *differential privacy* [170]. The key in differential privacy is that the noisy channel releasing statistics guarantees bounded likelihood ratio between *neighboring samples*, which differ in only a single entry.

Definition 8.1 (Differential privacy)**.** *Let $M : \mathcal{P}_n \to \mathcal{Z}$ be a randomized mapping. Then M is* ε-differentially private *if for all (measurable) sets $S \subset \mathcal{Z}$ and all neighboring $P_n, P_n' \in \mathcal{P}_n$,*

$$\frac{\mathbb{P}(M(P_n) \in S)}{\mathbb{P}(M(P_n') \in S)} \le e^\varepsilon. \tag{8.1.1}$$

With this definition, an individual has little incentive to not participate in a study, as the individual's data has limited effect on the outcome.

The model (8.1.1) of differential privacy presumes a trusted curator, such as a hospital or researcher, who can collect all the data into a centralized location, so we call it the *centralized* model. The *local model* of privacy, in which data providers trust no one, not even the data collector, and privatize their individual data before the collector even sees it, is stronger:

Definition 8.2 (Local differential privacy)**.** *A channel Q from $\mathcal{X}$ to $\mathcal{Z}$ is* ε-locally differentially private *if for all measurable $S \subset \mathcal{Z}$ and all $x, x' \in \mathcal{X}$,*

$$\frac{Q(Z \in S \mid x)}{Q(Z \in S \mid x')} \le e^{\varepsilon}. \tag{8.1.2}$$

Definition 8.2 and the condition (8.1.2) are stronger than Definition 8.1: when samples $\{x_1, \ldots, x_n\}$ and $\{x_1', \ldots, x_n'\}$ differ in at most one observation, the local model (8.1.2) with independent randomization guarantees that the densities

$$\frac{dQ(Z_1^n \mid \{x_i\})}{dQ(Z_1^n \mid \{x_i'\})} = \prod_{i=1}^n \frac{dQ(Z_i \mid x_i)}{dQ(Z_i \mid x_i')} \le e^{\varepsilon},$$

where the inequality follows because only a single ratio may contain $x_i \neq x_i'$.

In the remainder of this introductory section, we provide a few of the basic mechanisms in differential privacy, then discuss the semantic connections to the three desiderata D.1–D.3. In the coming sections, we visit more advanced topics, in particular, the composition of multiple private mechanisms, weakenings of differential privacy, and more sophisticated examples.

8.1.1 Basic mechanisms

The basic mechanisms in differential privacy use noise addition to ensure privacy. We begin with Warner's classical randomized response for local privacy.

Example 8.1.1 (Randomized response)**:** We wish study participants to answer a yes/no question about a sensitive topic to estimate the proportion of the population with a characteristic. Call this group 1 and its complement 0. To provide privacy, we give them a spinner with a face painted into regions 1 and 0 of relative area $\frac{e^\varepsilon}{1+e^\varepsilon}$ and $\frac{1}{1+e^\varepsilon}$, respectively. When the participant spins the spinner, it lands in region 1 with probability $e^\varepsilon/(1+e^\varepsilon)$, and the participant answers "Yes" if they belong to the indicated group, "No" otherwise.

This randomized response mechanism provides ε-local differential privacy: we have

$$\frac{Q(\text{Yes} \mid x = 1)}{Q(\text{Yes} \mid x = 0)} = e^{\varepsilon} \quad \text{and} \quad \frac{Q(\text{No} \mid x = 0)}{Q(\text{No} \mid x = 1)} = e^{\varepsilon},$$

so that $Q(Z = z \mid x)/Q(Z = z \mid x') \in [e^{-\varepsilon}, e^{\varepsilon}]$ for all x, z. $\diamond$

We can use randomized response to estimate population proportions. Indeed, for $x \in \{0, 1\}$, let

$$Z = \begin{cases} x & \text{with probability } q_\varepsilon \\ 1 - x & \text{otherwise} \end{cases} \quad \text{where } q_\varepsilon = \frac{e^\varepsilon}{1 + e^\varepsilon}. \tag{8.1.3}$$

Then $\mathbb{E}[Z \mid x] = (2q_\varepsilon - 1)x + (1 - q_\varepsilon)$, so the debiased quantity $\widehat{x} = \frac{1}{2q_\varepsilon - 1}(z - (1 - q_\varepsilon))$ satisfies

$$\mathbb{E}[\widehat{x} \mid x] = x \quad \text{and} \quad \mathrm{Var}(\widehat{x}) \le \frac{1}{4(2q_\varepsilon - 1)^2} = \frac{1}{4}\left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)^2.$$

Assuming $X_i \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(p)$, given n observations Z_i drawn (conditionally on X_i) from randomized response, the estimate $\widehat{p} := \frac{1}{2q_\varepsilon - 1}(\overline{Z}_n - (1 - q_\varepsilon))$ satisfies

$$\mathbb{E}\left[(\widehat{p} - p)^2\right] \le O(1)\frac{1}{\varepsilon^2 \wedge 1} \cdot \frac{1}{n}.$$

Exercise 8.1 extends this to estimating multinomial (categorical) distributions.

While randomized response is essentially standard in locally private settings, in centralized privacy, adding Laplace noise is the default. Suppose we wish to release a function $f(P_n) \in \mathbb{R}^d$ of the sample P_n. If f is Lipschitz with respect to the Hamming metric, it is relatively straightforward to develop private mechanisms. To better reflect the nomenclature in the privacy literature and for easier use in our future development, for $p \in [1, \infty]$ we define the *ℓ_p-global sensitivity* of f by

$$\mathrm{GS}_p(f) := \sup_{P_n, P_n' \in \mathcal{P}_n} \left\{ \left\| f(P_n) - f(P_n') \right\|_p \mid d_{\rm ham}(P_n, P_n') \le 1 \right\},$$

which is the Lipschitz constant of f with respect to the Hamming metric. Adding noise scaling with the global sensitivity is then sufficient to hide any one individual's contributions to $f(P_n)$.

Example 8.1.2 (Laplace mechanisms)**:** The $\mathsf{Laplace}(\beta)$ distribution, $\beta > 0$, has density

$$p(w) = \frac{1}{2\beta} \exp(-|w|/\beta),$$

on $\mathbb{R}$. If $W \sim \mathsf{Laplace}(\beta)$, then $\mathbb{E}[W] = 0$ by symmetry, while $\mathbb{E}[W^2] = \frac{1}{\beta} \int_0^\infty w^2 e^{-w/\beta} = 2\beta^2$.

Let $f : \mathcal{P}_n \to \mathbb{R}^d$ have finite ℓ_1-global sensitivity

$$\mathrm{GS}_1(f) = \sup \left\{ \left\| f(P_n) - f(P_n') \right\|_1 \mid d_{\rm ham}(P_n, P_n') \le 1, P_n, P_n' \in \mathcal{P}_n \right\}.$$

The *Laplace mechanism* releases

$$Z = M(P_n) := f(P_n) + \frac{\mathrm{GS}_1(f)}{\varepsilon} W, \quad W_j \stackrel{\rm iid}{\sim} \mathsf{Laplace}(1). \tag{8.1.4}$$

Its ε-differential privacy is immediate: for neighboring samples P_n, P_n', Z has density ratio

$$\frac{q(z \mid P_n)}{q(z \mid P_n')} = \exp\left(\frac{\varepsilon(\|f(P_n') - z\|_1 - \|f(P_n) - z\|_1)}{\mathrm{GS}_1(f)} \right) \le \exp\left(\frac{\varepsilon \|f(P_n) - f(P_n')\|_1}{\mathrm{GS}_1(f)} \right) \le \exp(\varepsilon)$$

by the triangle inequality and definition of the global sensitivity. Moreover, we have

$$\mathbb{E}[\|Z - f(P_n)\|_2^2] = \frac{2d \, \mathrm{GS}_1(f)^2}{\varepsilon^2},$$

so that if the global sensitivity $\mathrm{GS}_1(f)$ is small, we may report the value of f accurately. $\diamond$

Mean and histogram estimation comprise the most common applications of the Laplace mechanism, and we typically view privacy as having little cost for estimating a statistic $f(P_n)$ when the "statistical" error dominates the privacy error, that is, when $\mathrm{GS}_1^2(f) \ll \mathrm{Var}(f(P_n))$. We demonstrate two more worked examples.

Example 8.1.3 (Private one-dimensional mean estimation)**:** Let the variables X_i take values in $[-b, b]$ for some $b < \infty$; we wish to estimate $\mathbb{E}[X]$. The sample mean $f(P_n) = \mathbb{E}_{P_n}[X] = \overline{X}_n$ has Lipschitz constant $2b/n$ with respect to the Hamming metric (i.e., $\mathrm{GS}_1(f) = 2b/n$) because for any two samples $x, x' \in [-b, b]^n$ differing in only entry i,

$$|f(x) - f(x')| = \frac{1}{n} |x_i - x_i'| \le \frac{2b}{n}.$$

The Laplace mechanism (8.1.4) thus yields

$$\mathbb{E}[(Z - \mathbb{E}[X])^2] = \mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] + \mathbb{E}[(Z - \overline{X}_n)^2] = \frac{1}{n}\mathrm{Var}(X) + \frac{8b^2}{n^2\varepsilon^2} \le \frac{b^2}{n} + \frac{8b^2}{n^2\varepsilon^2},$$

and the sample mean variance dominates the privacy penalty $\frac{b^2}{n^2\varepsilon^2}$ so long as $\varepsilon \gg n^{-1/2}$. $\diamond$

Example 8.1.4 (Private histogram (multinomial) release)**:** We wish to estimate a multinomial distribution, or, put differently, a histogram, so for observations $X \in \{1, \ldots, k\}$, where k may be large, we wish to estimate $p_j := \mathbb{P}(X = j)$ for $j = 1, \ldots, k$. For a sample x_1^n, the empirical count vector $\widehat{p}_n$ with coordinates $\widehat{p}_{n,j} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i = j\}$ satisfies

$$\mathrm{GS}_1(\widehat{p}_n) = \frac{2}{n}$$

because swapping a single example x_i for x_i' changes the counts for at most two coordinates j, j' by 1. Consequently, the Laplace noise addition mechanism

$$Z = M(P_n) := \widehat{p}_n + \frac{2}{n\varepsilon}W, \quad W_j \stackrel{\mathrm{iid}}{\sim} \mathsf{Laplace}(1)$$

is ε-differentially private and satisfies $\mathbb{E}[\|Z - \widehat{p}_n\|_2^2] = \frac{8k}{n^2\varepsilon^2}$, so

$$\mathbb{E}[\|Z - p\|_2^2] = \frac{8k}{n^2\varepsilon^2} + \frac{1}{n}\sum_{j=1}^k p_j(1 - p_j) \le \frac{8k}{n^2\varepsilon^2} + \frac{1}{n},$$

which tends to 0 as $n \to \infty$ and $k/n^2 \to 0$. $\diamond$

Example 8.1.4 shows one of the challenges of differentially private mechanisms: even in the case where the quantity of interest is quite stable, the resulting mechanism adds noise that introduces dimension-dependent scaling. In this case, we require $\varepsilon \gg \sqrt{k/n}$ to admit the non-private rate of convergence $\mathbb{E}[\|\widehat{p}_n - p\|_2^2] = \frac{1}{n}\sum_{j=1}^k p_j(1 - p_j) \le \frac{1}{n}$. Thus, when the number k of bins is large, the naive noise addition strategy cannot provide as much protection without sacrificing efficiency.

Example 8.1.5 (Example 8.1.4 continued, ℓ_∞ error)**:** Because $\mathbb{P}(\|W\|_\infty \ge t) \le k\exp(-t/b)$ for $W_j \stackrel{\mathrm{iid}}{\sim} \mathsf{Laplace}(b)$, the Laplace mechanism satisfies

$$\mathbb{P}(\|M(P_n) - \widehat{p}_n\|_\infty \ge t) \le k\exp\left(-\frac{tn\varepsilon}{2}\right) \quad \text{all } t \ge 0.$$

Using that each coordinate of $\widehat{p}_n$ is 1-sub-Gaussian, we have

$$\begin{aligned}\mathbb{E}[\|M(P_n) - p\|_\infty] \le \mathbb{E}[\|\widehat{p}_n - p\|_\infty] + \mathbb{E}[\|W\|_\infty] &\le \sqrt{\frac{2\log k}{n}} + \inf_{t \ge 0}\left\{t + \frac{2k}{n\varepsilon}\exp\left(-\frac{tn\varepsilon}{2}\right)\right\} \\ &\le \sqrt{\frac{2\log k}{n}} + \frac{2\log k}{n\varepsilon} + \frac{2}{n\varepsilon}.\end{aligned}$$

In this case, when $\varepsilon \gg (n/\log k)^{-1/2}$, we obtain expected error at worst $\sqrt{2\log k/n}$, which is somewhat more satisfying than the k-dependent penalty in Example 8.1.4. $\diamond$

8.1.2 Resilience to side information, Bayesian perspectives, and data processing

A major challenge in the definition of privacy is to protect against side information—as in point D.3 in our introduction—especially because in the future, other information about an individual may be compromised, allowing linkage attacks. With this in mind, we show how differential privacy addresses all three desiderata D.1–D.3. First, we note the following simple fact: if Z is a differentially private view of a sample P_n, then any downstream functions Y are also differentially private. That is, given the Markov chain $P_n \to Z \to Y$, for any neighboring $P_n, P_n' \in \mathcal{P}_n$ and set A,

$$\frac{\mathbb{P}(Y \in A \mid P_n)}{\mathbb{P}(Y \in A \mid P_n')} = \frac{\int P(Y \in A \mid z) q(z \mid P_n) d\mu(z)}{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)} \le e^\varepsilon \frac{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)}{\int P(Y \in A \mid z) q(z \mid P_n') d\mu(z)} = e^\varepsilon.$$

So post-processing (point D.2) cannot reduce privacy.

A Bayesian perspective allows us to more carefully consider side information. Consider the following thought experiment. An adversary has prior beliefs π on $\mathcal{X}^n$, and we consider the adversary's posterior $\pi(\cdot \mid Z)$ induced by observing the output Z of some mechanism M. *Bayes factors* measure how much prior and posterior distributions differ and connect immediately to privacy.

Proposition 8.1.6. *A mechanism $M : \mathcal{P}_n \to \mathcal{Z}$ is ε-differentially private if and only if for any prior distribution π on $\mathcal{P}_n$ and any observation $z \in \mathcal{Z}$, the posterior odds satisfy*

$$\frac{\pi(P_n \mid z)}{\pi(P_n' \mid z)} \le e^\varepsilon \frac{\pi(P_n)}{\pi(P_n')}$$

for all $P_n, P_n' \in \mathcal{P}_n$ with $d_{\rm ham}(P_n, P_n') \le 1$.

For example, consider the "most informed" adversary, with prior supported equally and only on two samples P_n and P_n' differing in a single element. Then $\frac{1}{1+e^\varepsilon} \le \pi(P_n \mid z) \le \frac{e^\varepsilon}{1+e^\varepsilon}$ for all z. We see that private channels guarantee that prior and posterior odds between two neighboring samples cannot change substantially, no matter the observation Z.

We can provide an alternative interpretation by considering testing-based variants of privacy. In this case, an adversary wishes to test the hypotheses

$$H_0 : X_1^n = x_1^n \quad \text{versus} \quad H_1 : X_1^n = (x_1^{i-1}, x_i', x_{i+1}^n),$$

so the samples under H_0 and H_1 differ only in the ith observation $X_i \in \{x_i, x_i'\}$. For a channel taking inputs in $\mathcal{X}^n$ and outputting $Z \in \mathcal{Z}$, we define *ε-conditional hypothesis testing privacy* by

$$\inf_\Psi \left\{Q(\Psi(Z) = 1 \mid H_0, Z \in A) + Q(\Psi(Z) = 0 \mid H_1, Z \in A)\right\} \ge e^{-\varepsilon} \tag{8.1.5}$$

for all sets $A \subset \mathcal{Z}$ satisfying $Q(A \mid H_0) > 0$ and $Q(A \mid H_1) > 0$. Roughly, no matter the value Z takes, the probability of error in a test of H_0 against H_1 is high. We have the following proposition.

Proposition 8.1.7. *Assume the channel Q is ε-differentially private. Then Q is also 2ε-conditional hypothesis testing private.*

Proof Let Ψ test H_0 versus H_1, and let $B = \{z \mid \Psi(z) = 1\}$ be its acceptance region. Then

$$\begin{aligned}
Q(B \mid H_0, Z \in A) + Q(B^c \mid H_1, Z \in A) &= \frac{Q(A, B \mid H_0)}{Q(A \mid H_0)} + \frac{Q(A, B^c \mid H_1)}{Q(A \mid H_1)} \\
&\ge e^{-2\varepsilon} \frac{Q(A, B \mid H_1)}{Q(A \mid H_1)} + \frac{Q(A, B^c \mid H_1)}{Q(A \mid H_1)} \\
&\ge e^{-2\varepsilon} \frac{Q(A, B \mid H_1) + Q(A, B^c \mid H_1)}{Q(A \mid H_1)},
\end{aligned}$$

by ε-differential privacy. Note that $Q(A, B \mid H_1) + Q(A, B^c \mid H_1) = Q(A \mid H_1)$. □

So we see that (roughly), even conditional on the output of the channel, we still cannot test whether the initial dataset was x or x' whenever x, x' differ in only a single observation. (Desideratum D.3.)

8.2 Weakenings of differential privacy

Sometimes, satisfying differential privacy requires the addition of more noise to a desired statistic than is practical. Moreover, the privacy considerations interact in different ways with geometry: as we saw in Example 8.1.4, the Laplace mechanism adds noise that introduces dimension-dependent scaling, which we discuss more in Example 8.2.9. Consequently, we wish to develop weaker notions that—at least hopefully—still provide satisfactory privacy protections. To that end, we develop two additional privacy notions that allow more sophisticated, lower-noise mechanisms than standard differential privacy; their protections are necessarily somewhat weaker.

We begin with a definition that allows (very rare) catastrophic privacy breaches—as long as the probability of this event is small, these may be acceptable.

Definition 8.3. *Let $\varepsilon, \delta \geq 0$. A mechanism $M : \mathcal{P}_n \to \mathcal{Z}$ is* (ε, δ)-differentially private *if for all (measurable) sets $S \subset \mathcal{Z}$ and neighboring samples P_n, P_n',*

$$\mathbb{P}(M(P_n) \in S) \leq e^{\varepsilon} \mathbb{P}(M(P_n') \in S) + \delta. \tag{8.2.1}$$

One typically thinks of δ in the definition above as satisfying $\delta = \delta_n$, where $\delta_n \ll n^{-k}$ for some $k \geq 2$. Some practitioners contend that real-world differentially private algorithms are all (ε, δ)-differentially private: even using cryptographically secure random number generators, there is some possibility (say δ) that a cryptographic key may leak or an encoding may be broken in the future.

An alternative definition of privacy reposes on Rényi divergences. These are monotonically transformed f divergences (recall Section 2.2) with a structure more amenable to analysis.

Definition 8.4. *Let P and Q be distributions on a space $\mathcal{X}$ with densities p and q (with respect to a measure μ). For $\alpha \in [0, \infty]$, the* Rényi-α-divergence *between P and Q is*

$$D_\alpha(P \| Q) := \frac{1}{\alpha - 1} \log \int \left(\frac{p(x)}{q(x)} \right)^\alpha q(x) d\mu(x).$$

We define values of $D_\alpha(P\|Q)$ for $\alpha \in \{0, 1, \infty\}$ in terms of their limits (see Exercise 8.2):

Proposition 8.2.1 (Basic facts on Rényi divergence)**.** *Rényi divergences satisfy the following.*

i. The divergence $D_\alpha(P\|Q)$ is non-decreasing in α.

ii. The Rényi divergence has limits

$$\begin{aligned} D_0(P\|Q) &= \lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = -\log Q(p(X) > 0) \\ D_1(P\|Q) &= \lim_{\alpha \to 1} D_\alpha(P\|Q) = D_{\rm kl}(P\|Q) \\ D_\infty(P\|Q) &= \lim_{\alpha \uparrow \infty} D_\alpha(P\|Q) = \sup_t \{\log t \mid Q(p(X)/q(X) \geq t) > 0\}. \end{aligned}$$

iii. Let $K(\cdot \mid x)$ be a Markov kernel from $\mathcal{X} \to \mathcal{Z}$ as in Proposition 2.2.13, and let K_P and K_Q be the induced marginals of P and Q under K, respectively. Then $D_\alpha(K_P\|K_Q) \leq D_\alpha(P\|Q)$.

Rényi divergences also admit generalizations of the tensorization properties of KL-divergence and information of Chapter 2 (recall the chain rule (2.1.6) for KL-divergence), and we return to this later. As a preview, however, these tensorization properties allow us to prove that the composition of multiple private data releases remains appropriately private.

With these preliminaries in place, we can then provide

Definition 8.5 (Rényi-differential privacy)**.** *Let $\varepsilon \geq 0$ and $\alpha \in [1, \infty]$. A channel Q from $\mathcal{P}_n$ to output space $\mathcal{Z}$ is* (α, ε)-Rényi private *if for all neighboring samples $P_n, P_n' \in \mathcal{P}_n$,*

$$D_\alpha\big(Q(\cdot \mid P_n) \| Q(\cdot \mid P_n')\big) \leq \varepsilon. \tag{8.2.2}$$

Clearly, any ε-differentially private channel is also (α, ε)-Rényi private for any $\alpha \geq 1$.

8.2.1 Basic mechanisms

We describe basic mechanisms that provide (ε, δ)-differential privacy and (α, ε)-Rényi privacy. These settings allow mechanisms that more naturally handle vectors in ℓ_2, and smoothness with respect to Euclidean norms, than with respect to ℓ_1, which is most natural for pure ε-differential privacy. A starting point is the following example, which we will leverage frequently.

Example 8.2.2 (Rényi divergence between Gaussian distributions)**:** Consider normal distributions $\mathsf{N}(\mu_0, \Sigma)$ and $\mathsf{N}(\mu_1, \Sigma)$. Then

$$D_\alpha(\mathsf{N}(\mu_0, \Sigma) \| \mathsf{N}(\mu_1, \Sigma)) = \frac{\alpha}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1). \tag{8.2.3}$$

To see this equality, let p and q be the densities of $\mathsf{N}(\mu_0, \Sigma)$ and $\mathsf{N}(\mu_1, \Sigma)$, respectively. Then letting $\mathbb{E}_{\mu_1}$ denote expectation over $X \sim \mathsf{N}(\mu_1, \Sigma)$, we have

$$\begin{aligned}
\int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x) dx &= \mathbb{E}_{\mu_1}\left[\exp\left(-\frac{\alpha}{2}(X - \mu_0)^\top \Sigma^{-1}(X - \mu_0) + \frac{\alpha}{2}(X - \mu_1)^\top \Sigma^{-1}(X - \mu_1)\right)\right] \\
&= \mathbb{E}_{\mu_1}\left[\exp\left(-\frac{\alpha}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1) + \alpha(\mu_0 - \mu_1)^\top \Sigma^{-1}(X - \mu_1)\right)\right] \\
&\overset{(\star)}{=} \exp\left(-\frac{\alpha}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1) + \frac{\alpha^2}{2}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)\right),
\end{aligned}$$

where equality $(\star)$ follows because $(\mu_0 - \mu_1)^\top \Sigma^{-1}(X - \mu_1) \sim \mathsf{N}(0, (\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 - \mu_0))$ under $X \sim \mathsf{N}(\mu_1, \Sigma)$. Using $-\alpha + \alpha^2 = \alpha(\alpha - 1)$ and taking logarithms gives the result. $\diamond$

Example 8.2.2 is the key to developing different privacy-preserving schemes under Rényi privacy. Let us reconsider the Laplace mechanism in Example 8.1.2, except that instead of assuming the function f is smooth with respect to ℓ_1 norm, we use the ℓ_2-norm:

Example 8.2.3 (Gaussian mechanisms)**:** Let $f : \mathcal{P}_n \to \mathbb{R}^d$ have global ℓ_2-sensitivity

$$L := \mathrm{GS}_2(f) = \sup\left\{\left\|f(P_n) - f(P_n')\right\|_2 \mid d_{\mathrm{ham}}(P_n, P_n') \leq 1\right\}.$$

Then, for any variance $\sigma^2 > 0$, the mechanism

$$Z = M(P_n) := f(P_n) + W, \quad W \sim \mathsf{N}(0, \sigma^2 I)$$

satisfies

$$D_\alpha\big(\mathsf{N}(f(P_n), \sigma^2 I) \| \mathsf{N}(f(P_n'), \sigma^2 I)\big) = \frac{\alpha}{2\sigma^2} \left\|f(P_n) - f(P_n')\right\|_2^2 \le \frac{\alpha}{2\sigma^2} L^2$$

for neighboring samples P_n, P_n'. To achieve (α, ε)-Rényi privacy, we take $\sigma^2 = \frac{L^2\alpha}{2\varepsilon}$, and the mechanism

$$Z = M(P_n) = f(P_n) + \mathrm{GS}_2(f)\sqrt{\frac{\alpha}{2\varepsilon}} W \quad \text{for} \quad W \sim \mathsf{N}(0, I_d) \tag{8.2.4}$$

satisfies (α, ε)-Rényi privacy. $\diamond$

Special cases can make this more concrete. The simplest arises for bounded mean estimation:

Example 8.2.4 (Mean estimation with Gaussian mechanisms)**:** Let X_i be i.i.d. and satisfy $\|X_i\|_2 \le r$, and consider estimating the mean $\mathbb{E}[X]$ via the sample mean $f(P_n) = \overline{X}_n$. We see immediately that

$$\mathrm{GS}_2(f) = \frac{2r}{n},$$

so the Gaussian mechanism (8.2.4) yields $\mathbb{E}[\|Z - \overline{X}_n\|_2^2] = \mathbb{E}[\|W\|_2^2] = \frac{2dr^2\alpha}{n^2\varepsilon}$. Then

$$\mathbb{E}[\|Z - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\|\overline{X}_n - \mathbb{E}[X]\|_2^2] + \mathbb{E}[\|Z - \overline{X}_n\|_2^2] \le \frac{r^2}{n} + \frac{2dr^2\alpha}{n^2\varepsilon}.$$

We will return to compare this quantity to the Laplace mechanism's performance in Example 8.1.2 once we develop more connections between privacy definitions. $\diamond$

8.2.2 Connections between privacy measures

An important consideration in our development of privacy definitions and mechanisms is to understand their relationships. We thus collect a few consequences of the definitions.

First, we see that ε-differential privacy implies stronger values of Rényi-differential privacy.

Proposition 8.2.5. *Let $\varepsilon \ge 0$ and let P and Q be distributions such that $e^{-\varepsilon} \le P(A)/Q(A) \le e^\varepsilon$ for all measurable sets A. Then for any $\alpha \in [1, \infty]$,*

$$D_\alpha(P\|Q) \le \min\left\{\frac{3\alpha}{2}\varepsilon^2, \varepsilon\right\}.$$

As an immediate corollary, we have

Corollary 8.2.6. *Let $\varepsilon \ge 0$ and assume that Q is ε-differentially private. Then for any $\alpha \ge 1$, Q is $(\alpha, \min\{\frac{3\alpha}{2}\varepsilon^2, \varepsilon\})$-Rényi private.*

Before proving the proposition, let us see its implications for Example 8.2.4 versus estimation under ε-differential privacy. Let $\varepsilon \le 1$, so to have "similar" privacy, we require that Rényi private channels satisfy $D_\alpha(Q(\cdot \mid x)\|Q(\cdot \mid x')) \le \varepsilon^2$. The ℓ_1-sensitivity of the mean satisfies $\|\overline{x}_n - \overline{x}_n'\|_1 \le \sqrt{d}\|\overline{x}_n - \overline{x}_n'\|_2 \le 2\sqrt{d}r/n$ for neighboring samples. Then the Laplace mechanism (8.1.4) satisfies

$$\mathbb{E}[\|Z_{\text{Laplace}} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + \frac{8r^2}{n^2\varepsilon^2} \cdot d^2,$$

while the Gaussian mechanism under (α, ε^2)-Rényi privacy will yield

$$\mathbb{E}[\|Z_{\text{Gauss}} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + \frac{2r^2}{n^2\varepsilon^2} \cdot d\alpha.$$

This improves the Laplace mechanism whenever $\alpha < d$.

Proof of Proposition 8.2.5 We assume w.l.o.g. that P and Q have densities p and q with respect to a base measure μ, whence $e^{-\varepsilon} \le p/q \le e^{\varepsilon}$ and $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int (p/q)^\alpha q d\mu$. We prove the result assuming that $\alpha \in (1,\infty)$, using continuity to extend to $\alpha \in [1,\infty]$.

First, it is clear that $D_\alpha(P\|Q) \le \varepsilon$ always. For the other term in the minimum, assume that $\alpha \le 1 + \frac{1}{\varepsilon}$ and $\varepsilon \le 1$. If either of these fails, the result is trivial, because for $\alpha > 1 + \frac{1}{\varepsilon}$ we have $\frac{3}{2}\alpha\varepsilon^2 \ge \frac{3}{2}\varepsilon \ge \varepsilon$, and similarly $\varepsilon \ge 1$ implies $\frac{3}{2}\alpha\varepsilon^2 \ge \varepsilon$.

By Taylor's theorem, for any $t > -1$ there is a $\widetilde{t}$ between 0 and t for which

$$(1+t)^\alpha = 1 + \alpha t + \frac{\alpha(\alpha-1)}{2}(1+\widetilde{t})^{\alpha-2}t^2.$$

If $1 + t \le c$, then $(1+t)^\alpha \le 1 + \alpha t + \frac{\alpha(\alpha-1)}{2}\max\{1, c^{\alpha-2}\}t^2$, so

$$\begin{aligned}
&\exp\left((\alpha-1)D_\alpha(P\|Q)\right) = \int \left(\frac{p(z)}{q(z)}\right)^\alpha q(z)d\mu(z) = \int \left(1 + \frac{p(z)}{q(z)} - 1\right)^\alpha q(z)d\mu(z) \\
&\le 1 + \alpha\int\left(\frac{p(z)}{q(z)} - 1\right)q(z)d\mu(z) + \frac{\alpha(\alpha-1)}{2}\max\{1, \exp(\varepsilon(\alpha-2))\}\int\left(\frac{p(z)}{q(z)} - 1\right)^2 q(z)d\mu(z) \\
&\le 1 + \frac{\alpha(\alpha-1)}{2}e^{\varepsilon[\alpha-2]_+}\cdot(e^\varepsilon - 1)^2.
\end{aligned}$$

By assumption, $\alpha - 2 \le 1/\varepsilon - 1$, so using that $\log(1+x) \le x$, we obtain

$$D_\alpha(P\|Q) \le \frac{\alpha}{2}(e^\varepsilon - 1)^2 \cdot \exp([1-\varepsilon]_+).$$

A numerical calculation yields that this quantity is at most $\frac{3\alpha}{2}\varepsilon^2$ for $\varepsilon \le 1$. □

We can also connect (α,ε)-Rényi privacy to (ε,δ)-differential privacy, and that to ε-differential privacy. To do so, we observe that whenever distributions P and Q have small Rényi divergence, only limited probability amplification is possible in moving from one to the other.

Proposition 8.2.7. *Let P and Q satisfy $D_\alpha(P\|Q) \le \varepsilon$. Then for any (measurable) set A,*

$$P(A) \le \exp\left(\frac{\alpha-1}{\alpha}\varepsilon\right)Q(A)^{\frac{\alpha-1}{\alpha}}.$$

Consequently, for any $\delta > 0$,

$$P(A) \le \max\left\{\exp\left(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}\right)Q(A), \delta\right\} \le \exp\left(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}\right)Q(A) + \delta.$$

As above, we have an immediate corollary to this result.

Corollary 8.2.8. *Assume that M is (α,ε)-Rényi private. Then it is also $(\varepsilon + \frac{1}{\alpha-1}\log\frac{1}{\delta}, \delta)$-differentially private for any $\delta > 0$.*

Before proving the proposition, we show how it can provide prototypical (ε,δ)-private mechanisms via Gaussian noise addition.

Example 8.2.9 (Gaussian mechanisms, continued)**:** Consider Example 8.2.3, where $f : \mathcal{P}_n \to \mathbb{R}^d$ has ℓ_2-sensitivity L. Then by Example 8.2.2, the Gaussian mechanism $Z = f(P_n) + W$ for $W \sim \mathsf{N}(0, \sigma^2 I)$ is $(\alpha, \frac{\alpha L^2}{2\sigma^2})$-Rényi private for all $\alpha \geq 1$. Combining this with Corollary 8.2.8, the Gaussian mechanism is also

$$\left(\frac{\alpha L^2}{2\sigma^2} + \frac{1}{\alpha - 1} \log \frac{1}{\delta}, \delta\right) \text{-differentially private}$$

for any $\delta > 0$ and $\alpha > 1$. Optimizing over α by taking $\alpha = 1 + \sqrt{2\sigma^2 \log \delta^{-1}/L^2}$, we see that the channel is $(\frac{L^2}{2\sigma^2} + \sqrt{2L^2 \log \delta^{-1}/\sigma^2}, \delta)$-differentially private. Thus the Gaussian mechanism

$$Z = M(P_n) := f(P_n) + W, \quad W \sim \mathsf{N}(0, \sigma^2 I) \text{ for } \sigma^2 = L^2 \max\left\{\frac{8 \log \frac{1}{\delta}}{\varepsilon^2}, \frac{1}{\varepsilon}\right\} \tag{8.2.5}$$

is (ε, δ)-differentially private.

Continuing ℓ_2-bounded mean-estimation in Example 8.2.4, assume that $\varepsilon < 8 \log \frac{1}{\delta}$, whence the Gaussian mechanism (8.2.5) with $L^2 = 4r^2/n^2$ achieves (ε, δ)-differential privacy, and

$$\mathbb{E}[\|Z_{\text{Gauss}} - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\left\|\overline{X}_n - \mathbb{E}[X]\right\|_2^2] + O(1)\frac{r^2}{n^2\varepsilon^2} \cdot d \log \frac{1}{\delta}.$$

Comparing to the Laplace mechanism (Example 8.1.2), we see an improvement whenever $\log \frac{1}{\delta} \ll d$, or that $\delta \gg e^{-d}$. $\diamond$

Proof of Proposition 8.2.7 The data processing inequality of Proposition 8.2.1.iii shows that

$$\varepsilon \geq D_\alpha(P\|Q) \geq \frac{1}{\alpha - 1} \log \left[\left(\frac{P(A)}{Q(A)}\right)^\alpha Q(A)\right].$$

Rearranging and taking exponentials, we immediately obtain the first claim of the proposition.

For the second, we require a bit more work. First, let us assume that $Q(A) > e^{-\varepsilon}\delta^{\frac{\alpha}{\alpha-1}}$. Then by the first claim of the proposition,

$$\begin{aligned} P(A) &\leq \exp\left(\frac{\alpha - 1}{\alpha}\varepsilon + \frac{1}{\alpha} \log \frac{1}{Q(A)}\right) Q(A) \\ &\leq \exp\left(\frac{\alpha - 1}{\alpha}\varepsilon + \frac{1}{\alpha}\varepsilon + \frac{1}{\alpha - 1} \log \frac{1}{\delta}\right) Q(A) = \exp\left(\varepsilon + \frac{1}{\alpha - 1} \log \frac{1}{\delta}\right) Q(A). \end{aligned}$$

On the other hand, when $Q(A) \leq e^{-\varepsilon}\delta^{\frac{\alpha}{\alpha-1}}$, and again using the first result of the proposition,

$$P(A) \leq \exp\left(\frac{\alpha - 1}{\alpha}(\varepsilon + \log Q(A))\right) \leq \exp\left(\frac{\alpha - 1}{\alpha}\left(\varepsilon - \varepsilon + \frac{\alpha}{\alpha - 1} \log \delta\right)\right) = \delta.$$

This gives the second claim of the proposition. $\square$

We finally develop our last set of connections, which relate (ε, δ)-private channels with ε-private channels. To provide this connection, we require one additional divergence notion, defining

$$D_\infty^\delta(P\|Q) := \sup_{S \subset \mathcal{X}} \left\{\log \frac{P(S) - \delta}{Q(S)} \mid P(S) > \delta\right\},$$

where the supremum is over measurable sets. Equivalently, $D_\infty^\delta(P\|Q) \le \varepsilon$ if and only if

$$P(S) \le e^\varepsilon Q(S) + \delta \quad \text{for all } S \subset \mathcal{X}.$$

Then we have the following lemma, which we will apply in Section 8.3 to develop guarantees for adaptive composition of private mechanisms.

Lemma 8.2.10. *Let $\varepsilon > 0$ and $\delta \in (0,1)$, and let P and Q be distributions on a space $\mathcal{X}$.*

(i) The distributions P and Q satisfy $D_\infty^\delta(P\|Q) \le \varepsilon$ if and only if there exists a probability distribution R on $\mathcal{X}$ such that $\|P - R\|_{\rm TV} \le \delta$ and $D_\infty(R\|Q) \le \varepsilon$.

(ii) The distributions P and Q satisfy $D_\infty^\delta(P\|Q) \le \varepsilon$ and $D_\infty^\delta(Q\|P) \le \varepsilon$ if and only if there exist distributions P_0 and Q_0 such that

$$\|P - P_0\|_{\rm TV} \le \frac{\delta}{1+e^\varepsilon}, \quad \|Q - Q_0\|_{\rm TV} \le \frac{\delta}{1+e^\varepsilon},$$

and

$$D_\infty\left(P_0\|Q_0\right) \le \varepsilon \quad \text{and} \quad D_\infty\left(Q_0\|P_0\right) \le \varepsilon.$$

The proof of the lemma is technical, so we defer it to Section 8.5.1. In words, however, it states that except with probability δ, we cannot test whether the release of a mechanism is ε-differentially private or (ε, δ)-differentially private.

8.2.3 Side information protections under weakened notions of privacy

We briefly discuss the side information protections these weaker notions of privacy protect (recall desideratum D.3), revisiting Proposition 8.1.6 and Bayes factors, as these are natural formulations of side information in terms of an adversary's probabilistic beliefs. Our first analogue of Proposition 8.1.6 applies to the (ε, δ)-private case.

Proposition 8.2.11. *Let M be (ε, δ)-differentially private and $P_n, P_n', P_n^{(0)} \in \mathcal{P}_n$ be pairwise neighboring. Then with probability at least $1 - \frac{\delta e^\varepsilon}{e^\varepsilon - 1}$ over $Z = M(P_n^{(0)})$, the posterior odds satisfy*

$$\frac{\pi(P_n \mid Z)}{\pi(P_n' \mid Z)} \le e^{2\varepsilon} \frac{\pi(P_n)}{\pi(P_n')}.$$

Deferring the proof, the result shows an adversary is unlikely to glean information distinguishing neighboring samples P_n and P_n'. This suggests the heuristic that one should take $\delta \ll 1/n$ to limit disclosure probability: by a union bound, for each individual $i \in \{1, \ldots, n\}$, we can guarantee that the posterior odds of P_n against P_n', where individual i's data is swapped, change little.

We can also give posterior update bounds for Rényi differential privacy. Though the connection to (ε, δ)-approximate differential privacy in Corollary 8.2.8 makes high probability bounds possible, we show that moments of the odds ratio do not change significantly.

Proposition 8.2.12. *Let M be (α, ε)-Rényi private, where $\alpha \in (1, \infty)$. Then for any neighboring $P_n, P_n', P_n^{(0)} \in \mathcal{P}_n$ and $Z = M(P_n^{(0)})$,*

$$\mathbb{E}_0\left[\left(\frac{\pi(P_n \mid Z)}{\pi(P_n' \mid Z)}\right)^{\alpha-1}\right]^{\frac{1}{\alpha-1}} \le e^\varepsilon \frac{\pi(P_n)}{\pi(P_n')}.$$

Proposition 8.2.12 communicates a similar message to our previous results on desideratum D.3: if an adversary receives the output of the private mechanism on a sample $P_n^{(0)} \in \mathcal{P}_n$ near the samples P_n, P_n', it is impossible to update beliefs by much. The parameter α controls this "impossibility."

We turn to the promised proofs of Propositions 8.2.11 and 8.2.12. To prove the former, we require a definition analogizing the "divergence" $D_\infty^\delta(P\|Q)$ to pointwise density ratios.

Definition 8.6. *Distributions P and Q on a space $\mathcal{X}$ are (ε, δ)-*close *if for all measurable A*

$$P(A) \le e^\varepsilon Q(A) + \delta \quad \text{and} \quad Q(A) \le e^\varepsilon P(A) + \delta.$$

*Letting p and q denote their densities with respect to any shared base measure, they are (ε, δ)-*pointwise close *if the set*

$$A := \{x \in \mathcal{X} : e^{-\varepsilon} q(x) \le p(x) \le e^\varepsilon q(x)\} = \{x \in \mathcal{X} : e^{-\varepsilon} p(x) \le q(x) \le e^\varepsilon p(x)\}$$

satisfies $P(A) \ge 1 - \delta$ and $Q(A) \ge 1 - \delta$.

Closeness and approximate differential privacy obviously relate:

Lemma 8.2.13. *If P and Q are (ε, δ)-close, then for any $\beta > 0$, the sets*

$$A_+ := \{x : p(x) > e^{(1+\beta)\varepsilon} q(x)\} \quad \text{and} \quad A_- := \{x : p(x) \le e^{-(1+\beta)\varepsilon} q(x)\}$$

satisfy

$$\max\{P(A_+), Q(A_-)\} \le \frac{e^{\beta\varepsilon}\delta}{e^{\beta\varepsilon} - 1}, \quad \max\{P(A_-), Q(A_+)\} \le \frac{e^{-\varepsilon}\delta}{e^{\beta\varepsilon} - 1}.$$

Conversely, if P and Q are (ε, δ)-pointwise close, then they are (ε, δ)-close.

Proof Let $A = A_+ = \{x : p(x) > e^{(1+\beta)\varepsilon} q(x)\}$. Then

$$P(A) \le e^\varepsilon Q(A) + \delta \le e^{-\beta\varepsilon} P(A) + \delta,$$

so that $P(A) \le \frac{\delta}{1 - e^{-\beta\varepsilon}}$. Similarly,

$$Q(A) \le e^{-(1+\beta)\varepsilon} P(A) \le e^{-\beta\varepsilon} Q(A) + e^{-(1+\beta)\varepsilon}\delta,$$

so that $Q(A) \le e^{-(1+\beta)\varepsilon}\delta/(1 - e^{-\beta\varepsilon}) = e^{-\varepsilon}\delta/(e^{\beta\varepsilon} - 1)$. The set A_- satisfies the symmetric properties.

For the converse result, let $B = \{x : e^{-\varepsilon} q(x) \le p(x) \le e^\varepsilon q(x)\}$. Then for any set A we have

$$P(A) = P(A \cap B) + P(A \cap B^c) \le e^\varepsilon Q(A \cap B) + \delta \le e^\varepsilon Q(A) + \delta,$$

and the same inequalities yield $Q(A) \le e^\varepsilon P(A) + \delta$. □

That is, (ε, δ)-close distributions are $(2\varepsilon, \frac{e^\varepsilon + e^{-\varepsilon}}{e^\varepsilon - 1}\delta)$-pointwise close.

A minor extension of this lemma (taking $\beta = 1$) yields the following result.

Lemma 8.2.14. *Let P_0, P_1, P_2 be pairwise (ε, δ)-close distributions on $\mathcal{X}$. Then for any i, j, k,*

$$A_{jk} := \left\{x \in \mathcal{X} : \log \frac{p_j(x)}{p_k(x)} > 2\varepsilon\right\} \quad \text{satisfies} \quad P_i(A_{jk}) \le \frac{e^\varepsilon}{e^\varepsilon - 1}\,\delta.$$

With Lemma 8.2.14 in hand, we can prove Proposition 8.2.11:

Proof of Proposition 8.2.11 Let $P_n^{(0)} \in \mathcal{P}_n$ denote the sample for which $Z = M(P_n^{(0)})$. Consider the three channels Q_0, Q_1, Q_2, which represent the induced distributions of $M(P_n^{(0)}), M(P_n)$, and $M(P_n')$, respectively. Then by Lemma 8.2.14, with probability at least $1 - \frac{e^\varepsilon}{e^\varepsilon - 1}\delta$, $Z \sim Q_0$ belongs to the set $A = \{z \in \mathcal{Z} \mid e^{-2\varepsilon} q_1(z) \le q_2(z) \le e^{2\varepsilon} q_1(z)\}$. Calculating the odds ratios $\pi(P_n \mid z)/\pi(P_n' \mid z) = \frac{q_1(z)\pi(P_n)}{q_2(z)\pi(P_n')}$ immediately gives the result. □

Finally, we provide the proof of Proposition 8.2.12.

Proof of Proposition 8.2.12 Let $r = \alpha - 1$ for shorthand, and let $p = \frac{\alpha}{r} = \frac{\alpha}{\alpha - 1} > 1$ and $p_* = \frac{p}{p-1} = \alpha$ be its conjugate. As in the proof of Proposition 8.2.11, let Q_0, Q_1, and Q_2 represent the distributions of $Z = M(P_n^{(0)})$, $Z = M(P_n)$, and $Z = M(P_n')$, respectively. We apply Hölder's inequality: letting q_i be the density of Q_i with respect to some base measure $d\mu$—which we know must exist by definition of Rényi differential privacy—we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\pi(P_n \mid Z)}{\pi(P_n' \mid Z)}\right)^r\right] &= \int \left(\frac{q_1(z)\pi(P_n)}{q_2(z)\pi(P_n')}\right)^r q_0(z) d\mu \\
&= \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \int \left(\frac{q_1(z)}{q_2(z)}\right)^r \frac{q_0(z)}{q_2(z)} q_2(z) d\mu \\
&\le \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \left(\int \left(\frac{q_1(z)}{q_2(z)}\right)^{pr} q_2(z) d\mu\right)^{\frac{1}{p}} \left(\int \left(\frac{q_0(z)}{q_2(z)}\right)^{p_*} q_2(z) d\mu\right)^{\frac{1}{p_*}} \\
&= \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \exp\left(\frac{(\alpha-1)^2}{\alpha} D_\alpha(Q_1 \| Q_2) + \frac{\alpha - 1}{\alpha} D_\alpha(Q_0 \| Q_2)\right) \\
&\le \left(\frac{\pi(P_n)}{\pi(P_n')}\right)^r \exp\left(\frac{(\alpha-1)^2 + \alpha - 1}{\alpha} \varepsilon\right)
\end{aligned}$$

as $pr = \alpha$ and $p_* = \alpha$. Taking everything to the $1/(\alpha - 1)$ power gives the result. □

8.3 Composition and privacy based on divergence

One of the major challenges in privacy is to understand what happens when an individual participates in multiple studies, each providing different privacy guarantees. In this case, we would like to understand and control privacy losses even when the mechanisms for information release may depend on one another, as in the desire D.1 for effective private mechanisms. Conveniently, all Rényi divergences provide strong guarantees on composition, essentially for free, and these then allow us to prove strong results on the composition of multiple private mechanisms.

8.3.1 Composition of Rényi-private channels

A natural idea to address composition is to attempt to generalize our chain rules for KL-divergence and related ideas to Rényi divergences. Unfortunately, this does not quite work, as there is no generally accepted definition of a conditional Rényi divergence, and associated chain rules do not sum naturally. In situations in which individual divergences of elements of a joint distribution have bounded Rényi divergence, however, we can provide some natural bounds.

Consider the following essentially arbitrary scheme for data generation: we have distributions P and Q on a space $\mathcal{Z}^n$, where $Z_1^n \sim P$ and $Z_1^n \sim Q$ may exhibit arbitrary dependence. Let $P_i(\cdot \mid z_1^{i-1})$ be the conditional probability of Z_i given $Z_1^{i-1} = z_1^{i-1}$ under P, and similarly for Q_i. Then the following theorem provides a tensorization guarantee for the Rényi divergence.

Theorem 8.3.1. *Let the conditions above hold, $\varepsilon_i < \infty$ for $i = 1, \ldots, n$, and $\alpha \in [1, \infty]$. Assume that conditional on z_1^{i-1}, we have $D_\alpha\big(P_i(\cdot \mid z_1^{i-1}) \| Q_i(\cdot \mid z_1^{i-1})\big) \le \varepsilon_i$. Then*

$$D_\alpha(P\|Q) \le \sum_{i=1}^n \varepsilon_i.$$

Proof It is enough to prove the claim for $\alpha \in (1, \infty)$; the endpoint cases follow by limits. We assume without loss of generality that the conditional distributions $P_i(\cdot \mid z_1^{i-1})$ and $Q_i(\cdot \mid z_1^{i-1})$ are absolutely continuous with respect to a base measure μ on $\mathcal{Z}$.[1] Then we have

$$\begin{aligned}
D_\alpha(P\|Q) &= \frac{1}{\alpha - 1} \log \int \prod_{i=1}^n \left(\frac{p_i(z_i \mid z_1^{i-1})}{q_i(z_i \mid z_1^{i-1})}\right)^\alpha q_i(z_i \mid z_1^{i-1}) d\mu^n(z_1^n) \\
&= \frac{1}{\alpha - 1} \log \int_{\mathcal{Z}_1^{n-1}} \left[\int \left(\frac{p_n(z_n \mid z_1^{n-1})}{q_n(z_n \mid z_1^{n-1})}\right)^\alpha q_n(z_n \mid z_1^{n-1}) d\mu(z_n)\right] \prod_{i=1}^{n-1} \left(\frac{p_i}{q_i}\right)^\alpha q_i d\mu^{n-1} \\
&\le \frac{1}{\alpha - 1} \log \int_{\mathcal{Z}_1^{n-1}} \exp((\alpha - 1)\varepsilon_n) \prod_{i=1}^{n-1} \left(\frac{p_i(z_i \mid z_1^{i-1})}{q_i(z_i \mid z_1^{i-1})}\right)^\alpha q_i(z_i \mid z_1^{i-1}) d\mu^{n-1}(z_1^{n-1}) \\
&= \varepsilon_n + D_\alpha\big(P_1^{n-1} \| Q_1^{n-1}\big).
\end{aligned}$$

Applying the obvious inductive argument gives the result. □

8.3.2 Privacy games and composition

To understand arbitrary composition of private channels, let us consider a privacy "game," where an adversary may sequentially choose a dataset and then observes a private release Z_i of some mechanism applied to the dataset and the dataset with one entry (observation) modified. The adversary may then select a new dataset, and repeat the game. We then ask whether the resulting sequence of (private) observations Z_1^k remains private. Figure 8.1 captures this.

Let $Z^{(0)} = (Z_1^{(0)}, \ldots, Z_k^{(0)})$ and $Z^{(1)} = (Z_1^{(1)}, \ldots, Z_k^{(1)})$ be the outputs of the privacy game, with respective marginal distributions $Q^{(0)}$ and $Q^{(1)}$. We then make the following definition.

Definition 8.7. *Let $\varepsilon \ge 0$, $\alpha \in [1, \infty]$, and $k \in \mathbb{N}$.*

(i) A collection $\mathcal{Q}$ of channels satisfies (α, ε)-Rényi privacy under k-fold adaptive composition *if, in the privacy game in Figure 8.1, the distributions $Q^{(0)}$ on $Z^{(0)}$ and $Q^{(1)}$ on $Z^{(1)}$ satisfy $D_\alpha\big(Q^{(0)} \| Q^{(1)}\big) \le \varepsilon$ and $D_\alpha\big(Q^{(1)} \| Q^{(0)}\big) \le \varepsilon$.*

(ii) Let $\delta > 0$. Then a collection $\mathcal{Q}$ of channels satisfies (ε, δ)-differential privacy under k-fold adaptive composition *if $D_\infty^\delta(Q^{(0)} \| Q^{(1)}) \le \varepsilon$ and $D_\infty^\delta(Q^{(1)} \| Q^{(0)}) \le \varepsilon$.*

[1] This is no loss of generality: the general definition of f-divergences is the supremum over finite rectangular partitions as in our discussion of KL-divergence in Section 2.2.2. One may thus prove the result for finite alphabets.

Input: Family of channels $\mathcal{Q}$ and bit $b \in \{0, 1\}$.
Repeat: for $k = 1, 2, \ldots$

i. Adversary chooses arbitrary space $\mathcal{X}$, $n \in \mathbb{N}$, and two datasets $x^{(0)}, x^{(1)} \in \mathcal{X}^n$ with $d_{\rm ham}(x^{(0)}, x^{(1)}) \le 1$.

ii. Adversary chooses private channel $Q_k \in \mathcal{Q}$.

iii. Adversary observes one sample $Z_k \sim Q_k(\cdot \mid x^{(b)})$.

Figure 8.1. The privacy game. In this game, the adversary may *not* directly observe the private bit $b \in \{0, 1\}$.

By considering a special case around a particular individual in the Game 8.1, we can gain intuition for the definition. Indeed, suppose that an individual has data x_0; in each round of the game the adversary generates two datasets, one containing x_0 and the other identical except that x_0 is removed. Then satisfying Definition 8.7 captures the intuition (cf. point D.1) that an individual's privacy remains protected, even in the face of multiple (private) accesses of the individual's data.

As an immediate corollary to Theorem 8.3.1, we then have the following.

Corollary 8.3.2. *Assume that each channel in the game in Figure 8.1 is (α, ε_i)-Rényi private. Then the arbitrary composition of k such channels remains $(\alpha, \sum_{i=1}^k \varepsilon_i)$-Rényi private.*

More sophisticated corollaries are possible using the connections between privacy measures we outline in Section 8.2.2. In this case, we can develop so-called *advanced composition* rules, which demonstrate that privacy degrades more slowly than might be expected under adaptive composition.

Corollary 8.3.3. *Assume that each channel in the game in Figure 8.1 is ε-differentially private. Then the composition of k such channels is $k\varepsilon$-differentially private as well as*

$$\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k \log \frac{1}{\delta}} \cdot \varepsilon, \delta\right)$$

differentially private for all $\delta > 0$.

Proof The first claim is immediate: for $Q^{(0)}, Q^{(1)}$ as in Definition 8.7, we know that $D_\alpha\left(Q^{(0)} \| Q^{(1)}\right) \le k\varepsilon$ for all $\alpha \in [1, \infty]$ by Theorem 8.3.1 coupled with Proposition 8.2.5 (or Corollary 8.2.6).

For the second claim, use the Rényi privacy bound in Corollary 8.2.6. Then for any $\alpha \ge 1$,

$$D_\alpha\left(Q^{(0)} \| Q^{(1)}\right) \le \frac{3k\alpha}{2}\varepsilon^2$$

by Theorem 8.3.1. Proposition 8.2.7 and Corollary 8.2.8 show (ε, δ)-differential privacy follows from Rényi privacy, so we set $\alpha = 1 + \eta$ in the preceding display to obtain that the composition is $(\frac{3k}{2}\varepsilon^2 + \frac{3k\eta}{2}\varepsilon^2 + \frac{1}{\eta}\log\frac{1}{\delta}, \delta)$-differentially private for $\eta > 0$ and $\delta > 0$. Optimize over η. □

It is possible to get sharper constants than those in Corollary 8.3.3; Exercise 8.9 shows how to achieve $(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta}}\varepsilon, \delta)$-differential privacy under adaptive composition.

A more sophisticated result, which shows adaptive composition for (ε, δ)-differentially private channels, is also possible using the probability modifications Lemma 8.2.10 provides.

Theorem 8.3.4. *Assume that each channel in the game in Figure 8.1 is (ε, δ)-differentially private. Then the composition of k such channels is $(k\varepsilon, k\delta)$-differentially private. Additionally, they are*

$$\left(\frac{3k}{2}\varepsilon^2 + \sqrt{6k \log \frac{1}{\delta_0}} \cdot \varepsilon, \delta_0 + \frac{k\delta}{1 + e^\varepsilon}\right)$$

differentially private for all $\delta_0 > 0$.

Proof Consider the channels Q_i in Fig. 8.1. As each satisfies $D_\infty^\delta(Q_i(\cdot \mid x^{(0)}) \| Q_i(\cdot \mid x^{(1)})) \le \varepsilon$ and $D_\infty^\delta(Q_i(\cdot \mid x^{(1)}) \| Q_i(\cdot \mid x^{(0)})) \le \varepsilon$, Lemma 8.2.10 guarantees the existence (at each sequential step, which may depend on the preceding $i-1$ outputs) of probability measures $Q_i^{(0)}$ and $Q_i^{(1)}$ such that $D_\infty(Q_i^{(1-b)} \| Q_i^{(b)}) \le \varepsilon$ and $\|Q_i^{(b)} - Q_i(\cdot \mid x^{(b)})\|_{\mathrm{TV}} \le \delta/(1 + e^\varepsilon)$ for $b \in \{0, 1\}$.

By construction and Theorem 8.3.1 we have $D_\alpha(Q_1^{(b)} \cdots Q_k^{(b)} \| Q_1^{(1-b)} \cdots Q_k^{(1-b)}) \le \min\{\frac{3k\alpha}{2}\varepsilon^2, k\varepsilon\}$, where $Q^{(b)}$ denotes the joint distribution on $Z_1, \ldots, Z_k$ under bit b. The triangle inequality implies that $\|Q_1^{(b)} \cdots Q_k^{(b)} - Q^{(b)}\|_{\mathrm{TV}} \le k\delta/(1+e^\varepsilon)$ for $b \in \{0, 1\}$. (See Exercise 2.19.) As a consequence, we see as in the proof of Corollary 8.3.3 that the composition is $(\frac{3k}{2}\varepsilon^2 + \frac{3k\eta}{2}\varepsilon^2 + \frac{1}{\eta}\log\frac{1}{\delta_0}, \delta_0 + k\delta/(1+e^\varepsilon))$-differentially private for all $\eta > 0$ and δ_0. Optimizing gives the result. □

These results show that whenever the privacy parameter $\varepsilon \le 1$, it is possible to compose multiple privacy mechanisms together with privacy penalty scaling as the larger of $\sqrt{k}\varepsilon$ and $k\varepsilon^2$, which is better than the "naive" $k\varepsilon$ bound. A challenge here—infrequently discussed in the privacy literature—is that in frequent practical deployments $\varepsilon \ge 1$. Then these bounds are worse than a naive bound that k-fold composition of ε-differentially private algorithms is $k\varepsilon$-differentially private.

8.4 Additional mechanisms and privacy-preserving algorithms

Since the introduction of differential privacy, a substantial literature has grown to provide mechanisms for different estimation, learning, and data release problems, so we describe a few mechanisms beyond the noise addition schemes. One challenge with the naive approaches is that they rely on *global* sensitivity of the function to be estimated—rather than local sensitivities—a worst-case notion that sometimes forces unnecessary noise. In Section 8.4.2, we give one approach to this problem, which we further develop in exercises and revisit in optimality guarantees later. Our view is circumscribed, but the results here can form a starting point for further work in this area.

8.4.1 The exponential mechanism

In many statistical learning problems, there is a loss (or utility) to measure the quality of a potentially noisy result of some computation. We abstract this by considering the input space $\mathcal{P}_n$ of samples of size n (that is, empirical distributions) and output space $\mathcal{Z}$ along with a loss function $\ell : \mathcal{P}_n \times \mathcal{Z} \to \mathbb{R}$, where $\ell(P_n, z)$ measures the loss of z on an input $P_n \in \mathcal{P}_n$. For example, if we wish to compute a function $f : \mathcal{P}_n \to \mathbb{R}$, the loss $\ell(P_n, z) = |f(P_n) - z|$ for $z \in \mathbb{R}$ is natural.

In scenarios employing such a loss, the *exponential mechanism* provides an attractive approach. Assume that for each $z \in \mathcal{Z}$, the loss $\ell(\cdot, z)$ has (global) sensitivity L, i.e., $|\ell(P_n, z) - \ell(P_n', z)| \le L$ for all neighboring $P_n, P_n' \in \mathcal{P}_n$. Assuming we have a base measure μ on $\mathcal{Z}$, we define the exponential mechanism by

$$\mathbb{P}(M(P_n) \in A) = \frac{1}{\int \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z)} \int_A \exp\left(-\frac{\varepsilon}{L}\ell(P_n, z)\right) d\mu(z), \tag{8.4.1}$$

assuming $\int e^{-\frac{\varepsilon}{L}\ell(x,z)}d\mu(z)$ is finite for each $P_n \in \mathcal{P}_n$. (Frequently one assumes ℓ takes on values in $\mathbb{R}_+$ and μ is a finite measure, making the last assumption trivial.) That is, the exponential mechanism M releases $Z = M(P_n)$ with probability proportional to

$$\exp\left(-\frac{\varepsilon}{L}\ell(P_n, z)\right).$$

The 2ε-differential privacy of the mechanism (8.4.1) is immediate: for neighboring P_n, P_n', we have

$$\begin{aligned}\frac{Q(A \mid P_n)}{Q(A \mid P_n')} &= \frac{\int \exp(-\frac{\varepsilon}{L}\ell(P_n', z))d\mu(z)}{\int \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z)} \frac{\int_A \exp(-\frac{\varepsilon}{L}\ell(P_n, z))d\mu(z)}{\int_A \exp(-\frac{\varepsilon}{L}\ell(P_n', z))d\mu(z)} \\ &\le \sup_{z\in\mathcal{Z}}\left\{\exp\left(\frac{\varepsilon}{L}[\ell(P_n, z) - \ell(P_n', z)]\right)\right\} \cdot \sup_{z\in A}\left\{\exp\left(\frac{\varepsilon}{L}[\ell(P_n', z) - \ell(P_n, z)]\right)\right\} \le \exp(2\varepsilon).\end{aligned}$$

It also trivially recovers the Laplace mechanism of Example 8.1.2:

Example 8.4.1 (The Laplace mechanism)**:** Suppose $f : \mathcal{P}_n \to \mathbb{R}^d$ has $\mathrm{GS}_1(f) \le L$. Then taking $\ell(P_n, z) = \|f(P_n) - z\|_1$ for $z \in \mathbb{R}^d$ and the usual Lebesgue measure on $\mathbb{R}^d$, the exponential mechanism uses density

$$q(z \mid P_n) \propto \exp\left(-\frac{\varepsilon}{L}\|f(P_n) - z\|_1\right),$$

which is the Laplace mechanism. $\diamond$

The exponential mechanism (8.4.1) can be challenging to compute, as it requires evaluating an often high-dimensional integral. Yet it provides an elegant abstract mechanism with reasonable utility guarantees. For the moment, we defer further examples and give utility guarantees when $\mu(\mathcal{Z})$ is finite, giving bounds based on the measure of "bad" solutions. For notational convenience, define the optimal value

$$\ell^\star(P_n) = \inf_{z\in\mathcal{Z}} \ell(P_n, z),$$

assuming tacitly that it is finite, and the sublevel sets

$$S_t := \{z \in \mathcal{Z} \mid \ell(P_n, z) \le \ell^\star(P_n) + t\}.$$

With these definitions, we have the following proposition.

Proposition 8.4.2. *Let $t \ge 0$ and $Z = M(P_n)$ for the exponential mechanism (8.4.1). Then*

$$\ell(P_n, Z) \le \ell^\star(P_n) + 2t$$

with probability at least $1 - \exp\left(-\frac{\varepsilon t}{L} + \log\frac{\mu(\mathcal{Z})}{\mu(S_t)}\right)$.

Proof Assume without loss of generality that the global sensitivity constant of ℓ is $L = 1$. Then for $Z \sim Q(\cdot \mid P_n)$, we have

$$\begin{aligned}\mathbb{P}\left(\ell(P_n, Z) \ge \ell^\star(P_n) + 2t\right) &= \frac{\int_{S_{2t}^c} \exp(-\varepsilon\ell(P_n, z))d\mu(z)}{\int \exp(-\varepsilon\ell(P_n, z))d\mu(z)} = \frac{\int_{S_{2t}^c} \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)}{\int \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)} \\ &\le \frac{\int_{S_{2t}^c} \exp(-2\varepsilon t)d\mu(z)}{\int_{S_t} \exp(-\varepsilon(\ell(P_n, z) - \ell^\star(P_n)))d\mu(z)} \le \exp(-\varepsilon t)\frac{\mu(S_{2t}^c)}{\mu(S_t)},\end{aligned}$$

where the last inequality uses that $\ell(P_n, z) - \ell^\star(P_n) \le t$ on S_t. $\square$

We can simplify this result in special cases. For example, if $\mathcal{Z}$ is finite, then Proposition 8.4.2 implies that taking μ to be the counting measure on $\mathcal{Z}$ we have

Corollary 8.4.3. *In addition to the conditions of Proposition 8.4.2, assume that* $\mathrm{card}(\mathcal{Z})$ *is finite. Then for any* $u \in (0,1)$*, with probability at least* $1-u$*,*

$$\ell(P_n, Z) \le \ell^\star(P_n) + \frac{2L}{\varepsilon} \log \frac{\mathrm{card}(\mathcal{Z})}{u}.$$

That is, with high probability, the loss of the exponential mechanism is at most logarithmic in $\mathrm{card}(\mathcal{Z})$ and grows only linearly with the global sensitivity L. A second corollary allows us to bound the expected loss of the exponential mechanism, assuming we have some control over the measure of the sublevel sets S_t. (Exercise 8.3 asks for a proof.)

Corollary 8.4.4. *Let* $t \ge 0$ *be the smallest scalar such that* $t \ge \frac{L}{\varepsilon} \log \frac{\mu(\mathcal{Z})}{\mu(S_t)}$ *and* $t \ge \frac{L}{\varepsilon}$*. Then* Z *drawn from the exponential mechanism* (8.4.1) *satisfies*

$$\mathbb{E}[\ell(P_n, Z)] \le \ell^\star(P_n) + 2t + \frac{2L}{\varepsilon} \le \ell^\star(P_n) + 4t \le \ell^\star(P_n) + O(1)\frac{L}{\varepsilon} \log\left(1 + \frac{\mu(\mathcal{Z})}{\mu(S_t)}\right).$$

Corollary 8.4.4 may seem circular, as we require the ratio $\mu(\mathcal{Z})/\mu(S_t)$ to be controlled, but it is relatively straightforward to use it (and Proposition 8.4.2) with standard volume bounds.

Example 8.4.5 (Empirical risk minimization via the exponential mechanism)**:** Consider the empirical risk minimization problem with losses $\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+$, and we seek

$$\widehat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ L(\theta, P_n) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, x_i) \right\}$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$. We make a few simplifying assumptions: first, that n is large enough that $\frac{n}{d} \ge \varepsilon$. We also assume that $\Theta \subset \mathbb{R}^d$ is an ℓ_2-ball of radius R, that $\theta \mapsto \ell(\theta, x_i)$ is M-Lipschitz for all x_i, and therefore w.l.o.g. that $\ell(\theta, x_i) \in [0, 2MR]$ for $\theta \in \Theta$.

Take the empirical loss $L(\theta, P_n)$ as our criterion for the exponential mechanism, which satisfies $|L(\theta, P_n) - L(\theta, P_n')| \le \frac{2MR}{n}$ whenever $d_{\mathrm{ham}}(P_n, P_n') \le 1$, so we release θ with density

$$q(\theta \mid x) \propto \exp\left(-\frac{n\varepsilon}{2MR} L(\theta, P_n)\right).$$

Let $\widehat{\theta}_n$ be the empirical minimizer as above; then by the Lipschitz continuity of ℓ, the sublevel set S_t satisfies

$$S_t \supset \left\{ \theta \in \Theta \mid \|\theta - \widehat{\theta}_n\|_2 \le \frac{t}{M} \right\}.$$

A volume calculation (with the factor of 2 necessary because we may have $\widehat{\theta}_n$ on the boundary of Θ) yields that for the Lebesgue measure μ,

$$\frac{\mu(S_t)}{\mu(\mathcal{Z})} \ge \left(\frac{t}{2MR}\right)^d.$$

Corollary 8.4.4 implies that if $t \ge O(1)\frac{MR}{n\varepsilon} \cdot d \log \frac{MR}{t}$, we have $\mathbb{E}[L(\theta, P_n) \mid P_n] \le L(\widehat{\theta}_n, P_n) + 4t$. The choice $t = O(1)\frac{MRd}{n\varepsilon}$ suffices whenever $\frac{\varepsilon}{d} \le 1$, so we obtain

$$\mathbb{E}[L(\theta, P_n) \mid P_n] \le L(\widehat{\theta}_n, P_n) + O(1)\frac{MRd}{n\varepsilon} \log \frac{n\varepsilon}{d}.$$

Risk minimization (peeking ahead to Chapter 17) admits convergence rates roughly of $MR/\sqrt{n}$, so privacy penalty of the exponential mechanism is lower order when $\frac{d}{\sqrt{n\varepsilon}} \ll 1$. $\diamond$

8.4.2 Local sensitivities and the inverse sensitivity mechanism

A particular instantiation of the exponential mechanism (8.4.1) can provide strong optimality guarantees for 1-dimensional quantities, and appears to be the "right" mechanism when one wishes to estimate a scalar-valued functional $f(P_n)$. Rather than adding noise scaling with the global sensitivity $\mathrm{GS}(f) = \sup |f(P_n) - f(P_n')|$, a natural idea is to use a *local* notion of sensitivity: we are only concerned with adding noise commensurate with the changes of f near $P_n \in \mathcal{P}_n$. With this in mind, define the *modulus of continuity* of f at P_n by

$$\omega_f(k; P_n) := \sup \left\{ |f(P_n') - f(P_n)| \mid d_{\rm ham}(P_n, P_n') \le k \right\}, \tag{8.4.2}$$

which measures the amount that changing k observations in P_n can change the function f. In the privacy literature, the particular choice $k = 1$ yields the *local sensitivity*

$$\mathrm{LS}(f, P_n) := \sup \left\{ |f(P_n) - f(P_n')| \mid d_{\rm ham}(P_n', P_n) = 1 \right\} = \omega_f(1; P_n). \tag{8.4.3}$$

A naive strategy, then, would be to analogize the Laplace mechanism (8.1.4) and release

$$Z = f(P_n) + \frac{\mathrm{LS}(f, P_n)}{\varepsilon} \cdot W \quad \text{for } W \sim \mathsf{Laplace}(1),$$

where the noise scales with the local sensitivity of f at P_n. The next example makes clear, however, that the scale of this noise can compromise privacy.

Example 8.4.6 (The sensitivity of the sensitivity)**:** Let $f(P_n) = \mathrm{med}(P_n)$ be the median, where the data $x \in [0, 1]$, and we take $n = 2m + 1$ for simplicity (to make the median unique). Consider two samples P_n and P_n', where P_n consists of m points $x_i = 0$ and $m + 1$ points $x_i = 1$, and P_n' consists of $m - 1$ points $x_i = 0$ and $m + 2$ points $x_i = 1$. Then the sensitivity $\mathrm{LS}(f, P_n) = 1$, the maximal value. On the other hand, $\mathrm{LS}(f, P_n') = 0$, because changing a single example cannot move the (unique) median from $f(P_n') = 1$. $\diamond$

Instead of using the inherently unstable quantity LS, we can instead use an inverse: define the *inverse sensitivity*

$$d_f(t, P_n) := \inf \left\{ d_{\rm ham}(P_n', P_n) \mid f(P_n') = t \right\}, \tag{8.4.4}$$

where $d_f(t, P_n) = +\infty$ if no P_n' yields $f(P_n') = t$. So $d_f(t, P_n)$ counts the number of examples that must be changed in the sample P_n to move $f(P_n)$ to a target t, and by inspection, always satisfies

$$|d_f(t, P_n) - d_f(t, P_n')| \le d_{\rm ham}(P_n, P_n').$$

Then the *inverse sensitivity mechanism* releases a value t with probability density proportional to

$$q(t \mid P_n) \propto \exp\left(-\frac{\varepsilon}{2} d_f(t, P_n)\right). \tag{8.4.5}$$

Implicit in the definition (8.4.5) is a base measure μ, typically one of Lebesgue measure or counting measure on a discrete set. Then a quick calculation (or recognition that the density (8.4.5) is a particular instance of the exponential mechanism) gives the following proposition.

Proposition 8.4.7. *Let M be the inverse sensitivity mechanism with density (8.4.5). Then M is ε-differentially private.*

As in the general exponential mechanism (8.4.1), efficiently sampling from the density (8.4.5) can be challenging. Some cases admit easier reformulations so long as computing d_f is easy. As a general approach, consider sampling strategies where we construct concentric shells S_k around $f(P_n)$ and sample from these with geometrically decaying probabilities $e^{-k\varepsilon/2}$. Define the "shells"

$$S_k := \{t \mid d_f(t, P_n) = k\}$$

of potential values of $f(P_n')$ when $d_{\rm ham}(P_n, P_n') = k$. When the output space is a subset of $\mathbb{R}^d$, sampling from the density (8.4.5) is then equivalent to drawing an index $I \in [n]$ with probability

$$\mathbb{P}(I = k \mid P_n) = \frac{\mathrm{Vol}(S_k)}{C(P_n)} e^{-\frac{\varepsilon}{2}k} \quad \text{for} \quad C(P_n) := \sum_{k=1}^{n} \mathrm{Vol}(S_k) e^{-\frac{\varepsilon}{2}k}, \tag{8.4.6}$$

then choosing t uniformly at random in S_k. We specialize this in the next two examples.

Example 8.4.8 (Mean estimation with bounded data)**:** Suppose the data $x \in [a, b]$ and we wish to estimate the sample mean $f(P_n) = \mathbb{E}_{P_n}[X] = \overline{x}_n$. Changing a single observation can move the mean by at most $\frac{b-a}{n}$ (replace $x_i = a$ with $x_i' = b$). Thus, while discretization issues make precisely computing d_f tedious, the approximation

$$d_{\rm mean}(t, P_n) = \left\lceil \frac{n|t - \overline{x}_n|}{b - a} \right\rceil,$$

where we define $d_{\rm mean}(t, P_n) = +\infty$ for $t \notin [a, b]$, is both Lipschitz (with respect to the Hamming metric) in the sample P_n, and approximates $d_f(t, P_n)$. (See Exercise 8.10 for a more general approach justifying this particular approximation.) The approximation

$$q(t \mid P_n) := \frac{\exp(-\frac{\varepsilon}{2} d_{\rm mean}(t, P_n))}{\int_a^b \exp(-\frac{\varepsilon}{2} d_{\rm mean}(s, P_n)) ds} \tag{8.4.7}$$

to the density (8.4.5) is thus ε-differentially private.

Sampling from the density (8.4.7) is equivalent to sampling from (8.4.6) with the shells

$$S_k = \left\{ \left[\overline{x}_n - k\frac{b-a}{n}, \overline{x}_n - (k-1)\frac{b-a}{n} \right] \cup \left[\overline{x}_n + (k-1)\frac{b-a}{n}, \overline{x}_n + k\frac{b-a}{n} \right] \right\} \cap [a, b],$$

which correspond to the amount the mean may change if we modify k examples. $\diamond$

Example 8.4.9 (Median estimation)**:** For the median, the inverse sensitivity takes a particularly clean form. In this case, for a sample $P_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i}$, where $x_i \in \mathbb{R}$, we have

$$d_f(t, P_n) = \mathrm{card}\left\{i \in [n] \mid x_i \text{ lies between } f(P_n) \text{ and } t\right\},$$

the number of examples between the median $f(P_n)$ and putative target t (we may ignore endpoint issues, as we will sample with a density). If the data lie in a range $x \in [a, b]$, then the sampling scheme (8.4.6) is easy. In this case, we have intervals

$$S_k^- := [a_k^-, a_{k-1}^-] \text{ and } S_k^+ := [a_{k-1}^+, a_k^+], \quad k = 1, \ldots, \lfloor n/2 \rfloor,$$

where

$$a_k^- = \inf\left\{f(P_n') \mid d_{\rm ham}(P_n', P_n) \le k\right\} \text{ and } a_k^+ = \sup\left\{f(P_n') \mid d_{\rm ham}(P_n', P_n) \le k\right\}.$$

That is, a_k^- and a_k^+ correspond to the $\frac{1}{2} - \frac{k}{n}$ and $\frac{1}{2} + \frac{k}{n}$ quantiles of the sample P_n, where the 0 quantile is a and 1 quantile is b. Then $S_k = S_k^- \cup S_k^+$, and we may sample (8.4.6). $\diamond$

Heuristic guarantees for the inverse sensitivity mechanism on $\mathbb{R}$. Define the shorthand $\omega(k) = \omega_f(k, P_n)$ for the modulus (8.4.2) and recall the shells $S_k = \{t \in \mathbb{R} \mid d_f(t, P_n) = k\}$. Then the values $t \in S_k$ satisfy $|f(P_n) - t| \le \omega(k)$, so the inverse sensitivity mechanism M guarantees

$$\mathbb{E}[|M(P_n) - f(P_n)|] \le \sum_{k=1}^{n} \mathbb{P}(M(P_n) \in S_k)\omega(k).$$

Temporarily, our calculations become heuristic, though later we obtain rigorous guarantees. Suppose that the interval lengths $\mathrm{Vol}(S_k)$ are of the same order for $k \lesssim \frac{1}{\varepsilon}$, and grow only polynomially quickly for $k \gg \frac{1}{\varepsilon}$. Then $C(P_n) := \sum_{k=1}^{n} \mathrm{Vol}(S_k) e^{-k\varepsilon/2} \gtrsim \mathrm{Vol}(S_1) \sum_{k=1}^{n} e^{-k\varepsilon/2} \asymp \varepsilon^{-1} \mathrm{Vol}(S_1)$, while

$$\mathbb{E}[|M(P_n) - f(P_n)|] \le \sum_{k=1}^{n} \frac{\mathrm{Vol}(S_k) e^{-k\varepsilon/2}}{\sum_{i=1}^{n} \mathrm{Vol}(S_i) e^{-i\varepsilon/2}} \omega(k) \overset{\text{heuristic}}{\lesssim} \sum_{k=1}^{n} \varepsilon e^{-k\varepsilon/2} \omega(k)$$

where the heuristic inequality uses our bound on the normalizing constant $C(P_n)$. The final sum places exponentially small weight on $\omega(k)$ for $k \gg \frac{1}{\varepsilon}$, and $\sum_k \varepsilon e^{-k\varepsilon/2} \lesssim 1$. Thus—non-rigorously—we expect that only the first $O(1)/\varepsilon$ terms should contribute to the sum $\sum_{k=1}^{n} \varepsilon e^{-k\varepsilon/2} \omega(k)$, yielding

$$\mathbb{E}[|M(P_n) - f(P_n)|] \overset{\text{heuristic}}{\lesssim} \omega_f\left(\frac{c}{\varepsilon}, P_n\right), \tag{8.4.8}$$

where c is some numerical constant.

To gain some intuition for the claims of optimality we have made, let us revisit the equivalent definitions of privacy that repose on testing, as in equation (8.1.5) and Proposition 8.1.7. By the definition of differential privacy, any ε-private mechanism satisfies

$$\mathbb{P}(M(P_n) \in A) \le e^{k\varepsilon} \mathbb{P}(M(P_n') \in A)$$

for any samples P_n, P_n' satisfying $d_{\mathrm{ham}}(P_n, P_n') \le k$. So for $k \le \frac{1}{\varepsilon}$, we have

$$\mathbb{P}(M(P_n) \in A) \le \exp(1) \mathbb{P}(M(P_n') \in A),$$

so any procedure testing whether the sample is P_n or P_n' has summed type-I and type-II error probability at least e^{-2} by Proposition 8.1.7. Thus *no* procedure can reliably distinguish the outputs of $M(P_n)$ from those of $M(P_n')$ when P_n and P_n' differ in only $1/\varepsilon$ examples. We cannot expect to estimate $f(P_n)$ to accuracy better than $\omega_f(\frac{1}{\varepsilon}, P_n)$, and so for any ε-differentially private mechanism M and P_n, there exists $P_n' \in \mathcal{P}_n$ with $d_{\mathrm{ham}}(P_n, P_n') \le \frac{1}{\varepsilon}$ for which

$$\max_{\widehat{P} \in \{P_n, P_n'\}} \mathbb{E}\left[\left|M(\widehat{P}) - f(\widehat{P})\right|\right] \gtrsim \omega_f\left(\frac{1}{\varepsilon}, P_n\right), \tag{8.4.9}$$

which the heuristic calculation (8.4.8) achieves.

Guarantees for the inverse sensitivity mechanism. To provide more rigorous guarantees requires restrictions on the function f. The simplest is that the function $f : \mathcal{P}_n \to \mathbb{R}$ obey an ordering property, where larger changes in the sample distribution P_n beget larger changes in f.

Definition 8.8. *A function $f : \mathcal{P}_n \to \mathbb{R}$ is* sample monotone *if for each $s, t \in f(\mathcal{P}_n)$ satisfying $f(P_n) \le s \le t$ or $t \le s \le f(P_n)$, we have $d_f(s, P_n) \le d_f(t, P_n)$.*

The mean and median (Examples 8.4.8 and 8.4.9) are both sample monotone. So, too, are appropriately continuous functions f. For this, we make the obvious identification of $f : \mathcal{P}_n \to \mathbb{R}$ with the induced function on $\mathcal{X}^n$ by defining $f_{\mathcal{X}}(x_1^n) := f(n^{-1} \sum_{i=1}^n \mathbf{1}_{x_i})$. Then we say $f : \mathcal{P}_n \to \mathbb{R}$ is continuous if the induced function $f_{\mathcal{X}}$ is.

Observation 8.4.10. *Let $f : \mathcal{P}_n \to \mathbb{R}$ be continuous and $\mathcal{X}$ convex. Then f is sample monotone.*

Proof Identify f with its induced function $f_{\mathcal{X}}$ for notational simplicity, and let $x \in \mathcal{X}^n$, $f(x) \le s \le t$, and $P_n = n^{-1} \sum_{i=1}^n \mathbf{1}_{x_i}$ be the empirical distribution associated with x. We show that $d_f(s, P_n) \le d_f(t, P_n)$. If $d_f(t, P_n) = +\infty$, then the desired inequality holds trivially. Otherwise, let $x' \in \mathcal{X}^n$ satisfy $f(x') = t$ and $d_{\rm ham}(x, x') = d_f(t, P_n)$. Then the function $g(\lambda) := f((1 - \lambda)x + \lambda x')$ is continuous in λ and satisfies $g(0) = f(x) \le g(1) = f(x') = t$. By the intermediate value theorem, there exists $\lambda_s \in [0, 1]$ with $g(\lambda_s) = s$, and as $\mathcal{X}$ is convex the vector $x_s = (1 - \lambda_s)x + \lambda_s x' \in \mathcal{X}^n$ satisfies $f(x_s) = g(\lambda_s) = s$. That x_s is a convex combination of x and x' then implies $d_f(s, P_n) \le d_{\rm ham}(x, x_s) \le d_{\rm ham}(x, x') = d_f(t, P_n)$. □

With Definition 8.8 in place, we can provide a few stronger guarantees for the inverse sensitivity mechanism. To avoid pathological sampling issues, one replaces the inverse sensitivity $d_f(t, P_n)$ with a "smoothed" version, where for $\rho \ge 0$ we define

$$d_{f,\rho}(t, P_n) := \inf \left\{ d_{\rm ham}(P_n, P_n') \mid |f(P_n') - t| \le \rho \right\}.$$

(Pathological cases include estimating the median where the sample P_n consists of a single point repeated n times, which would make the density (8.4.5) uniform.) Then instead of the density (8.4.5), we define the *continuous inverse sensitivity mechanism* $M_{\rm cont}$ to have density

$$q(t \mid P_n) = \frac{\exp(-\frac{\varepsilon}{2} d_{f,\rho}(t, P_n))}{\int \exp(-\frac{\varepsilon}{2} d_{f,\rho}(s, P_n)) ds}. \tag{8.4.10}$$

Setting ρ to be very small (say, $\rho = \frac{1}{n^2}$) is a reasonable practical default.

The next two propositions provide prototypical results on the error guarantees of the inverse sensitivity mechanism. (Exercises 8.13 and 8.14 outline proofs.) The first shows that the inverse sensitivity mechanism is never worse than the Laplace mechanism (8.1.4) when $\varepsilon \lesssim 1$.

Proposition 8.4.11. *Let f be sample monotone (Definition 8.8) and have finite global sensitivity $\mathrm{GS}(f) < \infty$. Then taking $\rho = 0$,*

$$\mathbb{E}\left[|M_{\rm cont}(P_n) - f(P_n)|\right] \le \frac{1}{1 - e^{-\varepsilon/2}} \mathrm{GS}(f).$$

As Example 8.1.2 shows, the standard Laplace mechanism M has error

$$\mathbb{E}\left[|M(P_n) - f(P_n)|\right] = \frac{\mathrm{GS}(f)}{\varepsilon},$$

the same scaling Proposition 8.4.11 guarantees, because $1 - e^{-\varepsilon/2} = \varepsilon/2 + O(\varepsilon^2)$.

For the next proposition, which provides a more nuanced guarantee, we require local sensitivities for samples P_n' near P_n, and so we define the largest local sensitivity within Hamming distance K of the sample P_n by

$$L(K) := \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\rm ham}(P_n, P_n') \le K \right\} = \sup_{P_n' \in \mathcal{P}_n} \left\{ \omega_f(1, P_n') \mid d_{\rm ham}(P_n, P_n') \le K \right\},$$

where we recall the definition (8.4.3) of the local sensitivity of f. Then we have the following.

Proposition 8.4.12. *Let f be sample monotone (Definition 8.8) and have finite global sensitivity* $\mathrm{GS}(f) < \infty$. *Then for $\rho \ge 0$ and $K_n = \left\lceil \frac{4 \log(2n \mathrm{GS}(f)/\rho)}{\varepsilon} \right\rceil$,*

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \le 2\rho + \frac{1}{1 - e^{-\varepsilon/2}} L(K_n).$$

Unpacking Proposition 8.4.12 a bit, let us make the default substitution $\rho = \frac{1}{n^2}$. Then because $1 - e^{-\varepsilon/2} = \varepsilon/2 + O(\varepsilon^2)$, for $\varepsilon \lesssim 1$ this yields

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \lesssim \frac{1}{\varepsilon} \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\mathrm{ham}}(P_n', P_n) \le K_n \right\} + \frac{1}{n^2},$$

where $K_n = \frac{4 \log \mathrm{GS}(f) + 12 \log n}{\varepsilon} \lesssim \frac{1}{\varepsilon} \log n$ for large sample sizes n. Comparing this to the sketched lower bound (8.4.9), these quantities are of the same order whenever the moduli of continuity $\omega_f(k; P_n)$ are roughly additive and comparable near P_n, so that for $k \lesssim \frac{1}{\varepsilon}$ there is a chain $P_n^{(1)}, P_n^{(2)}, \ldots, P_n^{(k)}$ with $d_{\mathrm{ham}}(P_n^{(i)}, P_n^{(i+1)}) = 1$, $\omega_f(k; P_n) \gtrsim \sum_{i=1}^k \mathrm{LS}(f, P_n^{(i)})$, and $\mathrm{LS}(f, P_n) \asymp \mathrm{LS}(f, P_n')$ for P_n' satisfying $d_{\mathrm{ham}}(P_n, P_n') \lesssim \frac{\log n}{\varepsilon}$. Under these conditions—which require care to check, but which hold, for example, for mean estimation—we obtain

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \lesssim \omega_f\left(\frac{1}{\varepsilon}, P_n\right) + \frac{1}{n^2}.$$

8.5 Deferred proofs

8.5.1 Proof of Lemma 8.2.10

We prove the first statement (i) of the lemma first. Assume there exists R such that $\|P - R\|_{\mathrm{TV}} \le \delta$ and $D_\infty(R \| Q) \le \varepsilon$. Then for any set S we have

$$P(S) \le R(S) + \delta \le e^\varepsilon Q(S) + \delta, \quad \text{i.e.} \quad \log \frac{P(S) - \delta}{Q(S)} \le \varepsilon,$$

which is equivalent to $D_\infty^\delta(P \| Q) \le \varepsilon$. Now, let us assume that $D_\infty^\delta(P \| Q) \le \varepsilon$, whence we must construct the distribution R.

We assume w.l.o.g. that P and Q have densities p, q, and define the sets

$$S := \{x : p(x) > e^\varepsilon q(x)\} \quad \text{and} \quad T := \{x : p(x) < q(x)\}.$$

On these sets, we have $0 \le P(S) - e^\varepsilon Q(S) \le \delta$ by assumption, and we then define a distribution R with density that we partially specify via

$$\begin{aligned} x \in S \quad &\text{implies} \quad r(x) := e^\varepsilon q(x) < p(x) \\ x \in (T \cup S)^c \quad &\text{implies} \quad r(x) := p(x) \le e^\varepsilon q(x) \quad \text{and} \quad r(x) \ge q(x). \end{aligned}$$

Now, we note that $e^\varepsilon q(x) \ge p(x) \ge q(x)$ for $x \in (S \cup T)^c$, and thus

$$\begin{aligned} Q(S) + Q(S^c \cap T^c) &\le e^\varepsilon Q(S) + P(S^c \cap T^c) \\ &= R(S) + R(S^c \cap T^c) \\ &= e^\varepsilon Q(S) + P(S^c \cap T^c) < P(S) + P(S^c \cap T^c). \end{aligned} \tag{8.5.1}$$

In particular, when $x \in T$, we may take the density r so that $p(x) \le r(x) \le q(x)$, as

$$R(S) + R(S^c \cap T^c) + P(T) < 1 \quad \text{and} \quad R(S) + R(S^c \cap T^c) + Q(T) > 1$$

by the inequalities (8.5.1), and so that $R(\mathcal{X}) = 1$. With this, we evidently have $r(x) \le e^\varepsilon q(x)$ by construction, and because $S \subset T^c$, we have

$$R(T) - P(T) = P(T^c) - R(T^c) = P(S \cap T^c) - R(S \cap T^c) + P(S^c \cap T^c) - R(S^c \cap T^c) = P(S) - R(S),$$

where we have used that $r = p$ on $(T \cup S)^c$ by construction. Thus we find that

$$\begin{aligned} \|P - R\|_{\rm TV} &= \frac{1}{2} \int_S |r - p| + \frac{1}{2} \int_T |r - p| = \frac{1}{2}(P(S) - R(S)) + \frac{1}{2}(R(T) - P(T)) \\ &= P(S) - R(S) = P(S) - e^\varepsilon Q(S) \le \delta \end{aligned}$$

by assumption.

We turn to the second statement (ii) of the lemma. We start with the easy direction, where we assume that P_0 and Q_0 satisfy $D_\infty(P_0 \| Q_0) \le \varepsilon$ and $D_\infty(Q_0 \| P_0) \le \varepsilon$ as well as $\|P - P_0\|_{\rm TV} \le \frac{\delta}{1+e^\varepsilon}$ and $\|Q - Q_0\|_{\rm TV} \le \frac{\delta}{1+e^\varepsilon}$. Then for any set S we have

$$P(S) \le P_0(S) + \frac{\delta}{1+e^\varepsilon} \le e^\varepsilon Q_0(S) + \frac{\delta}{1+e^\varepsilon} \le e^\varepsilon Q(S) + e^\varepsilon \delta + \frac{\delta}{1+e^\varepsilon},$$

or $D_\infty^\delta(P \| Q) \le \varepsilon$. The other direction is similar.

We consider the converse direction, where we have both $D_\infty^\delta(P \| Q) \le \varepsilon$ and $D_\infty^\delta(Q \| P) \le \varepsilon$. Let us construct P_0 and Q_0 as in the statement of the lemma. Define the sets

$$S := \{x : p(x) > e^\varepsilon q(x)\} \quad \text{and} \quad S' := \{x : q(x) > e^\varepsilon p(x)\}$$

as well as the sets

$$T := \{x : e^\varepsilon q(x) \ge p(x) \ge q(x)\} \quad \text{and} \quad T' := \{x : e^{-\varepsilon} q(x) \le p(x) < q(x)\},$$

so that S, S', T, T' are all disjoint and partition $\mathcal{X}$. We first construct intermediate measures P_1 and Q_1, which need not be probability measures, which we modify to construct P_0 and Q_0. On S, set

$$q_1(x) := \frac{1}{1+e^\varepsilon}(p(x) + q(x)), \quad p_1(x) := e^\varepsilon q_1(x)$$

so $p_1 \le p$ on S, and on S', set

$$p_1(x) := \frac{1}{1+e^\varepsilon}(p(x) + q(x)), \quad q_1(x) := e^\varepsilon p_1(x),$$

so $q_1 \le q$ on S'. On the remaining points $T \cup T'$, set $p_1 = p$ and $q_1 = q$. Then

$$\sup_{x \in \mathcal{X}} \left| \log \frac{p_1(x)}{q_1(x)} \right| \le \varepsilon. \tag{8.5.2}$$

Moreover, defining

$$\alpha := Q_1(S) - Q(S) = P(S) - P_1(S) = \frac{P(S) - e^\varepsilon Q(S)}{1 + e^\varepsilon} \le \frac{\delta}{1 + e^\varepsilon},$$
$$\alpha' := P_1(S') - P(S') = Q(S') - Q_1(S') = \frac{Q(S') - e^\varepsilon P(S')}{1 + e^\varepsilon} \le \frac{\delta}{1 + e^\varepsilon},$$

we have $P_1(\mathcal{X}) = 1 - \alpha + \alpha'$ and $Q_1(\mathcal{X}) = 1 + \alpha - \alpha'$

We assume w.l.o.g. that $\alpha \ge \alpha'$, as the argument in the reversed case that $\alpha' < \alpha$ is symmetric, and set $\beta = \alpha - \alpha'$. Then $P_1(\mathcal{X}) = 1 - \beta$ and $Q_1(\mathcal{X}) = 1 + \beta$. Let

$$R := \{x : p_1(x) \le q_1(x)\} = S \cup T = \{x : p(x) \ge q(x)\}.$$

Then

$$P_1(R) - Q_1(R) \ge P_1(\mathcal{X}) - Q_1(\mathcal{X}) = 2\beta.$$

Now, (roughly) we increase the density q_1 to q_0 on R and decrease p_1 to p_0 on R, while guaranteeing that $|\log \frac{q_0}{p_0}| \le \varepsilon$ on $R^c = S' \cup T'$. We choose the densities $q_0 = q_1$ and $p_0 = p_1$ on $T' \cup S'$. There exists a measurable h satisfying

$$0 \le h(x) \le \frac{p_1(x) - q_1(x)}{2} \mathbf{1}\{x \in R\} \quad \text{and} \quad \int h = \beta,$$

because $P_1(R) \ge Q_1(R) + 2\beta$. For each x, define

$$p_0(x) = p_1(x) - h(x) \quad \text{and} \quad q_0(x) = q_1(x) + h(x).$$

With these densities, Q_0 and P_0 are probability measures as $Q_0(\mathcal{X}) = P_0(\mathcal{X}) = 1$, while on $R = T \cup S$, inequality (8.5.2) implies

$$e^{-\varepsilon} \le \frac{q_1(x)}{p_1(x)} \le \frac{q_0(x)}{p_0(x)} = \frac{q_1(x) + h(x)}{p_1(x) - h(x)} = \frac{p_1(x) + q_1(x)}{p_1(x) + q_1(x)} \le 1.$$

On $R^c = T' \cup S'$, $p_0 = p_1$ and $q_0 = q_1$, so inequality (8.5.2) implies that $|\log \frac{p_0}{q_0}| \le \varepsilon$ everywhere, that is, $D_\infty(P_0 \| Q_0) \le \varepsilon$ and $D_\infty(Q_0 \| P_0) \le \varepsilon$.

It remains to bound the variation distances. As $p_0 = p_1 \ge p$ on S' and $p_0 = p_1 = p$ on T', while $p_0 = p_1 - h \le p$ on $S \cup T$, we have

$$\|P - P_0\|_{\rm TV} = \int [p_0 - p]_+ = P_0(S') - P(S') = P_1(S') - P(S') = \alpha' \le \frac{\delta}{1 + e^\varepsilon}.$$

Similarly, $q_0 = q_1 \le q$ on S' while $q_0 = q_1 = q$ on T', while $q_0 \ge q$ on $S \cup T$, so

$$\|Q - Q_0\|_{\rm TV} = \int [q - q_0]_+ = Q(S') - Q_0(S') = Q(S') - Q_1(S') = \alpha' \le \frac{\delta}{1 + e^\varepsilon}.$$

8.6 Bibliography and further reading

Given the broad focus of this book, our treatment of privacy is necessarily somewhat brief, and there is substantial material that we do not cover. The initial development of randomized response began with Warner [427], who proposed it in survey sampling as a way to collect sensitive data.

This elegant idea remained in use for many years, and in 2003 Evfimievski et al. [182] generalized it to data release mechanisms with bounded likelihood ratios, essentially, the local differential privacy Definition 8.2. Dwork, McSherry, Nissim, and Smith [170] and subsequently Dwork et al. [169] defined differential privacy and its (ε, δ)-approximate relaxation. A small industry of research has built out of these papers, with numerous extensions and developments, and the book [167] provides a computer scientist's survey of the field through the mid 2010s. Lemma 8.2.10 extends the approximate-max-divergence decomposition of Dwork et al. [171], which applies to discrete distributions, to general probability distributions, and uses a similar proof; see also [167, Lemma 3.17].

One thread of work within the differential privacy literature has been on relaxations of differential privacy—because it places strong restrictions on mechanisms—that still provide protections as in points D.1–D.3. Approximate (ε, δ)-privacy provides the first such benchmark [169]. Dwork and Rothblum's concentrated differential privacy [168] (extended by Bun and Steinke [93]) focuses on moments of the *privacy loss* random variable—the log-likelihood ratio $\log \frac{q(z|P_n)}{q(z|P_n')}$ for releases from P_n versus P_n'—and they provide definitions based on bounding this quantity. Mironov [315] extends these ideas to Rényi-differential privacy, which captures the essentiae of the approaches.

The perspective of investigating privacy via the likelihood ratios of data releases also motivates a view of privacy by bounding errors in hypothesis tests [428, 144]. Dong et al. [144] provide a comprehensive approach here, giving a broad definition of privacy related to the tradeoffs between power and level in hypothesis tests. An equivalent version of their results (implicit in their paper, and made precise by the Blackwell-Sherman-Stein comparison of experiments theorem, cf. [73, Thm. 10] or [282, Thm. 3.4]) follows. Let Z and Z' be random variables. Then a mechanism is as private as testing Z versus Z' if, for any neighboring samples P_n, P_n', there exists a function g for which

$$M(P_n) \stackrel{\text{dist}}{=} g(Z, U) \quad \text{and} \quad M(P_n') \stackrel{\text{dist}}{=} g(Z', U), \quad \text{where } U \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, 1].$$

Dong et al. focus specifically on Gaussian distributions, saying that a mechanism satisfies μ-Gaussian differential privacy if $Z \sim \mathsf{N}(0, 1)$ and $Z' \sim \mathsf{N}(\mu, 1)$ above. Equivalently, a mechanism satisfies μ-Gaussian differential privacy if testing the null that a response is generated from $M(P_n)$ against the alternative $M(P_n')$ is at least as hard—at all false-discovery levels—as testing $H_0 : Z \sim \mathsf{N}(0, 1)$ versus $H_1 : Z \sim \mathsf{N}(\mu, 1)$. This perspective also provides a transparent view of (adaptive) composition: by the Neyman-Pearson lemma, a sequence of $\mu_1, \mu_2, \ldots, \mu_k$-Gaussian differentially private mechanisms is immediately $\sqrt{\sum_{i=1}^k \mu_i^2}$-Gaussian differentially private.

A second major thread throughout the work on privacy is the development of various mechanisms satisfying privacy and work to make them more effective. We cannot cite all these papers, but point out a few that have influenced the treatment here. Nissim et al. [329] pioneered the use of the local sensitivity (8.4.3) via "smooth sensitivity" mechanisms. Dwork and Lei [166] introduce the "propose-test-release" framework, which shows that privately testing whether a particular statistic $f(P_n)$ is stable, then releasing $f(P_n)$ with noise commensurate to this stability, can be an effective approach, making connections with robust statistics that have continued to be fruitful [31, 241]. McSherry and Talwar [313] introduce the exponential mechanism, and Asi and Duchi [27, 30] introduce and analyze the inverse sensitivity approaches in Section 8.4.2.

These mechanisms and designs have begun to impact industry and government work, with the US Census adopting differential privacy in the 2020s [277], and deployments at tech companies. Particularly important for these have been *privacy amplification* approaches—which we only address in the exercises—but which show how certain best practices can increase privacy protections. As particular examples, subsampling data contributors can substantially improve privacy guarantees [39, 1], as can so-called "shuffle privacy", where individuals using locally differentially private

mechanisms are securely permuted before transmitting data to a collector [180, 40, 181, 185]. We explore a few of these ideas in Exercises 8.7 and 8.8. Xu et al. [432] detail an approach to privacy sufficient to train keyboard language models (for mobile devices) at industrial scale using careful matrix mechanisms [327, 288, 343]. (See also Exercise 8.15.)

We also note the long history of work on the *philosophy* of privacy, much of which motivates the definitions here and in the field. Pressly [348] provides a recent and quite readable view, arguing that the intellectual starting point of privacy should begin with what he terms *oblivion*, in that data does not exist at all; once data exists, it becomes a secret in a way distinct from the private sphere of life. Differential privacy approaches allow protections that can mesh with this perspective, but precisely implementing the philosophical approaches remains a challenge.

8.7 Exercises

Exercise 8.1 (Randomized response for multinomials)**:** Let $x \in \{1, \ldots, k\}$, $\varepsilon > 0$, and consider the randomized response mechanism

$$Z = \begin{cases} x & \text{with probability } \frac{e^\varepsilon}{k-1+e^\varepsilon} \\ \mathsf{Uniform}([k] \setminus \{x\}) & \text{with probability } \frac{k-1}{k-1+e^\varepsilon}. \end{cases}$$

(a) Show that Z is ε-locally differentially private (Definition 8.2).

Now let $p \in \mathbb{R}_+^k$ be the vector of probabilities $p_j = \mathbb{P}(X = j)$ for a variable $X \in [k]$, let X_i be i.i.d. p, and Z_i be drawn as above conditional on X_i.

(b) Let $\widehat{c}_n \in \mathbb{R}_+^k$ denote the empirical proportion of Z observations in a sample of size n, that is, $[\widehat{c}_n]_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i = j\}$. Provide a scalar a and vector b so that $a\widehat{c}_n - b$ is unbiased for p.

(c) Give $\mathbb{E}[\|\widehat{p}_n - p\|_2^2]$ for your estimator.

(d) Show that $\mathbb{E}[\|\widehat{p}_n - p\|_2^2] = O(1)\frac{k^2}{n\varepsilon^2}$ when $\varepsilon \le 1$, and $\mathbb{E}[\|\widehat{p}_n - p\|_2^2] = O(1)\frac{1}{n}$ when $\varepsilon \ge \log k$.

Exercise 8.2: Prove Proposition 8.2.1. *Hint.* Hölder's inequality implies property i, L'Hôpital's rule implies ii, and the f-divergence data processing inequality (Proposition 2.2.13) implies iii.

Exercise 8.3: Prove Corollary 8.4.4. *Hint.* If $W \ge 0$ is nonnegative, then $\mathbb{E}[W] = \int_0^\infty \mathbb{P}(W \ge u)du \le t + \int_t^\infty \mathbb{P}(W \ge u)du$.

Exercise 8.4: Prove Proposition 8.4.7.

Exercise 8.5 (Laplace mechanisms versus randomized response)**:** This question investigates Laplace and randomized response mechanisms, as in Examples 8.1.1 and 8.1.2 for *locally* private mean estimation of a mean $\mathbb{E}[X]$ for X_i drawn i.i.d. in $[0, 1]$.

(a) The Laplace mechanism sets $Z_i = X_i + W_i$ for $W_i \stackrel{\text{iid}}{\sim} \mathsf{Laplace}(b)$ for some b. What choice of b guarantees ε-local differential privacy?

(b) For your choice of b, let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$. Give $\mathbb{E}[(\overline{Z}_n - \mathbb{E}[X])^2]$.

(c) To define randomized response, first, randomly round X_i to $\{0, 1\}$, by setting

$$\widetilde{X}_i = \begin{cases} 1 & \text{with probability } X_i \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on $\widetilde{X}_i = x$, draw Z_i according to the randomized response (8.1.3). Give $\mathbb{E}[Z_i]$.

(d) For the randomized response Z_i above, give constants a and b so that $aZ_i - b$ is unbiased for $\mathbb{E}[X]$, that is, $\mathbb{E}[aZ_i - b] = \mathbb{E}[X]$. Let $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n (aZ_i - b)$ be your mean estimator. What is $\mathbb{E}[(\widehat{\theta}_n - \mathbb{E}[X])^2]$? Does this converge to the mean-square error of the sample mean $\mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] = \mathrm{Var}(X)/n$ as $\varepsilon \uparrow \infty$?

Exercise 8.6: This exercise compares the simple randomized response estimator from Exercise 8.5.(d) with the Laplace mechanism from Exercise 8.5.(b). For each of the following distributions, generate samples of size $N = 10, 100, 1000, 10000$, and then for $T = 25$ tests, compute the two estimators, both with $\varepsilon = 1$. Then plot the mean-squared error and confidence intervals for each of the two methods as well as the sample mean without any privacy.

i. Uniform distribution: $X \sim \mathsf{Uniform}[0, 1]$, with $\mathbb{E}[X] = 1/2$.

ii. Bernoulli distribution: $X \sim \mathsf{Bernoulli}(p)$, where $p = .1$.

iii. Uniform distribution: $X \sim \mathsf{Uniform}[.49, .51]$, with $\mathbb{E}[X] = 1/2$.

Do you prefer the Laplace or randomized response mechanism? In one sentence, why?

Exercise 8.7 (Subsampling via divergence measures (Balle et al. [39])): The *hockey stick* divergence functional, defined for $\alpha \geq 1$, is $\phi_\alpha(t) = [1 - \alpha t]_+$. It is straightforward to relate this to (ε, δ)-differential privacy via Definition 8.6: two distributions P and Q are (ε, δ)-close if and only their ϕ_{e^ε}-divergences are less than δ, i.e., if and only if

$$D_{\phi_{e^\varepsilon}}(P\|Q) \leq \delta \quad \text{and} \quad D_{\phi_{e^\varepsilon}}(Q\|P) \leq \delta.$$

(a) Let P_0, P_1, Q_1 be any three distributions, and for some $q \in [0, 1]$ and $\alpha \geq 1$, define $P = (1 - q)P_0 + qP_1$ and $Q = (1 - q)P_0 + qQ_1$. Let $\alpha' = 1 + q(\alpha - 1) = (1 - q) + q\alpha$ and $\theta = \alpha'/\alpha \leq 1$. Show that

$$D_{\phi_{\alpha'}}(P\|Q) = qD_{\phi_\alpha}((1 - \theta)P_0 + \theta P_1\|Q_1).$$

(b) Let $\varepsilon > 0$ and define $\varepsilon(q) = \log(1 + q(e^\varepsilon - 1))$. Show that

$$D_{\phi_{e^{\varepsilon(q)}}}(P\|Q) \leq q \max\left\{D_{\phi_{e^\varepsilon}}(P_0\|Q_1), D_{\phi_{e^\varepsilon}}(P_1\|Q_1)\right\}.$$

Exercise 8.8 (Subsampling and privacy amplification (Balle et al. [39])): Consider the following subsampling approach to privacy. Assume that we have a private (randomized) algorithm, represented by $\mathcal{A}$, that acts on samples of size m and guarantees (ε, δ)-differential privacy. The subsampling mechanism is then defined as follows: given a sample X_1^n of size $n > m$, choose a subsample X_{sub} of size m uniformly at random from X_1^n, and then release $Z = \mathcal{A}(X_{\mathrm{sub}})$.

(a) Use the results of parts (a) and (b) in Exercise 8.7 to show that Z is $(\varepsilon(q), \delta q)$-differentially private, where $q = m/n$ and $\varepsilon(q) = \log(1 + q(e^\varepsilon - 1))$.

(b) Show that if $\varepsilon \le 1$, then Z is $((e-1)q\varepsilon, q\delta)$-differentially private, and if $\varepsilon \le \frac{1}{2}$, then Z is $(2(\sqrt{e}-1)q\varepsilon, q\delta)$-differentially private. *Hint:* Argue that for any $T > 0$, one has $e^t - 1 \le (e^T - 1)\frac{t}{T}$ for all $t \in [0, T]$.

Exercise 8.9 (Concentration and privacy composition)**:** In this question, we use the Azuma-Hoeffding inequality to give an alternative to the Rényi privacy composition approaches we exploit in Section 8.3.2. Consider an identical scenario to that in Figure 8.1, and begin by assuming that each channel Q_i is ε-differentially private with density q_i, and let $Q^{(b)}$ be shorthand for $Q(\cdot \mid x^{(b)})$. Define the log-likelihood ratio

$$L^{(b)}(Z_1^k) := \sum_{i=1}^k \log \frac{q_i^{(b)}(Z_i)}{q_i^{(1-b)}(Z_i)}.$$

(a) Let P, Q be any two distributions satisfying $D_\infty(P\|Q) \le \varepsilon$ and $D_\infty(Q\|P) \le \varepsilon$, i.e., that $\log \frac{P(A)}{Q(A)} \in [-\varepsilon, \varepsilon]$ for all sets A. Show that

$$D_{\rm kl}(P\|Q) \le \varepsilon(e^\varepsilon - 1).$$

(b) Let $Q^{(b)}$ denote the joint distribution of $Z_1, \ldots, Z_k$ when bit b holds in the privacy game in Figure 8.1. Show that

$$\mathbb{E}_b[L^{(b)}(Z_1^k)] \le k\varepsilon(e^\varepsilon - 1)$$

where $\mathbb{E}_b$ denotes expectation under $Q^{(b)}$, and that for all $t \ge 0$,

$$Q^{(b)}\left(L^{(b)}(Z_1^k) \ge k\varepsilon(e^\varepsilon - 1) + t\right) \le \exp\left(-\frac{t^2}{2k\varepsilon^2}\right).$$

Conclude that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $Z_1^k \sim Q^{(b)}$,

$$L^{(b)}(Z_1^k) \le k(e^\varepsilon - 1)\varepsilon + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon.$$

(c) Argue that for any (measurable) set A,

$$Q^{(b)}(Z_1^k \in A) \le e^{\varepsilon(k,\delta)} \cdot Q^{(1-b)}(Z_1^k \in A) + \delta$$

for all $\delta \in [0, 1]$, where $\varepsilon(k, \delta) = k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon$.

(d) Conclude the following tighter variant of Corollary 8.3.3: if each channel in Figure 8.1 is ε-differentially private, then the composition of k such channels is

$$\left(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta}} \cdot \varepsilon, \delta\right)$$

differentially private for all $\delta > 0$.

As an aside, a completely similar derivation yields the following tighter analogue of Theorem 8.3.4: if each channel is (ε, δ)-differentially private, then their composition is

$$\left(k\varepsilon(e^\varepsilon - 1) + \sqrt{2k \log \frac{1}{\delta_0}} \cdot \varepsilon, \delta_0 + \frac{k\delta}{1 + e^\varepsilon}\right)$$

differentially private for all $\delta_0 > 0$.

Exercise 8.10 (One-dimensional minimization with inverse sensitivity)**:** Consider the private minimization of the one dimensional loss $\ell(\theta, x)$ (for $\theta \in \Theta \subset \mathbb{R}$), where we wish to estimate

$$\widehat{\theta}(P_n) \in \operatorname*{argmin}_\theta \{P_n \ell(\theta, X) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i)\},$$

where we recall the notation from Chapters 4 and 6. Assume that the loss ℓ is convex, differentiable in θ, and that it satisfies the Lipschitz-type guarantees that there exist constants $0 < L_0 \le L_1 < \infty$

$$[-L_0, L_0] \subset \{\ell'(\theta, x)\}_{x \in \mathcal{X}} \subset [-L_1, L_1] \tag{8.7.1}$$

for all $\theta \in \Theta$ and that $\{\ell'(\theta, x)\}_{x \in \mathcal{X}}$ is an interval. (That is, the set of potential derivatives $\ell'(\theta, x)$ as x varies includes $[-L_0, L_0]$, is convex, and $|\ell'(\theta, x)| \le L_1$ for all $\theta \in \Theta, x \in \mathcal{X}$.)

(a) Let the loss ℓ be the Huber loss $\ell(\theta, x) = h_u(\theta - x)$ for some fixed $u > 0$, where

$$h_u(t) = \begin{cases} \frac{1}{2u} t^2 & \text{if } |t| \le u \\ |t| - \frac{u}{2} & \text{if } |t| \ge u. \end{cases}$$

When $\mathcal{X} = \mathbb{R}$, show that ℓ satisfies the containment (8.7.1) with $L_0 = L_1 = 1$.

(b) Let the loss ℓ be the absolute value $\ell(\theta, x) = |\theta - x|$, where we abuse notation to call $\{\ell'(\theta, x)\}_{x=\theta} = [-1, 1]$ (the subdifferential). When $\mathcal{X} = \mathbb{R}$, show that ℓ satisfies the containment (8.7.1) with $L_0 = L_1 = 1$.

(c) Let $d_{\widehat{\theta}}$ be the inverse sensitivity (8.4.4) for the minimizer $\widehat{\theta}(P_n)$, which is the solution (in θ) to $P_n \ell'(\theta, X) = 0$. Assuming inequality (8.7.1) holds, show that

$$\left\lceil \frac{n|P_n \ell'(\theta, X)|}{2L_1} \right\rceil \le d_{\widehat{\theta}}(\theta, P_n) \le \left\lceil \frac{n|P_n \ell'(\theta, X)|}{L_0} \right\rceil.$$

(d) Show that the function

$$d(\theta, P_n) := \left\lceil \frac{n|P_n \ell'(\theta, X)|}{2L_1} \right\rceil$$

is 1-Lipschitz with respect to the Hamming metric in P_n.

The Lipschitz behavior of $d(\theta, P_n)$ in part (d) makes this a computationally attractive alternative to the pure inverse sensitivity (8.4.4) and associated mechanism with density (8.4.5).

Exercise 8.11 (Estimating means with inverse sensitivity mechanisms)**:** In this question, we compare behavior of mean estimation under differential privacy with the Laplace mechanism and the inverse sensitivity-type mechanism in Example 8.4.8. Let $\mathcal{X} = [-1, 1]$ be the data space and consider estimating the mean $\overline{x}_n$ of $x_1^n \in \mathcal{X}^n$.

(a) Implement the Laplace mechanism (8.1.4) for this problem. Fix $n = 200$ and repeat the following experiment 50 times. For $\varepsilon = .1, .5, 1, 2$, generate a sample $x_1^n \in \mathcal{X}^n$ (from whatever distribution you like), then estimate $\overline{x}_n$ using the Laplace mechanism. Give a table of the mean squared errors $(\overline{x}_n - M(x_1^n))^2$.

(b) Implement the inverse sensitivity mechanism using the approximation in Example 8.4.8. Repeat the experiment in part (a).

(c) Compare the results.

Exercise 8.12 (Estimating medians with the inverse sensitivity mechanism)**:** The data at `https://ee377.stanford.edu/data/salaries.txt` contains approximately 250,000 salaries from the University of California Schools between 2011 and 2014. Assuming that the maximum salary is $3 \cdot 10^6$ and minimum is 0 (so the data $x \in [0, 3 \cdot 10^6]$), implement the inverse sensitivity mechanism for the median as in Example 8.4.9. Repeat the following 20 times: for each of $\varepsilon = .0625, .125, .25, .5, 1, 2$, estimate the median using the inverse sensitivity mechanism with ε-differential privacy. Compute the mean absolute errors across the 20 experiments for each ε.

Exercise 8.13 (Shells and accuracy in inverse sensitivity)**:** Let $f : \mathcal{P}_n \to \mathbb{R}$ be sample monotone (Definition 8.8) and $\rho \geq 0$. Let $M = M_{\rm cont}$ be the continuous inverse sensitivity mechanism with density (8.4.10). Define the upper and lower shells

$$S_{k^+} = \{t > f(P_n) \mid d_{f,\rho}(t, P_n) = k\} \quad \text{and} \quad S_{k^-} = \{t < f(P_n) \mid d_{f,\rho}(t, P_n) = k\},$$

and the upper and lower moduli of continuity (values in the shells $S_{k^\pm}$)

$$\omega^+(k) := \sup\{t \in S_{k^+}\} - f(P_n) \quad \text{and} \quad \omega^-(k) := f(P_n) - \inf\{t \in S_{k^-}\}.$$

Let $S_0 = \{t \in \mathbb{R} \mid |f(P_n) - t| \leq \rho\}$.

(a) Justify the inequality

$$\begin{aligned}
&\mathbb{E}[|M(P_n) - f(P_n)|] \\
&\leq \mathbb{P}(M(P_n) \in S_0)\rho + \sum_{k=1}^n \mathbb{P}(M(P_n) \in S_{k^+})(\omega^+(k) + \rho) + \sum_{k=1}^n \mathbb{P}(M(P_n) \in S_{k^-})(\omega^-(k) + \rho).
\end{aligned}$$

(b) Bound $\mathbb{P}(M(P_n) \in S_{k^+})$ and $\mathbb{P}(M(P_n) \in S_{k^-})$, and using these bounds demonstrate that

$$\begin{aligned}
&\mathbb{E}\left[|M(P_n) - f(P_n)|\right] \\
&\leq \rho + \frac{\sum_{k=1}^n \omega^+(k) \cdot (\omega^+(k) - \omega^+(k-1)) e^{-k\varepsilon/2}}{\rho + \sum_{k=1}^n (\omega^+(k) - \omega^+(k-1)) e^{-k\varepsilon/2} + \sum_{k=1}^n (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}} \\
&\quad + \frac{\sum_{k=1}^n \omega^-(k) \cdot (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}}{\rho + \sum_{k=1}^n (\omega^+(k) - \omega^+(k-1)) e^{-k\varepsilon/2} + \sum_{k=1}^n (\omega^-(k) - \omega^-(k-1)) e^{-k\varepsilon/2}}
\end{aligned}$$

(c) Show that

$$\sum_{k=1}^n \left[(\omega^-(k) - \omega^-(k-1)) + \omega^+(k) - \omega^+(k-1)\right] e^{-k\varepsilon/2} \geq (1 - e^{-\varepsilon/2}) \sum_{k=1}^n \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}.$$

Exercise 8.14 (Accuracy of the inverse sensitivity mechanism)**:** This question outlines the proof of Propositions 8.4.11 and 8.4.12. Let the conditions of Exercise 8.13 hold. Recall the definition

$$L(K) := \sup_{P_n' \in \mathcal{P}_n} \left\{ \mathrm{LS}(f, P_n') \mid d_{\mathrm{ham}}(P_n, P_n') \le K \right\}.$$

(a) Use Exercise 8.13.(b) and (c) to show that for any $K \in \mathbb{N}$,

$$\mathbb{E}\left[|M_{\mathrm{cont}}(P_n) - f(P_n)|\right] \le \rho + \frac{L(K)}{1 - e^{-\varepsilon/2}} \cdot \frac{\sum_{k=1}^{K} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}}{\sum_{k=1}^{n} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}} + \frac{\mathrm{GS}(f)}{\rho} \sum_{k=K+1}^{n} \left(\omega^+(k) + \omega^-(k)\right) e^{-k\varepsilon/2}.$$

(b) Choose ρ and K to show Proposition 8.4.11 that $\mathbb{E}[|M_{\mathrm{cont}}(P_n) - f(P_n)|] \le \frac{1}{1-e^{-\varepsilon/2}} \mathrm{GS}(f)$.

(c) Prove Proposition 8.4.12.

Exercise 8.15 (Matrix mechanisms [343])**:** Let $X = [x_1 \; \cdots \; x_n]^\top \in \mathbb{R}^{n \times d}$ be a matrix with rows $x_i^\top$, where $\|x_i\|_2 \le 1$. The private linear estimation problem is to compute $y = AX$ for a matrix $A \in \mathbb{R}^{m \times n}$ privately. Naively, one may use the mechanism

$$M(X) = AX + \sigma W \quad \text{for } W_{ij} \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0, 1).$$

Given a factorization $A = BC$ of the matrix, one may instead consider a factorized (matrix) mechanism

$$M_{\mathrm{fact}}(X) = B(CX + \tau W) \quad \text{for } W_{ij} \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0, 1).$$

We say that M_{fact} provides as strong a privacy guarantee as M if

$$\sup_{X, X'} D_{\mathrm{kl}}\left(M(X) \| M(X')\right) \ge \sup_{X, X'} D_{\mathrm{kl}}\left(M_{\mathrm{fact}}(X) \| M_{\mathrm{fact}}(X')\right),$$

where the supremum is over matrices X, X' differing in at most a single row and whose rows satisfy $\|x_i\|_2 \le 1$. Let a_i and c_i, respectively, denote the columns of A and C.

(a) Show that M_{fact} provides as strong a privacy guarantee as M if $\max_i \|c_i\|_2^2 / \tau^2 \le \max_i \|a_i\|_2^2 / \sigma^2$.

For the remainder of the question, consider the maximum row error

$$\mathrm{err}(M) := \sup_X \mathbb{E}\left[\max_{i \le m} \left\| (M(X) - AX)^\top e_i \right\|_2\right],$$

where the supremum is over X with rows in the ℓ_2-ball. Define the $\ell_1 \to \ell_2$ and $\ell_2 \to \ell_\infty$ norms

$$\|A\|_{2 \to \infty} := \sup_{\|v\|_2 \le 1} \|Av\|_\infty \quad \text{and} \quad \|A\|_{1 \to 2} := \sup_{\|v\|_1 \le 1} \|Av\|_2,$$

that is, the maximum row and column ℓ_2-norm of A.

(b) Show that the naive mechanism M satisfies

$$\mathrm{err}(M) = K_{m,d} \cdot \sigma \left\|A\right\|_{2\to\infty}$$

while if $A = BC$ for $C \in \mathbb{R}^{m' \times n}$ and $B \in \mathbb{R}^{m \times m'}$, the factorized mechanism M_{fact} satisfies

$$\mathrm{err}(M_{\mathrm{fact}}) = K_{m',d} \cdot \tau \left\|B\right\|_{2\to\infty},$$

where $K_{m,d}$ is a dimension-dependent constant (depending only on m, d).

(c) Show that if M and M_{fact} achieve the privacy level corresponding to $\sigma = 1$, then

$$\mathrm{err}(M) = K_{m,d} \left\|A\right\|_{2\to\infty} \quad \text{while} \quad \mathrm{err}(M_{\mathrm{fact}}) = K_{m',d} \left\|C\right\|_{1\to 2} \left\|B\right\|_{2\to\infty}.$$

(d) Let A be the all-ones matrix $A = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n\times n}$. Give a factorization of $A = BC$ into B and C so that $\left\|B\right\|_{2\to\infty} = 1$ and $\left\|C\right\|_{1\to 2} = 1$, and conclude that $\mathrm{err}(M_{\mathrm{fact}}) = O(1/\sqrt{n})\mathrm{err}(M)$.

(e) Let $A \in \{0,1\}^{n\times n}$ be the lower triangular matrix with all ones on its diagonal and below, so that AX computes the running sum of rows of X. Give a factorization of $A = BC$ into B and C so that $\left\|B\right\|_{2\to\infty} = O(\log n)$ and $\left\|C\right\|_{1\to 2} = O(\log n)$. What can you conclude about the errors of the two mechanisms? (You may assume that n is any appropriate power of 2 for simplicity in your construction; ours uses $n = 2^k - 1$ and inducts on k.)

Part II

Fundamental limits and optimality

Précis

The second major part of this book brings information theory to bear in proving fundamental limits for statistical problems. With its strong emphasis on converse results in communication and coding, information theory provides a natural set of tools for developing these lower bounds. In the broader perspective of the book, these limits repose on results that guarantee that for a parameter $\theta = \theta(P)$ of a distribution P, distributions P and Q may be close enough together that it is hard to test whether the observed data follow distribution P or Q, while the associated parameters $\theta(P)$ and $\theta(Q)$ are far apart. Once again, we see the ways in which closeness of distributions impacts estimation, though in a way distinct from Part I, where we controlled forms of closeness between empirical and population quantities.

We will investigate several techniques for lower bounds as well as different areas—statistics, machine learning, communication, hypothesis testing—in which we may apply them. We will develop three of the fundamental methods for lower bounds in the first chapter of this part of the book: the Le Cam, Fano, and Assouad methods. By connecting these methods with quantitative data processing inequalities or other information constraints, we will extend the results to constrained problems, where estimators must reflect real-world constraints, such as privacy or communication limits. Some special cases, especially when we are interested in squared error and parameter estimation, admit special techniques for achieving (asymptotically) exact optimality guarantees, and we will treat these along the way. Though some of the techniques are classical, they also illuminate modern problems in estimation with privacy or how much information a learning procedure must retain about its data. We conclude with testing and functional estimation, where the object of interest may be a scalar functional of a much larger model.

Chapter 9

Minimax lower bounds: the Le Cam, Fano, and Assouad methods

Understanding the fundamental limits of estimation and optimization procedures can give complementary insights to those we obtain in developing procedures. By exhibiting such lower bounds, it is possible to guarantee that an algorithm we have developed is optimal, so that searching for estimators with better statistical performance will have limited returns, though searching for estimators with better performance in other metrics may remain interesting. Exhibiting refined lower bounds can suggest avenues for developing alternative, new optimal estimators; developing converses need not be a fully pessimistic exercise.

In this chapter, we define and then discuss techniques for lower bounding the minimax risk, giving three standard recipes that have proven fruitful in a variety of estimation problems [435]. In addition to reviewing these standard Le Cam, Fano, and Assouad methods, we present a few simplifications and extensions to make them more user friendly. In this chapter, our lower bounds use "local" minimax risk, showing that estimation is difficult in the neighborhood of a particular point, while subsequent chapters build from these ideas to more global optimality notions.

9.1 Basic framework and minimax risk

We first establish the minimax frameworks we use. When we study classical estimation problems, we use a standard version of minimax risk; we will also show how minimax bounds can be used to study optimization problems, in which case we use a specialization of the general minimax risk that we call minimax *excess* risk (it is important enough that we define additional notation).

Let us begin by defining the standard minimax risk. We let $\mathcal{P}$ denote a class of distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \Theta$ denote a function defined on $\mathcal{P}$, that is, a mapping $P \mapsto \theta(P)$. The goal is to estimate the parameter $\theta(P)$ from observations X_i drawn from the (unknown) distribution P. In certain cases, the parameter $\theta(P)$ uniquely determines the underlying distribution; for example, the mean $\theta(P) = \mathbb{E}_P[X]$ from a normal family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with known variance σ^2 uniquely determines distributions in $\mathcal{P}$. In other families, a mean $\theta(P) = \mathbb{E}_P[X]$ obviously need not uniquely determine the distribution. We therefore take a slightly broader viewpoint of estimating *functions* of distributions.

The space Θ in which the parameter $\theta(P)$ takes values depends on the underlying statistical problem; for example, for the univariate mean $\theta(P) = \mathbb{E}_P[X]$ we have $\Theta \subset \mathbb{R}$. To evaluate the quality of an estimator $\widehat{\theta}$, we let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ denote a (semi)metric on Θ, which we use to

measure the error of an estimator for the parameter θ, and let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (for example, $\Phi(t) = t^2$).

For a distribution $P \in \mathcal{P}$, we assume we receive i.i.d. observations $X_i \sim P$, and based on these $\{X_i\}$, the goal is to estimate the parameter $\theta(P) \in \Theta$. For an estimator $\widehat{\theta}$—a measurable function $\widehat{\theta} : \mathcal{X}^n \to \Theta$—we assess the quality of the estimate $\widehat{\theta}(X_1, \ldots, X_n)$ in terms of the risk

$$\mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right].$$

For instance, for a univariate mean problem with $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$, this risk is the mean-squared error. As the distribution P varies, we obtain the *risk functional* for the problem, which gives the risk of any estimator $\widehat{\theta}$ for the family $\mathcal{P}$.

For any fixed distribution P, there is always a trivial estimator of $\theta(P)$: simply return $\theta(P)$, which will have minimal risk. This "estimator" is poor in any real sense, and it is thus important to consider the risk functional as P varies. One approach to this is Bayesian: we place a prior π on the set of possible distributions $\mathcal{P}$, viewing $\theta(P)$ as a random variable, and evaluate the risk of an estimator $\widehat{\theta}$ taken in expectation with respect to this prior on P. Another approach, first suggested by Wald [424, 425], is to choose the estimator $\widehat{\theta}$ minimizing the maximum risk

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right].$$

An optimal estimator for this metric then gives the *minimax risk*,

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X_1, \ldots, X_n), \theta(P))\big) \right], \tag{9.1.1}$$

where we take the supremum (worst-case) over distributions $P \in \mathcal{P}$ and the infimum over estimators $\widehat{\theta}$. Here the notation $\theta(\mathcal{P})$ indicates that we consider parameters $\theta(P)$ for distributions $P \in \mathcal{P}$.

In some scenarios, we study a specialized notion of risk appropriate for optimization problems (or prediction-focused statistical problems). In these settings, we assume a loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$, where the value $\ell(\theta; x)$ measures the instantaneous loss associated with using θ as a predictor on an observation $x \in \mathcal{X}$. In this case, we define the population loss

$$L_P(\theta) := \mathbb{E}_P[\ell(\theta; X)] = \int_{\mathcal{X}} \ell(\theta; x) dP(x). \tag{9.1.2}$$

We have seen this in the context of M-estimation in Section 5.3, and we revisit it more seriously in Chapter 17. For an estimator $\widehat{\theta} : \mathcal{X}^n \to \Theta$, we define the maximum *excess risk* for the family $\mathcal{P}$ by

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[L_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} L_P(\theta) \right\},$$

where the expectation is taken over $X_i \stackrel{\text{iid}}{\sim} P$ and randomness in the procedure $\widehat{\theta}$. This captures the difference between the expected risk of $\widehat{\theta}$ and the best possible risk under the distribution P. The *minimax excess risk* with respect to the loss ℓ, domain Θ, and family $\mathcal{P}$ of distributions is then

$$\mathfrak{M}_n(\Theta, \mathcal{P}, \ell) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P \left[L_P(\widehat{\theta}(X_1, \ldots, X_n)) \right] - \inf_{\theta \in \Theta} L_P(\theta) \right\}, \tag{9.1.3}$$

where the infimum is taken over all estimators $\widehat{\theta} : \mathcal{X}^n \to \Theta$ and the loss ℓ implicitly defines the population loss L_P. The techniques for providing lower bounds for the minimax (9.1.1) or the excess (9.1.3) risk essentially coincide; we focus in this chapter on providing lower bounds on the minimax risk, returning in Chapter 17 to specialize the techniques to the excess risk.

9.2 Preliminaries on methods for lower bounds

Techniques for lower bounding the minimax risk (9.1.1) typically transform the maximum risk by lower bounding it via a Bayesian problem, then proving a lower bound on the performance of all possible estimators for the Bayesian problem. (It is often the case that the worst case Bayesian problem is equivalent to the original minimax problem [282], but we do not pursue this here). In particular, let $\{P_v\} \subset \mathcal{P}$ be a collection of distributions in $\mathcal{P}$ indexed by v and π be any probability mass function over v. Then for any estimator $\widehat{\theta}$, the maximum risk has lower bound

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P)))\right] \geq \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1^n), \theta(P_v)))\right]. \tag{9.2.1}$$

This trivial inequality serves as the departure point for each of the subsequent techniques for lower bounding the minimax risk. Because it is a Bayesian quantity, it allows us with no loss of generality to consider only deterministic estimators, a powerful reduction principle (Exercise 9.1.)

9.2.1 From estimation to testing

A standard first step in proving minimax bounds is to "reduce" the estimation problem to a testing problem. The idea is to show that the probability of error in testing problems lower bounds the estimation risk, and we can develop tools for the former. We use two types of testing problems: one a multiple hypothesis test and the second based on multiple binary hypothesis tests.

Given an index set $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v \in \mathcal{V}}$ contained within $\mathcal{P}$. This family induces a collection of parameters $\{\theta(P_v)\}_{v \in \mathcal{V}}$; recalling Section 5.1, the collection is a 2δ-packing in the ρ-semimetric if

$$\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta \quad \text{for all } v \neq v'.$$

We use this family to define the *canonical hypothesis testing problem*:

- first, nature chooses V according to the uniform distribution over $\mathcal{V}$;
- second, conditioned on the choice $V = v$, draw the random sample $X_1^n = (X_1, \ldots, X_n)$ from the n-fold product distribution P_v^n.

Given the observed sample X_1^n, the goal is to determine the value of the underlying index v. We refer to any measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$ as a test with associated error probability $\mathbb{P}(\Psi(X_1^n) \neq V)$, where $\mathbb{P}$ denotes the joint distribution over the random index V and X_1^n. In particular, if $\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v^n$ denotes the mixture distribution, then the hypothesis testing problem is to determine the random index V given a sample from $\overline{P}$.

With this setup, we obtain the classical reduction from estimation to testing.

Proposition 9.2.1. *The minimax error* (9.1.1) *has lower bound*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X_1, \ldots, X_n) \neq V), \tag{9.2.2}$$

where the infimum ranges over all test functions.

Proof Fix an arbitrary estimator $\widehat{\theta}$. Suppressing dependence on X throughout the derivation, first note that it is clear that for any fixed θ, we have

$$\mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \mathbb{E}\left[\Phi(\delta) \mathbf{1}\left\{\rho(\widehat{\theta}, \theta) \geq \delta\right\}\right] = \Phi(\delta) \mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta),$$

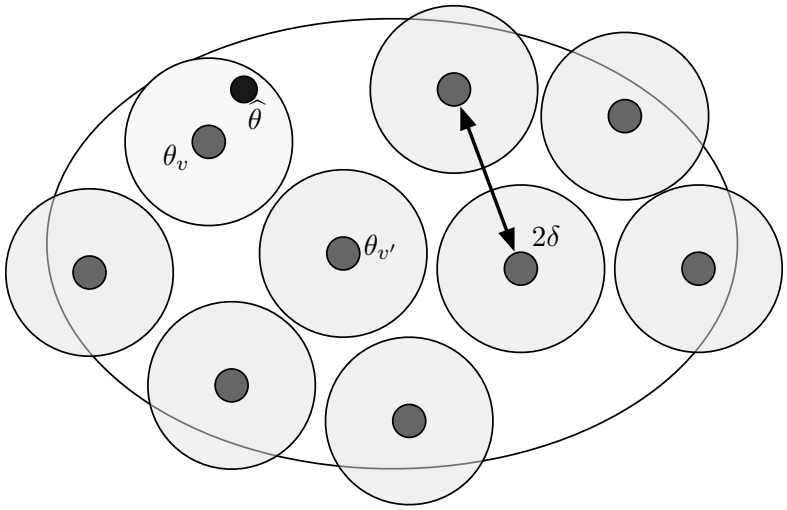


Figure 9.1: The estimate $\widehat{\theta}$ is contained in at most one of the δ-balls around the points θ_v.

where the final inequality follows because Φ is non-decreasing. Now, let us define $\theta_v = \theta(P_v)$, so that $\rho(\theta_v, \theta_{v'}) \geq 2\delta$ for $v \neq v'$. By defining the test function

$$\Psi(\widehat{\theta}) := \operatorname*{argmin}_{v \in \mathcal{V}} \{\rho(\widehat{\theta}, \theta_v)\},$$

breaking ties arbitrarily, $\rho(\widehat{\theta}, \theta_v) < \delta$ implies that $\Psi(\widehat{\theta}) = v$ because of the triangle inequality and 2δ-separation of the set $\{\theta_v\}_{v \in \mathcal{V}}$. Indeed, assume that $\rho(\widehat{\theta}, \theta_v) < \delta$; then for any $v' \neq v$, we have

$$\rho(\widehat{\theta}, \theta_{v'}) \geq \rho(\theta_v, \theta_{v'}) - \rho(\widehat{\theta}, \theta_v) > 2\delta - \delta = \delta.$$

The test must thus return v as claimed. Equivalently, for $v \in \mathcal{V}$, the inequality $\Psi(\widehat{\theta}) \neq v$ implies $\rho(\widehat{\theta}, \theta_v) \geq \delta$. (See Figure 9.1.) By averaging over $\mathcal{V}$, we find that

$$\sup_P \mathbb{P}(\rho(\widehat{\theta}, \theta(P)) \geq \delta) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\rho(\widehat{\theta}, \theta(P_v)) \geq \delta \mid V = v) \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{P}(\Psi(\widehat{\theta}) \neq v \mid V = v).$$

Taking an infimum over all tests $\Psi : \mathcal{X}^n \to V$ gives inequality (9.2.2). □

The remaining challenge is to lower bound the probability of error in the underlying multi-way hypothesis testing problem, which we do by choosing the separation δ to trade off between the loss $\Phi(\delta)$ (large δ increases the loss) and the probability of error (small δ, and hence separation, makes the hypothesis test harder). Usually, one attempts to choose the largest separation δ that guarantees a constant probability of error. We present three techniques for this: Le Cam's method, Fano's method, and Assouad's method, including extensions of the latter two to enhance their applicability. Before continuing, however, we recall the essential inequalities between divergence measures defined on probabilities and concepts related to packings and covering.

9.2.2 Inequalities between divergences and product distributions

Our lower bounds require showing distributions, especially product distributions, are close in different metrics, so we recall the relationships between f-divergences that Section 2.2.4 establishes. We

first recall the definitions of the total variation, Hellinger, and KL divergence. Take distributions P and Q on a set $\mathcal{X}$, and assume (without loss of generality) that they have densities p, q with respect to a base measure μ. The total variation distance (2.2.6) is

$$\|P - Q\|_{\rm TV} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int |p(x) - q(x)| d\mu(x).$$

The Hellinger distance (2.2.7) is

$$d_{\rm hel}(P, Q)^2 := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x).$$

We also recall the Kullback-Leibler (KL) divergence

$$D_{\rm kl}\left(P \| Q\right) := \int p(x) \log \frac{p(x)}{q(x)} d\mu(x). \tag{9.2.3}$$

As we note in Section 2.2.3, Proposition 2.2.9, these divergences have the following relationships.

Corollary 9.2.2 (Propositions 2.2.8 and 2.2.9)**.** *The total variation distance satisfies the following.*

(a) For the Hellinger distance,

$$d_{\rm hel}(P, Q)^2 \le \|P - Q\|_{\rm TV} \le d_{\rm hel}(P, Q) \sqrt{2 - d_{\rm hel}(P, Q)^2}.$$

(b) Pinsker's inequality: for any distributions P, Q,

$$\|P - Q\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P \| Q\right).$$

Corollary 9.2.2 proves useful because KL-divergence and Hellinger distance both tensorize under product distributions. Specifically, consider $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$. Then the KL-divergence satisfies the tensorization

$$D_{\rm kl}\left(P \| Q\right) = \sum_{i=1}^n D_{\rm kl}\left(P_i \| Q_i\right), \tag{9.2.4}$$

while because $d_{\rm hel}^2(P, Q) = 1 - \int \sqrt{dP dQ}$, the Hellinger distance tensorizes as well:

$$d_{\rm hel}^2(P, Q) = 1 - \prod_{i=1}^n \int \sqrt{p_i(x_i) q_i(x_i)} d\mu(x_i) = 1 - \prod_{i=1}^n \left(1 - d_{\rm hel}^2(P_i, Q_i)\right). \tag{9.2.5}$$

So for product distributions P^n and Q^n, Corollary 9.2.2 implies that

$$\|P^n - Q^n\|_{\rm TV}^2 \le \frac{1}{2} D_{\rm kl}\left(P^n \| Q^n\right) = \frac{n}{2} D_{\rm kl}\left(P \| Q\right)$$

and

$$\|P^n - Q^n\|_{\rm TV} \le d_{\rm hel}(P^n, Q^n) \sqrt{2 - d_{\rm hel}^2(P^n, Q^n)} = \sqrt{1 - (1 - d_{\rm hel}^2(P, Q))^n} \sqrt{1 + (1 - d_{\rm hel}^2(P, Q))^n}.$$

As a consequence, if we can guarantee that $D_{\rm kl}\left(P \| Q\right) \le 1/n$ or $d_{\rm hel}(P, Q) \le 1/\sqrt{n}$, then we have $\|P^n - Q^n\|_{\rm TV} \le 1 - c$ for a fixed numerical constant $c > 0$, for any n. We will see how proofs of minimax bounds leverage these types of guarantees in the following sections.

9.2.3 Metric entropy and packing numbers

The second step in proving our lower bounds involves constructing the packing set in Section 9.2.1. We therefore recapitulate the results we develop in Section 5.1.3 on covering and packing numbers. The most central construction relies on volume bounds, which we recall from Lemma 5.1.10: the covering and packing numbers of a norm ball $\mathbb{B}$ in its own norm $\|\cdot\|$ scale exponentially in the dimension. In particular, for $\delta < 1$, there is a packing $\mathcal{V}$ of $\mathbb{B}$ such that $\|v - v'\| \geq \delta$ for all distinct $v, v' \in \mathcal{V}$ and $|\mathcal{V}| \geq (1/\delta)^d$. We record that we can construct exponentially large packings of arbitrary norm-balls where the points have constant distance from one another.

Corollary 9.2.3. *Let $\mathbb{B}^d = \{v \in \mathbb{R}^d \mid \|v\| \leq 1\}$ be the unit ball for the norm $\|\cdot\|$. Then there exists $\mathcal{V} \subset \mathbb{B}^d$ with $|\mathcal{V}| \geq 2^d$ and $\|v - v'\| \geq \frac{1}{2}$ for each $v \neq v' \in \mathcal{V}$.*

Another common packing arises from coding theory, where the technique is to construct well-separated code-books ($\{0,1\}$-valued bit strings associated to individual symbols to be communicated). In our lower bounds, we show that codewords in a well-separated code-book may be hard to estimate. We now demonstrate a well-separated codebook, i.e., an $O(d)$-separated but exponentially large packing of the d-dimensional hypercube.

Lemma 9.2.4 (Gilbert-Varshamov bound)**.** *Let $d \geq 1$. There is a subset $\mathcal{V}$ of the d-dimensional hypercube $\mathcal{H}_d = \{-1, 1\}^d$ of size $|\mathcal{V}| \geq \exp(d/8)$ such that*

$$\left\|v - v'\right\|_1 = 2\sum_{j=1}^{d} \mathbf{1}\left\{v_j \neq v'_j\right\} \geq \frac{d}{2} \quad \text{for } v \neq v' \in \mathcal{V}.$$

Proof We follow Guntuboyina [221]. Consider a maximal subset $\mathcal{V}$ of $\mathcal{H}_d = \{-1, 1\}^d$ satisfying

$$\left\|v - v'\right\|_1 \geq d/2 \quad \text{for all distinct } v, v' \in \mathcal{V}. \tag{9.2.6}$$

That is, the addition of any vector $w \in \mathcal{H}_d, w \notin \mathcal{V}$ to $\mathcal{V}$ breaks the constraint (9.2.6). This means that the closed balls $B(v, d/2) := \{w \in \mathcal{H}_d : \|v - w\|_1 \leq d/2\}$ must satisfy

$$\bigcup_{v \in \mathcal{V}} B(v, d/2) = \mathcal{H}_d \quad \text{so} \quad |\mathcal{V}| \cdot |B(0, d/2)| = \sum_{v \in \mathcal{V}} |B(v, d/2)| \geq 2^d. \tag{9.2.7}$$

We upper bound the cardinality of $B(v, d/2)$ using the probabilistic method, which will imply the desired result. Let $S_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}\{0, 1\}$, $i = 1, \ldots, d$. By uniformity, for any $v \in \mathcal{H}_d$,

$$\begin{aligned} 2^{-d}|B(v, d/2)| &= \mathbb{P}(S_1 + S_2 + \ldots + S_d \leq d/4) = \mathbb{P}(S_1 + S_2 + \ldots + S_d \geq 3d/4) \\ &\leq \mathbb{E}\left[\exp(\lambda S_1 + \ldots + \lambda S_d)\right] \exp(-3\lambda d/4) \end{aligned}$$

for $\lambda \geq 0$, by Markov's inequality. As $\mathbb{E}[\exp(\lambda S_1)] = \frac{1}{2}(1 + e^\lambda)$, we obtain

$$2^{-d}|B(v, d/2)| \leq \inf_{\lambda \geq 0} \left\{2^{-d}(1 + e^\lambda)^d \exp(-3\lambda d/4)\right\}$$

Choosing $\lambda = \log 3$, we have

$$|B(v, d/2)| \leq 4^d \exp(-(3/4)d \log 3) = 3^{-3d/4} 4^d.$$

Recalling inequality (9.2.7), we have $|\mathcal{V}| 3^{-3d/4} 4^d \geq |\mathcal{V}| \cdot |B(0, d/2)| \geq 2^d$, or

$$|\mathcal{V}| \geq \frac{3^{3d/4}}{2^d} = \exp\left(d\left[\frac{3}{4}\log 3 - \log 2\right]\right) \geq \exp(d/8),$$

as claimed. □

9.3 Le Cam's method

Le Cam's method, in its simplest form, provides lower bounds on the error in simple binary hypothesis testing problems. We explore this connection, revisiting the relationship between hypothesis testing and total variation distance, and we then show how this can yield lower bounds on minimax error (or the optimal Bayes' risk) for simple—often one-dimensional—estimation problems.

Section 2.2.3 (see also Exercise 2.1) considers several representations of the total variation distance, including its relationships with optimal testing. We begin again with this strand of thought, recalling the general testing problem in Section 9.2.1. Suppose we have a Bayesian hypothesis testing problem where V is chosen with equal probability to be 1 or 2, and given $V = v$, the sample X is drawn from the distribution P_v. Denoting by $\mathbb{P}$ the joint distribution of V and X, we have for any test $\Psi : \mathcal{X} \to \{1, 2\}$ that the probability of error is

$$\mathbb{P}(\Psi(X) \neq V) = \frac{1}{2}P_1(\Psi(X) \neq 1) + \frac{1}{2}P_2(\Psi(X) \neq 2).$$

Proposition 2.3.1 gives an exact representation of the testing error using total variation distance:

Proposition (Proposition 2.3.1, restated)**.** *For any distributions P_1 and P_2 on $\mathcal{X}$, we have*

$$\inf_{\Psi} \left\{P_1(\Psi(X) \neq 1) + P_2(\Psi(X) \neq 2)\right\} = 1 - \|P_1 - P_2\|_{\rm TV}, \tag{9.3.1}$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$.

Returning to the setting in which we receive n i.i.d. observations $X_i \sim P$, when $V = 1$ with probability $\frac{1}{2}$ and 2 with probability $\frac{1}{2}$, we have

$$\inf_{\Psi} \mathbb{P}\left(\Psi(X_1, \ldots, X_n) \neq V\right) = \frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{\rm TV}. \tag{9.3.2}$$

The representations (9.3.1) and (9.3.2), in conjunction with our reduction of estimation to testing in Proposition 9.2.1, imply the following lower bound on minimax risk. For any family $\mathcal{P}$ of distributions for which there exists a pair $P_1, P_2 \in \mathcal{P}$ satisfying $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, then the minimax risk after n observations has lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2}\|P_1^n - P_2^n\|_{\rm TV}\right]. \tag{9.3.3}$$

The lower bound (9.3.3) suggests the following strategy: we find distributions P_1 and P_2, which we choose as a function of δ, that guarantee $\|P_1^n - P_2^n\|_{\rm TV} \leq \frac{1}{2}$. In this case, so long as $\rho(\theta(P_1), \theta(P_2)) \geq 2\delta$, we have the lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)\left[\frac{1}{2} - \frac{1}{2} \cdot \frac{1}{2}\right] = \frac{1}{4}\Phi(\delta).$$

The next example illustrates this idea.

Example 9.3.1 (Bernoulli mean estimation)**:** Consider estimating the mean $\theta \in [-1, 1]$ of a $\{\pm 1\}$-valued Bernoulli distribution under the squared error loss $(\theta - \widehat{\theta})^2$, and let $\mathcal{B}$ be the collection of such distributions. Fixing $\delta > 0$ and $\mathcal{V} = \{-1, 1\}$, we define P_v so that

$$P_v(X = 1) = \frac{1 + v\delta}{2} \quad \text{and} \quad P_v(X = -1) = \frac{1 - v\delta}{2},$$

whence we see that the mean $\theta(P_v) = \delta v$. Using the metric $\rho(\theta, \theta') = |\theta - \theta'|$ and loss $\Phi(\delta) = \delta^2$, we have separation 2δ of $\theta(P_{-1})$ and $\theta(P_1)$. Thus, via Le Cam's method (9.3.3),

$$\mathfrak{M}_n(\mathcal{B}, (\cdot)^2) \geq \frac{1}{2}\delta^2 \left(1 - \left\|P_{-1}^n - P_1^n\right\|_{\rm TV}\right).$$

We therefore wish to upper bound $\|P_{-1}^n - P_1^n\|_{\rm TV}$ as a function of the separation δ and sample size n; here we use Pinsker's inequality (Proposition 2.2.9(a)) and the tensorization identity (9.2.4) that makes KL-divergence so useful. Indeed, we have

$$\left\|P_{-1}^n - P_1^n\right\|_{\rm TV}^2 \leq \frac{1}{2} D_{\rm kl}\left(P_{-1}^n \| P_1^n\right) = \frac{n}{2} D_{\rm kl}\left(P_{-1} \| P_1\right) = \frac{n}{2}\delta \log \frac{1+\delta}{1-\delta}.$$

Noting that $\delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, 1/2]$, we obtain that $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \delta\sqrt{3n/2}$ for $\delta \leq 1/2$. In particular, we can guarantee constant probability of error in the associated hypothesis testing problem (recall inequality (9.3.2)) by taking $\delta = 1/\sqrt{6n}$; this guarantees $\|P_{-1}^n - P_1^n\|_{\rm TV} \leq \frac{1}{2}$. We thus have the minimax lower bound

$$\mathfrak{M}_n(\mathcal{B}, (\cdot)^2) \geq \frac{1}{2}\delta^2 \left(1 - \frac{1}{2}\right) = \frac{1}{24n}.$$

While the factor $1/24$ is smaller than necessary, this bound is optimal to within constant factors, as the sample mean $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$ achieves $\mathbb{E}[(\overline{X}_n - \theta)^2] = (1-\theta^2)/n$. Exercise 9.2 asks for a proof that this dependence on $(1-\theta^2)$ is optimal using Hellinger distances. $\diamond$

This example shows the "usual" rate of convergence in parametric estimation problems, that is, that we can estimate a parameter θ at a rate (in squared error) scaling as $1/n$. Mean estimation is, in some sense, the prototypical example of such regular problems. In some irregular scenarios—including estimating the support of a uniform random variable, which the exercises study—faster rates are possible. In Chapter 13, we revisit Le Cam's method in more complex problems, where by elaborating the technique to work with not just two-point problems but mixtures of distributions, we can extend the ideas to more sophisticated scenarios.

9.4 Fano's method

Fano's method, originally proposed by Has'minskii [228] for providing lower bounds in nonparametric estimation, gives a more general technique than Le Cam's method, and it applies when the packing set $\mathcal{V}$ has cardinality larger than two.

9.4.1 The classical (local) Fano method

We begin by recalling Fano's inequality, which provides a lower bound on the error in a multiway hypothesis testing problem. Let V be a random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$. If we let the function $h_2(p) = -p\log p - (1-p)\log(1-p)$ denote the entropy of the Bernoulli random variable with parameter p, Fano's inequality (Proposition 2.3.3 from Chapter 2) takes the following form:

Proposition 9.4.1 (Fano inequality)**.** *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(\mathbb{P}(\widehat{V} \neq V)) + \mathbb{P}(\widehat{V} \neq V)\log(|\mathcal{V}| - 1) \geq H(V \mid \widehat{V}). \tag{9.4.1}$$

Restating the results in Chapter 2, we also have the following convenient rewriting of Fano's inequality when V is uniform in $\mathcal{V}$ (recall Corollary 2.3.4).

Corollary 9.4.2. *Assume that V is uniform on $\mathcal{V}$ and $V \to X$. Then*

$$\inf_{\Psi} \mathbb{P}(\Psi(X) \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log |\mathcal{V}|}, \tag{9.4.2}$$

where the infimum is taken over all testing procedures Ψ.

By combining Corollary 9.4.2 with the reduction from estimation to testing in Proposition 9.2.1, we obtain the following result.

Proposition 9.4.3. *Let $\{\theta(P_v)\}_{v \in \mathcal{V}}$ be a 2δ-packing in the ρ-semimetric. Assume that V is uniform on the set $\mathcal{V}$, and conditional on $V = v$, we draw a sample $X \sim P_v$. Then the minimax risk has lower bound*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(V;X) + \log 2}{\log |\mathcal{V}|}\right).$$

To gain intuition for Proposition 9.4.3, we think of the lower bound as a function of the separation $\delta > 0$. As $\delta \downarrow 0$, the separation condition between the distributions P_v is relaxed and we expect the distributions P_v to be closer to one another. In this case, the hypothesis testing problem of distinguishing the P_v becomes more challenging, and the information $I(V;X)$ shrinks. Thus, we choose our packing $\theta(P_v)$ as a function of δ, finding the largest $\delta > 0$ making the mutual information small enough that

$$\frac{I(V;X) + \log 2}{\log |\mathcal{V}|} \leq \frac{1}{2}. \tag{9.4.3}$$

That is, we make the probability of error in the hypothesis testing lower bound (9.4.2) a constant, so the minimax lower bound at least $\Phi(\delta)/2$. We now explore techniques for achieving such results.

Mutual information and KL-divergence

Many techniques for upper bounding mutual information rely on its representation as the KL-divergence between multiple distributions. Given random variables V and X as in the preceding sections, if we let $P_{V,X}$ denote their joint distribution and P_V and P_X their marginals, then

$$I(V;X) = D_{\rm kl}\left(P_{X,V} \| P_X \times P_V\right),$$

where $P_X \times P_V$ denotes the distribution of (X, V) when the random variables are independent. By manipulating this definition, we can rewrite it into a form more convenient for our purposes.

Focusing on our setting of testing, let us assume that V is drawn from a prior distribution π (this may be a discrete or arbitrary distribution, though for simplicity we focus on the case when π is discrete). Let P_v denote the distribution of X conditional on $V = v$, as in Proposition 9.4.3. Then marginally, we know that X is drawn from the mixture distribution

$$\overline{P} := \sum_v \pi(v) P_v.$$

With this definition of the mixture, via algebraic manipulations, we have

$$I(V;X) = \sum_v \pi(v) D_{\rm kl}\left(P_v \| \overline{P}\right), \tag{9.4.4}$$

a representation that plays an important role in our subsequent derivations. To see equality (9.4.4), let μ be a base measure over $\mathcal{X}$ (assume w.l.o.g. that X has density $p(\cdot \mid v) = p_v(\cdot)$ conditional on $V = v$), and note that

$$I(V;X) = \sum_v \int_{\mathcal{X}} p(x \mid v)\pi(v) \log \frac{p(x \mid v)}{\sum_{v'} p(x \mid v')\pi(v')} d\mu(x) = \sum_v \pi(v) \int_{\mathcal{X}} p(x \mid v) \log \frac{p(x \mid v)}{\overline{p}(x)} d\mu(x).$$

Representation (9.4.4) makes it clear that if the distributions of the sample X conditional on V are all similar, then there is little information content. Returning to the discussion after Proposition 9.4.3, we have in this uniform setting that

$$\overline{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \quad \text{and} \quad I(V;X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right).$$

The mutual information is small if the typical conditional distribution P_v is difficult to distinguish—has small KL-divergence—from $\overline{P}$.

The local Fano method

The local Fano method is based on a weakening of the mixture representation of mutual information (9.4.4), then giving a uniform upper bound on divergences between all pairs of the conditional distributions P_v and $P_{v'}$. (This method is known in the statistics literature as the "generalized Fano method," a poor name, as it is based on a weak upper bound on mutual information.) Focusing on the case when V is uniform, the convexity of $-\log$ implies that

$$I(V;X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v \| \overline{P}\right) \le \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{\rm kl}\left(P_v \| P_{v'}\right). \tag{9.4.5}$$

In the local Fano method, we construct a *local packing*. This local packing approach is based on constructing a family of distributions P_v for $v \in \mathcal{V}$ defining a 2δ-packing (recall Section 9.2.1), meaning $\rho(\theta(P_v), \theta(P_{v'})) \ge 2\delta$ for all $v \ne v'$, while satisfying the uniform upper bound

$$D_{\rm kl}\left(P_v \| P_{v'}\right) \le \kappa^2 \delta^2 \quad \text{for all } v, v' \in \mathcal{V}, \tag{9.4.6}$$

for a fixed problem-dependent constant $\kappa > 0$. If we have the inequality (9.4.6), then so long as we can find a packing $\mathcal{V}$ such that

$$\log |\mathcal{V}| \ge 2(\kappa^2 \delta^2 + \log 2),$$

we are guaranteed the testing error condition (9.4.3), and hence the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \ge \frac{1}{2}\Phi(\delta).$$

Constructing local packings The main difficulty in using Fano's method is in the construction of these so-called "local" packings. The typical approach is to construct a packing $\mathcal{V}$ of a fixed set (in a vector space, say $\mathbb{R}^d$) with constant radius and constant distance. Then we scale elements of the packing by $\delta > 0$, which leaves the cardinality $|\mathcal{V}|$ identical, but allows us to choose δ in the separation in the packing and the uniform divergence bound (9.4.6). Corollary 9.2.3 and Lemma 9.2.4, for example, show we can construct exponentially large packings of fixed radius with cardinality $\log |\mathcal{V}| \gtrsim d$, giving dimension dependence.

We illustrate these techniques via two examples.

Example 9.4.4 (Normal mean estimation)**:** Consider the d-dimensional normal location family $\mathcal{N}_d = \{\mathsf{N}(\theta, \sigma^2 I_d) \mid \theta \in \mathbb{R}^d\}$; we wish to estimate the mean $\theta = \theta(P)$ in mean-squared error $\|\widehat{\theta} - \theta\|_2^2$. Let $\mathcal{V}$ be a 1/2-packing of the unit ℓ_2-ball with cardinality at least 2^d, as guaranteed by Corollary 9.2.3. (We assume for simplicity that $d \geq 2$.)

We construct our local packing. Fix $\delta > 0$, and for $v \in \mathcal{V}$, set $\theta_v = \delta v \in \mathbb{R}^d$. Then

$$\|\theta_v - \theta_{v'}\|_2 = \delta \left\|v - v'\right\|_2 \geq \frac{\delta}{2}$$

for each distinct pair $v, v' \in \mathcal{V}$. By applying the Fano minimax bound of Proposition 9.4.3, we see that for n normal observations $X_i \stackrel{\text{iid}}{\sim} P$,

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \left(\frac{1}{2} \cdot \frac{\delta}{2}\right)^2 \left(1 - \frac{I(V; X_1^n) + \log 2}{\log |\mathcal{V}|}\right) = \frac{\delta^2}{16}\left(1 - \frac{I(V; X_1^n) + \log 2}{d \log 2}\right).$$

Now note that for any pair v, v', if P_v is the normal distribution $\mathsf{N}(\theta_v, \sigma^2 I_d)$ we have

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = n \cdot D_{\rm kl}\left(\mathsf{N}(\delta v, \sigma^2 I_d) \| \mathsf{N}(\delta v', \sigma^2 I_d)\right) = n \cdot \frac{\delta^2}{2\sigma^2}\left\|v - v'\right\|_2^2,$$

where we have used Example 2.1.9 to compute the KL-divergence between normal distributions. As $\|v - v'\|_2 \leq 2$, we have the KL-divergence bound (9.4.6) with $\kappa^2 = 2n/\sigma^2$.

Combining our derivations, we have the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \frac{\delta^2}{16}\left(1 - \frac{2n\delta^2/\sigma^2 + \log 2}{d \log 2}\right). \tag{9.4.7}$$

Then by taking $\delta^2 = d\sigma^2 \log 2/(8n)$, we see that

$$1 - \frac{2n\delta^2/\sigma^2 + \log 2}{d \log 2} = 1 - \frac{1}{d} - \frac{1}{4} \geq \frac{1}{4}$$

by assumption that $d \geq 2$, and inequality (9.4.7) implies the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}_d), \|\cdot\|_2^2) \geq \frac{d\sigma^2 \log 2}{128n} \cdot \frac{1}{4} \gtrsim \frac{d\sigma^2}{n}.$$

While the constants are loose, we obtain the right scaling in d, n, and the variance σ^2; the sample mean attains the same risk. $\diamond$

Example 9.4.5 (Linear regression)**:** We show how local packings can give rate-optimal minimax bounds for linear regression problems. For a fixed matrix $X \in \mathbb{R}^{n \times d}$, we observe

$$Y = X\theta + \varepsilon,$$

where $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$ and $\theta \in \mathbb{R}^d$ may vary over $\mathbb{R}^d$. Let $\mathcal{P}$ denote the family of such linear regression problems, and assume for simplicity that $d \geq 32$.

The Gilbert-Varshamov bound (Lemma 9.2.4) provides a packing $\mathcal{V}$ of $\{-1, 1\}^d$ such that $\|v - v'\|_1 \geq d/2$ for distinct elements of $\mathcal{V}$, with $|\mathcal{V}| \geq \exp(d/8)$, which we use to develop the minimax lower bound. For $\delta > 0$, setting $\theta_v = \delta v$ yields separation

$$\|\theta_v - \theta_{v'}\|_2^2 = \delta^2 \sum_{j=1}^d (v_j - v_j')^2 = 2\delta^2 \left\|v - v'\right\|_1 \geq d\delta^2 \quad \text{for } v \neq v'.$$

Moreover, we have the upper bound

$$D_{\rm kl}\left(\mathsf{N}(X\theta_v, \sigma^2 I_n) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_n)\right) = \frac{1}{2\sigma^2}\left\|X(\theta_v - \theta_{v'})\right\|_2^2 \\ \le \frac{\delta^2}{2\sigma^2}\gamma_{\max}^2(X)\left\|v - v'\right\|_2^2 \le \frac{2d}{\sigma^2}\gamma_{\max}^2(X)\delta^2,$$

where $\gamma_{\max}(X)$ denotes the maximum singular value of X. Consequently, the bound (9.4.6) holds with $\kappa^2 \le 2d\gamma_{\max}^2(X)/\sigma^2$, and we have the minimax lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{4}\left(1 - \frac{I(V;Y) + \log 2}{\log|\mathcal{V}|}\right) \ge \frac{d\delta^2}{4}\left(1 - \frac{\frac{2d\gamma_{\max}^2(X)}{\sigma^2}\delta^2 + \log 2}{d/8}\right).$$

Now, if we choose

$$\delta^2 = \frac{\sigma^2}{64\gamma_{\max}^2(X)}, \quad \text{then} \quad 1 - \frac{8\log 2}{d} - \frac{16d\gamma_{\max}^2(X)\delta^2}{d} \ge 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2},$$

by assumption that $d \ge 32$. In particular, we obtain the lower bound

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|_2^2) \ge \frac{1}{512}\frac{\sigma^2 d}{\gamma_{\max}^2(X)} = \frac{1}{512}\frac{\sigma^2 d}{n}\frac{1}{\gamma_{\max}^2(\frac{1}{\sqrt{n}}X)},$$

for a convergence rate (roughly) of $\sigma^2 d/n$ after rescaling the singular values of X by $1/\sqrt{n}$. $\diamond$

This bound is sharp in terms of the dimension, dependence on n, and the variance σ^2, but it does not fully capture the dependence on X, as it depends only on the maximum singular value. When $X^\top X$ is nonsingular, an exact calculation (cf. [286]) shows that the problem has minimax value $\sigma^2 \operatorname{tr}((X^\top X)^{-1})$. Letting $\lambda_j(A)$ be the jth eigenvalue of a matrix A, we have

$$\sigma^2 \operatorname{tr}((X^\top X)^{-1}) = \frac{\sigma^2}{n}\operatorname{tr}((n^{-1}X^\top X)^{-1}) = \frac{\sigma^2}{n}\sum_{j=1}^d \frac{1}{\lambda_j(\frac{1}{n}X^\top X)} \\ \ge \frac{\sigma^2 d}{n}\min_j \frac{1}{\lambda_j(\frac{1}{n}X^\top X)} = \frac{\sigma^2 d}{n}\frac{1}{\gamma_{\max}^2(\frac{1}{\sqrt{n}}X)}.$$

Thus, the local Fano method captures most—but not all—of the difficulty of the problem.

9.4.2 A distance-based Fano method

While the testing lower bound (9.4.2) is sufficient for proving lower bounds for many estimation problems, for the sharpest results it can sometimes require a delicate construction of a well-separated packing (e.g. [100, 159]). Extensions of inequalities (9.4.1) and (9.4.2) that more directly yield bounds on estimation error allow simpler proofs of a variety of minimax lower bounds. Specifically, let $\rho_{\mathcal{V}}$ be a distance on $\mathcal{V}$, consider the estimation error $\rho_{\mathcal{V}}(\widehat{V}, V)$. We can then provide analogues of the lower bounds (9.4.1) and (9.4.2) that replace the testing error with the tail probability $\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$, which directly lower bounds the expectation $\mathbb{E}[\rho_{\mathcal{V}}(\widehat{V}, V)] \ge t\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$.

We begin with the distance-based analogue of the discrete Fano inequality in Proposition 9.4.1. Let V be a random variable supported on a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \ge 2$, and let $\rho : \mathcal{V} \times \mathcal{V} \to$

$\mathbb{R}$ be a function defined on $\mathcal{V} \times \mathcal{V}$. In the usual setting, the function ρ is a metric on the space $\mathcal{V}$, but our theory applies to general functions. For a given scalar $t \geq 0$, the maximum and minimum *neighborhood sizes at radius* t are given by

$$N_t^{\max} := \max_{v \in \mathcal{V}} \left\{ \mathrm{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\} \quad \text{and} \quad N_t^{\min} := \min_{v \in \mathcal{V}} \left\{ \mathrm{card}\{v' \in \mathcal{V} \mid \rho(v, v') \leq t\} \right\}. \tag{9.4.8}$$

Defining the error probability $P_t = \mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t)$, we then have the following generalization of Fano's inequality:

Proposition 9.4.6. *For any Markov chain $V \to X \to \widehat{V}$, we have*

$$h_2(P_t) + P_t \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log N_t^{\max} \geq H(V \mid \widehat{V}). \tag{9.4.9}$$

The proof is similar to our earlier proof of Proposition 2.3.3, so we defer it to Section 9.6.1. (See also exercise 9.3, which extends the proposition to an even more general distance-based Bayesian Fano inequality.) It reduces to the standard form of Fano's inequality (9.4.1) in a special case: take $\rho_{\mathcal{V}}$ to be the 0-1 metric, meaning that $\rho_{\mathcal{V}}(v, v') = 0$ if $v = v'$ and 1 otherwise. Then setting $t = 0$ in Proposition 9.4.6, we have $P_0 = \mathbb{P}(\widehat{V} \neq V)$ and $N_0^{\min} = N_0^{\max} = 1$, whence inequality (9.4.9) reduces to inequality (9.4.1). Other weakenings allow cleaner statements (see Section 9.6.2 for a proof):

Corollary 9.4.7. *If V is uniform on $\mathcal{V}$ and $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$, then*

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}. \tag{9.4.10}$$

Inequality (9.4.10) analogizes the classical mutual-information based form of Fano's inequality (Corollary 2.3.4). The main difference is the replacement of the usual cardinality $|\mathcal{V}|$ by the ratio $|\mathcal{V}|/N_t^{\max}$. This quantity measures the number of distinguishable "regions" in the space $\mathcal{V}$. While this construction is similar in spirit to the usual construction of packing sets in the reduction from estimation to testing (cf. Section 9.2.1), the bound (9.4.10) allows skipping the packing set construction. We can directly compute $I(V; X)$ where V is uniform over the full space, as opposed to the mutual information $I(V'; X)$ for V' uniform over a packing set inside $\mathcal{V}$. In some cases, the former calculation can be much simpler, as we illustrate in examples to follow.

We provide a few minimax lower bounds that follow from Proposition 9.4.6 and Corollary 9.4.7. The new reduction from estimation to testing generalizes the classical one, as we map the estimate to a test that allows errors. Consider as in the standard reduction of estimation to testing in Section 9.2.1 a family of distributions $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by a finite set $\mathcal{V}$. This family induces an associated collection of parameters $\{\theta_v := \theta(P_v)\}_{v \in \mathcal{V}}$. Given a function $\rho_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and a scalar t, we define the separation $\delta(t)$ of this set relative to the metric ρ on Θ via

$$\delta(t) = \inf \left\{ \rho(\theta_v, \theta_{v'}) \mid \rho_{\mathcal{V}}(v, v') > t \right\}. \tag{9.4.11}$$

When $t = 0$ and $\rho_{\mathcal{V}}$ is the discrete metric, this definition reduces to that of a packing set: we are guaranteed that $\rho(\theta_v, \theta_{v'}) \geq \delta(0)$ for all distinct pairs $v \neq v'$, as in the classical approach to minimax lower bounds. Allowing $t > 0$ lends greater flexibility, since only certain pairs θ_v and $\theta_{v'}$ need to be well-separated.

Given a set $\mathcal{V}$ and associated separation function (9.4.11), we assume the canonical estimation setting (Section 9.2.1): nature chooses $V \in \mathcal{V}$ uniformly at random, and conditioned on $V = v$, a sample X is drawn from the distribution P_v. We then have the following corollary of Proposition 9.4.6, whose argument is completely identical to that for inequality (9.2.2):

Corollary 9.4.8. *Given V uniformly distributed over $\mathcal{V}$ with separation function $\delta(t)$, we have*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \ge \Phi\left(\frac{\delta(t)}{2}\right)\left[1 - \frac{I(X;V) + \log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}\right] \quad \text{for all } t. \tag{9.4.12}$$

Using the discrete metric $\rho_{\mathcal{V}}(v, v') = \mathbf{1}\{v \neq v'\}$ and taking $t = 0$ in the lower bound (9.4.12) gives the classical Fano lower bound on the minimax risk based on constructing a packing.

We turn to two examples to illustrate the corollary.

Example 9.4.9 (Normal regression model)**:** We revisit the linear regression model (Example 9.4.5), showing how the distance-based method gives a sharper lower bound that depends on the Frobenius norm of the design matrix rather than its maximum singular value. We have $Y = X\theta + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2 I_n)$ and $X \in \mathbb{R}^{n \times d}$ is known. Our family of distributions is

$$\mathcal{P}_X := \left\{Y \sim \mathsf{N}(X\theta, \sigma^2 I_n) \mid \theta \in \mathbb{R}^d\right\} = \left\{Y = X\theta + \varepsilon \mid \varepsilon \sim \mathsf{N}(0, \sigma^2 I_n), \theta \in \mathbb{R}^d\right\}.$$

Then there is a universal (numerical) constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge c\frac{\sigma^2 d^2}{\|X\|_{\mathrm{Fr}}^2} \ge \frac{c}{\gamma_{\max}(X/\sqrt{n})^2} \cdot \frac{\sigma^2 d}{n}, \tag{9.4.13}$$

where $\gamma_{\max}$ denotes the maximum singular value. This inequality is nearly the sharpest known bound proved via Fano inequality-based methods [100], but the technique is essentially direct.

To see inequality (9.4.13), let $\mathcal{V} = \{-1, 1\}^d$ be the d-dimensional hypercube and $\theta_v = \delta v$ for a fixed $\delta > 0$. Taking $\rho_{\mathcal{V}}$ to be the Hamming metric on $\mathcal{V}$ and ρ the usual ℓ_2-norm, the separation (9.4.11) satisfies $\delta(t) > \max\{\sqrt{t}, 1\}\delta$. For $t \le \lceil d/3 \rceil$, the neighborhood size satisfies

$$N_t^{\max} = \sum_{\tau=0}^{t} \binom{d}{\tau} \le 2\binom{d}{t} \le 2\left(\frac{de}{t}\right)^t.$$

Consequently, for $t \le d/6$ and $d \ge 12$, the ratio $|\mathcal{V}|/N_t^{\max}$ satisfies

$$\log \frac{|\mathcal{V}|}{N_t^{\max}} \ge d\log 2 - \log 2\binom{d}{t} \ge d\log 2 - \frac{d}{6}\log(6e) - \log 2 = d\log\frac{2}{2^{1/d}\sqrt[6]{6e}} > \max\left\{\frac{d}{6}, \log 4\right\}$$

(The case $2 \le d < 12$ follows directly.) By taking $t = \lfloor d/6 \rfloor$ we obtain via Corollary 9.4.8 that

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4}\left(1 - \frac{I(Y;V) + \log 2}{\max\{d/6, 2\log 2\}}\right).$$

But of course, for V uniform on $\mathcal{V}$, we have $\mathbb{E}[VV^\top] = I_d$, and thus for $V, V' \overset{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V})$,

$$\begin{aligned} I(Y;V) &\le \frac{1}{|\mathcal{V}|^2}\sum_{v \in \mathcal{V}}\sum_{v' \in \mathcal{V}} D_{\mathrm{kl}}\left(\mathsf{N}(X\theta_v, \sigma^2 I_n) \| \mathsf{N}(X\theta_{v'}, \sigma^2 I_n)\right) \\ &= \frac{\delta^2}{2\sigma^2}\mathbb{E}\left[\left\|XV - XV'\right\|_2^2\right] = \frac{\delta^2}{\sigma^2}\|X\|_{\mathrm{Fr}}^2. \end{aligned}$$

Substituting this into the preceding minimax bound, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_X), \|\cdot\|_2^2) \ge \frac{\max\{\lfloor d/6 \rfloor, 2\}\delta^2}{4}\left(1 - \frac{\delta^2\|X\|_{\mathrm{Fr}}^2/\sigma^2 + \log 2}{\max\{d/6, 2\log 2\}}\right).$$

Choosing $\delta^2 \asymp d\sigma^2/\|X\|_{\mathrm{Fr}}^2$ gives the result (9.4.13). $\diamond$

As a second example, we can revisit the general M-estimation results in Section 5.3. The construction here extends that in Example 9.4.9 to settings where there is no "true" parameter; we leave working out the details as exercises. (See Exercises 9.8 9.10, and 9.11.)

Example 9.4.10 (Local minimax lower bounds for M-estimation)**:** Let $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ be a convex loss as in Section 5.3, and let it and the population loss $L(\theta) := \mathbb{E}_{P_0}[\ell(\theta, Z)]$ satisfy the conditions of Assumption A.5.1. Let $\theta_0 = \operatorname{argmin} L(\theta)$ have Hessian $H = \nabla^2 L(\theta_0)$ and covariance $\Sigma = \operatorname{Cov}_0(\nabla \ell(\theta_0; Z))$, so that the results in Section 5.3 (Corollary 5.3.10 specifically) show that the empirical risk minimizer satisfies

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta_0\right\|_2^2\right] \lesssim \frac{\operatorname{tr}(H^{-1}\Sigma H^{-1})}{n} \quad \text{and} \quad \mathbb{E}\left[\left\|\widehat{\theta}_n - \theta_0\right\|_{(H^{-1}\Sigma H^{-1})^{-1}}^2\right] \lesssim \frac{d}{n},$$

where we use the notation $\|x\|_A^2 = x^\top A x$. We can show these convergence results are sharp by constructing an appropriate packing of the parameter space, induced by distributions P on $\mathcal{Z}$, and showing that the information the sample Z_1^n carries about the particular distribution is limited, allowing us to apply the local Fano method.

We first develop the underlying distribution family. Let $\delta \geq 0$ to be chosen. For any bounded mean-zero function $g : \mathcal{Z} \to \mathbb{R}^d$, Exercise 9.8 shows how to "tilt" P_0 by $u^\top g$ to construct a collection of distributions $\{P_u\}_{u \in \mathbb{R}^d}$, indexed by $u \in \mathbb{R}^d$, such that

$$D_{\rm kl}\left(P_u \| P_0\right) = \frac{1}{2} u^\top \operatorname{Cov}_0(g(Z)) u + o(\|u\|^2)$$

for u near 0. Then as usual drawing a random element $V \sim \mathsf{Uniform}(\{\pm 1\}^d)$ and drawing $Z_1^n \stackrel{\rm iid}{\sim} P_{\delta v}$ conditional on $V = v$, we have the mutual information bound

$$I(Z_1^n; V) \leq \frac{n\delta^2}{2} \operatorname{tr}(\operatorname{Cov}_0(g(Z))) + n \cdot o(\delta^2)$$

as $\delta \to 0$. For the separation of the induced parameters, let

$$\theta_v = \operatorname*{argmin}_\theta \mathbb{E}_{P_{\delta v}}[\ell(\theta, Z)].$$

Then (see Exercise 9.11), these parameters satisfy

$$\theta_v = \theta(P_0) + \delta H^{-1} \operatorname{Cov}_0(\nabla \ell(\theta_0, Z), g(Z)) v + O(\delta^2).$$

Lastly we choose the function g by taking

$$g(Z) = -(H^{-1}\Sigma H^{-1})^{-1/2} H^{-1} \nabla \ell(\theta_0, Z),$$

which is mean-zero and bounded so long as ℓ is Lipschitz near θ_0 and satisfies $\operatorname{Cov}_0(g(Z)) = I_d$ and $\operatorname{tr}(\operatorname{Cov}_0(g(Z))) = d$. We therefore have $I(Z_1^n; V) \lesssim n\delta^2 d$ and $\theta_v - \theta_0 = \delta(H^{-1}\Sigma H^{-1})^{1/2} v + O(\delta^2)$. This gives a separation in the metric induced by the matrix $H\Sigma^{-1}H$, so applying Corollary 9.4.7, there exists a numerical constant $c > 0$ such that for all suitably large n

$$\mathbb{P}\left(\left\|\widehat{\theta}_n - \theta_V\right\|_{H\Sigma^{-1}H} \geq c\sqrt{d/n}\right) \geq \frac{1}{4}.$$

Exercises 9.10 and 9.11 work through the remaining details. $\diamond$

9.5 Assouad's method

Assouad's method provides a somewhat different technique for proving lower bounds. Instead of reducing the estimation problem to a multiple hypothesis test or simpler estimation problem, as with the Le Cam and Fano methods, here we transform the original estimation problem into multiple binary hypothesis testing problems, using the structure of the problem in an essential way. Assouad's method applies only problems where the loss can be related to identification of individual points on a hypercube. In simple or standard problems, Assouad's method rarely provides stronger lower bounds than the local Fano methods; its power lies in its applications to *adaptive* problems, where one may choose points at which to query a statistical model. We develop this idea in the exercises (see Exercise 9.6), and also leverage it in Chapter 18, Section 18.5.2 to prove lower bounds for bandit problems, where one must balance adaptive exploration of a function with performance.

9.5.1 Well-separated problems

To describe the method, we begin by encoding a notion of separation and loss, similar to what we did in the classical reduction of estimation to testing. For some $d \in \mathbb{N}$, let $\mathcal{V} = \{-1, 1\}^d$, and let us consider a family $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ indexed by the hypercube. We say that the family P_v induces a 2δ-*Hamming separation* for the loss $\Phi \circ \rho$ if there exists a function $\widehat{\mathsf{v}} : \Theta \to \{-1, 1\}^d$ satisfying

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathsf{v}}(\theta)]_j \neq v_j\right\}. \tag{9.5.1}$$

That is, we can take the parameter θ and test the individual indices via $\widehat{\mathsf{v}}$.

Example 9.5.1 (Estimation in ℓ_1-error)**:** Suppose we have a family of multivariate Laplace distributions on $\mathbb{R}^d$ with density proportional to $p(x) \propto \exp(-\left\|x - \mu\right\|_1)$, and we wish to estimate the mean in ℓ_1-distance. For $v \in \{-1, 1\}^d$ and a $\delta > 0$ let p_v be the density

$$p_v(x) = \frac{1}{2^d} \exp\left(-\left\|x - \delta v\right\|_1\right),$$

which has mean $\theta(P_v) = \delta v$. Under the ℓ_1-loss, we have for any $\theta \in \mathbb{R}^d$ that

$$\left\|\theta - \theta(P_v)\right\|_1 = \sum_{j=1}^{d} \left|\theta_j - \delta v_j\right| \geq \delta \sum_{j=1}^{d} \mathbf{1}\left\{\operatorname{sign}(\theta_j) \neq v_j\right\},$$

so that this family induces a δ-Hamming separation for the ℓ_1-loss. $\diamond$

9.5.2 From estimation to multiple binary tests

As in the standard reduction from estimation to testing, we consider the following random process: nature chooses a vector $V \in \{-1, 1\}^d$ uniformly at random, then draws the sample $X \sim P_v$ conditional on $V = v$. If we let $\mathbb{P}_{\pm j}$ denote the joint distribution over the random index V and X conditional on the jth coordinate $V_j = \pm 1$, we obtain the following sharper version of Assouad's lemma [32]; we provide a proof in Section 9.6.3 to follow.

Theorem 9.5.2. *Under the conditions of the previous paragraph, we have*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\Psi} \left[\mathbb{P}_{+j}(\Psi(X) \neq +1) + \mathbb{P}_{-j}(\Psi(X) \neq -1)\right].$$

While Theorem 9.5.2 requires conditions on the loss Φ and metric ρ for the separation condition (9.5.1) to hold, it is sometimes easier to apply than Fano's method. It is closely related to Le Cam's method, discussed previously, as we see that if we define $P_{+j} = 2^{1-d} \sum_{v:v_j=1} P_v$ (and similarly for $-j$), Theorem 9.5.2 is equivalent to

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right]. \tag{9.5.2}$$

There are standard weakenings of the lower bound (9.5.2). We give one such weakening. Because the total variation is convex, if we define $P_{v,+j}$ to be the distribution obtained from P_v by forcing the jth coordinate to $+1$ and leaving the other coordinates inherited from v (and similarly for $P_{v,-j}$), we have

$$P_{+j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} \quad \text{and} \quad P_{-j} = \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,-j}.$$

Thus, by the triangle inequality,

$$\begin{aligned} \|P_{+j} - P_{-j}\|_{\mathrm{TV}} &= \left\| \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} P_{v,+j} - P_{v,-j} \right\|_{\mathrm{TV}} \\ &\leq \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}} \leq \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}. \end{aligned}$$

Then as long as the loss satisfies the per-coordinate separation (9.5.1), we obtain the following:

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq d\delta \left(1 - \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}\right). \tag{9.5.3}$$

This common version of Assouad's lemma sometimes too brutally controls $\|P_{+j} - P_{-j}\|_{\mathrm{TV}}$, so other versions are useful. For example, by the Cauchy-Schwarz inequality and convexity of the variation-distance, we have

$$\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}} \leq \sqrt{d} \left(\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \right)^{1/2} \leq \sqrt{d} \left(\sum_{j=1}^{d} \frac{1}{2^d} \sum_{v} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}^2 \right)^{\frac{1}{2}},$$

and consequently we have a "quadratic" version of inequality (9.5.2):

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta d \left[1 - \left(\frac{1}{2^d d} \sum_{j=1}^{d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}^2 \right)^{\frac{1}{2}}\right]. \tag{9.5.4}$$

Regardless of whether we use the sharper version (9.5.2) or weakened versions (9.5.3) or (9.5.4), the technique is essentially the same. We seek distributions P_v so that the probability of making a mistake in each of the implied hypothesis tests of Theorem 9.5.2 is of constant order, equivalently, the variation distance is small enough that (say) $\|P_{+j} - P_{-j}\|_{\mathrm{TV}} \leq 1/2$ for all j. Once this is satisfied, we obtain a minimax lower bound of the form

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \left[1 - \frac{1}{2}\right] = \frac{d\delta}{2}.$$

9.5.3 Example applications of Assouad's method

We provide two example applications of Assouad's method. The first revisits Example 9.4.4, and for the second, we consider estimation in a logistic regression problem.

Example 9.5.3 (Normal mean estimation)**:** For some $\sigma^2 > 0$ and $d \in \mathbb{N}$, consider estimation of the mean parameter θ for the normal location family $\mathcal{N} := \{\mathsf{N}(\theta, \sigma^2 I_d)\}_{\theta \in \mathbb{R}^d}$. Assouad's method will imply the (to constant factors) sharp lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{16n}. \tag{9.5.5}$$

We proceed in the usual way. Fix $\delta > 0$ and define $\theta_v = \delta v$ for $v \in \mathcal{V} = \{-1, 1\}^d$, taking $P_v = \mathsf{N}(\theta_v, \sigma^2 I_d)$ to be the normal distribution with mean θ_v. The loss function decomposes on coordinates: we have $\|\theta - \theta_v\|_2^2 \geq \delta^2 \sum_{j=1}^d \mathbf{1}\left\{\mathrm{sign}(\theta_j) \neq v_j\right\}$. The family P_v thus induces a δ^2-Hamming separation for $\|\cdot\|_2^2$, so Assouad's method (9.5.2) implies

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}\right],$$

where $P_{\pm j}^n = 2^{1-d} \sum_{v : v_j = \pm 1} P_v^n$. It remains to provide upper bounds on $\|P_{+j}^n - P_{-j}^n\|_{\mathrm{TV}}$. By the convexity of $\|\cdot\|_{\mathrm{TV}}^2$ and Pinsker's inequality, we have

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}^2 \leq \max_{d_{\mathrm{ham}}(v,v') \leq 1} \|P_v^n - P_{v'}^n\|_{\mathrm{TV}}^2 \leq \frac{1}{2} \max_{d_{\mathrm{ham}}(v,v') \leq 1} D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right).$$

But of course, for any v and v' differing in only 1 coordinate,

$$D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right) = \frac{n}{2\sigma^2} \|\theta_v - \theta_{v'}\|_2^2 = \frac{2n}{\sigma^2}\delta^2,$$

giving the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \sqrt{n\delta^2/\sigma^2}\right].$$

Choosing $\delta^2 = \sigma^2/4n$ gives the claimed lower bound (9.5.5). $\diamond$

Example 9.5.4 (Logistic regression)**:** Consider the logistic regression model, where we have known regressors $X_i \in \mathbb{R}^d$ and unknown $\theta \in \mathbb{R}^d$; we observe $Y_i \in \{-1, 1\}$ with distribution

$$P(Y_i = y \mid X_i, \theta) = \frac{1}{1 + \exp(-y X_i^\top \theta)} \quad \text{for } y \in \{-1, 1\}.$$

Denote this family by $\mathcal{P}_{\mathsf{log}}$, and for $P \in \mathcal{P}_{\mathsf{log}}$, let $\theta(P)$ be the parameter vector θ. We measure squared ℓ_2 error. As in Example 9.5.3, if we choose some $\delta > 0$ and for each $v \in \{-1, 1\}^d$ set $\theta_v = \delta v$, then we have the Hamming separation $\|\theta - \theta_v\|_2^2 \geq \delta^2 \sum_{j=1}^d \mathbf{1}\left\{\mathrm{sign}(\theta_j) \neq v_j\right\}$. Let P_v^n denote the distribution of the n independent observations Y_i when $\theta = \theta_v$. Then by Assouad's lemma (and its weakening (9.5.4)),

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \geq \frac{d\delta^2}{2}\left[1 - \left(\frac{1}{d}\sum_{j=1}^d \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}}\right]. \tag{9.5.6}$$

It remains to bound $\|P_{v,+j}^n - P_{v,-j}^n\|_{\rm TV}^2$. Use the shorthands $p_v(x) = 1/(1+\exp(\delta x^\top v))$ and let $D_{\rm kl}\left(p\|q\right)$ be the binary KL-divergence between $\mathsf{Bernoulli}(p)$ and $\mathsf{Bernoulli}(q)$ distributions. Then Pinsker's inequality (Corollary 9.2.2) implies that for any v, v',

$$\begin{aligned}\|P_v^n - P_{v'}^n\|_{\rm TV}^2 &\le \frac{1}{4}[D_{\rm kl}\left(P_v^n\|P_{v'}^n\right) + D_{\rm kl}\left(P_{v'}^n\|P_v^n\right)] \\ &= \frac{1}{4}\sum_{i=1}^n \left[D_{\rm kl}\left(p_v(X_i)\|p_{v'}(X_i)\right) + D_{\rm kl}\left(p_{v'}(X_i)\|p_v(X_i)\right)\right].\end{aligned}$$

To upper bound the final KL-divergence, when $p_c = 1/(1+e^c)$, we claim that

$$D_{\rm kl}\left(p_a\|p_b\right) + D_{\rm kl}\left(p_b\|p_a\right) \le (a-b)^2 \tag{9.5.7}$$

for any $a, b \in \mathbb{R}$. Deferring the proof of claim (9.5.7), we immediately see that

$$\|P_v^n - P_{v'}^n\|_{\rm TV}^2 \le \frac{\delta^2}{4}\sum_{i=1}^n \left(X_i^\top(v-v')\right)^2.$$

Using inequality (9.5.6) for motivation, the preceding display implies

$$\frac{1}{2^d d}\sum_{j=1}^d \sum_{v\in\{-1,1\}^d} \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\rm TV}^2 \le \frac{\delta^2}{4d}\frac{1}{2^d}\sum_{v\in\{-1,1\}^d}\sum_{j=1}^d\sum_{i=1}^n (2X_{ij})^2 = \frac{\delta^2}{d}\|X\|_{\rm Fr}^2,$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix with rows $X_i^\top$. Thus

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{2}\left[1 - \left(\frac{\delta^2}{d}\|X\|_{\rm Fr}^2\right)^{\frac{1}{2}}\right].$$

Setting $\delta^2 = d/(4\|X\|_{\rm Fr}^2)$, we obtain

$$\mathfrak{M}_n(\theta(\mathcal{P}_{\mathsf{log}}), \|\cdot\|_2^2) \ge \frac{d\delta^2}{4} = \frac{d^2}{16\|X\|_{\rm Fr}^2} = \frac{d}{n}\cdot\frac{1}{16\frac{1}{dn}\sum_{i=1}^n\|X_i\|_2^2}.$$

That is, we have a minimax lower bound scaling roughly as d/n for logistic regression.

We return to prove the claim (9.5.7). By a straightforward expansion,

$$D_{\rm kl}\left(p_a\|p_b\right) + D_{\rm kl}\left(p_b\|p_a\right) = (p_a - p_b)\log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right).$$

Because $p_a/(1-p_a) = e^{-a}$ and $(1-p_b)/p_b = e^b$, we obtain

$$D_{\rm kl}\left(p_a\|p_b\right) + D_{\rm kl}\left(p_b\|p_a\right) = \left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right)\log\left(e^{b-a}\right) = (b-a)\left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right).$$

Assume without loss of generality that $b \ge a$. Noting that $e^x \ge 1+x$ by convexity, we have

$$\frac{1}{1+e^a} - \frac{1}{1+e^b} = \frac{e^b - e^a}{(1+e^a)(1+e^b)} \le \frac{e^b - e^a}{e^b} = 1 - e^{a-b} \le 1 - (1+(a-b)) = b - a,$$

yielding claim (9.5.7). $\diamond$

9.6 Deferred proofs

9.6.1 Proof of Proposition 9.4.6

Our argument for proving the proposition parallels that of the classical Fano inequality. Letting E be a $\{0, 1\}$-valued indicator variable for the event $\rho_{\mathcal{V}}(\widehat{V}, V) \le t$, we compute the entropy $H(E, V \mid \widehat{V})$ in two different ways. On one hand, by the chain rule for entropy, we have

$$H(E, V \mid \widehat{V}) = H(V \mid \widehat{V}) + \underbrace{H(E \mid V, \widehat{V})}_{=0}, \tag{9.6.1}$$

where the final term vanishes since E is $(V, \widehat{V})$-measurable. On the other hand, we also have

$$H(E, V \mid \widehat{V}) = H(E \mid \widehat{V}) + H(V \mid E, \widehat{V}) \le H(E) + H(V \mid E, \widehat{V}),$$

using the fact that conditioning reduces entropy. Applying the definition of conditional entropy yields

$$H(V \mid E, \widehat{V}) = \mathbb{P}(E = 0) H(V \mid E = 0, \widehat{V}) + \mathbb{P}(E = 1) H(V \mid E = 1, \widehat{V}),$$

and we upper bound each of these terms separately. For the first term, we have

$$H(V \mid E = 0, \widehat{V}) \le \log(|\mathcal{V}| - N_t^{\min}),$$

since conditioned on the event $E = 0$, the random variable V may take values in a set of size at most $|\mathcal{V}| - N_t^{\min}$. For the second, we have

$$H(V \mid E = 1, \widehat{V}) \le \log N_t^{\max},$$

since conditioned on $E = 1$, or equivalently on the event that $\rho_{\mathcal{V}}(\widehat{V}, V) \le t$, we are guaranteed that V belongs to a set of cardinality at most $N_t^{\max}$.

Combining the pieces and noting $\mathbb{P}(E = 0) = P_t$, we have proved that

$$H(E, V \mid \widehat{V}) \le H(E) + P_t \log \left(|\mathcal{V}| - N_t^{\min}\right) + (1 - P_t) \log N_t^{\max}.$$

Combining this inequality with our earlier equality (9.6.1), we see that

$$H(V \mid \widehat{V}) \le H(E) + P_t \log(|\mathcal{V}| - N_t^{\min}) + (1 - P_t) \log N_t^{\max}.$$

Since $H(E) = h_2(P_t)$, the claim (9.4.9) follows.

9.6.2 Proof of Corollary 9.4.7

By the information-processing inequality, we have $I(V; \widehat{V}) \le I(V; X)$, and hence $H(V \mid X) \le H(V \mid \widehat{V})$. Since $h_2(P_t) \le \log 2$, inequality (9.4.9) implies that

$$H(V \mid X) - \log N_t^{\max} \le H(V \mid \widehat{V}) - \log N_t^{\max} \le \mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}} + \log 2.$$

Rearranging the preceding equations yields

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \ge \frac{H(V \mid X) - \log N_t^{\max} - \log 2}{\log \frac{|\mathcal{V}| - N_t^{\min}}{N_t^{\max}}}. \tag{9.6.2}$$

Note that this bound holds without any assumptions on the distribution of V.

By definition, we have $I(V;X) = H(V) - H(V \mid X)$. When V is uniform on $\mathcal{V}$, we have $H(V) = \log|\mathcal{V}|$, and hence $H(V \mid X) = \log|\mathcal{V}| - I(V;X)$. Substituting this relation into the bound (9.6.2) yields the inequality

$$\mathbb{P}(\rho_{\mathcal{V}}(\widehat{V}, V) > t) \geq \frac{\log \frac{|\mathcal{V}|}{N_t^{\max}}}{\log \frac{|\mathcal{V}|-N_t^{\min}}{N_t^{\max}}} - \frac{I(V;X)+\log 2}{\log \frac{|\mathcal{V}|-N_t^{\min}}{N_t^{\max}}} \geq 1 - \frac{I(V;X)+\log 2}{\log \frac{|\mathcal{V}|}{N_t^{\max}}}.$$

9.6.3 Proof of Theorem 9.5.2

Fix an (arbitrary) estimator $\widehat{\theta}$. By assumption (9.5.1), we have

$$\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\left\{[\widehat{\mathrm{v}}(\theta)]_j \neq v_j\right\}.$$

Taking expectations, we see that

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\Phi(\rho(\widehat{\theta}(X), \theta(P)))\right] &\geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}\left[\Phi(\rho(\widehat{\theta}(X), \theta_v))\right] \\ &\geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} 2\delta \sum_{j=1}^{d} \mathbb{E}_{P_v}\left[\mathbf{1}\left\{[\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right\}\right]\end{aligned}$$

as the average is smaller than the maximum of a set and using the separation assumption (9.5.1). Recalling the definition of the mixtures $\mathbb{P}_{\pm j}$ as the joint distribution of V and X conditional on $V_j = \pm 1$, we swap the summation orders to see that

$$\begin{aligned}\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) &= \frac{1}{|\mathcal{V}|}\sum_{v:v_j=1} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{|\mathcal{V}|}\sum_{v:v_j=-1} P_v\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) \\ &= \frac{1}{2}\mathbb{P}_{+j}\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right) + \frac{1}{2}\mathbb{P}_{-j}\left([\widehat{\mathrm{v}}(\widehat{\theta})]_j \neq v_j\right).\end{aligned}$$

This gives the statement claimed in the theorem, while taking an infimum over all testing procedures $\Psi : \mathcal{X} \to \{-1, +1\}$ gives the claim (9.5.2).

9.7 Bibliography and further reading

The program of measuring optimality of optimization and estimation procedures via their risk dates back to Wald [424, 425, 426] in the late 1930s and 1940s, who began by developing Bayesian criteria for minimum risk in hypothesis testing problems, and then built off of von Neumann's approach [415] to develop the minimax criterion. From these foundations, research in statistics, fairly independently of information theoretic techniques, developed a number of tools for fundamental limits of a quite distinct flavor from those we present here. In particular, often in (asymptotic) statistics, one shows that solving an estimation problem is asymptotically (as $n \to \infty$) equivalent to estimating a normal mean, for which there are precise optimality results [278, 279, 226, 280, 282, 407, 287]. Our focus here, of course, is on the information-theoretic approaches, and those we have developed have been influential for some forty years.

Fano's method in particular has played a central role in minimax theory, beginning with the pioneering work of Has'minskii and Ibragimov [228, 246]. The general recipe for proving information-theoretic minimax lower bounds, which begins by reducing estimation to testing, then proving a lower bound on the testing error, originates out of this work. Work following this initial push continues [71, 435, 433, 356, 221, 9, 100, 421]. Yu [435] provides an overview of the techniques, and motivates our delineation of the Le Cam, Fano, and Assouad methods, and Wainwright [422, Chapter 14] also provides an excellent overview of the techniques for minimax lower bounds. More recent work on fundamental limits often considers problems with constraints on the estimators—rather than the unrestricted estimators that this section investigates—which we discuss in subsequent chapters.

Assouad [32] introduced the eponymous Assouad method (others, of course, named it after him). The particular form we present here slightly extends of the method there, though other work prefigures it [23, 161, 156]. The recognition that the Assouad method is particularly well-suited to adaptive estimation problems appears in the work of Arias-Castro, Candés, and Davenport [21, 23], who show the difficulties of applying the Fano method in adaptive linear regression problems, developing applications and extensions of the Assouad method to address the challenge.

9.8 Exercises

Exercise 9.1: In the context of inequality (9.2.1), show that if $U \in \mathcal{U}$ is a source of external randomness (independent of all quantities in the problem), then

$$\inf_{\widehat{\theta}} \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1^n, U), \theta(P_v)))\right] = \inf_{\widehat{\theta}} \sum_v \pi(v) \mathbb{E}_{P_v} \left[\Phi(\rho(\widehat{\theta}(X_1^n, U), \theta(P_v)))\right],$$

where the first infimum is over (measurable) functions $\widehat{\theta} : \mathcal{X}^n \times \mathcal{U} \to \Theta$, and the second is over functions of $\mathcal{X}^n$. What does this say about the risk of randomized procedures?

Exercise 9.2: In the context of Example 9.3.1, let $\epsilon > 0$ and $\theta \in [-1, 1]$, and $\mathcal{B}_{\theta,\epsilon}$ consist of the distributions on $\{-1, 1\}$ satisfying $|\mathbb{E}[X] - \theta| \le \epsilon$. Use Hellinger distances to show that

$$\liminf_n \sqrt{n} \cdot \mathfrak{M}_n \left(\mathcal{B}_{\theta,\epsilon}, |\cdot|\right) \gtrsim \sqrt{1 - \theta^2}.$$

Exercise 9.3 (A generalized version of Fano's inequality; cf. Proposition 9.4.6)**:** Let $\mathcal{V}$ and $\widehat{\mathcal{V}}$ be arbitrary sets, and suppose that π is a (prior) probability measure on $\mathcal{V}$, where V is distributed according to π. Let $V \to X \to \widehat{V}$ be Markov chain, where V takes values in $\mathcal{V}$ and $\widehat{V}$ takes values in $\widehat{\mathcal{V}}$. Let $\mathcal{N} \subset \mathcal{V} \times \widehat{\mathcal{V}}$ denote a measurable subset of $\mathcal{V} \times \widehat{\mathcal{V}}$ (a collection of neighborhoods), and for any $\widehat{v} \in \widehat{\mathcal{V}}$, denote the slice

$$\mathcal{N}_{\widehat{v}} := \{v \in \mathcal{V} : (v, \widehat{v}) \in \mathcal{N}\}. \tag{9.8.1}$$

That is, $\mathcal{N}$ denotes the neighborhoods of points v for which we do not consider a prediction $\widehat{v}$ for v to be an error, and the slices (9.8.1) index the neighborhoods. Define the "volume" constants

$$p^{\max} := \sup_{\widehat{v}} \pi(V \in \mathcal{N}_{\widehat{v}}) \quad \text{and} \quad p^{\min} := \inf_{\widehat{v}} \pi(V \in \mathcal{N}_{\widehat{v}}),$$

the error probability $P_{\mathsf{error}} = \mathbb{P}[(V, \widehat{V}) \notin \mathcal{N}]$, and entropy $h_2(p) = -p \log p - (1 - p) \log(1 - p)$.

(a) Prove that for any Markov chain $V \to X \to \widehat{V}$, we have

$$h_2(P_{\mathsf{error}}) + P_{\mathsf{error}} \log \frac{1 - p^{\min}}{p^{\max}} \geq \log \frac{1}{p^{\max}} - I(V; \widehat{V}). \tag{9.8.2}$$

(b) Conclude from inequality (9.8.2) that

$$\mathbb{P}[(V, \widehat{V}) \notin \mathcal{N}] \geq 1 - \frac{I(V; X) + \log 2}{\inf_{\widehat{v}} \log \frac{1}{\pi(\mathcal{N}_{\widehat{v}})}}.$$

(c) Let $\mathcal{V} \subset \mathbb{R}^d$ and define $\mathcal{N} = \{(v, v') : \|v - v'\| \leq \delta\}$ to be the points within δ of one another. Let $\mathbb{B}_v$ denote the $\|\cdot\|$-ball of radius 1 centered at v. Conclude that for any prior π on $\mathbb{R}^d$ that

$$\mathbb{P}\left(\|V - \widehat{V}\|_2 \geq \delta\right) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{1}{\sup_v \pi(\delta \mathbb{B}_v)}}.$$

Exercise 9.4: This questions develops the minimax rate of estimation for the parameter of a uniform distribution in squared error. Assume that $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[\theta, \theta + 1]$, meaning that X_i have densities $p(x) = \mathbf{1}\{x \in [\theta, \theta + 1]\}$. Let $X_{(1)} = \min_i\{X_i\}$ denote the first order statistic.

(a) Prove that

$$\mathbb{E}[(X_{(1)} - \theta)^2] = \frac{2}{(n+1)(n+2)}.$$

Hint. That $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t) dt$ for any positive Z may be useful.

(b) Show that the minimax rate for estimating θ in the uniform family $\mathcal{U} = \{\mathsf{Uniform}[\theta, \theta + 1]\}_{\theta \in \mathbb{R}}$ in squared error has lower bound c/n^2, where $c > 0$ is a numerical constant.

Exercise 9.5 (Sign identification in sparse linear regression)**:** We investigate *sign recovery* in sparse linear regression, where we seek the signs $\mathrm{sign}(\theta_j^*)$ for $j = 1, \ldots, d$, where $\mathrm{sign}(0) = 0$, of the parameter θ^*. Assume the following process: fix a signal threshold $\theta_{\min} > 0$. Choose $S \in \{-1, 0, 1\}^d$ uniformly at random from the k-sparse vectors $\mathcal{S}_k := \{s \in \{-1, 0, 1\}^d : \|s\|_1 = k\}$. Define vectors θ^s so that $\theta_j^s = \theta_{\min} s_j$, and conditional on $S = s$, observe

$$Y = X\theta^s + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2 I_n).$$

(a) Use Fano's inequality to show that for any estimator $\widehat{S}$ of S, we have

$$\mathbb{P}(\widehat{S} \neq S) \geq \frac{1}{2} \quad \text{unless} \quad n \geq c \frac{\frac{d}{k} \log \binom{d}{k}}{\left\|n^{-1/2} X\right\|_{\mathrm{Fr}}^2} \frac{\sigma^2}{\theta_{\min}^2},$$

where c is a numerical constant. You may assume that $k \geq 4$ or $\log \binom{d}{k} \geq 4 \log 2$.

(b) Assume that $X \in \{-1, 1\}^{n \times d}$. Give a lower bound on how large n must be for sign recovery. Give a one sentence interpretation of $\sigma^2/\theta_{\min}^2$.

Exercise 9.6: This question investigates whether adaptivity can allow better estimation performance for linear regression. For $i = 1, \ldots, n$, assume we observe variables

$$Y_i = \langle X_i, \theta \rangle + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2), \tag{9.8.3}$$

where $\theta \in \mathbb{R}^d$ is unknown. Based on observing $Y_1^{i-1} = \{Y_1, \ldots, Y_{i-1}\}$, we allow an adaptive choice of the next predictor variables $X_i \in \mathbb{R}^d$. Let $\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2)$ denote the family of linear regression problems under this adaptive setting (with n observations) where we constrain the Frobenius norm of the design matrix $X^\top = [X_1 \; \cdots \; X_n]$, $X \in \mathbb{R}^{n \times d}$, to have bound $\|X\|_{\text{Fr}}^2 \le \mathsf{F}^2$. We use Assouad's method to show that the minimax mean-squared error satisfies

$$\mathfrak{M}(\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2), \|\cdot\|_2^2) := \inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \ge \frac{d\sigma^2}{n} \cdot \frac{1}{16 \frac{1}{dn} \mathsf{F}^2}. \tag{9.8.4}$$

Here the infimum is taken over all adaptive procedures satisfying $\|X\|_{\text{Fr}}^2 \le \mathsf{F}^2$. When we choose X_i based on the observations Y_1^{i-1}, we take $X_i = F_i(Y_1^{i-1}, U_1^i)$, where U_i is independent of ε_i and Y_1^{i-1} and F_i is a function. Justify the following steps in the proof of inequality (9.8.4):

(i) Assume that nature chooses $v \in \mathcal{V} = \{-1, 1\}^d$ uniformly at random and, conditionally on v, let $\theta = \theta_v$. Justify

$$\mathfrak{M}(\mathcal{L}^n_{\text{ada}}(\mathsf{F}^2), \|\cdot\|_2^2) \ge \inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2].$$

Argue it is no loss of generality to assume that the choices for X_i are deterministic based on the Y_1^{i-1}. Thus, throughout we assume that $X_i = F_i(Y_1^{i-1}, u_1^i)$, where u_1^n is a fixed sequence, or, for simplicity, that X_i is a function of Y_1^{i-1}.

(ii) Fix $\delta > 0$. Let $v \in \{-1, 1\}^d$, and for each such v, define $\theta_v = \delta v$. Let P_v^n denote the joint distribution (over adaptively chosen X_i) of the observed variables Y_1^n, and define $P_{\pm j}^n = \frac{1}{2^{d-1}} \sum_{v : v_j = \pm 1} P_v^n$, so that $P_{\pm j}^n$ denotes the distribution of the Y_i when $v \in \{-1, 1\}^d$ is chosen uniformly at random but conditioned on $v_j = \pm 1$. Then

$$\inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{\theta_v}[\|\widehat{\theta} - \theta_v\|_2^2] \ge \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right].$$

(iii) We have

$$\frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}\right] \ge \frac{\delta^2 d}{2} \left[1 - \left(\frac{1}{d} \sum_{j=1}^d \left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}^2\right)^{\frac{1}{2}}\right].$$

(iv) Let $P_{+j}^{(i)}$ be the distribution of the random variable Y_i conditioned on $v_j = +1$ (with the other coordinates of v chosen uniformly at random), and let $P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)$ denote the distribution of Y_i conditioned on $v_j = +1$, $Y_1^{i-1} = y_1^{i-1}$, and x_i. Justify

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\text{TV}}^2 \le \frac{1}{2} \sum_{i=1}^n \int D_{\text{kl}}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) dP_{+j}^{i-1}(y_1^{i-1}, x_i).$$

(v) Then we have

$$\sum_{j=1}^{d} D_{\rm kl}\left(P_{+j}^{(i)}(\cdot \mid y_1^{i-1}, x_i) \| P_{-j}^{(i)}(\cdot \mid y_1^{i-1}, x_i)\right) \le \frac{2\delta^2}{\sigma^2} \|x_i\|_2^2 .$$

(vi) We have

$$\sum_{j=1}^{d} \left\|P_{+j}^n - P_{-j}^n\right\|_{\rm TV}^2 \le \frac{\delta^2}{\sigma^2} \mathbb{E}[\|X\|_{\rm Fr}^2],$$

where the final expectation is over V drawn uniformly in $\{-1, 1\}^d$ and all Y_i, X_i.

(vii) Show how to choose δ appropriately to conclude the minimax bound (9.8.4).

Exercise 9.7: Suppose under the setting of Question 9.6 that we may no longer be adaptive, meaning that the matrix $X \in \mathbb{R}^{n \times d}$ must be chosen ahead of time (without seeing any data). Assuming $n \ge d$, is it possible to attain (within a constant factor) the risk (9.8.4)? If so, give an example construction, if not, explain why not.

Exercise 9.8 (Tilting and information bounds)**:** Let P_0 be a distribution on $X \in \mathcal{X}$, and assume w.l.o.g. that P_0 has a density p_0. For a bounded function $g : \mathcal{X} \to \mathbb{R}^k$ with $\mathbb{E}_{P_0}[g(X)] = 0$, define the *tilted density*

$$p_{t,g}(x) := (1 + \langle t, g(x)\rangle) p_0(x), \tag{9.8.5}$$

for $t \in \mathbb{R}^k$, with induced distribution $P_{t,g}$.

(a) Show that if t is near enough to 0, then $p_{t,g}$ is indeed a valid density.

(b) Show that for $t \in \mathbb{R}^k$ near 0,

$$D_{\rm kl}\left(P_{t,g} \| P_0\right) = \frac{1}{2} t^\top {\rm Cov}_0(g(X)) t + O(\|t\|^3) \quad \text{and} \quad D_{\rm kl}\left(P_0 \| P_{t,g}\right) = \frac{1}{2} t^\top {\rm Cov}_0(g(X)) t + O(\|t\|^3).$$

(c) Let $\mathcal{V} \subset \mathbb{R}^k$ be a bounded set and $V \sim \mathsf{Uniform}(\mathcal{V})$. Let $\delta \ge 0$. Suppose that conditional on $V = v$, we draw $X_i \stackrel{\rm iid}{\sim} P_{\delta v, g}$. Show the mutual information bound

$$I(X_1^n; V) \le \frac{n\delta^2}{2} {\rm tr}\left({\rm Cov}_0(g(X)) \mathbb{E}[VV^\top]\right) + O(n\delta^3).$$

Exercise 9.9* (General tilting)**:** Replace the tilting (9.8.5) with

$$p_{t,g}(x) = \frac{1}{C_t} \phi(\langle t, g(x)\rangle) p_0(x), \quad C_t := \int \phi(\langle t, g(x)\rangle) p_0(x) d\mu(x)$$

where $\phi : \mathbb{R} \to \mathbb{R}_+$ is a bounded nonnegative function, differentiable at 0 with $\phi(0) = \phi'(0) = 1$, with bounded second derivatives. (For example, $\phi(t) = \frac{2}{1+e^{-2t}}$.) Show that if $\mathbb{E}_0[\|g(X)\|^2] < \infty$ then the following analogues of the results (a), (b), and (c) from Exercise 9.8 hold:

(a) The normalization $C_t := \mathbb{E}_0[\phi(\langle t, g(X)\rangle)]$ satisfies $C_t = 1 + \frac{\phi''(0)}{2} t^\top {\rm Cov}_0(g(X)) t + o(\|t\|^2)$.

(b) For t near 0,

$$D_{\rm kl}\left(P_{t,g} \| P_0\right) = \frac{1}{2} t^\top {\rm Cov}_0(g(X)) t + o(\|t\|^2).$$

(c) Let $\mathcal{V} \subset \mathbb{R}^k$ be a bounded set and $V \sim \mathsf{Uniform}(\mathcal{V})$. Let $\delta \geq 0$. Suppose that conditional on $V = v$, we draw $X_i \stackrel{\text{iid}}{\sim} P_{\delta v, g}$. Then

$$I(X_1^n; V) \leq \frac{n\delta^2}{2} \operatorname{tr}\left(\operatorname{Cov}_0(g(X))\mathbb{E}[VV^\top]\right) + n \cdot o(\delta^2).$$

Exercise 9.10: Let $\mathcal{P}$ be a collection of distributions on a set $\mathcal{X}$, and let $\theta : \mathcal{P} \to \mathbb{R}^d$ be a parameter of the distribution. Let $\mathcal{G}$ be a collection of mean-zero functions mapping $\mathcal{X} \to \mathbb{R}^k$ with $\mathbb{E}_0[g(X)] = 0$ for each $g \in \mathcal{G}$. For $t \in \mathbb{R}^k$ define the distribution $P_{t,g}$ by the tilting (9.8.5), and assume $P_{t,g} \in \mathcal{P}$ for small enough t. We say that θ is *differentiable at P_0 relative to $\mathcal{G}$* with derivative $\dot{\theta}_{P_0} : \mathcal{X} \to \mathbb{R}^d$ if

$$\theta(P_{t,g}) = \theta(P_0) + \mathbb{E}_{P_0}[\dot{\theta}_{P_0}(X)\langle g(X), t\rangle] + o(\|t\|)$$

as $t \to 0$ for each $g \in \mathcal{G}$, where $\mathbb{E}_{P_0}[\dot{\theta}_{P_0}(X)] = 0$. We call such a function $\dot{\theta}_{P_0}$ the *influence function* of θ. (See also Section 11.3.1 to come.)

(a) Show that the mean is differentiable relative to $\mathcal{G}$ for any collection of mean-zero bounded functions, and give the influence function $\dot{\theta}_{P_0}$.

For the remainder of the problem, assume that for any set $\mathcal{G}$ of bounded functions, $\theta : \mathcal{P} \to \mathbb{R}^d$ is differentiable relative to $\mathcal{G}$ with bounded influence $\dot{\theta}_{P_0}$. Let $\Sigma = \operatorname{Cov}_0(\dot{\theta}_{P_0})$ and for $A \succ 0$ define the Mahalanobis norm $\|v\|_A^2 = v^\top A v$.

(b) For $\delta \geq 0$, show how to construct a collection of distributions $\{Q_v\}$ indexed by $v \in \mathcal{V} = \{-1, 1\}^d$ such that as $\delta \to 0$, for any pair v, v'

$$\|\theta(Q_v) - \theta(Q_{v'})\|_A = \delta \left\|v - v'\right\|_A + o(\delta).$$

(c) Show that (for large enough dimension d) there exist numerical constants $c_0, c_1 > 0$ and a collection of distributions $\{Q_v\} \subset \mathcal{P}$, indexed by $v \in \mathcal{V} = \{-1, 1\}^d$ such that if we choose $V \sim \mathsf{Uniform}(\mathcal{V})$ then draw $X_1^n \stackrel{\text{iid}}{\sim} Q_v$ conditional on $V = v$,

$$\mathbb{P}\left(\left\|\widehat{\theta}(X_1^n) - \theta(Q_V)\right\|_{\Sigma^{-1}} \geq c_0\sqrt{d/n}\right) \geq c_1.$$

Exercise 9.11 (A minimax lower bound for M-estimation)**:** Let ℓ be a convex loss and $L(\theta) = \mathbb{E}_{P_0}[\ell(\theta, Z)]$ satisfy all the conditions of Assumption A.5.1 in Section 5.3, where P_0 is a distribution on $Z \in \mathcal{Z}$ that we assume w.l.o.g. has a density p_0. Define the tilted densities p_t as in Exercise 9.8, equation (9.8.5). Let $\theta(P) = \operatorname{argmin}_\theta \mathbb{E}_P[\ell(\theta, Z)]$ be the population minimizer and define the Hessian $H = \nabla^2 L(\theta_0)$ and covariance $\Sigma := \operatorname{Cov}_{P_0}(\nabla\ell(\theta_0, Z)) \succ 0$. Previewing Section 11.3.2, the implicit function theorem implies that

$$\theta(P_{t,g}) = \theta(P_0) - H^{-1}\mathbb{E}_{P_0}\left[\nabla\ell(\theta_0, Z)g(Z)^\top\right] t + o(\|t\|)$$

for $t \in \mathbb{R}^k$ near 0. Show that by choosing an appropriate tilting function $g : \mathcal{Z} \to \mathbb{R}^d$, there exist numerical constants $c_0, c_1, c_2 > 0$ such that for large enough n, the choice $\delta_n := c_0/\sqrt{n}$ yields

$$\inf_{\widehat{\theta}} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P^n_{\delta_n v, g}\left(\left(\widehat{\theta}(Z_1^n) - \theta(P_{\delta_n v, g})\right)^\top (H^{-1}\Sigma H^{-1})^{-1}\left(\widehat{\theta}(Z_1^n) - \theta(P_{\delta_n v, g})\right) \geq c_1 \frac{d}{n}\right) \geq c_2.$$

How does this lower bound compare to the guarantee in Corollary 5.3.10?

Chapter 10

Constrained risk inequalities

This chapter revisits our minimax bounds under *constraints* on the procedures we allow. While the minimax risk of Chapter 9 provides a first approach to fundamental limits on procedures, its reliance on the collection of *all* measurable functions as potential estimators is somewhat limiting. In most statistical and machine learning problems, procedures face further constraints: they must be efficiently computable, communicate little across a network, be robust, or protect the privacy of the data providers.

With this as context, we replace the minimax risk of Chapter 9 with the *constrained minimax risk*, which, given a collection $\mathcal{C}$ of possible procedures, defines

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) := \inf_{\widehat{\theta} \in \mathcal{C}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi\big(\rho(\widehat{\theta}(X), \theta(P))\big) \right]. \tag{10.0.1}$$

As in the original defining equation (9.1.1) of the minimax risk, $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing loss, ρ is a semimetric on the space Θ, and the expectation is taken over $X \sim P$. The class $\mathcal{C}$ may encode computation, memory, communication, privacy, or other restrictions on the estimator. We focus on settings in $\mathcal{C}$ consists of estimators that process each point X_i into an observation Z_i; this representation covers, for example, communication complexity, memory complexity, and local privacy.

10.1 Strong data processing inequalities

Our results begin by considering *strong data processing inequalities*, which provide more quantitative versions of the standard data processing inequalities for divergences. The setting is straightforward: we have distributions P_0 and P_1 on a space $\mathcal{X}$ and a channel (Markov kernel) Q from $\mathcal{X}$ to $\mathcal{Z}$. When Q contracts on the space of distributions, we have a strong data processing inequality.

Definition 10.1 (Strong data processing)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ be convex with $f(1) = 0$. For distributions P_0, P_1 on $\mathcal{X}$ and a channel Q from $\mathcal{X}$ to $\mathcal{Z}$, define the marginal distribution $M_v(A) := \int Q(A \mid x) dP_v(x)$. Then Q satisfies a* strong data processing *inequality with constant $\alpha \le 1$ for the f-divergence if*

$$D_f\left(M_0 \| M_1\right) \le \alpha D_f\left(P_0 \| P_1\right)$$

*for any P_0, P_1 on $\mathcal{X}$. For any such f, we define the f-*strong data processing constant

$$\alpha_f(Q) := \sup_{P_0 \neq P_1} \frac{D_f\left(M_0 \| M_1\right)}{D_f\left(P_0 \| P_1\right)}.$$

Perhaps the most frequent use of such inequalities is in providing conditions for the fast mixing of Markov chains. Indeed, suppose the Markov kernel Q satisfies a strong data processing inequality with constant α with respect to variation distance. If π denotes the stationary distribution of the Markov kernel Q and the operator $\circ$ denotes one step of the Markov kernel, $Q \circ P := \int Q(\cdot \mid x) dP(x)$, then for any initial distribution π_0 on the space $\mathcal{X}$ we have

$$\|\underbrace{Q \circ \cdots \circ Q}_{k \text{ times}} \circ \pi_0 - \pi\|_{\mathrm{TV}} = \|Q \circ \cdots \circ Q \circ \pi_0 - Q \circ \cdots \circ Q \circ \pi\|_{\mathrm{TV}} \le \alpha^k \|\pi_0 - \pi\|_{\mathrm{TV}} .$$

Thus, the Markov chain enjoys geometric mixing (see also Exercise 10.4). We will see that the *Dobrushin* coefficient provides a way to calculate this mixing constant:

Definition 10.2. *The* Dobrushin coefficient *of a channel or Markov kernel Q is*

$$\alpha_{\mathrm{TV}}(Q) := \sup_{x,y} \|Q(\cdot \mid x) - Q(\cdot \mid y)\|_{\mathrm{TV}} .$$

The Dobrushin coefficient coincides with the strong data processing constant for variation distance:

Proposition 10.1.1. *For any channel Q, we have*

$$\alpha_{\mathrm{TV}}(Q) = \sup_{P_0 \neq P_1} \frac{\|Q \circ P_0 - Q \circ P_1\|_{\mathrm{TV}}}{\|P_0 - P_1\|_{\mathrm{TV}}}.$$

Exercise 10.1 outlines one of the standard proofs of this equality.

Substantially, the Dobrushin coefficient upper bounds *every* strong data processing constant.

Theorem 10.1.2. *Let $f : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ be convex and satisfy $f(1) = 0$. Then for any channel Q,*

$$\alpha_{\mathrm{TV}}(Q) \ge \alpha_f(Q).$$

Exercise 10.3 outlines a proof, which we sketch. Without loss of generality we may assume that $f \ge 0$ by replacing f with the function $f_+(t) = f(t) - f'(1)t$, so that $f_+ \ge 0$ and $D_{f_+} = D_f$. Then functions of the form $h(t) = \sum_{i=1}^k a_i [t - c_i]_+ + \sum_{i=1}^k b_i [d_i - t]_+$, where $c_i \ge 1$ and $d_i \le 1$, can approximate nonnegative convex f with $f(1) = 0$ arbitrarily accurately. For such h, Lemma 2.2.5 connects variation distance and divergences with positive part functions $[\cdot]_+$ to imply

$$D_h(Q \circ P_0 \| Q \circ P_1) \le \alpha_{\mathrm{TV}}(Q) D_h(P_0 \| P_1).$$

Though we do not prove this, the strong data processing constants for χ^2-divergences provide the opposite extremal bound to Theorem 10.1.2: they are the smallest constants. In particular, for any twice continuously differentiable f on $\mathbb{R}_{>0}$ with $f''(1) > 0$, the χ^2 contraction coefficient

$$\alpha_{\chi^2}(Q) = \sup_{P_0 \neq P_1} \frac{D_{\chi^2}(Q \circ P_0 \| Q \circ P_1)}{D_{\chi^2}(P_0 \| P_1)}.$$

satisfies

$$\alpha_{\chi^2}(Q) \le \alpha_f(Q), \tag{10.1.1}$$

and we also have $\alpha_{\chi^2}(Q) = \alpha_{\mathrm{kl}}(Q)$, so that the strong data processing inequalities for KL-divergence and χ^2-divergence coincide. (See the bibliographic section for more on this.)

In our context of constrained minimax lower bounds, such data processing inequalities immediately imply somewhat sharper lower bounds than the unconstrained applications previously. Let us revisit the situation in the local packing lower bounds, where the KL-divergence has a Euclidean structure as in the bound (9.4.6), meaning that $D_{\rm kl}\left(P_0 \| P_1\right) \le \kappa^2 \delta^2$ when our parameters of interest $\theta_v = \theta(P_v)$ satisfy $\rho(\theta_0, \theta_1) \le \delta$. We assume the constraints $\mathcal{C}$ impose that the data X_i is passed through a channel Q with KL-data processing constant $\alpha_{\rm KL}(Q) \le 1$. In this case, in Le Cam's method (9.3.2), an application of Pinsker's inequality yields that whenever $\rho(\theta_0, \theta_1) \ge 2\delta$ then

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) \ge \frac{\Phi(\delta)}{2}\left[1 - \sqrt{\frac{n}{2} D_{\rm kl}\left(M_0 \| M_1\right)}\right] \ge \frac{\Phi(\delta)}{2}\left[1 - \sqrt{n\kappa^2 \alpha_{\rm KL}(Q)\delta^2/2}\right].$$

The "recipe" choice of δ to make the probability of error constant yields $\delta^2 = (2n\kappa^2\alpha_{\rm KL}(Q))^{-1}$, or the minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{C}) \ge \frac{1}{4}\Phi\left(\frac{1}{\sqrt{2n\kappa^2\alpha_{\rm KL}(Q)}}\right),$$

which suggests an effective sample size degradation of $n \mapsto n \cdot \alpha_{\rm KL}(Q)$. Similarly, in the local Fano method in Section 9.4.1, we see identical behavior and an effective sample size degradation of $n \mapsto n\alpha_{\rm KL}(Q)$, that is, if without constraints a sample size of $n(\epsilon)$ is required to achieve some desired accuracy ϵ, with the constraint a sample size of at least $n(\epsilon)/\alpha_{\rm KL}(Q)$ is necessary.

10.2 Local privacy

In Chapter 8 on differential privacy, we define *locally private mechanisms* (Definition 8.2) as those for which there is no trust: individuals randomize their own data, and no central curator collects or analyzes and then privatizes the resulting statistics. With such privacy mechanisms, we can directly develop strong data processing inequalities, after which we can prove strong lower bounds on estimation. In this section, we focus on one-dimensional quantities and Le Cam's two-point method for lower bounds, as they allow the most direct application of strong data processing. We will later develop more sophisticated techniques using communication complexity.

In this setting, we have a ε-differentially private channel Q taking inputs $x \in \mathcal{X}$ and outputting Z. We allow *sequential interactivity*, meaning that the ith private variable Z_i may depend on both X_i and Z_1^{i-1} (see the graphical model in Figure 10.1), so that instead of the basic constraint in Definition 8.2 that $Q(A \mid x) \le e^\varepsilon Q(A \mid x')$ for all x, x', local differential privacy means

$$\frac{Q(Z_i \in A \mid X_i = x, z_1^{i-1})}{Q(Z_i \in A \mid X_i = x', z_1^{i-1})} \le e^\varepsilon \tag{10.2.1}$$

for all (measurable) sets A and inputs x, x', z_1^{i-1}. Any such ratio bound implies strong contraction on the space of probabilities:

Theorem 10.2.1. *Let Q be an ε-locally differentially private channel from $\mathcal{X}$ to $\mathcal{Z}$. Then for any distributions P_0, P_1 inducing marginal distributions $M_v(\cdot) = \int Q(\cdot \mid x) dP_v(x)$,*

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) \le 4(e^\varepsilon - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2.$$

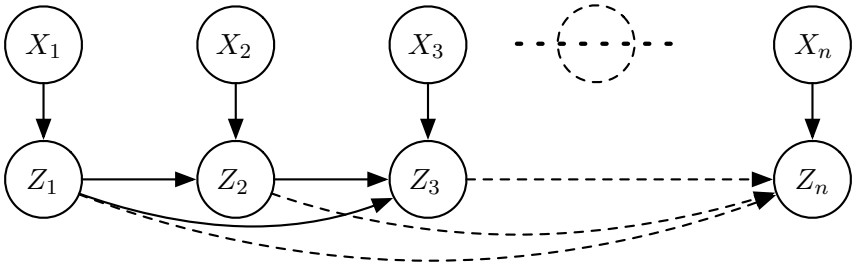


Figure 10.1. The sequentially interactive private observation model: the ith output Z_i may depend on X_i and the previously released Z_1^{i-1}.

Proof Without loss of generality, we assume that the output space $\mathcal{Z}$ is finite (by definition (2.2.3)), let $m_v(z)$ and $q(z \mid x)$ be the p.m.f.s of M and Q, respectively, and let P_0 and P_1 have densities p_0 and p_1 with respect to a measure μ. Then

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) = \sum_z (m_0(z) - m_1(z)) \log \frac{m_0(z)}{m_1(z)}$$

For any $a, b > 0$, we have $\log \frac{a}{b} = \log(1 + \frac{a}{b} - 1) \le \frac{a}{b} - 1$, and similarly, $\log \frac{b}{a} \le \frac{b}{a} - 1$. That is, $|\log \frac{a}{b}| \le \frac{|a-b|}{\min\{a,b\}}$. Substituting above, we obtain

$$D_{\rm kl}\left(M_0 \| M_1\right) + D_{\rm kl}\left(M_1 \| M_0\right) \le \sum_z \frac{(m_0(z) - m_1(z))^2}{\min\{m_0(z), m_1(z)\}}.$$

To control the difference $m_0(z) - m_1(z)$, note that for any fixed $x_0 \in \mathcal{X}$ we have

$$\int_{\mathcal{X}} q(z \mid x_0)(p_0(x) - p_1(x)) d\mu(x) = 0.$$

Thus

$$m_0(z) - m_1(z) = \int_{\mathcal{X}} (q(z \mid x) - q(z \mid x_0))(p_0(x) - p_1(x)) d\mu(x),$$

and so

$$\begin{aligned} |m_0(z) - m_1(z)| &\le \sup_{x \in \mathcal{X}} |q(z \mid x) - q(z \mid x_0)| \int_{\mathcal{X}} |p_0(x) - p_1(x)| d\mu(x) \\ &= 2q(z \mid x_0) \sup_{x \in \mathcal{X}} \left(\frac{q(z \mid x)}{q(z \mid x_0)} - 1 \right) \left\| P_0 - P_1 \right\|_{\rm TV}. \end{aligned}$$

By definition of local differential privacy, $\frac{q(z|x)}{q(z|x_0)} - 1 \le e^\varepsilon - 1$, and as x_0 was arbitrary we obtain

$$|m_0(z) - m_1(z)| \le 2(e^\varepsilon - 1) \inf_x q(z \mid x) \left\| P_0 - P_1 \right\|_{\rm TV}.$$

Noting that $\inf_x q(z \mid x) \le \min\{m_0(z), m_1(z)\}$ we obtain the theorem. □

To apply this result to obtain minimax lower bounds for estimation, we need to address samples drawn from product distributions with the potential interaction (10.2.1). In this case, we consider sequential samples $Z_i \sim Q(\cdot \mid X_i, Z_1^{i-1})$ and define $M_v^n = \int Q(\cdot \mid x_1^n) dP_v(x_1^n)$ to be the marginal distribution over Z_1^n. Then we have the following corollary.

Corollary 10.2.2. *Assume that each channel $Q(\cdot \mid X_i, Z_1^{i-1})$ is ε_i-differentially private. Then*

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) \le 4 \sum_{i=1}^n (e^{\varepsilon_i} - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2 .$$

Proof Recalling the chain rule (2.1.6) for the KL-divergence, we have

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) = \sum_{i=1}^n \mathbb{E}_{M_0}\left[D_{\rm kl}\left(M_{0,i}(\cdot \mid Z_1^{i-1}) \| M_{1,i}(\cdot \mid Z_1^{i-1})\right)\right],$$

where the outer expectation is taken over Z_1^{i-1} drawn marginally from M_0^n, and $M_{v,i}(\cdot \mid z_1^{i-1})$ denotes the conditional distribution on Z_i given $Z_1^{i-1} = z_1^{i-1}$ when $X_1^n \stackrel{\rm iid}{\sim} P_v$. Writing this distribution out, we note that Z_i is conditionally independent of $X_{\setminus i}$ given X_i and Z_1^{i-1} by construction, so for any set A

$$\begin{aligned} M_{v,i}(A \mid z_1^{i-1}) &= \int Q(Z_i \in A \mid x_1^n, z_1^{i-1}) dP_v(x_1^n \mid z_1^{i-1}) = \int Q(Z_i \in A \mid x_i, z_1^{i-1}) dP_v(x_1^n \mid z_1^{i-1}) \\ &= \int Q(Z_i \in A \mid x_i, z_1^{i-1}) dP_v(x_i). \end{aligned}$$

Because $Q(Z_i \in \cdot \mid x_i, z_1^{i-1})$ is ε_i-differentially private by assumption, Theorem 10.2.1 gives

$$D_{\rm kl}\left(M_{0,i}(\cdot \mid z_1^{i-1}) \| M_{1,i}(\cdot \mid z_1^{i-1})\right) \le 4(e^{\varepsilon_i} - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2$$

for any realization z_1^{i-1} of Z_1^{i-1}. Iterating this gives the result. □

Local privacy is such a strong condition on the channel Q that it actually "transforms" the KL-divergence into a variation distance, so that even if two distributions P_0 and P_1 have infinite divergence $D_{\rm kl}\left(P_0 \| P_1\right) = +\infty$—for example, if their supports do not completely overlap—their induced marginals have the much smaller (finite) divergence $D_{\rm kl}\left(M_0 \| M_1\right) \le 4(e^\varepsilon - 1)^2 \left\|P_0 - P_1\right\|_{\rm TV}^2 \le 4(e^\varepsilon - 1)^2$. This transformation into a different metric means that even estimation problems that appear easy can become challenging under local privacy constraints; for example, minimax squared error for estimating the mean of a random variable with finite variance scales as $1/\sqrt{n}$ rather than the typical $1/n$ scaling in non-private cases (see Exercise 10.5).

Let us demonstrate applications of Corollary 10.2.2. Our main object of interest is the private analogue of the minimax risk (9.1.1), where for a parameter $\theta : \mathcal{P} \to \Theta$, semimetric ρ, and loss Φ, for a family of channels $\mathcal{Q}$ we define the *channel-constrained minimax risk*

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \mathcal{Q}) := \inf_{\widehat{\theta}_n} \inf_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[\Phi(\rho(\widehat{\theta}_n(Z_1^n), \theta(P)))\right]. \tag{10.2.2}$$

When we take $\mathcal{Q} = \mathcal{Q}_\varepsilon$ to be the collection of ε-locally differentially private (interactive) channels (10.2.1), we obtain the ε-locally private minimax risk.

A few examples showing lower (and upper) bounds for the private minimax risk (10.2.2) in mean estimation follow.

Example 10.2.3 (Bounded mean estimation)**:** Let $\mathcal{P}$ be the collection of distributions supported on $[-b, b]$, where $0 < b < \infty$. Then for any $\varepsilon \ge 0$, the minimax squared error satisfies

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \mathcal{Q}_\varepsilon) \gtrsim \min\left\{b^2, \frac{b^2}{(e^\varepsilon - 1)^2 n}\right\} + \frac{b^2}{n}.$$

The second term is the classic minimax rate for this collection of distributions. To see the first term, let $0 \le \delta \le 1$ to be chosen, and take distributions P_0 and $P_1 \in \mathcal{P}$ with $P_0(X = b) = \frac{1-\delta}{2} = P_1(X = -b)$ and $P_1(X = b) = \frac{1+\delta}{2} = P_0(X = -b)$. Then $\|P_0 - P_1\|_{\rm TV} = \delta$, $\mathbb{E}_1[X] - \mathbb{E}_0[X] = 2b\delta$, and by Le Cam's method (9.3.3), for any ε-locally private channel Q and induced marginals M_0^n, M_1^n as in Corollary 10.2.2, we have

$$\begin{aligned}\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \{Q\}) &\ge \frac{b^2\delta^2}{2}\left(1 - \sqrt{\frac{1}{2}D_{\rm kl}\left(M_0^n \| M_1^n\right)}\right) \\ &\ge \frac{b^2\delta^2}{2}\left(1 - \sqrt{2(e^\varepsilon - 1)^2 n \left\|P_0 - P_1\right\|_{\rm TV}^2}\right) = \frac{b^2\delta^2}{2}\left(1 - \sqrt{2(e^\varepsilon - 1)^2 n\delta^2}\right).\end{aligned}$$

Setting $\delta^2 = \min\{1, \frac{1}{8n(e^\varepsilon - 1)^2}\}$ gives the claimed minimax bound. $\diamond$

We see a reduction in the effective sample size: when ε is large, there is no change, but for $\varepsilon < 1$, the estimation error is similar to that when we observe a sample of size $n\varepsilon^2$.

Example 10.2.4 (Estimating the parameter of a uniform distribution)**:** Exercise 9.4 shows that estimating the parameter θ of a $\mathsf{Uniform}(\theta, \theta + 1)$ distribution has minimax mean squared error scaling as $1/n^2$. Under local differential privacy, this rate is unachievable. Let $\mathcal{P} = \{\mathsf{Uniform}(\theta, \theta+1), \theta \in [0, 1]\}$ be the collection of uniform distributions with the given supports. Letting P_0 and P_1 be $\mathsf{Uniform}(0, 1)$ and $\mathsf{Uniform}(\delta, 1 + \delta)$, respectively, with $\delta \ge 0$ to be chosen, we have $\|P_0 - P_1\|_{\rm TV} = \delta$, while for any ε-differentially private channel Q and induced marginals M_0 and M_1,

$$D_{\rm kl}\left(M_0^n \| M_1^n\right) \le 4(e^\varepsilon - 1)^2 n \left\|P_0 - P_1\right\|_{\rm TV}^2 = 4(e^\varepsilon - 1)^2 n\delta^2.$$

Applying Le Cam's method (9.3.3) and taking $\delta^2 \asymp 1/n(e^\varepsilon - 1)^2$, the collection $\mathcal{Q}_\varepsilon$ of ε-locally differentially private (sequentially interacting) channels has

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \mathcal{Q}_\varepsilon) \gtrsim \min\left\{1, \frac{1}{(e^\varepsilon - 1)^2 n}\right\}.$$

When $\varepsilon \lesssim 1$, the best attainable rate thus scales as $\frac{1}{n\varepsilon^2}$ instead of $\frac{1}{n^2}$. $\diamond$

In both the preceding examples, many simple estimators achieve the minimax rates. The Laplace mechanism (Example 8.1.2) suffices for both: let $W_i \stackrel{\rm iid}{\sim} \mathsf{Laplace}(1)$, and set $Z_i = X_i + \frac{2b}{\varepsilon}W_i$ in Example 10.2.3 and $Z_i = X_i + \frac{2}{\varepsilon}W_i$ in Example 10.2.4. In the former, define $\widehat{\theta}_n = \overline{Z}_n$ to be the mean; in the latter, $\mathbb{E}[\overline{Z}_n] = \frac{\theta+1}{2}$, so $\widehat{\theta}_n = 2\overline{Z}_n - 1$ achieves the minimax rate. More extreme rate degradations from local privacy occur frequently; Exercises 10.5 and 10.6 exhibit examples.

10.3 Communication complexity

Communication complexity, which originally arose to answer questions about the number of bits necessary to communicate to compute different functions, has grown to provide tools to establish fundamental limits in a diversity of domains, including in streaming and online algorithms, memory-limited procedures, and (of course) in minimal communication. Recent connections between communication complexity and information-theoretic techniques have increased its applicability in statistical problems, which is our main motivation here, and to which we return in force

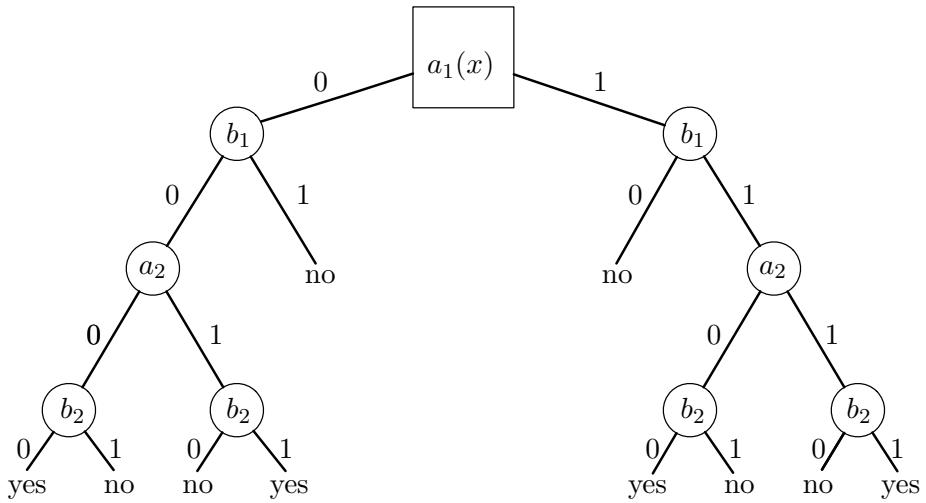


Figure 10.2. A communication tree representing testing equality for 2-dimensional bit strings $x, y \in \{0,1\}^2$. Internal nodes labeled a_j communicate the jth bit $a_j(x) = x_j$ of x, while internal nodes labeled b_j communicate the jth bit $b_j(y) = y_j$ of y. The maximum number of messages is 4.

in Section 10.4 to come. We give a (necessarily limited) overview of communication complexity, along with some of the basic techniques and approaches and several exercises, which then extend to statistical problems.

10.3.1 Classical communication complexity problems

The most basic problems in communication complexity ask a simple question: two entities (always named Alice and Bob) have inputs x, y and wish to jointly compute a function $f(x, y)$. How many bits must Alice and Bob communicate to compute this value? Less abstractly, Alice and Bob have input domains $\mathcal{X}$ and $\mathcal{Y}$ (often, these are $\{0,1\}^n$), and Alice receives $x \in \mathcal{X}$ and Bob $y \in \mathcal{Y}$, each unknown to the other, and they jointly exchange messages until they can successfully evaluate $f(x, y)$. We assume each has infinite computational power. To formulate this as communication, we consider a *protocol* Π, which specifies the messages that each of Alice and Bob send to one another. We view this as a series of rounds, where at each round, the protocol allows one $\{0,1\}$-valued bit to be sent and determines who sends this bit, and, at termination time, can compute $f(x, y)$ based on the communicated message. Then the communication cost of Π is the maximum number of messages sent to (correctly) compute f over all inputs x, y.

For analysis, we typically formulate protocols via their associated binary tree:

Definition 10.3. *A* protocol Π *over a domain $\mathcal{X} \times \mathcal{Y}$ with output space $\mathcal{Z}$ is a binary tree, where each internal node v is labeled with a mapping $a_v : \mathcal{X} \to \{0,1\}$ or $b_v : \mathcal{Y} \to \{0,1\}$ and each leaf is labeled with a value $z \in \mathcal{Z}$.*

To execute a communication protocol Π on input (x, y), we walk down the tree: beginning at the root node, for each internal node v labeled a_v (an Alice node) we walk left if $a_v(x) = 0$ and right if $a_v(x) = 1$, and each node v labeled b_v (a Bob node) we walk left if $b_v(y) = 0$ and right if $b_v(y) = 1$. The *communication cost* of the protocol Π is the depth $\mathrm{depth}(\Pi)$ of the tree. Figure 10.2 shows an example for testing the equality $x = y$ of two 2-dimensional bit strings $x, y \in \{0,1\}^2$.

In classical communication complexity, the main questions center around the *communication complexity* of $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, which is the length of the shortest protocol that computes f correctly on all inputs: letting $\Pi_{\text{out}}(x, y)$ denote the final output of the protocol Π on inputs (x, y), this is

$$\mathsf{CC}(f) := \inf \left\{ \text{depth}(\Pi) \mid \Pi_{\text{out}}(x, y) = f(x, y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

It is useful to allow randomized communication protocols tolerating errors; in this case, we without loss of generality let Alice and Bob access independent $U_a, U_b \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, 1]$. The nodes a_v and b_v in Definition 10.3 become mappings $a_v : \mathcal{X} \times [0, 1] \to \{0, 1\}$ and $b_v : \mathcal{Y} \times [0, 1] \to \{0, 1\}$, calculating $a_v(\cdot, U_a)$ and $b_v(\cdot, U_b)$, respectively. Abusing notation by leaving this randomness implicit, the *randomized communication complexity* for accuracy δ is the length of the shortest randomized protocol that calculates $f(x, y)$ correctly with probability at least $1 - \delta$, that is,

$$\mathsf{RCC}_\delta(f) := \inf \left\{ \text{depth}(\Pi) \mid \mathbb{P}(\Pi_{\text{out}}(x, y) \neq f(x, y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}. \tag{10.3.1}$$

In the definition (10.3.1), we leave the randomization in Π implicit, and we require that the tree it induces still have finite depth. The choice of $\delta > 0$ is almost immaterial (see Exercise 10.8), as

$$\mathsf{RCC}_\delta(f) \leq O(1) \log \frac{1}{\delta} \cdot \mathsf{RCC}_{1/3}(f) \quad \text{for } 0 < \delta \leq \frac{1}{3}.$$

Variants of randomized complexity allow public randomness, which can yield simpler algorithms, but the improvement in communication complexity is limited, as Alice and Bob can nearly simulate public randomness (see Exercise 10.9). Letting $\mathfrak{P}_{\text{pub}}$ be the collection of protocols in which both parties have access to a shared $U \sim \mathsf{Uniform}[0, 1]$, we make the obvious extension

$$\mathsf{RCC}_\delta^{\text{pub}}(f) := \inf_{\Pi \in \mathfrak{P}_{\text{pub}}} \left\{ \text{depth}(\Pi) \mid \mathbb{P}(\Pi_{\text{out}}(x, y, U) \neq f(x, y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}.$$

Finally, we have notions of communication complexity that mimic the Bayesian lower bounds underpinning all of Chapter 9, which admit the most powerful algorithms and so provide the strongest lower bounds. For a probability measure μ on inputs $\mathcal{X} \times \mathcal{Y}$, *distributional* communication complexity is the depth of the shortest protocol that succeeds with a given μ-probability:

$$\mathsf{DCC}_\delta^\mu(f) := \inf \left\{ \text{depth}(\Pi) \mid \mu(\Pi_{\text{out}}(X, Y) \neq f(X, Y)) \leq \delta \right\}, \tag{10.3.2}$$

where the infimum is taken over *deterministic* protocols. As in Section 9.2, once we reduce to distributional complexity, it is no loss of generality to assume Π is deterministic.

The final notion we consider is the *information complexity*. In this case, we require again that for each input pair x, y, the protocol $\Pi(x, y)$ compute $f(x, y)$ correctly with probability at least $1 - \delta$, but instead of measuring the depth of the tree, we let X, Y be drawn randomly from some distribution and measure the mutual information $I_2(X, Y; \Pi(X, Y))$. (We use base-2 logarithms to reflect bit communication.) In this case, we define

$$\mathsf{IC}_\delta(f) := \sup \inf_\Pi \left\{ I_2(X, Y; \Pi(X, Y)) \mid \mathbb{P}(\Pi_{\text{out}}(x, y) \neq f(x, y)) \leq \delta \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y} \right\}, \tag{10.3.3}$$

where the supremum is taken over joint distributions on (X, Y), the infimum over randomized protocols Π, and the probability $\mathbb{P}$ is over randomness in Π. There is a subtlety in this definition: we require Π to be accurate on *all* inputs (x, y), not just with high probability over the distribution on (X, Y) in the information measure $I(X, Y; \Pi(X, Y))$. Relaxations to distributional variants

of the information complexity (10.3.3) are also natural, as in the definition (10.3.2). Thus we sometimes consider the distributional information complexity

$$\mathsf{IC}_\delta^\mu(f) := \inf_\Pi \left\{I_2(X, Y; \Pi(X, Y)) \mid \mu(\Pi_{\text{out}}(X, Y) \neq f(X, Y)) \leq \delta\right\},$$

where the infimum can be taken (with no loss of generality) over deterministic protocols.

The different notions of communication complexity satisfy a natural ordering. We record the standard inequalities, which essentially follow immediately from the operational interpretation of entropy as the average length of the best encoding of a signal (Section 2.4.1).

Proposition 10.3.1. *For any function f, $\delta \in (0, 1)$, and probability measure μ on $\mathcal{X} \times \mathcal{Y}$,*

$$\mathsf{CC}(f) \geq \mathsf{RCC}_\delta(f) \geq \mathsf{RCC}_\delta^{\text{pub}}(f) \geq \mathsf{DCC}_\delta^\mu(f) \geq \mathsf{IC}_\delta^\mu(f)$$

and

$$\mathsf{RCC}_\delta(f) \geq \mathsf{IC}_\delta(f).$$

Proof The first two inequalities are immediate. By Theorem 2.4.3, we have

$$\text{depth}(\Pi) \geq H_2(\Pi) \geq H_2(\Pi) - H_2(\Pi \mid X, Y) = I_2(X, Y; \Pi(X, Y)),$$

and so for all $\delta \in (0, \frac{1}{2})$ we have both

$$\mathsf{RCC}_\delta(f) \geq \mathsf{IC}_\delta(f) \quad \text{and} \quad \mathsf{DCC}_\delta^\mu(f) \geq \mathsf{IC}_\delta^\mu(f).$$

It remains to demonstrate $\mathsf{RCC}_\delta^{\text{pub}}(f) \geq \mathsf{DCC}_\delta^\mu(f)$. Let Π be any protocol with public randomness U such that $\mathbb{P}(\Pi_{\text{out}}(x, y, U) \neq f(x, y)) \leq \delta$ for all x, y. Taking an expectation over $(X, Y) \sim \mu$,

$$\delta \geq \mathbb{E}_\mu \left[\mathbb{P}(\Pi_{\text{out}}(X, Y, U) \neq f(X, Y) \mid X, Y)\right] \geq \inf_u \mu\left(\Pi_{\text{out}}(X, Y, u) \neq f(X, Y)\right),$$

that is, there must be at least some u achieving the average error of Π, and the protocol Π is deterministic given u. So any protocol Π using public randomness to achieve probability of error δ can be modified into a deterministic protocol $\Pi(\cdot, \cdot, u)$ that achieves μ-probability of error δ.[1] □

Frequently, the first inequality in Proposition 10.3.1 is strict—even exponentially large—while the randomized complexity and information complexity are roughly the same order. Understanding these differences is one of the major goals in communication complexity research.

10.3.2 Deterministic communication: lower bounds and structure

Deterministic communication complexity lower bounds often admit fairly elegant and somewhat elementary arguments, and the gaps between them and the randomized complexity highlight that we expect providing lower bounds on randomized communication (10.3.1) or information (10.3.3) complexity to be quite challenging. The starting point, to which we return when we consider randomized protocols, is to understand structural aspects of a protocol tree.

A set $R \subset \mathcal{X} \times \mathcal{Y}$ is a *rectangle* if it has the form $R = A \times B$ for some $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$. Rectangular sets provide a key to understanding communication complexity:

[1] This is one direction of Yao's minimax theorem [434], which states that communication complexity with public (shared) randomness and worst-case distributional complexity are identical: $\mathsf{RCC}_\delta^{\text{pub}}(f) = \sup_\mu \mathsf{DCC}_\delta^\mu(f)$.

Proposition 10.3.2. *Let v be a node in a deterministic protocol Π and R_v be those pairs (x, y) reaching node v. Then R_v is a rectangle.*

Proof We prove the result by induction. For the root node v, $R_v = \mathcal{X} \times \mathcal{Y}$ is a rectangle. Let v be a (non-root) node and w its parent; assume w.l.o.g. that v is the left child of w and Alice speaks at w (that is, we use $a_w : \mathcal{X} \to \{0, 1\}$). Then $R_w = A \times B$ by the inductive assumption, and

$$R_v = \{\{x\} \times B \mid a_w(x) = 0, x \in A\} = \{x \in A \mid a_w(x) = 0\} \times B$$

by definition of R_v, which is a rectangle. □

The structure of rectangles for correct protocols thus determines the communication complexity of a function f. For a set $R \subset \mathcal{X} \times \mathcal{Y}$, we say R is f-constant if $f(x, y) = f(x', y')$ for all $(x, y) \in R$ and $(x', y') \in R$. Thus, any correct protocol Π partitions $\mathcal{X} \times \mathcal{Y}$ into a collection of f-constant rectangles, where we identify the rectangles with the leaves l of the protocol tree. In particular, Proposition 10.3.2 implies the following corollary.

Corollary 10.3.3. *Let N be the size of the minimal partition of $\mathcal{X} \times \mathcal{Y}$ into f-constant rectangles. Then $\mathsf{CC}(f) \geq \log_2 N$.*

Proof Any correct protocol Π partitions $\mathcal{X} \times \mathcal{Y}$ into the f-constant rectangles $\{R_l\}$ indexed by its leaves l. The minimal depth of a binary tree with at least N leaves is $\log_2 N$. □

A related corollary follows by considering *fooling sets*, which are basically sets that rectangles cannot contain.

Definition 10.4 (Fooling sets)**.** *A set $S \subset \mathcal{X} \times \mathcal{Y}$ is a* fooling set *for f if for any two pairs $(x_0, y_0) \in S$ and $(x_1, y_1) \in S$ satisfying $f(x_0, y_0) = f(x_1, y_1)$, at least one of the inequalities $f(x_0, y_1) \neq f(x_0, y_0)$ or $f(x_1, y_0) \neq f(x_0, y_0)$ holds.*

With this definition, the next corollary is almost immediate.

Corollary 10.3.4. *Let f have a fooling set S of size N. Then $\mathsf{CC}(f) \geq \log_2 N$.*

Proof By definition, no f-constant rectangle contains more than a single element of S. So the tree associated with any correct protocol Π has a single leaf for each element of S. □

This corollary provides a lower bound that will inform our coming development:

Example 10.3.5 (Equality)**:** Consider the testing equality of n-bit strings $x, y \in \{0, 1\}^n$, letting $f = \text{EQ}$ be $f(x, y) = 1$ if $x = y$ and 0 otherwise. Define the set $S = \{(x, x) \mid x \in \{0, 1\}^n\}$, which has cardinality 2^n, and satisfies $f(x, x) = 1$ for all $(x, x) \in S$. For any (x, x) and $(x', x') \in S$, if $x \neq x'$, then certainly $(x, x') \notin S$, so S is a fooling set and

$$n \leq \mathsf{CC}(\text{EQ}) \leq n + 1,$$

where the upper bound follows by letting Alice communicate x and Bob check if $x = y$. ◇

An extension of the fooling set idea is the *rectangle measure* method, which proves that (for some probability measure P) the "size" of f-constant rectangles is small. By judicious choice of the probability, we can then demonstrate lower bounds.

Proposition 10.3.6. *Let P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. If all f-constant rectangles R have probability at most $P(R) \le \delta$, then $\mathsf{CC}(f) \ge \log_2 \frac{1}{\delta}$.*

Proof By the union bound, any f-constant partition of $\mathcal{X} \times \mathcal{Y}$ into rectangles $\{R_l\}_{l=1}^N$ satisfies $1 \le \sum_{l=1}^N P(R_l) \le N\delta$. So $N \ge \frac{1}{\delta}$, and the result follows by Corollary 10.3.3. □

This proposition allows us to lower bound the communication complexity of inner products on $\mathbb{F}_2$, the field of arithmetic on the integers modulo 2 (that is, with bit strings).

Example 10.3.7 (Inner products)**:** Consider the inner product $\mathrm{IP}_2(x, y) = \langle x, y \rangle \mod 2$ for n-bit strings $x, y \in \{0,1\}^n$. Rather than constructing a fooling set directly, we use Proposition 10.3.6 and let P be the uniform distribution on $\{0,1\}^n \times \{0,1\}^n$. Let $R = A \times B$ be a rectangle with $\langle x, y \rangle = 0$ for all $x \in A$ and $y \in B$. The linearity of the inner product guarantees that $\langle x, y \rangle = 0$ for all $x \in \mathrm{span}(A)$ and $y \in \mathrm{span}(B)$, the (linear) spans of A and B in $\mathbb{F}_2^n$, respectively. Recognize that $\mathrm{span}(A), \mathrm{span}(B) \subset \mathbb{F}_2^n$ are orthogonal subspaces of $\mathbb{F}_2^n$, so their dimensions $d_0 = \dim(\mathrm{span}(A))$ and $d_1 = \dim(\mathrm{span}(B))$ satisfy $d_0 + d_1 \le n$.

If $d_0 = \dim(\mathrm{span}(A))$ then $|A| \le 2^{d_0}$ in $\mathbb{F}_2^n$, so $|R| \le |A| \cdot |B| \le 2^n$, which (under the uniform measure P) satisfies $P(R) \le \frac{2^n}{2^{2n}} = 2^{-n}$. By Proposition 10.3.6, we thus have

$$n \le \mathsf{CC}(\mathrm{IP}_2) \le n + 1,$$

because Alice can communicate $x \in \{0,1\}^n$ and Bob can output $\langle x, y \rangle \mod 2$. ◇

10.3.3 Randomization, conditional information complexity, and direct sums

When we allow randomization, the complexity bounds may drastically change. Consider again the equality function in Example 10.3.5, which admits $O(\log n)$ randomized complexity because nonzero nth degree polynomials have at most n roots.

Example 10.3.8 (Equality with randomization)**:** Let $x, y \in \{0,1\}^n$ and p be a prime number satisfying $n^2 \le p \le 2n^2$ (the Prime Number Theorem guarantees the existence of such a p). Alice chooses $U \sim \mathsf{Uniform}\{0, \ldots, p-1\}$ and computes the polynomial

$$a(U) = x_1 + x_2 U + x_3 U^2 + \cdots + x_n U^{n-1} \mod p.$$

Bob checks whether

$$b(U) = y_1 + y_2 U + y_3 U^2 + \cdots + y_n U^{n-1} \mod p$$

satisfies $b(U) = a(U)$. If so, Bob outputs "Yes" (equality), and otherwise, Bob outputs "No." This protocol requires communicating at most $O(1) \log_2 n$ bits, is always correct when $x = y$, and satisfies that $P(\Pi(x, y) \text{ fails}) < \frac{1}{n}$ when $x \ne y$ (see Exercise 10.10). In summary, $\mathsf{RCC}_{1/n}(\mathrm{EQ}) \le O(1) \log n$, exponentially improving over deterministic complexity. ◇

Example 10.3.8 makes clear that any lower bounds on randomized communication complexity, or, relatedly, information complexity, will necessarily be more subtle than those we have presented for CC. We develop a few of the main ideas here. Because our focus is on information theoretic techniques, we pass over several standard tools for proving lower bounds, though we mention them in the bibliographic notes at the end of the chapter. One of our main goals will be to show that the information complexity of the inner product is indeed $\Omega(n)$, a much stronger result than

Example 10.3.7. In contrast to the lower bounds we provide for minimax risk in most of this book, the focus in communication complexity is to take an *a priori* accurate estimator and demonstrate that it *requires* a certain amount of information to be communicated, rather than the contrapositive result that limited information yields inaccurate estimators. While these are clearly equivalent, it can be fruitful to use the perspective most relevant for the problem at hand.

Two main ideas form the basis for information complexity lower bounds: first, *direct sum* inequalities, which loosely analogize the tensorization and chain rules of Chapter 2, and show that computing a function on n inputs requires roughly order n more communication than computing it on one. The second important insight is to provide lower bounds on the information necessary to compute different primitives, and the particular structure of randomized communication protocols makes this possible. For the remainder of Section 10.3.3, we address the first of these, returning to the information complexity of primitives in Section 10.3.4*.

Direct sum bounds and decomposition

Direct sum inequalities demonstrate that computing a function on n inputs requires roughly n times the communication of single-input computation. In general, we consider functions f of the form

$$f(x_1^n, y_1^n) = g(h(x_1, y_1), h(x_2, y_2), \ldots, h(x_n, y_n)), \tag{10.3.4}$$

where g is the global function of the n primitives h, calling such functions decomposable with primitive h. Many problems have the structure (10.3.4); focusing on the case that the inputs $x, y \in \{0,1\}^n$ and $f(x,y) \in \{0,1\}$, we have the following three immediate examples (some authors define these functions to be $1 - f$ instead of f in our notation, but this is immaterial).

Example 10.3.9 (Decomposition of equality)**:** The equality function $f(x,y) = 1$ if $x = y$ and $f(x,y) = 0$ otherwise satisfies the decomposition (10.3.4), where $h(x_i, y_i) = \mathbf{1}\{x_i \neq y_i\}$ and $g(z) = 1 - \mathbf{1}\{\langle \mathbf{1}, z\rangle > 0\}$. $\diamond$

Example 10.3.10 (Decomposition of inner product)**:** The inner product in $\mathbb{F}_2$, $f(x,y) = \langle x, y\rangle \mod 2$, where $h(x_i, y_i) = x_i y_i$, and $g(z) = \langle \mathbf{1}, z\rangle \mod 2$ satisfies $g(z) = 0$ if $\sum_{i=1}^n z_i$ is even and $g(z) = 1$ otherwise. $\diamond$

While Example 10.3.8 makes clear that the decomposition (10.3.4) is not sufficient to guarantee a randomized complexity lower bound of order n, the structure (10.3.4) will be useful.

To develop the main information complexity direct sum theorem showing that the information complexity of f is at least the sum of the complexities of its constituent primitives, we leverage what we term *plantable inputs*:

Definition 10.5. *Let $f : \mathcal{X}^n \times \mathcal{Y}^n \to \{0,1\}$ have the decomposition (10.3.4), where the primitive h is $\{0,1\}$-valued. The pair $(x,y) \in \mathcal{X}^n \times \mathcal{Y}^n$ admits a* planted solution *if for each $i \in \{1, \ldots, n\}$, all x_i', y_i', and vectors*

$$x' = (x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \quad \text{and} \quad y' = (y_1, \ldots, y_{i-1}, y_i', y_{i+1}, \ldots, y_n),$$

we have $f(x', y') = h(x_i', y_i')$.

The binary inner product in Examples 10.3.7 and 10.3.10 has many plantable inputs: any of the 3^n pairs of vectors $x, y \in \{0,1\}^n$ with $x_i y_i = 0 \mod 2$ for each i admit planted solutions with $\langle x, y\rangle = 0 \mod 2$. For the equality function, only the 2^n pairs $x = y$ admit planted solutions.

We outline the key idea to our direct sum lower bounds. Because we define information complexity for protocols Π that are correct on all inputs with high probability, we can choose any distribution on inputs $(x_1^n, y_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n$. Thus (recalling the fooling set idea in Definition 10.4), we choose a *fooling distribution* μ for f, meaning that for $(X_i, Y_i) \stackrel{\text{iid}}{\sim} \mu$ the pair $(X_1^n, Y_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ always admits a planted solution (Definition 10.5). The next definition says this slightly differently.

Definition 10.6. *A distribution μ on $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a* fooling distribution *if all (x_1^n, y_1^n) in the support of the product μ^n admit planted solutions (Definition 10.5).*

Typically, fooling distributions μ require some dependence between X_i and Y_i—for example, in the inner product, we require $X_i Y_i = 0$, so that if $X_i = 1$ then $Y_i = 0$ and vice versa:

Example 10.3.11 (A fooling distribution for inner products)**:** Define the distribution μ on pairs $(x, y) \in \{0, 1\} \times \{0, 1\}$ as follows: let V be uniform on $\{0, 1\}$, and conditional on $V = 0$, set $X = 0$ and let $Y \sim \mathsf{Uniform}\{0, 1\}$; conditional on $V = 1$, set $Y = 0$ and let $X \sim \mathsf{Uniform}\{0, 1\}$. Then $XY = 0$, and any set of pairs $(X_i, Y_i) \stackrel{\text{iid}}{\sim} \mu$ satisfy that the binary inner product $\text{IP}_2(X_1^n, Y_1^n) = \langle X_1^n, Y_1^n \rangle \mod 2 = 0$. For any i, replacing (X_i, Y_i) with (x_i, y_i) allows the product $x_i y_i$ to determine $\text{IP}_2(X_1^n, Y_1^n)$. $\diamond$

Fooling distributions, as in Example 10.3.11, make conditioning natural in information complexity. If $(X, Y) \sim \mu$, there is always a random variable V such that $X \perp\!\!\!\perp Y \mid V$, that is, X and Y are conditionally independent given V (trivially, we can take $V = X$). Thus, for function $h : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, we define the *conditional information complexity*

$$\mathsf{CIC}_\delta^\mu(h) := \sup_V \inf_\Pi \left\{ I_2(X, Y; \Pi(X, Y) \mid V) \quad \text{s.t.} \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(\Pi_{\text{out}}(x, y) \neq h(x, y)) \leq \delta \right\}, \quad (10.3.5)$$

where the infimum is over all randomized protocols and the supremum is over all random variables making X and Y conditionally independent with joint distribution $(X, Y) \sim \mu$. Equivalently, to lower bound $\mathsf{CIC}_\delta^\mu(h)$ it suffices to exhibit conditioning variable V for which every δ-accurate protocol reveals the desired amount of conditional information.

With this, we obtain the direct sum theorem for information complexity. (When n is clear from context, we write $\mathsf{IC}_\delta^\mu = \mathsf{IC}_\delta^{\mu^n}$ to indicate $(X_i, Y_i)_{i=1}^n \sim \mu^n$.)

Theorem 10.3.12. *Let μ be a fooling distribution on $\mathcal{X} \times \mathcal{Y}$ for a function f with primitive h. Then*

$$\mathsf{IC}_\delta^\mu(f) \geq n \cdot \mathsf{CIC}_\delta^\mu(h).$$

The theorem also explains why this route cannot prove strong lower bounds for equality: the only fooling distributions for equality have $X = Y$ with probability one, so $\mathsf{CIC}_\delta^\mu(h) = 0$ for any such μ.

Proof Fix $\epsilon > 0$ and choose a random variable V making $X \perp\!\!\!\perp Y \mid V$ with $(X, Y) \sim \mu$ such that every δ-accurate protocol for h satisfies

$$I_2(X, Y; \Pi(X, Y) \mid V) \geq \mathsf{CIC}_\delta^\mu(h) - \epsilon.$$

Let V_1^n be i.i.d. copies of this variable and, conditional on V_1^n, draw (X_i, Y_i) independently from $\mu(\cdot \mid V_i)$. It suffices to lower bound the information of an arbitrary δ-accurate protocol Π for f under this product input distribution, because $\mathsf{IC}_\delta(f)$ takes the supremum over input distributions.

Because $V_1^n \to (X_1^n, Y_1^n) \to \Pi$ is a Markov chain,

$$I_2(X_1^n, Y_1^n; \Pi) \geq I_2(X_1^n, Y_1^n; \Pi \mid V_1^n).$$

Using the chain rule for mutual information and conditional independence of the pairs (X_i, Y_i) given V_1^n, we have

$$I_2(X_1^n, Y_1^n; \Pi \mid V_1^n) \geq \sum_{i=1}^n I_2(X_i, Y_i; \Pi \mid V_1^n). \tag{10.3.6}$$

We reduce the global protocol Π to protocols for individual primitives. Fix i and a value $v_{\setminus i}$ of $V_{\setminus i}$. On inputs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, Alice and Bob independently generate $(X_j^*, Y_j^*) \sim \mu(\cdot \mid V_j = v_j)$ for $j \neq i$, which is possible because of the conditional independence given V_j. They then execute

$$\Pi\big((X_{\setminus i}^*, x), (Y_{\setminus i}^*, y)\big),$$

with x and y substituted into coordinate i; call the resulting simulated protocol $\Pi_{i,v}(x, y)$. Because the other coordinates follow the fooling distribution, they form a planted solution, so

$$f\big((X_{\setminus i}^*, x), (Y_{\setminus i}^*, y)\big) = h(x, y).$$

Thus $\Pi_{i,v}$ is a δ-accurate protocol for the primitive h.

Moreover, by construction,

$$(X_i, Y_i, V_i, \Pi_{i,v}(X_i, Y_i)) \stackrel{\text{dist}}{=} (X_i, Y_i, V_i, \Pi(X_1^n, Y_1^n)) \mid V_{\setminus i} = v_{\setminus i}. \tag{10.3.7}$$

Therefore, for every v,

$$I_2(X_i, Y_i; \Pi(X_1^n, Y_1^n) \mid V_i, V_{\setminus i} = v_{\setminus i}) = I_2(X_i, Y_i; \Pi_{i,v}(X_i, Y_i) \mid V_i) \geq \mathsf{CIC}_\delta^\mu(h) - \epsilon.$$

Averaging over $V_{\setminus i}$ and substituting in (10.3.6) yields

$$I_2(X_1^n, Y_1^n; \Pi) \geq n \left(\mathsf{CIC}_\delta^\mu(h) - \epsilon\right).$$

As $\epsilon > 0$ was arbitrary, the theorem follows. □

With Theorem 10.3.12 in hand, we have our desired direct sum result, so that proving information complexity lower bounds reduces to providing lower bounds on the (conditional) information complexity of 1-bit primitives. The following corollary highlights the theorem's applications to inner product (Example 10.3.10).

Corollary 10.3.13. *Let f be the binary inner product $f(x, y) = \langle x, y \rangle \mod 2$. Let μ be the fooling distribution in Example 10.3.11. Then*

$$\mathsf{IC}_\delta^\mu(f) \geq n \cdot \mathsf{CIC}_\delta^\mu(h)$$

where $h(a, b) = ab$ is the product (or AND) function.

Exercise 10.11 explores similar results for the set-disjointness function, and Exercise 10.12 shows that the entrywise less-than-or-equal function admits similar complexity lower bounds.

10.3.4* The structure of randomized communication and communication complexity of primitives

Theorem 10.3.12 provides a powerful direct sum result that demonstrates that, at least if a problem admits planted solutions for (nearly) i.i.d. sampling, then the information complexity must scale at least linearly in the complexity of the primitives making up the function f. Thus, we turn to providing information lower bounds for computing different primitive functions. Our main tool will be to show that even randomized communication protocols essentially partition the input space $\mathcal{X} \times \mathcal{Y}$ into rectangles—in analogy with Proposition 10.3.2 in the deterministic case—which allows us to provide lower bounds. The broad idea is simple: if we have an accurate protocol for computing a function h, we must necessarily be able to distinguish between the distribution of Π on different inputs (x, y), as the fundamental connection between tests and variation distance (Proposition 2.3.1) reveals.

Our main goal now is to prove the following proposition, which gives a lower bound on the (conditional) information complexity of computing the AND of two bits.

Proposition 10.3.14. *Let $h(x, y) = xy$ for inputs $x, y \in \{0, 1\}$. Let μ be the fooling distribution in Example 10.3.11. Then*

$$\mathsf{CIC}_\delta^\mu(h) \geq \frac{1}{4}\left(1 - 2\sqrt{\delta(1-\delta)}\right).$$

As an immediate corollary, we obtain the following lower bounds on the communication complexity of binary inner product.

Corollary 10.3.15. *Let f be the binary inner product $f(x, y) = \langle x, y \rangle \mod 2$. Then*

$$\mathsf{IC}_\delta(f) \geq \frac{n}{4}\left(1 - 2\sqrt{\delta(1-\delta)}\right).$$

The remainder of this section proves Proposition 10.3.14. To control the complexity of computing individual primitives, it proves easier to use metrics tied more directly to testing. The key will be to leverage that for *binary* random variables, the mutual information and Hellinger distance become, essentially, the same to within numerical constants, which then allows us to directly control the Hellinger information using an analogue of the rectangular structures we develop in non-randomized communication complexity (recall Proposition 10.3.2). The next lemma shows the equivalence of information and Hellinger distance in the binary case:

Lemma 10.3.16. *Let $V \in \{0, 1\}$ be a uniform binary random variable, and conditional on $V = v$, let $X \sim P_v$. Then*

$$\log 2 \cdot d_{\mathrm{hel}}^2(P_0, P_1) \leq I(X; V) \leq 2 \cdot d_{\mathrm{hel}}^2(P_0, P_1).$$

Proof We assume without loss of generality that P_0 and P_1 have densities p_0 and p_1, and let $\overline{P} = \frac{1}{2}(P_0 + P_1)$ and $\overline{p} = \frac{1}{2}(p_0 + p_1)$.

For the upper bound, we have

$$I(X; V) = \frac{1}{2} D_{\mathrm{kl}}\left(P_0 \| \overline{P}\right) + \frac{1}{2} D_{\mathrm{kl}}\left(P_1 \| \overline{P}\right).$$

As a consequence, observe that by Proposition 2.2.10,

$$\begin{aligned} 2I(X; V) &\leq D_{\chi^2}\left(P_0 \| \overline{P}\right) + D_{\chi^2}\left(P_1 \| \overline{P}\right) \\ &= \int \left(\frac{p_0 - p_1}{p_0 + p_1}\right)^2 (p_0 + p_1) = \int \frac{(\sqrt{p_0} - \sqrt{p_1})^2 (\sqrt{p_0} + \sqrt{p_1})^2}{p_0 + p_1} \leq 2 \int (\sqrt{p_0} - \sqrt{p_1})^2, \end{aligned}$$

where we have used that $(\sqrt{a}+\sqrt{b})^2 \le 2(a+b)$ for any $a, b \ge 0$.

For the lower bound, we use the analytic inequality $h_2(p) \le 2\log 2\sqrt{p(1-p)}$ on the binary entropy $h_2(p) = -p\log p - (1-p)\log(1-p)$ (recall Exercise 2.21). Then by symmetry,

$$I(X;V) = H(V) - H(V \mid X) = \log 2 - \mathbb{E}[h_2(\mathbb{P}(V=1 \mid X))],$$

and so

$$\frac{1}{\log 2} I(X;V) \ge 1 - 2\mathbb{E}\left[\sqrt{\mathbb{P}(V=1 \mid X)(1-\mathbb{P}(V=1 \mid X))}\right].$$

As $\mathbb{P}(V=v \mid X=x) = \frac{p_v(x)}{p_0(x)+p_1(x)}$ by Bayes' rule, we rewrite this via

$$\frac{1}{\log 2} I(X;V) \ge 1 - \int (p_0+p_1)\sqrt{\frac{p_0}{p_0+p_1}\cdot\frac{p_1}{p_0+p_1}} = 1 - \int \sqrt{p_0 p_1} = d_{\rm hel}^2(P_0, P_1),$$

giving the result. □

Rescaling Lemma 10.3.16 to work with base-2 logarithms, for a uniform bit $B \in \{0,1\}$ and drawing $Z \sim P_b$ conditional on $B=b$, we have

$$I_2(Z;B) \ge d_{\rm hel}^2(P_0, P_1).$$

To apply the inequality to prove Proposition 10.3.14, recall the fooling distribution μ for inner products in Example 10.3.11, where $V \sim \mathsf{Uniform}\{0,1\}$ and conditional on $V=0$ we set $X=0$ and draw $Y \sim \mathsf{Uniform}\{0,1\}$, and otherwise $Y=0$ and $X \sim \mathsf{Uniform}\{0,1\}$. Then for $V \to (X,Y)$ from this distribution, we have

$$I_2(X,Y;\Pi(X,Y) \mid V) = \frac{1}{2} I_2(Y;\Pi(0,Y) \mid V=0) + \frac{1}{2} I_2(X;\Pi(X,0) \mid V=1).$$

Letting Q_{xy} denote the (conditional) distribution over Π on input bits $x, y \in \{0,1\}$ and noting that X and Y above are each uniform on $\{0,1\}$, Lemma 10.3.16 applies to each of $I_2(Y;\Pi(0,Y) \mid V=0)$ and $I_2(X;\Pi(X,0) \mid V=1)$, and so

$$I_2(X,Y;\Pi(X,Y) \mid V) \ge \frac{1}{2} d_{\rm hel}^2(Q_{01}, Q_{00}) + \frac{1}{2} d_{\rm hel}^2(Q_{10}, Q_{00}).$$

Applying the triangle inequality that $(a-b)^2 \le (|a-c|+|c-b|)^2 \le 2(a-c)^2 + 2(b-c)^2$, we obtain the following lemma.

Lemma 10.3.17. *Let Π be any protocol acting on two bit inputs $x, y \in \{0,1\}$, and let μ be the fooling distribution in Example 10.3.11. Let Q_{xy} be the distribution of $\Pi(x,y)$ on inputs x, y. Then*

$$I_2(X,Y;\Pi(X,Y) \mid V) \ge \frac{1}{4} d_{\rm hel}^2(Q_{01}, Q_{10}).$$

The last step in the proof of Proposition 10.3.14 is to demonstrate a property of (randomized) protocols Π analogous to the rectangular property of deterministic communication that Propositions 10.3.2 and 10.3.6 demonstrate. In analogy with the output leaf in the tree for deterministic communication complexity, let τ be the *transcript* of the communication protocol, that is, its entire communication trace. Then we claim the following analog of Proposition 10.3.2 that the set of inputs resulting in a particular output in deterministic complexity is a rectangle in $\mathcal{X} \times \mathcal{Y}$.

Lemma 10.3.18. *Let Π be any randomized protocol with inputs in $\mathcal{X} \times \mathcal{Y}$ using independent private randomness for Alice and Bob. Then there exist functions q_x and q_y such that for any transcript τ,*

$$\mathbb{P}(\Pi(x, y) = \tau) = q_x(\tau) \cdot q_y(\tau).$$

Proof Conditioning on Alice's and Bob's private randomness, we may view the randomized protocol as a deterministic protocol $\Pi(\cdot, \cdot, u_a, u_b)$, where $u_a, u_b \in [0, 1]$ are realizations of the randomness available to Alice and Bob, respectively, inducing a particular binary communication tree. By Proposition 10.3.2, for any leaf l, the set

$$R_l(u_a, u_b) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid \Pi(x, y, u_a, u_b) \text{ reaches } l\}$$

is a rectangle, that is, $R_l(u_a, u_b) = A_l(u_a) \times B_l(u_b)$ for sets $A_l(u) \subset \mathcal{X}$ and $B_l(u) \subset \mathcal{Y}$. Of course, the leaves l of the tree are in bijection with the entire transcript τ, so that if τ ends in leaf l, then

$$\mathbb{P}(\Pi(x, y) = \tau) = \mathbb{P}((x, y) \in R_l(U_a, U_b)) = \mathbb{P}(x \in A_l(U_a), y \in B_l(U_b))$$

where $U_a, U_b \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, 1]$ are the randomness Alice and Bob use, respectively.

Expanding this as an integral gives

$$\begin{aligned}\mathbb{P}(x \in A_l(U_a), y \in B_l(U_b)) &= \int_0^1 \int_0^1 \mathbf{1}\left\{x \in A_l(u_a)\right\} \mathbf{1}\left\{y \in B_l(u_b)\right\} du_a du_b \\ &= \mathbb{P}(x \in A_l(U_a))\mathbb{P}(y \in B_l(U_b)).\end{aligned}$$

Set $q_x(\tau) = \mathbb{P}(x \in A_l(U_a))$ and $q_y(\tau) = \mathbb{P}(y \in B_l(U_b))$. □

We thus have the following key *cut and paste* property, which shows that in some sense, Hellinger distances respect the "rectangular" structure of communication protocols.

Lemma 10.3.19. *Let Π be any protocol acting on inputs in $\mathcal{X} \times \mathcal{Y}$ and let $Q_{x,y}$ be the distribution of $\Pi(x, y)$ on inputs x, y. Then*

$$d_{\rm hel}(Q_{x,y}, Q_{x',y'}) = d_{\rm hel}(Q_{x,y'}, Q_{x',y}).$$

Proof Let $\mathcal{T}$ be the collection of all possible transcripts the protocol outputs. By Lemma 10.3.18 we have

$$\begin{aligned}d_{\rm hel}^2(Q_{x,y}, Q_{x',y'}) &= \frac{1}{2} \sum_{\tau \in \mathcal{T}} \left(\sqrt{Q_{x,y}(\tau)} - \sqrt{Q_{x',y'}(\tau)}\right)^2 \\ &= \frac{1}{2} \sum_{\tau \in \mathcal{T}} \left(\sqrt{q_x(\tau) q_y(\tau)} - \sqrt{q_{x'}(\tau) q_{y'}(\tau)}\right)^2 = 1 - \sum_{\tau} \sqrt{q_x(\tau) q_y(\tau) q_{x'}(\tau) q_{y'}(\tau)}.\end{aligned}$$

Rearranging by the trivial modification $q_x q_y q_{x'} q_{y'} = q_x q_{y'} q_{x'} q_y$, we have the result. □

We now finalize the proof of Proposition 10.3.14. Substituting this cutting and pasting in Lemma 10.3.17 we have

$$I_2(X, Y; \Pi(X, Y) \mid V) \ge \frac{1}{4} d_{\rm hel}^2(Q_{01}, Q_{10}) = \frac{1}{4} d_{\rm hel}^2(Q_{00}, Q_{11}).$$

Then a simple lemma recalling the testing inequalities in Section 2.3.1 completes the proof of the proposition, because it guarantees that $4I_2(X, Y; \Pi(X, Y) \mid V) \ge 1 - 2\sqrt{\delta(1-\delta)}$ no matter the choice of protocol Π, and so

$$\mathsf{CIC}_\delta^\mu(h) \ge \inf_\Pi I_2(X, Y; \Pi(X, Y) \mid V) \ge \frac{1}{4}\left(1 - 2\sqrt{\delta(1-\delta)}\right).$$

Lemma 10.3.20. *Let Π be any δ-accurate protocol for computing $h(x, y) = xy$ and Q_{xy} be its distribution on inputs (x, y). Then $d_{\rm hel}^2(Q_{00}, Q_{11}) \ge 1 - 2\sqrt{\delta(1-\delta)}$.*

Proof Assume that Π computes the product $xy \in \{0, 1\}$ correctly with probability at least $1 - \delta$, that is, $\mathbb{P}(\Pi_{\rm out}(x, y) \ne xy) \le \delta$ for all $x, y \in \{0, 1\}$. By Le Cam's testing lower bounds (Proposition 2.3.1), we know that

$$\begin{aligned} 2\delta \ge \mathbb{P}(\Pi_{\rm out}(0, 0) \ne 0) + \mathbb{P}(\Pi_{\rm out}(1, 1) \ne 1) &\ge 1 - \|Q_{00} - Q_{11}\|_{\rm TV} \\ &\overset{(\star)}{\ge} 1 - d_{\rm hel}(Q_{00}, Q_{11})\sqrt{2 - d_{\rm hel}^2(Q_{00}, Q_{11})}, \end{aligned}$$

where inequality $(\star)$ follows from the inequalities in Proposition 2.2.8 relating Hellinger and total-variation distance. Let $d = d_{\rm hel}^2(Q_{00}, Q_{11})$ for shorthand. Then rearranging gives $d(2 - d) \ge (1 - 2\delta)^2$. Solving for d in $0 \ge d^2 - 2d + (1 - 2\delta)^2$ yields $d \ge 1 - \sqrt{1 - (1 - 2\delta)^2}$. Recognize that $1 - (1 - 2\delta)^2 = 4(\delta - \delta^2)$. □

10.4 Communication complexity in estimation

We have developed communication complexity and strong data processing inequalities for a major application: to understand the information complexity of statistical estimation itself. In this context, we limit the amount of information—or bits—that a procedure may send about individual examples, and then ask to what extent this constrains the estimator. This has applications in situations in which the memory available to an estimator is limited, to private estimators and of course, when we restrict the number of bits different machines storing distributed data may send.

We consider the following setting: m machines, or agents, have data X_i, $i = 1, \ldots, m$. Communication proceeds in rounds $t = 1, 2, \ldots, T$, where in each round t machine i sends datum $Z_i^{(t)}$. To allow for powerful protocols with little restriction except that each machine i may send only a certain amount of information, $Z_i^{(t)}$ may depend arbitrarily on the previous messages $Z_1^{(t)}, \ldots, Z_{i-1}^{(t)}$ as well as $Z_k^{(\tau)}$ for all $k \in \{1, \ldots, m\}$ and $\tau < t$. We visualize this as a public blackboard B, where in each round t each $Z_i^{(t)}$ is collected into $B^{(t)}$, along with the previous public blackboards $B^{(\tau)}$ for $\tau < t$, and all machines may read these public blackboards. We use the notation $Z_{<i} := (Z_1, \ldots, Z_{i-1})$, and will use $Z_{\le i} := (Z_1, \ldots, Z_i)$ and similarly for superscripts throughout, including the shorthand $Z_{\to i}^{(t)} = (B^{(t-1)}, Z_{<i}^{(t)})$ to denote all the messages coming into communication of $Z_i^{(t)}$. Thus in round t, individual i generates the communicated variable $Z_i^{(t)}$ according to the channel

$$Q_{Z_i^{(t)}}(\cdot \mid X_i, Z_{<i}^{(t)}, B^{(t-1)}) = Q_{Z_i^{(t)}}(\cdot \mid X_i, Z_{\to i}^{(t)}).$$

Figure 10.3 illustrates two rounds of this communication scheme.

Our approach to the lower bounds, which we provide in Sections 10.4.1 and 10.4.2, is roughly as follows. First, we develop another *direct sum* bound, in analogy with Theorem 10.3.12, meaning that

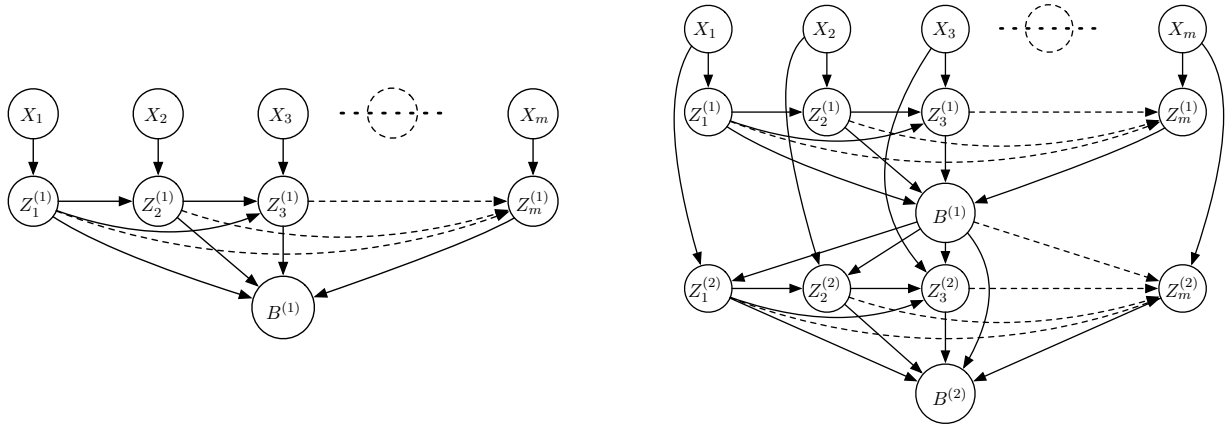


Figure 10.3. Left: single round of communication of variables, writing to public blackboard $B^{(1)}$. Right: two rounds of communication of variables, writing to public blackboards $B^{(1)}$ and $B^{(2)}$.

the difficulty of solving a d-dimensional problem is roughly d-times that of solving a 1-dimensional version of the problem; thus, any lower bounds on the error in 1-dimensional problems imply lower bounds for d-dimensional problems. Second, we provide an extension of the data processing inequalities we have developed thus far to apply to particular communication scenarios.

The key to our reductions is that we consider families of distributions where the coordinates of X are independent, which dovetails with Assouad's method. We thus index our distributions by $v \in \{0,1\}^d$, and in proving our lower bounds, assume the typical Markov structure

$$V \to (X_1, \ldots, X_m) \to \Pi(X_1^m),$$

where V is chosen uniformly at random from $\{-1,1\}^d$, and $T = \Pi(X_1^m)$ denotes the transcript of the entire communication, i.e., the blackboard messages

$$T = (B^{(1)}, \ldots, B^{(T)}),$$

(this also encodes the message order). We assume that X follows a d-dimensional product distribution, so that conditional on $V = v$ we have

$$X \sim P_v = P_{v_1} \otimes P_{v_2} \otimes \cdots \otimes P_{v_d}. \tag{10.4.1}$$

The generation strategy (10.4.1) guarantees that conditional on the jth coordinate $V_j = v_j$, the coordinates $X_{i,j}$ are i.i.d. and independent of $V_{\setminus j} = (V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_d)$ as well as independent of $X_{i',j}$ for data points $i' \neq i$.

10.4.1 Direct sum communication bounds

Our first step is to argue that, for data of the form (10.4.1), lower bounds on the information complexity of one-dimensional estimation imply lower bounds on d-dimensional problems that scale with the dimension. To accomplish this reduction, let $X_{\le m,j} = (X_{i,j})_{i=1}^m$ be the jth coordinate of the data, and let $X_{\le m,\setminus j}$ be the remaining $d-1$ coordinates across all $i = 1, \ldots, m$. Then by the construction (10.4.1), we have the Markov structure

$$V_j \to X_{\le m,j} \to \Pi(X_1^m) \leftarrow X_{\le m,\setminus j} \leftarrow V_{\setminus j}.$$

Viewing $X_{\le m,\setminus j}$ as extraneous randomness, we have the simpler Markovian structure

$$V_j \to X_{\le m,j} \to T \tag{10.4.2}$$

where we think of Π as acting only on $X_{\le m,j}$ while simulating $X_{\le m,\setminus j}$. Now, define M_{-j} and M_{+j} to be the marginal distributions over the output transcript $T = \Pi(X_1^m)$ conditional on $V_j = \pm j$, the one-variable model (10.4.2). Le Cam's testing inequality (Proposition 2.3.1), and the equivalence between Hellinger and variation distance (Proposition 2.2.8) imply

$$\begin{aligned}\inf_{\widehat{V}} 2\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\tau) \neq V_j) \ge \sum_{j=1}^d (1 - \|M_{-j} - M_{+j}\|_{\rm TV}) &\ge \sum_{j=1}^d (1 - \sqrt{2} d_{\rm hel}(M_{-j}, M_{+j})) \\ &\ge d\left(1 - \left(\frac{2}{d}\sum_{j=1}^d d_{\rm hel}^2(M_{-j}, M_{+j})\right)^{1/2}\right)\end{aligned}$$

by Cauchy-Schwarz. Summarizing, we have the following

Proposition 10.4.1 (Assouad's method in communication)**.** *Let M_{+j} be the marginal distribution over $T = \Pi(X_1^m)$ conditional on $V_j = 1$ and M_{-j} be the marginal T conditional on $V_j = -1$ in Markov structure (10.4.2), and assume X_i follow the product distribution (10.4.1). Then*

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(T) \neq V_j) \ge \frac{d}{2}\left(1 - \sqrt{\frac{2}{d}\sum_{j=1}^d d_{\rm hel}^2(M_{-j}, M_{+j})}\right).$$

Recalling Assouad's method (Theorem 9.5.2) of Section 9.5, we see that any time we have a problem with separation with respect to the Hamming metric (9.5.1), we have a lower bound on its error in estimation problems. This proposition analogizes Theorem 10.3.12, in that small Hellinger distance between the individual marginals $M_{\pm j}$ necessarily makes the testing and estimation problems hard.

10.4.2 Communication data processing

We now revisit the data processing inequalities in Section 10.1, where we consider a variant that allows us to prove lower bounds for estimation problems with limited communication. It will be more notationally convenient in this section to use $V \in \{0, 1\}$ rather than $\{-1, 1\}$, so we state this change explicitly. We begin with a revised strong data processing inequality.

Definition 10.7. *Let P_0, P_1 be arbitrary distributions on a space $\mathcal{X}$, let $V \in \{0, 1\}$ uniformly at random, and conditional on $V = v$, draw $X \sim P_v$. Consider the Markov chain $V \to X \to Z$. The* mutual information strong data processing constant $\beta(P_0, P_1)$ *is*

$$\beta(P_0, P_1) := \sup_{X \to Z} \frac{I(V; Z)}{I(X; Z)},$$

where the supremum is taken over all conditional distributions (Markov kernels) from X to Z.

In contrast to Definition 10.1, in this definition we have a contraction over the "beginning" of the chain $V \to X$ rather than the distribution $X \to Z$. Identifying Z with a communication protocol $\Pi(X_1^m)$, this makes it possible to develop lower bounds on estimation and testing that then depend on the information $I(X; \Pi(X))$.

Distributions with bounded likelihood ratios provide one way to demonstrate a strong data processing inequality of the form in Definition 10.7, where in analogy with Theorem 10.2.1 we obtain a contraction inequality involving the total variation distance.

Proposition 10.4.2. *Let $V \to X \to Z$, where $X \sim P_v$ conditional on $V = v$. Let P_X and $P_X(\cdot \mid Z)$ denote the marginal and conditional distributions on X given Z, respectively. If $|\log \frac{dP_v}{dP_{v'}}| \le \alpha$ for all v, v', then*

$$I(V; Z) \le 4(e^\alpha - 1)^2 \mathbb{E}_Z \left[\|P_X(\cdot \mid Z) - P_X\|_{\rm TV}^2 \right] \le 2(e^\alpha - 1)^2 I(X; Z).$$

We leave the proof of this proposition as Exercise 10.14, as it follows by adapting the techniques we use to prove Theorem 10.2.1, with the main difference being the random variables with bounded likelihood ratios ($X \to Z$ versus $V \to X$). A brief example illustrates Proposition 10.4.2.

Example 10.4.3 (Bernoulli distributions)**:** Let $P_v = \mathsf{Bernoulli}(\frac{1+v\delta}{2})$ for $v \in \{-1, 1\}$. Then we have likelihood ratio bound

$$\left| \log \frac{dP_1}{dP_{-1}} \right| \le \log \frac{1+\delta}{1-\delta}$$

and so under the conditions of Proposition 10.4.2, for any Z we have

$$I(V; Z) \le 2 \left(\frac{1+\delta}{1-\delta} - 1 \right)^2 I(X; Z) = 2 \left(\frac{2\delta}{1-\delta} \right)^2 I(X; Z) \stackrel{(i)}{\le} 10\delta^2 I(X; Z),$$

where inequality (i) holds for $\delta \in [0, 1/10]$. $\diamond$

We now give the two main results connecting mutual information and the contraction-type bounds in Definition 10.7. To provide bounds using Proposition 10.4.1, we wish to control the Hellinger distance between individual marginals $M_{\pm j}$, so we consider single variables in the Markov chain

$$V \to (X_1, \ldots, X_m) \to \Pi,$$

where $V \in \{0, 1\}$. To state the coming theorems, we make a restriction on the data generation $V \to X$, calling distributions P_0 and P_1 (c, β)-*contractive* if

$$\beta(P_0, P_1) \le \beta \le 1 \quad \text{and} \quad \max\left\{ D_\infty\left(P_0 \| P_1\right), D_\infty\left(P_1 \| P_0\right) \right\} \le \log c, \tag{10.4.3}$$

where $D_\infty\left(\cdot \| \cdot\right)$ denotes the Rényi-∞-divergence. Proposition 10.4.2 shows that whenever such a c exists we certainly have $\beta(P_0, P_1) \le 2(c-1)^2$.

The next theorem then provides the basic information contraction inequality for single-variable communication.

Theorem 10.4.4. *Let $1 \le c < \infty$ and $\beta \le 1$. Let P_0 and P_1 be (c, β)-contractive (10.4.3) distributions on $\mathcal{X}$ and M_v, $v \in \{0, 1\}$ be the marginal distribution of the protocol Π conditional on $V = v$. Then*

$$d_{\rm hel}^2(M_0, M_1) \le \frac{7}{2}(c+1)\beta \cdot \min\left\{ I(X_1^m; \Pi(X_1^m) \mid V = 0), I(X_1^m; \Pi(X_1^m) \mid V = 1) \right\}.$$

The proof of Theorem 10.4.4 is quite complicated, so we defer it to Section 10.5*.

We can use Theorem 10.4.4 to obtain bounds on the probability of error—detection of d-dimensional signals—in higher dimensional problems based on mutual information alone. Because the theorem provides a bound involving the minimum of the conditional mutual informations, we have substantial freedom to massage the direct-sum lower bounds in Section 10.4.1 into the mutual information between the data X_1^m and the protocol $\Pi(X_1^m)$.

We thus recall the definition (10.4.1) of our product distribution signals, where we assume that each individual datum $X_i = (X_{i,1}, \ldots, X_{i,d}) = (X_{i,j})_{j=1}^d$ belongs to a d-dimensional set and, conditional on $V = v \in \{-1, 1\}^d$, has independent coordinates distributed as $X_{i,j} \sim P_{v_j}$. With this, we have the following theorem, which follows by a combination of Assouad's method (in the context of communication bounds, i.e. Proposition 10.4.1) and Theorem 10.4.4.

Theorem 10.4.5. *Let Π be the communication protocol in Figure 10.3, $V \in \{-1, 1\}^d$ be uniform, and generate $X_i \stackrel{\text{iid}}{\sim} P_v$, $i = 1, \ldots, m$ according to the independent coordinate distribution (10.4.1). Assume additionally that, for each coordinate $j = 1, \ldots, d$, the pair of coordinate distributions P_{-1} and P_1 is (c, β)-contractive (10.4.3). Then for any estimator $\widehat{V}$,*

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(\Pi(X_1^m)) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d} \cdot I(X_1, \ldots, X_m; \Pi(X_1^m) \mid V)}\right).$$

Proof Let $T = \Pi(X_1^m)$ be the (randomized) transcript. Under the given conditions, Proposition 10.4.1 and Theorem 10.4.4 immediately combine to give

$$\sum_{j=1}^d \mathbb{P}(\widehat{V}_j(T) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d}\sum_{j=1}^d \min_{v \in \{-1,1\}} I(X_{1,j}, \ldots, X_{m,j}; T \mid V_j = v)}\right).$$

Certainly

$$\min_{v \in \{-1,1\}} I(X_{1,j}, \ldots, X_{m,j}; T \mid V_j = v) \leq I(X_{1,j}, \ldots, X_{m,j}; T \mid V_j).$$

Then, using that w.l.o.g. we may assume the $X_{i,j}$ are discrete, we obtain

$$\begin{aligned}
\sum_{j=1}^d I((X_{i,j})_{i=1}^m; T \mid V_j) &= \sum_{j=1}^d \left[H((X_{i,j})_{i=1}^m \mid V_j) - H((X_{i,j})_{i=1}^m \mid T, V_j)\right] \\
&\stackrel{(i)}{=} \sum_{j=1}^d \left[H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, V) - H((X_{i,j})_{i=1}^m \mid T, V_j)\right] \\
&\leq \sum_{j=1}^d \left[H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, V) - H((X_{i,j})_{i=1}^m \mid (X_{i,j'})_{i \leq m, j' < j}, T, V)\right] \\
&= \sum_{j=1}^d I((X_{i,j})_{i=1}^m; T \mid V, (X_{i,j'})_{i \leq m, j' < j}) = I(X_1, \ldots, X_m; T \mid V),
\end{aligned}$$

where equality (i) used the independence of $X_{i,j}$ from $V_{\setminus j}$ and $X_{i,j'}$ for $j' \neq j$ given V_j, and the inequality that conditioning reduces entropy. This gives the theorem. □

10.4.3 Applications: communication and privacy lower bounds

Let us turn to a few applications of our lower bounds on communication complexity. To use them to develop estimation lower bounds, we evidently require two conditions: first, we must show that the data distributions satisfy a strong (mutual information) data processing inequality (Definition 10.7). Second, we must provide a (good enough) upper bound on the mutual information

$I(X_1^m; \Pi(X_1^m) \mid V)$ between the data points X_i and communication. We focus mainly on situations with bounded likelihood ratio, where Proposition 10.4.2 directly provides the type of strong data processing inequality we require.

Communication lower bounds

Our first set of examples considers direct communication bounds, where controlling $I(X_1^m; \Pi)$ is relatively straightforward. Assume the setting in the introduction to Section 10.4, where each machine $i = 1, \ldots, m$ may send at most B_i total bits of information throughout the entire communication protocol—that is, for each pair i, t, we have bounds

$$H(Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le B_{i,t} \quad \text{and} \quad \sum_t B_{i,t} \le B_i \tag{10.4.4}$$

on the messages from X_i across rounds t. (This is weaker than bounding $H(Z_i^{(t)}) \le B_{i,t}$.)

Consider estimating the parameters of d independent Bernoulli distributions in squared error. Let $\mathcal{P}_d$ be the family of d-dimensional Bernoulli distributions, where we let the parameter $\theta \in [0,1]^d$ be such that $P_\theta(X_j = 1) = \theta_j$. Then we have the following result.

Proposition 10.4.6. *Let $\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m)$ denote the minimax mean-square error for estimation of a d-dimensional Bernoulli under the information constraint (10.4.4). Then*

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c \min\left\{\frac{d}{m} \frac{d}{\frac{1}{m}\sum_{i=1}^m B_i}, d\right\},$$

where $c > 0$ is a numerical constant.

Proof By the standard Assouad reduction (Section 9.5), when we take coordinate distributions $P_{v_j} = \mathsf{Bernoulli}(\frac{1+\delta v_j}{2})$, we have a $c\delta^2$-separation in Hamming metric. Applying Theorem 10.4.5 and Example 10.4.3, we obtain the minimax lower bound, valid for $0 \le \delta \le \frac{1}{10}$, of

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c\delta^2 d \left(1 - \sqrt{C \frac{\delta^2}{d} I(X_1^m; \Pi(X_1^m) \mid V)}\right).$$

Now, we note that for any Markov chain $V \to X \to Z$,

$$I(X; Z \mid V) = H(Z \mid V) - H(Z \mid X, V) = H(Z \mid V) - H(Z \mid X) \le H(Z) - H(Z \mid X) = I(X; Z).$$

Thus we obtain

$$\begin{aligned} I(X_1, \ldots, X_m; \Pi \mid V) &\le I(X_1, \ldots, X_m; \Pi) \\ &= \sum_{i=1}^m \sum_{t=1}^T I(X_1, \ldots, X_m; Z_i^{(t)} \mid Z_{\to i}^{(t)}). \end{aligned}$$

As the message $Z_i^{(t)}$ satisfies the conditional independence $Z_i^{(t)} \perp\!\!\!\perp X_{\setminus i} \mid Z_{\to i}^{(t)}, X_i$, this final quantity equals $\sum_{i,t} I(X_i; Z_i^{(t)} \mid Z_{\to i}^{(t)})$. But of course $I(X_i; Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le H(Z_i^{(t)} \mid Z_{\to i}^{(t)}) \le B_{i,t}$, and so

$$\mathfrak{M}_m(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \{B_i\}_{i=1}^m) \ge c\delta^2 d \left(1 - \sqrt{C \frac{\delta^2}{d} \sum_{i,t} B_{i,t}}\right).$$

Choosing $\delta^2 = c_0 \min\{1, d/\sum_i B_i\}$ for a sufficiently small numerical constant c_0 gives the result. $\square$

This result deserves some discussion. It is sharp in the case that the number of bits is of order d or less from each machine: when we set $B_i = d$, the lower bound becomes

$$\sup_\theta \mathbb{E}_\theta[\|\widehat{\theta}(\Pi) - \theta\|_2^2] \gtrsim \min\left\{\frac{d}{m} \cdot \frac{d}{d}, d\right\} = \frac{d}{m},$$

which is certainly achievable (each machine simply sends its entire vector $X_i \in \{0,1\}^d$). When machines communicate fewer than d bits, we have a tighter result; for example, if only k/m machines send d bits, and the rest communicate little, we obtain

$$\sup_\theta \mathbb{E}_\theta[\|\widehat{\theta}(\Pi) - \theta\|_2^2] \gtrsim \min\left\{\frac{d}{m} \cdot \frac{md}{kd}, d\right\} = \frac{d}{k},$$

which is similarly intuitive. The extension of these ideas to the case when each machine has an individual sample of size n is more challenging, as it requires tensorized variants of the strong data processing inequality in Definition 10.7; we provide remarks in the bibliographical section.

Lower bounds in locally private estimation

We return to the local privacy setting in Section 10.2, except now we allow substantially more interaction. We treat local differential privacy in the communication model of Figure 10.3, where n individuals have data X_i they wish to privatize, and proceed in rounds, releasing data $Z_i^{(t)}$ from individual i in round t. We assume each data release $Z_i^{(t)}$ is $\varepsilon_{i,t}$-differentially private: instead of the sequentially interactive model (10.2.1), we have

$$Q(Z_i^{(t)} \in A \mid X_i = x, Z_{\to i}^{(t)} = z_{\to i}^{(t)}) \le \exp(\varepsilon_{i,t}) \cdot Q(Z_i^{(t)} \in A \mid X_i = x', Z_{\to i}^{(t)} = z_{\to i}^{(t)}) \tag{10.4.5}$$

for each i, t and all possible $x, x', z_{\to i}^{(t)}$. At a more abstract level, rather than a particular privacy guarantee on each individual data release $Z_i^{(t)}$, we can assume a more global stability guarantee akin to the (average) KL-stability in interactive data analysis (Definition 6.1). Thus, let $\Pi(x_1^n)$ be the entire collection of communicated information in the protocol in Figure 10.3 on input data $x_1, \ldots, x_n$. Abusing notation to let $D_{\rm kl}\left(Z_0 \| Z_1\right)$ be the KL-divergence between the distributions of Z_0 and Z_1, as in Definition 6.1, we make the following definition to capture arbitrary interactions.

Definition 10.8 (Average KL-privacy)**.** *Let the samples $x_{\le n} \in \mathcal{X}^n$ and $x_{\le n}^{(i)} \in \mathcal{X}^n$ differ only in example i. Then the data release Π is $\varepsilon_{\rm kl}$-*KL-locally-private on average *if*

$$\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(\Pi(x_{\le n}^{(i)}) \| \Pi(x_{\le n})\right) \le \varepsilon_{\rm kl}.$$

The following observation shows that for appropriate choices of $\varepsilon_{\rm kl}$, this is indeed weaker than the interactive guarantee (10.4.5).

Lemma 10.4.7. *Let the communication Q satisfy the interactive privacy guarantee (10.4.5) and Π be the induced communication protocol over rounds $t \le T$. Then*

$$\frac{1}{n}\sum_{i=1}^n D_{\rm kl}\left(\Pi(x_{\le n}^{(i)}) \| \Pi(x_{\le n})\right) \le \frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T \min\left\{\varepsilon_{i,t}, \frac{3}{2}\varepsilon_{i,t}^2\right\}.$$

Proof Using the chain rule for the KL-divergence, we have for any j that

$$D_{\rm kl}\left(\Pi(x_{\le n}^{(j)})\|\Pi(x_{\le n})\right) = \sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\left[D_{\rm kl}\left(Q(Z_i^{(t)}\in\cdot\mid x_i^{(j)}, Z_{\to i}^{(t)})\|Q(Z_i^{(t)}\in\cdot\mid x_i, Z_{\to i}^{(t)})\right)\right]$$
$$= \sum_{t=1}^{T}\mathbb{E}\left[D_{\rm kl}\left(Q(Z_j^{(t)}\in\cdot\mid x_j^{(j)}, Z_{\to i}^{(t)})\|Q(Z_j^{(t)}\in\cdot\mid x_j, Z_{\to i}^{(t)})\right)\right].$$

where the expectation is taken over $Z_{\to i}^{(t)}$ in the protocol $\Pi(x_{\le n}^{(j)})$. The second equality follows because, given the same incoming transcript, the conditional law of message $Z_i^{(t)}$ is identical for $i \neq j$; only the messages from individual j can change. Now let P_0 and P_1 be arbitrary distributions whose densities satisfy $e^{-\varepsilon} \le p_0(z)/p_1(z) \le e^{\varepsilon}$. Then

$$D_{\rm kl}\left(P_0\|P_1\right) \le \varepsilon \;\text{ and }\; D_{\rm kl}\left(P_0\|P_1\right) \le \log\left(1 + D_{\chi^2}(P_0\|P_1)\right) \le \log\left(1 + (e^{\varepsilon}-1)^2\right)$$

by Proposition 2.2.10. By inspection $\min\{\varepsilon, \log(1+(e^{\varepsilon}-1)^2)\} \le \min\{\varepsilon, \frac{3}{2}\varepsilon^2\}$ for all $\varepsilon \ge 0$. Returning to the initial KL-divergence sum, we thus obtain

$$\sum_{i=1}^{n} D_{\rm kl}\left(\Pi(x_{\le n}^{(i)})\|\Pi(x_{\le n})\right) \le \sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\left[\min\left\{\varepsilon_{i,t}, \frac{3}{2}\varepsilon_{i,t}^2\right\}\right],$$

as desired. □

Unsurprisingly, the average KL-local privacy guarantee is sufficient to provide a mutual information bound, thus allowing us to apply Theorem 10.4.5 as in the proof of Proposition 10.4.6.

Proposition 10.4.8. *Let Π be any $\varepsilon_{\rm kl}$-KL-locally-private on average protocol and assume that $X_1, \ldots, X_n$ are independent conditional on V. Then*

$$I(X_1, \ldots, X_n; \Pi(X_1^n) \mid V) \le n\varepsilon_{\rm kl}.$$

Proof The conditional independence of the X_i guarantees that

$$I(X_1^n; \Pi(X_1^n) \mid V) = \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}, V) - H(X_i \mid \Pi, X_1^{i-1}, V)$$
$$\le \sum_{i=1}^{n} H(X_i \mid X_{\setminus i}, V) - H(X_i \mid \Pi, X_{\setminus i}, V) = \sum_{i=1}^{n} I(X_i; \Pi(X_1^n) \mid V, X_{\setminus i}).$$

We abuse notation to let $\Pi^*(X_{\setminus i})$ be the marginal protocol (marginalizing over X_i). Then

$$I(X_i; \Pi(X_1^n) \mid V, X_{\setminus i}) = \mathbb{E}\left[D_{\rm kl}\left(\Pi(X_{\setminus i}, X_i)\|\Pi^*(X_{\setminus i})\right)\right] \le \mathbb{E}\left[D_{\rm kl}\left(\Pi(X_{\setminus i}, X_i)\|\Pi(X_{\setminus i}, X_i')\right)\right]$$

where the first expectation is taken over V and $X_j \stackrel{\rm iid}{\sim} P_v$ conditional on $V = v$ and the inequality uses convexity and draws X_i' independently. Summing over $i = 1, \ldots, n$, Definition 10.8 gives the result. □

Applying Theorem 10.4.5, we then obtain the following corollary.

Corollary 10.4.9. *Let the conditions of Theorem 10.4.5 hold. If the data release Π is $\varepsilon_{\rm kl}$-private on average, then*

$$\sum_{j=1}^{d} \mathbb{P}(\widehat{V}_j(\Pi(X_1^n)) \neq V_j) \geq \frac{d}{2}\left(1 - \sqrt{7(c+1)\frac{\beta}{d}n\varepsilon_{\rm kl}}\right).$$

Specializing to the case that we wish to estimate a d-dimensional Bernoulli vector, where $X \in \{\pm 1\}$ has coordinates with $\mathbb{P}(X_j = 1) = \theta_j$, Example 10.4.3 gives the following minimax lower bound.

Corollary 10.4.10. *Let $\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl})$ denote the minimax mean-square error for estimation of a d-dimensional Bernoulli under the $\varepsilon_{\rm kl}$-KL-locally-private-on-average constraint in Definition 10.8. Then there is a numerical constant $c > 0$ such that*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl}) \geq c \min\left\{d, \frac{d^2}{n\varepsilon_{\rm kl}}\right\}.$$

Proof By Corollary 10.4.9 and Example 10.4.3, we have minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon_{\rm kl}) \gtrsim d\delta^2\left(1 - \sqrt{C\frac{\delta^2}{d}n\varepsilon_{\rm kl}}\right)$$

for a numerical constant C, which is valid for $\delta \lesssim 1$. Choose δ^2 to scale as $\min\{1, \frac{d}{n\varepsilon_{\rm kl}}\}$. □

When instead of the average KL-privacy we use the pure local differential privacy constraint (10.4.5), Lemma 10.4.7 implies the following.

Corollary 10.4.11. *Let $\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon)$ denote the minimax mean-square error for estimation of a d-dimensional Bernoulli where each data release is $\varepsilon_{i,t}$-locally differentially private (10.4.5), and $\sum_{t=1}^{\infty} \varepsilon_{i,t} \leq \varepsilon$. Then*

$$\mathfrak{M}_n(\theta(\mathcal{P}_d), \|\cdot\|_2^2, \varepsilon) \geq c \min\left\{d, \frac{d^2}{n(\varepsilon \wedge \varepsilon^2)}\right\}. \tag{10.4.6}$$

To interpret the bound (10.4.6), we think of each private release Z_i as effectively releasing $\min\{\varepsilon, \varepsilon^2\}$ bits for each vector $X_i \in \{0,1\}^d$. So the "fraction" of information released—out of d bits necessary to describe all of X_i—is $\frac{\varepsilon \wedge \varepsilon^2}{d}$, and so we see the effective sample size reduction of $n \mapsto \frac{n(\varepsilon \wedge \varepsilon^2)}{d}$.

10.5* Proof of Theorem 10.4.4

The proof proceeds in stages. The basic ideas are as follows:

1. Relate the Hellinger distance between the marginal distributions M_0 and M_1 of Π conditional on $V = 0$ or 1 to a sum of Hellinger distances between the marginal M_0 and an alternative M_i' where $X_i \sim P_1$ and $X_{\setminus i} \stackrel{\rm iid}{\sim} P_0$.
2. Provide a data processing inequality to relate $d_{\rm hel}(M_0, M_i')$ and the mutual information $I(X_i; \Pi(X_1^m))$ between the individual observation X_i and the protocol Π.
3. Use the standard chain rules for mutual information to finalize the theorem.

Step 1: sequential modification of marginals

We begin by relating the marginal distributions M_0 and M_1 by a sequence of one-variable changes. For bit vectors $b \in \{0,1\}^m$ define M_b to be the marginal distribution over the protocol $\Pi(X_1^m)$ generated from $(X_1, \ldots, X_m)$, where for each i we generate X_i by independently sampling

$$X_i \mid b \sim P_{b_i}. \tag{10.5.1}$$

For the standard basis vectors $e_1, \ldots, e_m$, we expect M_0 to be close to M_{e_l}, and thus hope for some type of tensorization behavior, where we can relate M_0 and M_1 via one-step changes from M_0 to M_{e_l}. The next lemma realizes this promise.

Lemma 10.5.1. *Let M_0, M_1, and M_{e_l} be as above. Then*

$$d_{\rm hel}^2(M_0, M_1) \le 7 \sum_{l=1}^m d_{\rm hel}^2(M_0, M_{e_l}). \tag{10.5.2}$$

Proof The proof crucially relies on the Euclidean structures that the Hellinger distance induces along with analogues of the cut-and-paste (the "rectangular" structure of inputs in communication protocols) properties from deterministic and randomized two-player communication. We assume without loss of generality that Π is discrete, as the Hellinger distance is an f-divergence and so can be arbitrarily approximated by discrete random variables.

First, we analogize the "rectangular" probabilistic structure of two-player communication protocols in Lemmas 10.3.18 and 10.3.19, which yields a multi-way cut-and-paste lemma.

Lemma 10.5.2 (cutting and pasting)**.** *Let $a, b, c, d \in \{0,1\}^m$ be bit vectors satisfying $a_i + b_i = c_i + d_i$ for each $i = 1, \ldots, m$. Then*

$$d_{\rm hel}^2(M_a, M_b) = d_{\rm hel}^2(M_c, M_d).$$

Proof We claim the following analogue of Lemma 10.3.18: for any $X_1^m = x_1^m$ and any communication transcript τ, we may write

$$Q(\Pi(x_1^m) = \tau \mid x_1^m) = \prod_{i=1}^m f_{i,x_i}(\tau) \tag{10.5.3}$$

for some functions f_{i,x_i}. Indeed, letting $\tau = \{z_i^{(t)}\}_{i \le m, t \le T}$ we have

$$Q(\Pi(x_1^m) = \tau \mid x_1^m) = \prod_{i,t} Q(z_i^{(t)} \mid x_1^m, z_{\to i}^{(t)}) = \prod_{i=1}^m \underbrace{\prod_{t=1}^T Q(z_i^{(t)} \mid x_i, z_{\to i}^{(t)})}_{=: f_{i,x_i}(\tau)}$$

where we use that message $z_i^{(t)}$ depends only on x_i and $z_{\to i}^{(t)}$. Then we can write $M_b(\Pi(X_1^m) = \tau)$ as a product using Eq. (10.5.3): integrating over independent $X_i \sim P_{b_i}$, we have

$$M_b(\Pi(X_1^m) = \tau) = \int Q(\tau \mid x_1^m) dP_{b_1}(x_1) \cdots dP_{b_m}(x_m) = \prod_{i=1}^m \underbrace{\int f_{i,x_i}(\tau) dP_{b_i}(x_i)}_{:= g_{i,b_i}(\tau)} = \prod_{i=1}^m g_{i,b_i}(\tau).$$

Taking M_a, M_b, M_c, M_d as in the statement of the lemma,

$$d_{\rm hel}^2(M_a, M_b) = 1 - \sum_\tau \sqrt{\prod_{i=1}^m g_{i,a_i}(\tau) g_{i,b_i}(\tau)}.$$

But as $a_i + b_i = c_i + d_i$ and each is $\{0,1\}$-valued, we have $g_{i,a_i} g_{i,b_i} = g_{i,c_i} g_{i,d_i}$, so the lemma follows. □

The second result we require is due to Jayram [256], and is the following:

Lemma 10.5.3. *Let $\{P_b\}_{b\in\{0,1\}^m}$ be any collection of distributions satisfying the cutting and pasting property $d_{\rm hel}^2(P_a, P_b) = d_{\rm hel}^2(P_c, P_d)$ whenever $a, b, c, d \in \{0,1\}^m$ satisfy $a + b = c + d$. Let $N = 2^k$ for some $k \in \mathbb{N}$. Then for any collection of bit vectors $\{b^{(i)}\}_{i=1}^N \subset \{0,1\}^m$ with $\langle b^{(i)}, b^{(j)}\rangle = 0$ for all $i \neq j$ and $b = \sum_i b^{(i)}$,*

$$\prod_{l=1}^k (1 - 2^{-l}) d_{\rm hel}^2(P_{\mathbf{0}}, P_b) \le \sum_{i=1}^N d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)}}).$$

We defer the technical proof to Section 10.5.1.

A computation shows that $\prod_{l=1}^k (1 - 2^{-l}) > \frac{2}{7}$. Lemma 10.5.3 nearly gives us our desired result (10.5.2), except that it requires a power of 2. To that end, let k_0 be the largest $k \in \mathbb{N}$ such that $2^{k_0} \le m$, and construct bit vectors $b^{(1)}, \ldots, b^{(2^{k_0})}$ satisfying $\sum_i b^{(i)} = \mathbf{1}$ and $1 \le \left\|b^{(i)}\right\|_0 \le 2$ for each i. Then Lemma 10.5.3, via the cutting-pasting property of the marginals M, implies

$$\frac{2}{7} d_{\rm hel}^2(M_0, M_1) \le \sum_{i=1}^{2^{k_0}} d_{\rm hel}^2(M_0, M_{b^{(i)}}) \le 2 \sum_{i=1}^m d_{\rm hel}^2(M_0, M_{e_i}),$$

where the second inequality again follows from Lemma 10.5.3 as $b^{(i)} = e_j$ or $e_j + e_{j'}$ for some basis vectors e_j, e_j'. This gives Lemma 10.5.1. □

Step 2: from Hellinger to Shannon information

Now we relate the strong data processing processing constants for mutual information in Definition 10.7 to compare Hellinger distances with mutual information. We claim the following lemma.

Lemma 10.5.4. *Let the conditions of Theorem 10.4.4 hold. Let M_0 and M_{e_l} be the marginal distributions over $\Pi(X_1^m)$ when X_i have the sampling distribution (10.5.1). Then for $l \in \{1, \ldots, m\}$,*

$$d_{\rm hel}^2(M_{e_l}, M_0) \le \frac{c+1}{2} \beta I(X_l; \Pi(X_1^m) \mid V = 0).$$

Proof Consider the following alternative distributions. Let $W \sim \mathsf{Uniform}\{0,1\}$, and draw $X' \in \mathcal{X}^m$ with independent coordinates according to

$$X_i' \stackrel{\rm iid}{\sim} P_0 \text{ if } W = 0 \quad \text{or} \quad X_i' \sim \begin{cases} P_0 & \text{if } i \neq l \\ P_1 & \text{if } i = l \end{cases} \text{ if } W = 1.$$

Then we have the Markov chain $W \to X' \to \Pi(X')$, and moreover,

$$W \to X'_l \to \Pi(X') \leftarrow X'_{\setminus l},$$

so that additionally $W \to X'_l \to \Pi(X')$ is a Markov chain. As a consequence, Definition 10.7 of the strong data processing inequality gives

$$I(W; \Pi(X')) \le \beta I(X'_l; \Pi(X')).$$

Lemma 10.3.16 thus implies

$$d_{\rm hel}^2(M_{e_l}, M_0) \le I(W; \Pi(X')) \le \beta I(X'_l; \Pi(X')). \tag{10.5.4}$$

It remains to relate $I(X'_l; \Pi(X'))$ to $I(X_l; \Pi(X) \mid V = 0)$. Here we may leverage the bounded likelihood ratio between P_0 and P_1. By the condition (10.4.3),

$$P_0 \ge \frac{1}{c} P_1 \;\; \text{so} \;\; (c+1)P_0 \ge P_0 + P_1 \;\; \text{or} \;\; P_0 \ge \frac{2}{c+1} \frac{P_0 + P_1}{2}.$$

As a consequence, we have

$$\begin{aligned}
I(X_l; \Pi(X_1^m) \mid V = 0) &= \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| M_0\right) dP_0(x) \\
&\ge \frac{2}{c+1} \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| M_0\right) \frac{dP_0(x) + dP_1(x)}{2} \\
&\ge \frac{2}{c+1} \int D_{\rm kl}\left(Q(\cdot \mid X_l = x) \| \overline{M}\right) \frac{dP_0(x) + dP_1(x)}{2} \\
&= \frac{2}{c+1} I(X'_l; \Pi(X')),
\end{aligned}$$

where the second inequality uses that $\overline{M} = \int Q(\cdot \mid X_l = x) \frac{dP_0(x) + dP_1(x)}{2}$ minimizes the integrated KL-divergence (recall inequality (12.2.1)). Returning to inequality (10.5.4), we evidently have the result of the lemma. □

Step 3: Completing the proof of Theorem 10.4.4

Let $T = \Pi(X_1^m)$ be the (random) realized transcript. By combining the tensorization Lemma 10.5.1 with the information bound in Lemma 10.5.4, we obtain

$$d_{\rm hel}^2(M_0, M_1) \le 7 \sum_{i=1}^m d_{\rm hel}^2(M_0, M_{e_i}) \le \frac{7}{2}(c+1)\beta \sum_{i=1}^m I(X_i; T \mid V = 0).$$

By symmetry, we also have

$$d_{\rm hel}^2(M_0, M_1) \le 7 \sum_{i=1}^m d_{\rm hel}^2(M_0, M_{e_i}) \le \frac{7}{2}(c+1)\beta \sum_{i=1}^m I(X_i; T \mid V = 1).$$

As the X_i are independent conditional on V (and w.l.o.g. for the purposes of mutual information, we may assume they are discrete), for any $v \in \{0, 1\}$ we have

$$\begin{aligned}
\sum_{i=1}^m I(X_i; T \mid V = v) &= \sum_{i=1}^m \left[H(X_i \mid V = v) - H(X_i \mid T, V = v)\right] \\
&= \sum_{i=1}^m \left[H(X_i \mid X_1^{i-1}, V = v) - H(X_i \mid T, V = v)\right] \\
&\le \sum_{i=1}^m \left[H(X_i \mid X_1^{i-1}, V = v) - H(X_i \mid X_1^{i-1}, T, V = v)\right] \\
&= \sum_{i=1}^m I(X_i; T \mid X_1^{i-1}, V = v) = I(X_1, \ldots, X_m; T \mid V = v),
\end{aligned}$$

where the inequality used that conditioning decreases entropy. We thus obtain

$$d_{\rm hel}^2(M_0, M_1) \le \frac{7}{2}(c+1)\beta \min_{v \in \{0,1\}} I(X_1, \ldots, X_m; T \mid V = v).$$

10.5.1 Proof of Lemma 10.5.3

We prove the result by induction. It is trivially true for $N = 1 = 2^0$, that is, $k = 0$, so now we consider the inductive case, that is, it holds for $N = 1, \ldots, 2^{k-1}$ and we consider $N = 2^k$.

First, we observe that if $\{u_i\}_{i=1}^N$ are arbitrary vectors and $\overline{u} = \frac{1}{N}\sum_{i=1}^N u_i$ is their mean, then

$$\sum_{i,j} \|u_i - u_j\|_2^2 = \sum_{i,j} \|u_i - \overline{u} + \overline{u} - u_j\|_2^2 = \sum_{i,j} \|u_i - \overline{u}\|_2^2 + \sum_{i,j} \|\overline{u} - u_j\|_2^2 = 2N \sum_{i=1}^N \|\overline{u} - u_i\|_2^2 .$$

Thus, if u_0 is any other vector, that $\overline{u}$ minimizes $\sum_i \|u_i - u\|_2^2$ over all u gives

$$\frac{1}{N} \sum_{1 \le i < j \le N} \|u_i - u_j\|_2^2 \le \sum_{i=1}^N \|u_i - \overline{u}\|_2^2 \le \sum_{i=1}^N \|u_i - u_0\|_2^2 . \tag{10.5.5}$$

We return to the Hellinger distances. Evidently $2d_{\rm hel}^2(P_a, P_b) = \left\|\sqrt{p_a}(\cdot) - \sqrt{p_b}(\cdot)\right\|_2^2$, so that it is a Euclidean distance. As a consequence, for any pairwise disjoint collection of N bit vectors $b^{(i)}$,

$$\sum_{i=1}^N d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)}}) \ge \frac{1}{N} \sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{b^{(i)}}, P_{b^{(j)}}) = \frac{1}{N} \sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) \tag{10.5.6}$$

where the inequality follows from (10.5.5) and the equality by the assumed cut-and-paste property. Now, we apply the standard one-factorization theorem for complete graphs, which says that we may decompose any complete graph K_N, where N is even, into $N - 1$ perfect matchings $\mathcal{M}_i$ with $N/2$ edges—necessarily, as they form a perfect matching—where each $\mathcal{M}_i$ is edge disjoint. Identifying the pairs $i < j$ with the complete graph, we thus obtain

$$\sum_{1 \le i < j \le N} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) = \sum_{l=1}^{N-1} \sum_{(i,j) \in \mathcal{M}_l} d_{\rm hel}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}). \tag{10.5.7}$$

Now fix $n \in \{1, \ldots, N-1\}$ and a matching $\mathcal{M}_n$. By assumption we have $\langle b^{(i)} + b^{(j)}, b^{(i')} + b^{(j')} \rangle = 0$ for any distinct pairs $(i,j), (i',j') \in \mathcal{M}_n$, and moreover, $\sum_{(i,j) \in \mathcal{M}_n} (b^{(i)} + b^{(j)}) = b$. Thus, our induction hypothesis gives that for any $l \in \{1, \ldots, N-1\}$ and any of our matchings $\mathcal{M}_n$, we have

$$\sum_{(i,j) \in \mathcal{M}_n} d_{\mathrm{hel}}^2(P_{\mathbf{0}}, P_{b^{(i)} + b^{(j)}}) \geq d_{\mathrm{hel}}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k-1} (1 - 2^{-l}).$$

Substituting this lower bound into inequality (10.5.7) and using inequality (10.5.6), we obtain

$$\sum_{i=1}^{N} d_{\mathrm{hel}}^2(P_{\mathbf{0}}, P_{b^{(i)}}) \geq \frac{1}{N} \cdot (N-1) d_{\mathrm{hel}}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k-1} (1 - 2^{-l}) = d_{\mathrm{hel}}^2(P_{\mathbf{0}}, P_b) \prod_{l=1}^{k} (1 - 2^{-l}),$$

where we have used $N = 2^k$.

10.6 Bibliography and further reading

Data processing inequalities originate with Dobrushin's study of central limit theorems for Markov chains [141, 142]; Dobrushin first proved Proposition 10.1.1 (see [142, Sec. 3.1]). Cohen et al. [114] show that the strong data processing constant for variation distance is the largest of the strong data processing constants (Theorem 10.1.2) for finite state spaces using careful linear algebraic techniques, also showing the opposite extremality (inequality (10.1.1)) of the χ^2 contraction coefficient [114, Proposition II.6.15]. Del Moral et al. [138] and Polyanskiy and Wu [345] give approachable treatments for general alphabets. Exercises 10.2 and 10.3, which prove Theorem 10.1.2, that the Dobrushin coefficient is the largest strong data processing constant, follow [138]. Inequality (10.1.1), that the χ^2-divergence provides the smallest strong data processing constant, appears in Polyanskiy and Wu [345, Thms. 1 and 2]. Strong data processing inequalities arise in many applications in communication, estimation, and some functional analysis [350, 345].

Yao [434] is the reference most researchers cite as beginning communication complexity, which introduces the communication complexity setting we discuss in Section 10.3, making the connections between randomized complexities and public (shared) randomness. The standard classical reference for the subject is Kushilevitz and Nisan's book [270]. There are numerous techniques that we do not discuss, including so-called discrepancy lower bounds, which address both randomized and deterministic communication complexity; for example, these give the stronger lower bound that $\mathsf{DCC}_\delta(\mathrm{IP}_2) \geq n - O(1)$ [270, Example 3.29 and Exercise 3.30], as compared to our Example 10.3.7, which applies only to communication complexity. Communication complexity has uses far beyond the "standard" communication setting we have outlined, with more recent research showing how to use the techniques to provide lower bounds on the performance of algorithms in many computational models, such as streaming models and memory-limited computation [318, 354]. Our information complexity approach follows Bar-Yossef et al. [41], who first showed how to use conditional information complexity to obtain direct sum lower bounds [41, Sec. 5], a major building block throughout communication complexity. The particular leveraging of Hellinger distances in Section 10.3.4* follows Jayram [256]. Braverman and collaborators have also developed several important techniques in information complexity [84, 85, 86].

Our treatment of applications of communication complexity in estimation follows an approach Zhang et al. [438] originate. We adapt the use of direct sums and strong data processing inequalities in communication from Braverman et al. [86] and Garg et al. [206]. Our results apply most easily

to scenarios in which each machine or agent owns only a single data item, which allows application of Proposition 10.4.2; tensorizing this to multiple observations requires some care, but can be done with a truncation argument [438, 86] or more careful Sobolev inequalities [350]. (See also the papers [3, 4, 5].) Our extension to private estimation scenarios follows the paper [153], which also shows how to generalize to other variants of privacy, including (ε, δ)-local differential privacy. Recent work has shown how communication lower bounds and strong data processing inequalities can be used to show the necessity of "memorization" in some natural problems in machine learning, where any learning procedure with good enough performance necessarily encodes substantial irrelevant information about a dataset [89].

10.7 Exercises

Exercise 10.1 (Proposition 10.1.1 and the Dobrushin coefficient)**:** Recall Definition 10.2 of the Dobrushin coefficient $\alpha_{\mathrm{TV}}(Q) = \sup_{x,y} \|Q(\cdot \mid x) - Q(\cdot \mid y)\|_{\mathrm{TV}}$ for a channel Q from $\mathcal{X} \to \mathcal{Z}$.

(a) Show that $\sup_{P_0 \neq P_1} \|Q \circ P_0 - Q \circ P_1\|_{\mathrm{TV}} / \|P_0 - P_1\|_{\mathrm{TV}} \geq \alpha_{\mathrm{TV}}(Q)$.

(b) Define $Q_\star(A) = \inf_y Q(A \mid y)$. Justify that $|Q(A \mid x) - Q_\star(A)| \leq \alpha_{\mathrm{TV}}(Q)$.

(c) Let $M_v = Q \circ P_v$ for $v \in \{0, 1\}$. Use Lemma 2.2.5 to justify

$$M_0(A) - M_1(A) \leq \int Q(A \mid x) \left[dP_0(x) - dP_1(x)\right]_+ - \int Q_\star(A) \left[dP_0(x) - dP_1(x)\right]_+ .$$

(d) Prove Proposition 10.1.1.

Exercise 10.2: Let $f : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ be a closed, nonnegative convex function.

(a) Show there exists a sequence of piecewise linear functions f_n satisfying $f_{n-1} \leq f_n \leq f$ for all n and for which $f_n(x) \uparrow f(x)$ pointwise for all x. *Hint.* Let $\mathcal{L} = \{l \mid l(x) = a + bx, l \leq f\}$ be the collection of affine underestimators of f, and note that $f(x) = \sup_{l \in \mathcal{L}} l(x)$. (See Appendix C.2.) Restrict to l of the form $l(x) = f(x_0) + g(x - x_0)$, where $g \in \partial f(x_0)$ is a subderivative of f.

(b) Show that if for some $z_0 \in \mathbb{R}$ we have $f(z_0) = 0$, then one may take the functions f_n of the form $f_n(x) = \sum_{i=1}^n a_i \left[b_i - x\right]_+ + \sum_{i=1}^n c_i \left[x - d_i\right]_+$, where $b_i \leq z_0$, $d_i \geq z_0$, and $a_i, c_i \geq 0$.

(c) Conclude that for any measure μ on $\mathbb{R}_+$, $\int f_n d\mu \uparrow \int f d\mu$.

Exercise 10.3 (Proving Theorem 10.1.2)**:** Let $\alpha = \alpha_{\mathrm{TV}}(Q)$ be the Dobrushin coefficient of the channel Q and $f : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$, $f(1) = 0$, be closed convex. Justify the following claims.

(a) There exists a nondecreasing sequence f_n of piecewise linear functions, each of the form $f_n(x) = \sum_{i=1}^n a_i \left[b_i - x\right]_+ + \sum_{i=1}^n c_i \left[x - d_i\right]_+$, where $b_i \leq 1$, $d_i \geq 1$, and $a_i, c_i \geq 0$. *Hint:* Exercise 10.2.

(b) Let $M_v(A) = \int Q(A \mid x) dP_v(x)$ for $v \in \{0, 1\}$ be the induced marginal distributions. Show that for any function of the form $h(t) = \left[t - \Delta\right]_+$, where $\Delta > 1$,

$$D_h\left(M_0 \| M_1\right) \leq \alpha D_h\left(P_0 \| P_1\right) \tag{10.7.1}$$

by the following steps:

i. Define the set $\mathcal{X}(\Delta) := \{x \mid dP_0(x) \le \Delta dP_1(x)\}$. Argue that $\mathcal{X}(\Delta)$ must be non-null (i.e., have positive measure).

ii. Define the probability distribution P_Δ with density

$$dP_\Delta(x) = \frac{\Delta dP_1(x) - dP_0(x)}{\int [\Delta dP_1(x) - dP_0(x)]_+} \mathbf{1}\left\{x \in \mathcal{X}(\Delta)\right\}.$$

Argue that the measure

$$G = \Delta P_1 - (\Delta - 1)P_\Delta$$

is a probability distribution.

iii. Show that

$$D_h\left(P_0 \| P_1\right) = \|P_0 - G\|_{\rm TV}.$$

It may be useful to show that $dP_0 - dG \le 0$ on $\mathcal{X}(\Delta)$.

iv. Conclude that

$$D_h\left(P_0 \| P_1\right) \ge \frac{1}{\alpha}\left\|Q \circ P_0 - Q \circ G\right\|_{\rm TV} \ge \frac{1}{\alpha} D_h\left(Q \circ P_0 \| Q \circ P_1\right).$$

(c) Using the monotone convergence theorem, show that $D_f\left(M_0 \| M_1\right) \le \alpha D_f\left(P_0 \| P_1\right)$.

Exercise 10.4 (Markov chain mixing)**:** Consider a Markov chain $X_1, X_2, \ldots$ with transition distribution $P(\cdot \mid x)$ and stationary distribution π. Let $P^k(\cdot \mid x)$ denote the distribution of the Markov chain initialized in state x after k steps. Assume there exists some (finite) positive integer $k \in \mathbb{N}$ such that for any two initial states x_0, x_1, the Markov chain satisfies

$$\left\|P^k(\cdot \mid x_0) - P^k(\cdot \mid x_1)\right\|_{\rm TV} \le \beta < 1.$$

Show that the Markov chain enjoys *fast mixing* for any f divergence: if there is any n such that $D_f\left(P^n(\cdot \mid x) \| \pi\right) < \infty$, the Markov chain mixes exponentially quickly in that it satisfies

$$\limsup_n \frac{1}{n} \log D_f\left(P^n(\cdot \mid x) \| \pi\right) \le \frac{1}{k} \log \beta < 0.$$

Exercise 10.5: For $k \in [1, \infty]$, consider the collection of distributions

$$\mathcal{P}_k := \{P : \mathbb{E}_P[|X|^k]^{1/k} \le 1\}$$

supported on $\mathbb{R}$ with kth moment at most 1. We consider minimax estimation of the mean $\mathbb{E}[X]$ for these families under ε-local differential privacy. Let $\mathcal{Q}_\varepsilon$ denote the collection of all sequentially interactive ε-differentially private channels (10.2.1), and define the (locally) private minimax risk

$$\mathfrak{M}_n(\theta(\mathcal{P}), (\cdot)^2, \varepsilon) := \inf_{\widehat{\theta}_n} \inf_{Q \in \mathcal{Q}_\varepsilon} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}[(\widehat{\theta}_n(Z_1^n) - \theta(P))^2].$$

(a) Assume that $\varepsilon \le 1$. For $k \in [1, \infty]$, show that there exists a constant $c > 0$ such that

$$\mathfrak{M}_n(\theta(\mathcal{P}_k), (\cdot)^2, \varepsilon) \ge c \min\left\{\left(\frac{1}{n\varepsilon^2}\right)^{\frac{k-1}{k}}, 1\right\}.$$

(b) Give an ε-locally differentially private estimator achieving the minimax rate in part (a).

Exercise 10.6: Given n observations, we wish to test whether $P = P_0 = \mathsf{Uniform}[0,1]$ or $P = P_1 = \mathsf{Uniform}[\theta, 1]$ for some $\theta \in (0,1)$. We ask the rate at which we may take $\theta \downarrow 0$ with n while still achieving non-trivial testing power. Without privacy, a simple test Ψ checks whether any observation $X_i < \theta$, in which case we can trivially accept P_0 and reject P_1, otherwise accepting P_1.

(a) Show that $P_0^n(\Psi = 1) + P_1^n(\Psi = 0) = (1-\theta)^n$ and if $\theta_n = \theta_0/n$, then $\lim_n P_0^n(\Psi = 1) = e^{-\theta_0}$.

(b) Assume that Z_i are ε-locally differentially private views of X_i, generated sequentially interactively (10.2.1). Use the Bretagnolle-Huber inequality to show that for $c_\varepsilon = 4(e^\varepsilon - 1)^2$,

$$\inf_\Psi \left\{P_0(\Psi(Z_1^n) \neq 0) + P_1(\Psi(Z_1^n) \neq 1)\right\} \geq 1 - \sqrt{1 - \exp(-c_\varepsilon n \theta^2)}.$$

(c) Describe the gap between the private and non-private testing error.

Exercise 10.7: Show the strong data processing inequality in Theorem 10.2.1 is order optimal: there exist ε-differentially private channels Q_ε such that for any Bernoulli distributions P_0 and P_1 and induced marginal distributions $M_{v,\varepsilon} = Q(\cdot \mid X = 1)P_v(X = 1) + Q(\cdot \mid X = 0)P_v(X = 0)$,

$$\frac{D_{\rm kl}\left(M_{0,\varepsilon} \| M_{1,\varepsilon}\right)}{\|P_0 - P_1\|_{\rm TV}^2} = \frac{\varepsilon^2}{2} + O(\varepsilon^3) \quad \text{as } \varepsilon \downarrow 0.$$

Exercise 10.8: Show that the randomized communication complexity (10.3.1) satisfies $\mathsf{RCC}_\delta(f) \leq O(1) \log \frac{1}{\delta} \mathsf{RCC}_{1/3}(f)$ for any f and any $\delta < 1$.

Exercise 10.9 (From public to private randomness)**:** Consider the randomized complexity (10.3.1) and associated public-randomness complexity $\mathsf{RCC}_\delta^{\rm pub}$. Let $\mathcal{X} = \mathcal{Y} = \{0,1\}^n$ and $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$, and let Π be a protocol using public randomness U such that $\max_{x,y} \mathbb{P}(\Pi(x,y,U) \neq f(x,y)) \leq \epsilon$.

(a) Use Hoeffding's inequality to show that there are $k = \frac{\log 2}{\delta^2} \cdot n$ points $u_1, \ldots, u_k$ such that if $I \in [k]$ is chosen uniformly at random, then $\mathbb{P}(\Pi(x,y,u_I) \neq f(x,y)) \leq \epsilon + \delta$.

(b) Give a protocol that uses no public randomness but whose communication complexity is at most $\mathrm{depth}(\Pi) + O(1) \log \frac{n}{\delta}$.

(c) Conclude that $\mathsf{RCC}_\delta(f) \leq \mathsf{RCC}_\delta^{\rm pub}(f) + O(1) \log \frac{n}{\delta}$.

Exercise 10.10: Complete the details of Example 10.3.8, and additionally show that for any $k \geq 2$, $\mathsf{RCC}_{1/n^{k-1}}(\mathrm{EQ}) \leq 2k \log_2 n + O(1)$. *Hint.* A degree n polynomial has at most n roots.

Exercise 10.11 (Information complexity for set disjointness)**:** Define the set disjointness function $f(x,y) = \mathrm{DISJ}(x,y) := 1 - \mathbf{1}\{\langle x, y\rangle > 0\}$. This arises when x, y are characteristic vectors of subsets A, B of $[n]$, that is, $x_i = \mathbf{1}\{i \in A\}$ and $y_i = \mathbf{1}\{i \in B\}$, so $f(x,y) = \mathbf{1}\{A \cap B = \varnothing\}$.

(a) Give the decomposition (10.3.4) to write $f(x,y) = g(h(x_1,y_1), \ldots, h(x_n,y_n))$.

(b) Give the set of planted solutions $x, y \in \{0,1\}^n$ satisfying Definition 10.5.

(c) Why does the distribution in Example 10.3.11 work as a fooling distribution for disjointness?

(d) Use Corollary 10.3.13 and Proposition 10.3.14 to show that $\mathsf{IC}_\delta(f) \gtrsim n$.

Exercise 10.12 (Information complexity for entrywise less or equal)**:** Consider the entrywise less than or equal to function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ with $f(x,y) = \mathbf{1}\{x \preceq y\}$, so that $f(x,y) = 1$ if $x_i \le y_i$ for each i and 0 if there exists i such that $x_i > y_i$.

(a) Show that f has the decompositional structure (10.3.4). Give the functions g and h.

(b) Give a fooling distribution μ on $\mathcal{X} \times \mathcal{Y}$ for f.

(c) Use Theorem 10.3.12 and a modification of the proof of Proposition 10.3.14 to show that $\mathsf{IC}_\delta(f) \ge \frac{n}{4}(1 - 2\sqrt{\delta(1-\delta)})$. (This is order optimal, because $\mathsf{IC}_\delta(f) \le \mathsf{CC}(f) \le n+1$ trivially.)

Exercise 10.13* (Private logistic regression)**:** Consider the logistic regression model

$$p_\theta(y \mid x) = \frac{1}{1 + \exp(-y\langle \theta, x\rangle)}, \quad y \in \{\pm 1\}, \quad x \in \mathbb{R}^d.$$

For a distribution P on (X, Y), where $Y \mid X = x$ has logistic distribution, define the excess risk

$$r(\theta, P) := \mathbb{E}_P[\ell(\theta; X, Y)] - \inf_\theta \mathbb{E}_P[\ell(\theta; X, Y)]$$

where $\ell(\theta; x, y) = \log(1 + \exp(-y\langle x, \theta\rangle))$ is the log loss. Let $\mathcal{P}$ be the collection of such distributions with $X \in \{-1, 1\}^d$. Looking ahead to Section 17.3, for a channel Q mapping $(X, Y) \to Z$, define

$$\mathfrak{M}_n(\mathcal{P}, Q) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,Q}\left[r(\widehat{\theta}(Z_1^n), P)\right],$$

where the expectation is taken over $Z_i \sim Q(\cdot \mid X_i, Z_1^{i-1})$ and Q is (locally) ε-differentially private.

(a) Show that for all n large enough, there is a (numerical) constant $c > 0$ such that

$$\mathfrak{M}_n(\mathcal{P}, Q) \ge c \cdot \frac{d}{n} \cdot \frac{d}{\varepsilon \wedge \varepsilon^2}.$$

(b) Suppose we allow additional passes through the dataset (i.e. multiple rounds of communication), but still require that all data Z_i released from X_i be ε-differentially private. That is, assume we have the (sequential and interactive) release schemes of Fig. 10.3, and we guarantee that

$$Z_i^{(t)} \sim Q(\cdot \mid X_i, B^{(1)}, \ldots, B^{(t)}, Z_1^{(t)}, \ldots, Z_{i-1}^{(t)})$$

is $\varepsilon_{i,t}$-differentially private, where $\sum_t \varepsilon_{i,t} \le \varepsilon$ for all i. Does the lower bound of part (a) change?

Exercise 10.14: Let the notation of Proposition 10.4.2 hold.

(a) Show that if $p(v)$ and $p(v \mid x)$ denote the p.m.f.s of V and V conditional on $X = x$, then

$$e^{-\alpha} p(v) \le p(v \mid x) \le e^\alpha p(v).$$

(b) Show that

$$|p(v \mid z) - p(v)| \le 2(e^\alpha - 1) \left\|P_X(\cdot \mid z) - P_X(\cdot)\right\|_{\mathrm{TV}}.$$

(c) Complete the proof of the proposition.

Chapter 11

Asymptotically exact optimality guarantees and score correlation

The squared error takes a particularly central place in the theory of estimation, as its simple structure means it admits closed forms for many estimators, allows one to evaluate exact minimax or Bayesian risks, and, for lack of a better description, it plays nicely with differentiation. For example, in any Bayesian problem, where we draw a parameter θ from a prior π on θ and then observe a sample $X \sim P_\theta$, the posterior mean $\mathbb{E}[\theta \mid X]$ is always Bayes optimal: we have

$$\mathbb{E}\left[\left\|\widehat{\theta}(X) - \theta\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{\theta}(X) - \theta\right\|_2^2 \mid X\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbb{E}[\theta \mid X] - \theta\right\|_2^2\right]\right],$$

where we use that $\inf_t \mathbb{E}[\|Y - t\|_2^2] = \mathbb{E}[\|Y - \mathbb{E}[Y]\|_2^2]$ for any random vector Y. In this chapter, we develop elements of this theory, presenting some of the classical bounds on estimation accuracy.

While this explicit characterization of Bayesian optimality gives one avenue for (asymptotically) exact optimality guarantees, we will instead leverage connections between estimation error and correlation, which extend even beyond squared error. Essentially, we may boil the results down to the fact that for an accurate estimator, on "average" over parameters θ, the parameter θ necessarily correlates with the *Fisher score* (or score) vector

$$\dot{\ell}_\theta(x) := \nabla_\theta \log p_\theta(x) = \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}$$

when $X \sim P_\theta$ has density p_θ. The main results in this chapter provide lower bounds on the estimation error in terms of the *Fisher information matrix* of the parameter θ,

$$J(\theta) := \mathbb{E}_\theta\left[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^\top\right] = \mathrm{Cov}_\theta(\dot{\ell}_\theta(X)), \tag{11.0.1}$$

where the equality exploits that, so long as we may exchange integration and differentiation,

$$\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = \int p_\theta(x)\frac{\nabla p_\theta(x)}{p_\theta(x)}dx = \int \nabla p_\theta(x)dx \overset{(\star)}{=} \nabla_\theta \int p_\theta(x)dx = \nabla_\theta 1 = 0,$$

where equality $(\star)$ typically requires justification.

To see why we expect $J(\theta)$ to somehow be fundamental, let $L(\beta) := \mathbb{E}_{P_\theta}[-\log p_\beta(X)]$ be the negative log likelihood of a putative parameter β evaluated when $X \sim P_\theta$; then

$$L(\beta) - L(\theta) = \mathbb{E}_\theta\left[\log \frac{p_\theta(X)}{p_\beta(X)}\right] = D_{\rm kl}\left(P_\theta \| P_\beta\right),$$

while simultaneously

$$L(\beta) - L(\theta) = \langle \underbrace{\nabla_\theta L(\theta)}_{=0}, \beta - \theta \rangle + \frac{1}{2}(\beta - \theta)^\top \nabla_\theta^2 L_{P_\theta}(\theta)(\beta - \theta) + O(\|\beta - \theta\|^3).$$

Again assuming we can interchange expectation and differentiation,

$$\begin{aligned}\nabla_\theta^2 L(\theta) &= \int \left(-\frac{\nabla^2 p_\theta(x)}{p_\theta(x)} + \frac{\nabla p_\theta(x) \nabla p_\theta(x)^\top}{p_\theta(x)^2} \right) p_\theta(x) dx \\ &= \mathbb{E}_\theta[\dot\ell_\theta(X) \dot\ell_\theta(X)^\top] - \int \nabla^2 p_\theta(x) dx = \mathrm{Cov}_\theta(\dot\ell_\theta(X)) - \nabla^2 \int p_\theta(x) dx = \mathrm{Cov}_\theta(\dot\ell_\theta(X)).\end{aligned}$$

So we see that, to within higher order errors,

$$\frac{1}{2}(\beta - \theta)^\top \mathrm{Cov}_\theta(\dot\ell_\theta(X))(\beta - \theta) = D_{\rm kl}\left(P_\theta \| P_\beta\right) + O(\|\beta - \theta\|^3).$$

If we assume that—following the intuition of Fano's method in Chapter 9.4—we can estimate parameters to accuracy at best $D_{\rm kl}\left(P_\theta \| P_\beta\right) \approx \frac{d}{n}$ given a sample of size n, then taking $\beta = \theta + J(\theta)^{-1/2} u$ and allowing U to be chosen uniformly on the sphere $\|u\|_2 = \sqrt{d/n}$ with $\mathbb{E}[UU^\top] = \frac{1}{n} I_d$, we expect to obtain squared error

$$\mathbb{E}\left[\|\widehat\theta(X_1^n) - \theta\|_2^2\right] \gtrsim \mathbb{E}\left[\|J(\theta)^{-1/2} U\|_2^2\right] = \mathbb{E}[\langle J(\theta)^{-1}, UU^\top \rangle] = \frac{\mathrm{tr}(J(\theta)^{-1})}{n}.$$

The main consequences of the development here are ways to make these inequalities rigorous—and with sharp constants—by exploiting the connection between the score $\dot\ell_\theta(x)$ and parameters θ.

11.1 The Cramér-Rao inequality

To maintain some connection with historical approaches, we begin with the Cramér-Rao bound, which provides lower bounds on the mean-squared error of *unbiased* estimators. We suppress several regularity conditions here, as we will not use the bound to develop optimality results. Nonetheless, we provide the heuristic here: for a parameter $\theta \in \mathbb{R}$, assume that X has density p_θ with respect to some base measure μ, where $\dot p_\theta(x) := \frac{\partial}{\partial \theta} p_\theta(x)$ exists for all x and is appropriately integrable.

Proposition 11.1.1 (The Cramér-Rao bound)**.** *Assume that $\widehat\theta : \mathcal{X} \to \mathbb{R}$ is unbiased for θ. Then*

$$\mathbb{E}_\theta\left[\left(\widehat\theta(X) - \theta\right)^2\right] \geq \frac{1}{J(\theta)}$$

under appropriate regularity conditions on the density p_θ.

Proof We use integration by parts, that is, that $\int u dv = uv - \int v du$. Because $\mathbb{E}_\theta[\widehat\theta] = \theta$, we have

$$1 = \frac{\partial}{\partial \theta} \mathbb{E}_\theta[\widehat\theta] = \frac{\partial}{\partial \theta} \int \widehat\theta(x) p_\theta(x) d\mu(x) \stackrel{(\star)}{=} \int \widehat\theta(x) \dot p_\theta(x) d\mu(x),$$

where $(\star)$ exchanges integration and differentiation. Then as $\dot p_\theta(x) = \dot\ell_\theta(x) p_\theta(x)$, we obtain

$$1 = \int \widehat\theta(x) \dot\ell_\theta(x) p_\theta(x) d\mu(x) = \mathbb{E}_\theta[\widehat\theta(X) \dot\ell_\theta(X)] = \mathbb{E}_\theta\left[(\widehat\theta(X) - \theta) \dot\ell_\theta(X)\right]$$

because $\mathbb{E}_\theta[\dot\ell_\theta(X)] = 0$. Applying Cauchy-Schwarz,

$$\mathbb{E}_\theta[(\widehat\theta(X) - \theta)\dot\ell_\theta(X)]^2 \le \mathbb{E}_\theta[(\widehat\theta(X) - \theta)^2]\mathbb{E}_\theta[\dot\ell_\theta^2(X)],$$

which gives the result. □

While there is a long history of using the Cramér-Rao bound to claim some type of statistical efficiency for an estimator, it is a brittle certificate of efficiency. It applies only to unbiased estimators, while most practical estimators are biased because of regularization, constraints, or other choices. Even asymptotic unbiasedness, that $\mathbb{E}[\widehat\theta_n(X_1^n)] \to \theta$, is not enough for the finite-sample Cramér-Rao inequality to certify optimality. If the underlying parameter space Θ is compact, typically, no unbiased estimator that guarantees $\widehat\theta \in \Theta$ even *exists*.

11.1.1 Compact sets and the failure of the Cramér-Rao bound

The following examples illustrate the limitation. Consider estimation on an interval.

Example 11.1.2 (Estimation of a bounded parameter)**:** Let $\{P_\theta\}$ be a family of mutually absolutely continuous probability models, meaning that $P_{\theta_0}(A) > 0$ implies $P_{\theta_1}(A) > 0$ for $\theta_0, \theta_1 \in \Theta$. Suppose the parameter $\theta \in [a, b]$, and assume $\widehat\theta \in [a, b]$ is unbiased for θ. We claim that $\widehat\theta = a$ with probability 1 and $\widehat\theta = b$ with probability 1. Indeed, when $\theta = a$, we have

$$\mathbb{E}_a[\widehat\theta(X)] = a,$$

and as $\widehat\theta \ge a$, we must have $\widehat\theta(X) = a$ with probability 1. Similarly, for $\theta = b$, we have $\widehat\theta \le b$ and as $\mathbb{E}_b[\widehat\theta(X)] = b$ then $\widehat\theta(X) = b$ with probability 1. This is a contradiction unless $a = b$. ◇

Building off of Example 11.1.2, we can show that if for some convex body Θ we have the *a priori* guarantee that the parameter $\theta \in \Theta$, then we can reduce the mean-squared error of any estimator $\widehat\theta$ trivially: simply project $\widehat\theta$ onto Θ. Stronger guarantees hold. For example, Exercise 11.1 shows that if $\widehat\theta$ is unbiased for θ, then (except in pathological settings) the estimator

$$\widetilde\theta(X) := \mathrm{Proj}_\Theta(\widehat\theta) = \operatorname*{argmin}_{\theta\in\Theta} \left\|\theta - \widehat\theta\right\|_2^2$$

has strictly lower mean-squared error

$$\mathbb{E}_\theta\left[\left\|\widetilde\theta(X) - \theta\right\|_2^2\right] < \mathbb{E}_\theta\left[\left\|\widehat\theta(X) - \theta\right\|_2^2\right]$$

for *all* $\theta \in \Theta$: a rebuke to the idea that the Cramér-Rao inequality provides any fundamental limit.

11.1.2 Regularization and the failure of the Cramér-Rao bound

An alternative perspective arises when we consider regularized estimators, which can be analytically simpler than projected estimators. Ridge or ℓ_2-regularization provides the most common example; for example, for linear regression, this is

$$\widehat\theta_\lambda = \operatorname*{argmin}_\theta \left\{\|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2\right\}.$$

There always exists $\lambda > 0$ with smaller mean-squared error than the unregularized estimator $\widehat{\theta}_0$ (see Exercise 11.2).

In the simpler problem of estimating the mean of a standard normal random vector, James-Stein-type estimators highlight an even more striking guarantee. Letting $Z \sim \mathsf{N}(\theta, I_d)$, James and Stein [382, 253] define the estimator

$$\widehat{\theta}_{\rm JS} := \left(1 - \frac{[d-2]_+}{\|Z\|_2^2}\right) Z.$$

Famously, this achieves smaller mean-squared error than the sample mean: in dimensions $d > 2$,

$$\mathbb{E}_\theta\left[\left\|\widehat{\theta}_{\rm JS} - \theta\right\|_2^2\right] < \mathbb{E}_\theta\left[\left\|Z - \theta\right\|_2^2\right] = d. \tag{11.1.1}$$

So an adaptive estimator strictly outperforms the sample mean, which achieves the information bound. While we do not prove the result, it shows why we must move beyond the Cramér-Rao bound: pointwise unbiased information bounds capture no fundamental limits.

11.2 The Van Trees inequality: a Bayesian Cramér-Rao bound

While our discussion in the preceding sections casts a dim view on the Fisher information "efficiency" bound that the Cramér-Rao inequality implies, it turns out that its main failing was in attempting to provide a pointwise bound valid for any parameter θ. By a bit of Bayesian averaging—as we do in essentially all of our minimax lower bounds—we can address many of the issues and show that similar information bounds hold, at least as the sample size n grows. The key will be that if we put a prior π on the parameter θ of interest, then *on average* over π, the parameter θ must correlate with the score $\dot{\ell}_\theta = \nabla \log p_\theta$. To set notation, we will let T denote a random draw $T \sim \pi$, so we generate X conditional on $T = \theta$.

11.2.1 The Van Trees inequality in one dimension

We first develop the *Van Trees* inequality in one-dimensional problems. Let $\theta \in \mathbb{R}$ be the parameter of interest, and assume that $\{p_\theta\}$ is a family of densities with respect to some base measure μ. Assume that the prior density π has support $[a, b]$, where $\pi(a) = \pi(b) = 0$, and that it is differentiable with finite prior Fisher information

$$J(\pi) := \int_a^b \frac{\pi'(t)^2}{\pi(t)} dt = \int_a^b \left((\log \pi(t))'\right)^2 \pi(t) dt.$$

Then in this case, we can show that a Fisher information lower bound holds *on average* over π, with a small additional penalty to capture the information π contains about the parameter θ. For notational simplicity in the theorem, we assume that $X \sim P_\theta$, detailing the i.i.d. case afterward.

Theorem 11.2.1 (The Van Trees inequality)**.** *Let the above conditions hold. Draw $T \sim \pi$ and conditional on $T = \theta$, draw $X \sim P_\theta$. Then*

$$\mathbb{E}_\pi\left[\mathbb{E}\left[(\widehat{\theta}(X) - T)^2 \mid T\right]\right] \geq \frac{1}{\mathbb{E}_\pi[J(T)] + J(\pi)}.$$

Proof Define the augmented score

$$\dot{\ell}_{\theta,\pi}(x) := \frac{\partial}{\partial\theta}\log(p_\theta(x)\pi(\theta)) = \frac{(p_\theta(x)\pi(\theta))'}{p_\theta(x)\pi(\theta)} = \frac{\dot{p}_\theta(x)}{p_\theta(x)} + \frac{\dot{\pi}(\theta)}{\pi(\theta)}.$$

Then $\widehat{\theta}(X) - \theta$ and $\dot{\ell}_{\theta,\pi}(X)$ are correlated, which then implies a lower bound. Indeed, taking an expectation jointly over $T \sim \pi$ and $X \sim P_\theta$, we can apply integration-by-parts (with $u = \widehat{\theta}(x) - \theta$ and $dv = (\pi(\theta)p_\theta(x))'d\theta$ in the identity $\int udv = uv - \int vdu$) to obtain

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\theta}(X) - T)\dot{\ell}_{T,\pi}(X)\right] &= \iint (\widehat{\theta}(x) - \theta)\dot{\ell}_{\theta,\pi}(x)\pi(\theta)p_\theta(x)d\mu(x)d\theta \\
&= \iint (\widehat{\theta}(x) - \theta)(\pi(\theta)p_\theta(x))'d\theta d\mu(x) \\
&= \int \left(\left[(\widehat{\theta}(x) - \theta)\pi(\theta)p_\theta(x)\right]_a^b + \int \pi(\theta)p_\theta(x)d\theta\right) d\mu(x) \\
&= 1, \qquad (11.2.1)
\end{aligned}$$

because $\pi(a) = \pi(b) = 0$.

Applying the Cauchy-Schwarz inequality to the key correlation identity (11.2.1), we have

$$\mathbb{E}\left[(\widehat{\theta}(X) - T)^2\right]^{1/2}\mathbb{E}\left[\dot{\ell}_{T,\pi}(X)^2\right]^{1/2} \geq 1.$$

Finally, we note that

$$\mathbb{E}[\dot{\ell}_{T,\pi}(X)^2] = \mathbb{E}[(\dot{\ell}_T(X) + \pi'(T)/\pi(T))^2] = \mathbb{E}_\pi[J(T)] + J(\pi),$$

because $\mathbb{E}_\pi[\pi'(T)/\pi(T)] = 0$. □

When we sample $X_1^n \stackrel{\text{iid}}{\sim} P_\theta$, then the Fisher information becomes $J_n(\theta) = nJ(\theta) = n\mathbb{E}_\theta[\dot{\ell}_\theta(X)^2]$, where X is a single observation. Theorem 11.2.1 thus provides the bound

$$\mathbb{E}\left[(\widehat{\theta}(X_1^n) - T)^2\right] \geq \frac{1}{n\mathbb{E}_\pi[J(T)] + J(\pi)} = \frac{1}{n\mathbb{E}_\pi[J(T)]} - O(1)\frac{J(\pi)}{\mathbb{E}_\pi[J(T)]^2 n^2},$$

assuming $J(\pi) < \infty$. In principle it is possible to choose the prior π to maximize the lower bound, but typically this is rather immaterial.

Because we usually want a *local* lower bound—something akin to the Cramér-Rao bound that (almost) applies to a particular problem parameter θ_0 of interest—it is most common to localize the prior π to some shrinking neighborhood of θ_0. Let π be any prior supported on a neighborhood of 0; for simplicity, we can take $[-1, 1]$. Then for any $a > 0$ and $\theta_0 \in \mathbb{R}$, the density

$$f_a(\theta) := \frac{1}{a} \cdot \pi\left(\frac{\theta - \theta_0}{a}\right)$$

satisfies $\int f_a(\theta)d\theta = 1$, has support $[\theta_0 - a, \theta_0 + a]$, and information $J(f_a) = \frac{1}{a^2}J(\pi)$. (See also Exercise 11.3.) If $J(\theta)$ is continuous in θ near θ_0, we obtain the following corollary.

Corollary 11.2.2. *Let $\theta_0 \in \mathbb{R}$, the density π have compact support on $[-1, 1]$ and finite information $J(\pi) < \infty$, and the sequence $a_n > 0$ satisfy $n^{-1/2} \ll a_n \ll 1$. Define the prior densities*

$$\pi_n(\theta) := \frac{1}{a_n} \cdot \pi\left(\frac{\theta - \theta_0}{a_n}\right).$$

If $J(\theta)$ is continuous in a neighborhood of θ_0, then

$$\int_{\theta_0 - a_n}^{\theta_0 + a_n} \mathbb{E}_\theta\left[\left(\widehat{\theta}(X_1^n) - \theta\right)^2\right] \pi_n(\theta) d\theta \ge \frac{1}{nJ(\theta_0)} - O(1)\frac{1}{nJ(\theta_0)} \cdot \frac{J(\pi)}{J(\theta_0)a_n^2 n}.$$

The inverse Fisher information $J(\theta)^{-1}$ thus provides a nearly pointwise lower bound on estimation in mean-square error.

11.2.2 The Van Trees inequality in d-dimensions

We can extend Theorem 11.2.1 to higher-dimensional scenarios as well: more or less, we simply apply the preceding lower bound to each coordinate of the parameter. In this case, instead of a scalar Fisher information, we have information matrices

$$J(\theta) := \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^\top] \quad \text{and} \quad J(\pi) = \mathbb{E}_\pi[\nabla \log \pi(\theta) \nabla \log \pi(\theta)^\top].$$

With this notation, the multivariate Van Trees inequality follows almost exactly the same lines of proof as Theorem 11.2.1, where we show that estimator error $\widehat{\theta} - \theta$ correlates with $\dot{\ell}_\theta$.

Theorem 11.2.3 (Basic multivariate Van Trees)**.** *Let $\pi = \pi_1 \times \cdots \times \pi_d$ be a product distribution on $\theta \in \mathbb{R}^d$, where each coordinate density π_j has support $[a_j, b_j]$ with $\pi_j(a_j) = \pi_j(b_j) = 0$. Then for any matrix $A \succ 0$,*

$$\mathbb{E}_\pi\left[\mathbb{E}\left[(\widehat{\theta}(X) - T)^\top A^{-1}(\widehat{\theta}(X) - T) \mid T\right]\right] \ge \frac{d^2}{\mathbb{E}_\pi[\mathrm{tr}(AJ(T))] + \mathrm{tr}(AJ(\pi))}.$$

Proof Let $\pi(\theta) = \prod_{j=1}^d \pi_j(\theta_j)$, and let $\pi_{\setminus j}(\theta_{\setminus j}) = \prod_{k \neq j} \pi_k(\theta_k)$. For shorthand, let $\partial_j = \frac{\partial}{\partial \theta_j}$. Expanding in each coordinate, we have

$$\begin{aligned}
\langle \widehat{\theta}(x) - \theta, \dot{\ell}_{\theta,\pi}(x) \rangle &= \sum_{j=1}^d (\widehat{\theta}_j(x) - \theta_j) \frac{\partial}{\partial \theta_j} \log(p_\theta(x)\pi(\theta)) \\
&= \sum_{j=1}^d (\widehat{\theta}_j(x) - \theta_j)\left([\dot{\ell}_\theta(x)]_j + \frac{\pi_j'(\theta_j)}{\pi_j(\theta_j)}\right) = \sum_{j=1}^d (\widehat{\theta}_j(x) - \theta_j) \frac{\partial_j(p_\theta(x)\pi_j(\theta_j))}{p_\theta(x)\pi_j(\theta_j)}.
\end{aligned}$$

Each term satisfies integration-by-parts identities as in (11.2.1) in the proof of Theorem 11.2.1, with

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\theta}_j(X) - T_j) \cdot [\dot{\ell}_{T,\pi}(X)]_j\right] &= \iint (\widehat{\theta}_j(x) - \theta_j)\partial_j(p_\theta(x)\pi_j(\theta_j)) d\theta_j \cdot \pi_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j} d\mu(x) \\
&= \iint \pi(\theta) p_\theta(x) d\theta d\mu(x) = 1.
\end{aligned}$$

We therefore obtain

$$\mathbb{E}\left[\langle \widehat{\theta}(X) - T, \dot{\ell}_{T,\pi}(X)\rangle\right] = \iint \langle \widehat{\theta}(x) - \theta, \dot{\ell}_{\theta,\pi}(x)\rangle \pi(\theta) p_\theta(x) d\mu(x) d\theta = d.$$

For a matrix $A \succ 0$, let $\|v\|_A^2 = \langle v, Av\rangle$ for shorthand. Applying the Cauchy-Schwarz inequality to the inner product $\langle u, v\rangle = \langle A^{-1/2}u, A^{1/2}v\rangle$,

$$\langle \widehat{\theta} - \theta, \dot{\ell}_{\theta,\pi}\rangle = \langle \widehat{\theta} - \theta, \dot{\ell}_\theta\rangle + \langle \widehat{\theta} - \theta, \nabla \log \pi(\theta)\rangle \le \|\widehat{\theta} - \theta\|_{A^{-1}} \left\|\dot{\ell}_\theta + \nabla \log \pi(\theta)\right\|_A.$$

Taking expectations, we obtain

$$\begin{aligned} d = \mathbb{E}\left[\langle \widehat{\theta}(X) - T, \dot{\ell}_{T,\pi}(X)\rangle\right] &\le \mathbb{E}\left[\left\|\widehat{\theta}(X) - T\right\|_{A^{-1}} \left\|\dot{\ell}_T(X) + \nabla \log \pi(T)\right\|_A\right] \\ &\le \mathbb{E}\left[\left\|\widehat{\theta}(X) - T\right\|_{A^{-1}}^2\right]^{1/2} \mathbb{E}\left[\left\|\dot{\ell}_T(X) + \nabla \log \pi(T)\right\|_A^2\right]^{1/2}. \end{aligned}$$

Recognize that $\mathbb{E}_\pi[\|\dot{\ell}_T(X) + \nabla \log \pi(T)\|_A^2] = \mathbb{E}_\pi\left[\mathrm{tr}(AJ(T))\right] + \mathrm{tr}(AJ(\pi))$. □

To demonstrate that the Van Trees inequality gives the exact optimality we have been seeking, we choose A to relate to the Fisher information $J(\theta)$. As motivation, recall our (heuristic) development in Chapter 3, where we approximate (3.2.6) that

$$\widehat{\theta}_n - \theta^\star \dot{\sim} \mathsf{N}\left(0, n^{-1} J(\theta^\star)^{-1}\right).$$

Chapter 5.3.4 makes this rigorous, and a small modification of Corollary 5.3.11 and its application of Corollary 5.3.10 yields the following result. (Exercise 11.4 asks for a proof.)

Corollary 11.2.4. *Let the conditions of Corollary 5.3.11 hold, let π be any density supported on a convex set Θ_0, and let Θ be any convex body with $\Theta \supset \Theta_0$. Let $\widehat{\theta}_n$ be the maximum likelihood estimator, define the projected estimator*

$$\widetilde{\theta}_n = \mathrm{Proj}_\Theta(\widehat{\theta}_n),$$

and let the matrix $B(\theta) \succeq 0$ be continuous in θ. Then uniformly in $\theta \in \Theta_0$

$$\mathbb{E}_\theta\left[\left\|\widetilde{\theta}_n(X_1^n) - \theta\right\|_{B(\theta)}^2\right] \le \frac{1}{n}\mathrm{tr}(B(\theta)\, J(\theta)^{-1}) + o(1/n).$$

The most natural idea to link Theorem 11.2.3 with Corollary 11.2.4 is to make the estimator $\widehat{\theta}_n$ *pivotal*, meaning that the asymptotic error $\widehat{\theta}_n - \theta^\star$ is independent of $\theta^\star$. Then (heuristically)

$$J(\theta^\star)^{1/2}\left(\widehat{\theta}_n - \theta^\star\right) \dot{\sim} \mathsf{N}\left(0, n^{-1} I_d\right).$$

Now assume $J(\theta)$ is continuous in θ. In Corollary 11.2.4, setting $B(\theta) = J(\theta)$ we obtain

$$\mathbb{E}_\theta\left[(\widetilde{\theta}_n(X_1^n) - \theta)^\top J(\theta)(\widetilde{\theta}_n(X_1^n) - \theta)\right] \le \frac{d}{n} + o(1/n)$$

simultaneously for all $\theta \in \Theta_0$. To compare with Theorem 11.2.3, fix $\theta^\star$ and let $\delta > 0$ and Θ_0 be a neighborhood of $\theta^\star$ small enough that $(1-\delta)J(\theta) \preceq J(\theta^\star) \preceq (1+\delta)J(\theta)$ for $\theta \in \Theta_0$. Then

$$\mathbb{E}_\pi\left[\left\|\widehat{\theta}_n(X_1^n) - T\right\|_{A^{-1}}^2\right] \ge \frac{d^2}{n\,\mathrm{tr}(A\mathbb{E}_\pi[J(T)]) + \mathrm{tr}(AJ(\pi))} \ge \frac{d^2}{n(1+\delta)\,\mathrm{tr}(AJ(\theta^\star)) + \mathrm{tr}(AJ(\pi))}$$

for *any* estimator $\widehat{\theta}_n$. Take $A = J(\theta^\star)^{-1}$ to obtain

$$\mathbb{E}_\pi\left[(\widehat{\theta}_n(X_1^n) - T)^\top J(\theta^\star)(\widehat{\theta}_n(X_1^n) - T)\right] \ge \frac{d^2}{n(1+\delta)d + \mathrm{tr}(J(\theta^\star)^{-1}J(\pi))} = \frac{d}{(1+\delta)n} - O(1/n^2),$$

showing that the Van Trees inequality (Theorem 11.2.3) is sharp.

11.2.3 The Van Trees inequality for a function of the parameter

In many statistical problems, it is interesting to estimate a function of a parameter of the underlying distribution rather than a parameter identifying the distribution itself. Suppose we have a statistic $\psi : \mathbb{R}^d \to \mathbb{R}^p$ and wish to estimate the parameter $\psi(\theta)$ rather than θ itself. The next example highlights a simple case where $\psi(\theta) = e_j^\top \theta$ for a standard basis vector e_j.

Example 11.2.5 (Treatment effect estimation)**:** Consider a randomized controlled trial, where individuals are assigned to treatment (with an indicator variable $Z = 1$) or control ($Z = 0$). In the *potential outcomes* framework, the response Y of an individual is $Y(1)$ in treatment and $Y(0)$ under control, and for any individual we observe only one of $Y(0)$ and $Y(1)$. We model the response linearly by

$$Y = Y(Z) = \beta_0 + tZ + \beta^\top X + \varepsilon,$$

where β_0 is an intercept and $X \in \mathbb{R}^d$ are covariates, where $\theta = (\beta_0, t, \beta) \in \mathbb{R}^{d+2}$. Assuming that $Z \sim \mathsf{Uniform}\{0, 1\}$, independently of X and $(Y(0), Y(1))$, the t-component of the optimizer $\theta^\star = \operatorname{argmin}_\theta \mathbb{E}[(Y - \beta_0 - tZ - \beta^\top X)^2]$ is the *average treatment effect* $t^\star = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. (See Exercise 11.5 as well as Example 18.1.1 to come.) $\diamond$

Assume we wish to estimate $\psi(\theta)$ for some $\psi : \mathbb{R}^d \to \mathbb{R}^p$ whose derivative matrix $\dot\psi(\theta) \in \mathbb{R}^{p\times d}$ has entries $\dot\psi_{ij}(\theta) = \frac{\partial}{\partial\theta_j}\psi_i(\theta)$, so that $\psi(\theta+\Delta) = \psi(\theta)+\dot\psi(\theta)\Delta+o(\|\Delta\|)$. We can extend Theorem 11.2.3 to this more general setting.

Theorem 11.2.6. *Let the conditions of Theorem 11.2.3 hold, and let $C \in \mathbb{R}^{d\times p}$. Assume that ψ is continuously differentiable on the support of π. Then for any $A \succ 0$ and estimator $\widehat\psi$ of $\psi(\theta)$,*

$$\mathbb{E}_\pi\left[(\widehat\psi(X) - \psi(T))^\top A^{-1}(\widehat\psi(X) - \psi(T))\right] \ge \frac{\mathbb{E}_\pi[\operatorname{tr}(\dot\psi(T)C)]^2}{\operatorname{tr}(CAC^\top \mathbb{E}_\pi[J(T)]) + \operatorname{tr}(CAC^\top J(\pi))}.$$

Proof Let $C \in \mathbb{R}^{d\times p}$ have rows $c_j^\top$ and define the error vector $E(x, \theta) = \widehat\psi(x) - \psi(\theta)$, so

$$C^\top \dot\ell_{\theta,\pi}(x) = \sum_{j=1}^d c_j \partial_j \log\left(p_\theta(x)\pi(\theta)\right) = \sum_{j=1}^d c_j\left(\frac{\partial_j p_\theta(x)}{p_\theta(x)} + \frac{\partial_j \pi_j(\theta_j)}{\pi_j(\theta_j)}\right).$$

Then using integration by parts,

$$\begin{aligned}\mathbb{E}[\langle\widehat\psi(X) - \psi(T), C^\top \dot\ell_{T,\pi}(X)\rangle] &= \sum_{j=1}^d \iint \langle E(x,\theta), c_j\rangle \partial_j\left(p_\theta(x)\pi_j(\theta_j)\right) d\theta_j \pi_{\setminus j}(\theta_{\setminus j}) d\theta_{\setminus j} d\mu(x) \\ &= \mathbb{E}_\pi\left[\operatorname{tr}(\dot\psi(T)C)\right].\end{aligned}$$

The remainder of the proof mimics that of Theorem 11.2.3. We have

$$\mathbb{E}\left[\langle E(X,T), C^\top \dot\ell_{T,\pi}(X)\rangle\right] \le \mathbb{E}\left[\|E(X,T)\|_{A^{-1}}^2\right]^{1/2} \mathbb{E}\left[\left\|C^\top \dot\ell_{T,\pi}(X)\right\|_A^2\right]^{1/2},$$

and because the scores are mean zero, we also have

$$\mathbb{E}_\pi\left[\left\|C^\top \dot\ell_{T,\pi}(X)\right\|^2\right] = \mathbb{E}_\pi\left[\left\|C^\top(\dot\ell_T(X) + \nabla\log\pi(T))\right\|_A^2\right] = \mathbb{E}_\pi\left[\operatorname{tr}(CAC^\top J(T))\right] + \operatorname{tr}(CAC^\top J(\pi)).$$

Squaring and dividing through gives the result. $\square$

11.2.4 The full multivariate Van Trees inequality

In demonstrating the sharpness of the multivariate Van Trees inequality (Theorem 11.2.3), we fixed the matrices A and C and leveraged that $J(\theta)$ was continuous. Doing so is not necessary, which leads to the final version of the Van Trees, or Bayesian Cramér-Rao, inequality, which extends Theorem 11.2.3 to include terms to better reflect the geometry of the problem at hand, which eliminates the need for some of the local approximations we have made to consider particular parameters θ_0 of interest. In this case, we let the matrices $A \succ 0$ and $C \in \mathbb{R}^{d \times p}$ vary in θ in Theorem 11.2.6. For such a setting, we will require a prior Fisher information with

$$s_{C,\pi}(\theta) := \frac{1}{\pi(\theta)} \left[\sum_{i=1}^{d} \frac{\partial}{\partial \theta_i} \left(C_{ij}(\theta) \pi(\theta) \right) \right]_{j=1}^{p} \in \mathbb{R}^p$$

taking the place of the usual score vector $\nabla \log \pi(\theta)$ (these agree when $p = d$ and $C = I_d$, so that $s_{I,\pi}(\theta) = \nabla \log \pi(\theta)$) and

$$J(\pi \mid A, C) := \mathbb{E}_\pi \left[s_{C,\pi}(T)^\top A(T) s_{C,\pi}(T) \right]. \tag{11.2.2}$$

To state the most general extension of Theorem 11.2.6, we will present a few additional definitions and conditions. A function f on $\Theta \subset \mathbb{R}^d$ is *suitably continuous* if for each coordinate j and almost all values $\theta \in \Theta$, the coordinate function $h(t) := f(\theta + te_j)$ is absolutely continuous. The conditions (i)–(v) below relax the assumptions necessary for Theorem 11.2.6.

(i) The density $p_\theta(x)$ is measurable in (x, θ) and, for almost every x, is suitably continuous in θ.

(ii) $\psi : \Theta \to \mathbb{R}^p$ and $C : \Theta \to \mathbb{R}^{d \times p}$ are suitably continuous.

(iii) The Fisher information $J(\theta) := \mathbb{E}_\theta[\dot{\ell}_\theta(X) \dot{\ell}_\theta(X)^\top]$ exists and $\operatorname{diag}(J(\theta))^{1/2}$ is locally integrable.

(iv) $A(\theta)$ is positive definite and continuous in θ.

(v) The prior density π is suitably continuous, its domain Θ is compact with piecewise $\mathcal{C}^1$ boundary, and $\pi(\operatorname{int} \Theta) > 0$ and $\pi(\theta) = 0$ for $\theta \in \operatorname{bd} \Theta$.

Then Gill and Levit [210, Theorem 1] show the following result.

Theorem 11.2.7 (The multivariate Van Trees inequality)**.** *Let conditions* (i)–(v) *hold. Then for any estimator* $\widehat{\psi} : \mathcal{X} \to \mathbb{R}^p$*,*

$$\mathbb{E}_\pi \left[(\widehat{\psi}(X) - \psi(T))^\top A(T)^{-1} (\widehat{\psi}(X) - \psi(T)) \right] \geq \frac{\mathbb{E}_\pi[\operatorname{tr}(\dot{\psi}(T) C(T))]^2}{\mathbb{E}_\pi[\operatorname{tr}(C(T) A(T) C(T)^\top J(T))] + J(\pi \mid A, C)}.$$

Exercise 11.6 asks you to prove a slightly weaker version of Theorem 11.2.7, which follows from arguments similar to those we use to prove Theorem 11.2.6. Under the same conditions, when $X_1^n \stackrel{\text{iid}}{\sim} P_\theta$, tensorization of the Fisher information implies

Corollary 11.2.8. *Let the conditions of Theorem 11.2.7 hold and* $X_1^n \stackrel{\text{iid}}{\sim} P_\theta$*. Then*

$$\mathbb{E}_\pi \left[(\widehat{\psi}_n(X_1^n) - \psi(T))^\top A(T)^{-1} (\widehat{\psi}_n(X_1^n) - \psi(T)) \right] \geq \frac{\mathbb{E}_\pi[\operatorname{tr}(\dot{\psi}(T) C(T))]^2}{n \mathbb{E}_\pi[\operatorname{tr}(C(T) A(T) C(T)^\top J(T))] + J(\pi \mid A, C)}.$$

We provide corollaries by choosing the matrices A and C appropriately. We begin by taking ψ and C to be identity and let $A(\theta) = J(\theta)^{-1}$ weight squared error. So long as $J(\theta) \succ 0$ and is continuous, using $\frac{1}{n+c} \geq \frac{1}{n} - \frac{c}{n^2}$ yields the following corollary.

Corollary 11.2.9. *Let the conditions of Theorem 11.2.7 hold. Then for $A(\theta) = J(\theta)$ and $C(\theta) = I_d$, we have $J(\pi \mid A, C) = \mathbb{E}_\pi[\nabla \log \pi(T) \nabla \log \pi(T)^\top]$ and*

$$\mathbb{E}_\pi \left[(\widehat{\theta}_n(X_1^n) - T)^\top J(T) (\widehat{\theta}_n(X_1^n) - T) \right] \geq \frac{d}{n} - \frac{J(\pi \mid A, C)}{n^2}.$$

Taking ψ to be the identity once again but letting $A(\theta) = I_d$ and $C(\theta) = J(\theta)^{-1}$, we obtain

$$\mathbb{E}_\pi \left[\left\| \widehat{\theta}_n(X_1^n) - T \right\|_2^2 \right] \geq \frac{\mathbb{E}_\pi[\mathrm{tr}(J(T)^{-1})]}{n} - \frac{J(\pi \mid A, C)}{n^2}.$$

When $\psi : \mathbb{R}^d \to \mathbb{R}^p$ is smooth we can develop related results. We first heuristically develop an upper bound using the *delta method*: for $\widehat{\theta}_n$ obeying the asymptotics $\widehat{\theta}_n - \theta^\star \dot{\sim} \mathsf{N}(0, (1/n)\Sigma)$, we proceed heuristically to obtain

$$\psi(\widehat{\theta}_n) - \psi(\theta^\star) = \dot{\psi}(\theta^\star)(\widehat{\theta}_n - \theta^\star) + \underbrace{O(\|\widehat{\theta}_n - \theta^\star\|^2)}_{= O(1/n)} \dot{\sim} \mathsf{N}\left(0, \frac{1}{n} \dot{\psi}(\theta^\star) \Sigma \dot{\psi}(\theta^\star)^\top \right).$$

(Refer to any text on asymptotic statistics, for example, van der Vaart [407, Ch. 3], or Exercise 5.13.)

To develop a matching lower bound, assume that $p \leq d$ and that $\dot{\psi}(\theta) \in \mathbb{R}^{p \times d}$ is rank p, so that a change of variables yields Fisher information

$$J_\psi(\theta) := \left(\dot{\psi}(\theta) J(\theta)^{-1} \dot{\psi}(\theta)^\top \right)^{-1} \in \mathbb{R}^{p \times p}. \tag{11.2.3}$$

Then Exercise 11.7 shows how to prove the following corollary.

Corollary 11.2.10. *In addition to the conditions of Theorem 11.2.7, assume that $\dot{\psi}(\theta) J(\theta)^{-1} \dot{\psi}(\theta)^\top$ is continuously differentiable in θ and positive definite. Let $J_\psi(\theta)$ be the Fisher information (11.2.3) for ψ in the model $\{P_\theta\}$. Then there is a constant K depending on π and ψ such that*

$$\mathbb{E}_\pi \left[(\widehat{\psi}_n(X_1^n) - \psi(T))^\top J_\psi(T) (\widehat{\psi}_n(X_1^n) - \psi(T)) \right] \geq \frac{p}{n} - \frac{K}{n^2}.$$

As an example, Corollary 11.2.10 provides lower bounds for estimating single coordinates in a linear regression problem (recall also Example 11.2.5).

Example 11.2.11: Assume the linear regression model

$$Y_i = X_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0, \sigma^2)$$

with known design $X = [X_1 \ \cdots \ X_n]^\top$. Then $\psi(\theta) = e_j^\top \theta$ satisfies $\dot{\psi}(\theta) = e_j^\top$, and Corollary 11.2.10 gives

$$\mathbb{E}_\pi \left[(\widehat{\theta}_j - T_j)^2 \right] \geq \frac{\sigma^2}{n} \cdot e_j^\top \left(n^{-1} X^\top X \right)^{-1} e_j - O\left(1/n^2\right).$$

The least-squares estimator $\widehat{\theta} = (X^\top X)^{-1} X^\top Y$ achieves this risk. $\diamond$

11.3 Beyond parametric problems

We often have statistical problems in which no particular model underlies the problem, and we wish only to estimate a parameter defined in terms of the underlying distribution generating the data. For example, in the M-estimation problems in Chapter 5.3 we seek the function

$$\theta(P) := \operatorname*{argmin}_\theta \mathbb{E}_P[\ell(\theta, Z)]$$

of the distribution P. Treating the probability distribution P as the "parameter," $\theta(P)$ is a function of that parameter, which may seem hard to address with the techniques we have developed, as P is (typically) infinite dimensional. This is also the more abstract setting we adopt in Chapter 9.

To extend the techniques for estimation error lower bounds to this setting, we take a perspective originating with Stein [383]: estimation in a general problem should be at least as hard as estimation in any particular (model-based) sub-problem. Thus, we take the following two-phase approach to developing lower bounds for estimating $\theta(P)$ around a base distribution P_0:

1. For the base distribution, define a model family $\{P_t\}$ indexed by $t \in \mathbb{R}^k$
2. Write $\theta(P_t)$ as a function $\theta(t)$ of the parameter t, then apply Theorem 11.2.7 or its corollaries.

By choosing the "hardest" sub-family of models in Step 1, we obtain the largest lower bounds. Exercises 9.8, 9.10, and 9.11 explore this approach in the context of information-theoretic lower bounds. Here, we employ the ideas to obtain asymptotically exact results.

The approach becomes rather pedestrian given our development. Assume without loss of generality that P_0 has a density p_0 on $\mathcal{X}$ with respect to a base measure μ (we could always simply take $\mu = P_0$). Then for a bounded function $g : \mathcal{X} \to \mathbb{R}^k$ with mean $\mathbb{E}_0[g(X)] = 0$, define the *tilted density*

$$p_t(x) := (1 + \langle t, g(x)\rangle)p_0(x). \tag{11.3.1}$$

For any $t \in \mathbb{R}^k$, we have $\int p_t(x)d\mu(x) = \mathbb{E}_0[(1 + \langle t, g(X)\rangle)] = 1$, and because g is bounded, for t near enough 0 we are guaranteed that $p_t \geq 0$, so that p_t is indeed a probability density. Given such tilted densities, we assume that the parameter $\theta(P)$ of interest is locally differentiable:

Assumption A.11.1. *Let U be a neighborhood of 0. For the model family $\{P_t\}_{t\in U}$, the parameter $\theta(P_t)$ is continuously differentiable in t on U, with derivative matrix $\dot\theta(t) \in \mathbb{R}^{d\times k}$.*

Assumption A.11.1 is abstract, but frequently holds. We provide an example with mean estimation here, and after presenting the general lower bound leveraging the assumption, show how it applies to the M-estimation problems in Chapter 5.3.

Example 11.3.1 (Nonparametric mean estimation)**:** For distributions P on $X \in \mathbb{R}^d$, define $\theta(P) := \mathbb{E}_P[X]$. For the tilted family $\{P_t\}$ of (11.3.1),

$$\theta(P_t) = \mathbb{E}_0[(1 + \langle t, g(X)\rangle)X] = \mathbb{E}_0[X] + \operatorname{Cov}_0(X, g(X))t = \theta(P_0) + \operatorname{Cov}_0(X, g(X))t,$$

where we use that g is mean zero. Then $\theta(P_t)$ is affine in t and hence differentiable. $\diamond$

The tilted construction (11.3.1) allows us to compute the Fisher information matrix for the nuisance parameter t we have invented. Because $\log p_t(x) = \log(1 + \langle t, g(x)\rangle) + \log p_0(x)$, we obtain

$$\nabla \log p_t(x) = \frac{g(x)}{1 + \langle t, g(x)\rangle} \quad \text{and} \quad \mathbb{E}_t[\nabla \log p_t(X)] = \mathbb{E}_0\left[\frac{1 + \langle t, g(X)\rangle}{1 + \langle t, g(X)\rangle} g(X)\right] = \mathbb{E}_0[g(X)] = 0$$

by assumption that $\mathbb{E}_0[g] = 0$. At $t = 0$, the Fisher information

$$J(0) = \mathbb{E}_0[g(X)g(X)^\top] = \mathrm{Cov}_0(g(X)),$$

and because g is bounded, has the uniform expansion in t near 0 that

$$J(t) = \mathbb{E}_0\left[\frac{g(X)g(X)^\top}{1 + \langle t, g(X)\rangle}\right] = \mathbb{E}_0[g(X)g(X)^\top(1 - \langle t, g(X)\rangle)] + O(\|t\|^2) = J(0) + O(\|t\|). \quad (11.3.2)$$

The information J is also continuously differentiable and positive definite when $\mathrm{Cov}_0(g(X)) \succ 0$.

As such, whenever Assumption A.11.1 holds, so do the assumptions necessary for Theorem 11.2.7. We therefore obtain the following corollary:

Corollary 11.3.2. *Let Assumption A.11.1 hold and P_t be the tilted distributions (11.3.1), and let conditions (i)–(v) hold with t replacing θ. Let $T \sim \pi$. Then for any estimator $\widehat{\theta} : \mathcal{X}^n \to \mathbb{R}^d$,*

$$\int \mathbb{E}_{P_t}\left[(\widehat{\theta}(X_1^n) - \theta(P_t))^\top A(t)^{-1}(\widehat{\theta}(X_1^n) - \theta(P_t))\right]\pi(t)dt \geq \frac{\mathbb{E}_\pi[\mathrm{tr}(\dot{\theta}(T)C(T))]^2}{n\mathbb{E}_\pi[\mathrm{tr}(C(T)A(T)C(T)^\top J(T))] + J(\pi \mid A, C)}.$$

As in Section 11.2, we choose A and C to optimize the lower bound in Corollary 11.3.2. In analogy with Corollary 11.2.10, for example, if the Fisher information for the parameter $\theta = \theta(t)$,

$$J_\theta(t) := \left(\dot{\theta}(t)J(t)^{-1}\dot{\theta}(t)^\top\right)^{-1},$$

is continuously differentiable in t and positive definite, then we obtain

$$\mathbb{E}_\pi\left[(\widehat{\theta}(X_1^n) - \theta(P_T))^\top J_\theta(T)(\widehat{\theta}(X_1^n) - \theta(P_T))\right] \geq \frac{d}{n} - O(1/n^2).$$

11.3.1 Influence function lower bounds

One particular (abstract but common) setting is important: when $\theta(P_t)$ is not only differentiable in t, but there exists a mean-zero *influence function* $\dot{\theta}_{P_0} : \mathcal{X} \to \mathbb{R}^d$ for θ satisfying

$$\theta(P_t) = \theta(P_0) + \mathbb{E}_{P_0}[\dot{\theta}_{P_0}(X)g(X)^\top]t + o(\|t\|) \quad (11.3.3)$$

for all bounded, mean-zero g. Then $\dot{\theta}(0) = \mathbb{E}_{P_0}[\dot{\theta}_{P_0}(X)g(X)^\top]$ and the derivative in Assumption A.11.1 is linear in g. (The assumption that $\dot{\theta}_{P_0}$ is mean zero loses no generality, as $\mathbb{E}_{P_0}[g(X)] = 0$.) This linearity holds, for example, for the mean in Example 11.3.1 with no error term, where

$$\dot{\theta}_{P_0}(x) = x - \mathbb{E}_{P_0}[X].$$

We will also see it for general M-estimators in Section 11.3.2.

Heuristically, we think of the influence function as the effect a single example x (with infinitesimal weight) has on the parameter. Letting $\mathbf{1}_x$ denote a point mass at x and assuming

$$\dot{\theta}(x) := \lim_{t \downarrow 0} \frac{\theta((1-t)P_0 + t\mathbf{1}_x) - \theta(P_0)}{t},$$

exists, let P_t have density $p_t(x) = (1 + \langle t, g(x)\rangle)p_0(x)$. Then write

$$\begin{aligned}\theta(P_t) &\doteq \theta\left(P_0 + \int \langle t, g(x)\rangle \mathbf{1}_x dP_0(x)\right) \\ &\doteq \theta(P_0) + \int \dot\theta(x) g(x)^\top t dP_0(x) + o(t) = \theta(P_0) + \mathbb{E}_{P_0}[\dot\theta(X)g(X)^\top]t + o(t),\end{aligned}$$

where each of the $\doteq$ equalities indicates a heuristic approximate equality, so long as we assume the derivatives are appropriately linear. (One may make this rigorous by working with square-integrable g and using the Riesz representation theorem [407, Ch. 25].)

Let us assume the influence function expansion (11.3.3); we show the types of lower bounds its existence provides. For simplicity, we specialize to $k = d$ and consider product priors, and let π_0 be a smooth density supported on $[-1, 1]$ and $r_n \to \infty$ satisfy $1 \ll r_n \lesssim \sqrt{n}$. Then we take the prior

$$\pi_n(t) = \prod_{j=1}^d r_n \pi_0(r_n t_j)$$

consisting of i.i.d. coordinates on $[-1/r_n, 1/r_n]^d$. The prior information $J(\pi_n)$ then scales as $d \cdot r_n^2$ (recall Corollary 11.2.2), and for bounded $g : \mathcal{X} \to \mathbb{R}^d$, we have $J(t) = \mathrm{Cov}_0(g(X)) + O(\|t\|)$ by the expansion (11.3.2). Taking $C = I_d$ in Theorem 11.2.7, we obtain that for a random vector $T \sim \pi_n$,

$$\mathbb{E}_{\pi_n}\left[(\widehat\theta(X_1^n) - \theta(P_T))^\top A(T)^{-1}(\widehat\theta(X_1^n) - \theta(P_T))\right] \ge \frac{\mathrm{tr}(\mathbb{E}_{P_0}[\dot\theta_{P_0}(X)g(X)^\top])^2}{n \cdot \mathbb{E}_{\pi_n}[\mathrm{tr}(A(T)\mathrm{Cov}_0(g(X)))]} - O\left(\frac{r_n^2}{n^2}\right)$$

for any estimator $\widehat\theta$ and positive definite $A(t)$, where $O(\cdot)$ hides problem-dependent terms. Taking $A(t) = I_d$ gives

$$\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(X_1^n) - \theta(P_T)\right\|_2^2\right] \ge \frac{\mathbb{E}_{P_0}[\langle \dot\theta_{P_0}(X), g(X)\rangle]^2}{n\mathbb{E}_{P_0}[\|g(X)\|_2^2]} - O(r_n^2/n^2).$$

The influence function $\dot\theta_{P_0}$ allows us to determine the "hardest" sub-model g by maximizing the preceding display, and by Cauchy-Schwarz, this is $g(x) = \dot\theta_{P_0}(x)$, giving the lower bound

$$\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(X_1^n) - \theta(P_T)\right\|_2^2\right] \ge \frac{\mathbb{E}_{P_0}[\|\dot\theta_{P_0}(X)\|_2^2]}{n} - O(r_n^2/n^2). \tag{11.3.4}$$

(The bound (11.3.4) holds for unbounded $g = \dot\theta_{P_0}$ by approximating with bounded functions.)

Because these results involve only linear and quadratic terms, it is easy to obtain other evocative lower bounds. If we let $\Sigma = \mathrm{Cov}(\dot\theta_{P_0}(X))$, then setting $A = \Sigma$ we obtain

$$\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(X_1^n) - \theta(P_T)\right\|_{\Sigma^{-1}}^2\right] \ge \frac{\mathbb{E}[\langle \dot\theta_{P_0}(X), g(X)\rangle]^2}{n\,\mathrm{tr}(\Sigma \cdot \mathrm{Cov}_0(g(X)))} - O(r_n^2/n^2),$$

where $\|x\|_A = \sqrt{x^\top A x}$ is the Mahalonobis norm. Once again applying Cauchy-Schwarz to obtain the most difficult model, we take $g(x) = \Sigma^{-1}\dot\theta_{P_0}(x)$ and a rate $r_n \ll \sqrt{n}$ to obtain

$$\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(X_1^n) - \theta(P_T)\right\|_{\Sigma^{-1}}^2\right] \ge \frac{d}{n} - o(1/n).$$

These lower bounds highlight both the power and downsides of the approach we have adopted: in comparison with the bounds available via information-theoretic techniques, such as those in Exercises 9.10 and 9.11, we lose generality in that the results apply only to the squared error. At the same time, the constants are sharp—they are unimprovable.

11.3.2 An extended example: M-estimation lower bounds

We can apply our techniques to develop a fuller optimality theory for M-estimators. Recall the setting of Chapter 5.3, with population loss $L(\theta) := \mathbb{E}_{P_0}[\ell(\theta, Z)]$ for a loss function ℓ and distribution P_0 on Z, and define the parameter

$$\theta(P_0) := \operatorname*{argmin}_\theta L(\theta).$$

We provide lower bounds on estimating $\theta(P)$ by deriving an influence function expansion (11.3.3) under tilted (11.3.1) models P_t of P_0.

Let $L_t(\theta) = \mathbb{E}_{P_t}[\ell(\theta, Z)] = \mathbb{E}_0[(1 + \langle t, g(Z)\rangle)\ell(\theta, Z)]$ be the tilted population loss, and for shorthand let $\theta_t = \theta(P_t)$. Then applying the implicit function theorem, we have the following lemma, whose proof we defer to the end of the section; while it holds under weaker conditions (see the bibliographic discussion), we prefer its simplicity.

Lemma 11.3.3. *Let $g : \mathcal{X} \to \mathbb{R}^k$ be bounded and mean-zero and Assumption A.5.1 hold on the losses ℓ and population loss L_0. Then for $t \in \mathbb{R}^k$ near 0, θ_t exists, is unique, and is differentiable in t, with*

$$\theta_{t+v} - \theta_t = -\mathbb{E}_0\left[(\nabla^2_\theta L_t(\theta_t))^{-1}\nabla\ell(\theta_t, Z)g(Z)^\top\right] v + o(\|v\|).$$

The smoothness conditions in Assumption A.5.1 immediately imply the influence function expansion (11.3.3), where

$$\dot\theta_{P_0}(z) = -(\nabla^2 L_0(\theta_0))^{-1}\nabla\ell(\theta_0, z),$$

and moreover, we have derivative matrix

$$\mathsf{D}_t\theta_t = \mathbb{E}_{P_0}[\dot\theta_{P_0}(Z)g(Z)^\top] + O(\|t\|).$$

Applying Corollary 11.3.2, we therefore obtain the following lower bound on the squared error in an M-estimation problem.

Corollary 11.3.4. *Let $1 \ll r_n \ll \sqrt{n}$ and π_n be smooth enough and have support on the scaled ℓ_∞ ball $[-1/r_n, 1/r_n]^d$. Let $T \sim \pi_n$. Then for any matrix $A(t) \succ 0$, continuous in t, and estimator $\widehat\theta$,*

$$\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(Z_1^n) - \theta(P_T)\right\|^2_{A(T)^{-1}}\right] \ge \frac{\mathbb{E}_{P_0}\left[\langle\nabla^2 L_0(\theta_0)^{-1}\nabla\ell(\theta_0, Z), g(Z)\rangle\right]^2}{n \cdot \mathbb{E}_{\pi_n}[\operatorname{tr}(A(T)\operatorname{Cov}_0(g(Z)))]} - o(1/n).$$

The game becomes to choose g and A to obtain strong lower bounds. For example, the choice $g(z) = \nabla^2 L_0(\theta_0)^{-1}\nabla\ell(\theta_0, z)$ and $A = I$ yields lower bound

$$n\mathbb{E}_{\pi_n}\left[\left\|\widehat\theta(Z_1^n) - \theta(P_T)\right\|_2^2\right] \ge \operatorname{tr}\left(\nabla^2 L_0(\theta_0)^{-1}\operatorname{Cov}_0(\nabla\ell(\theta_0, Z))\nabla^2 L_0(\theta_0)^{-1}\right) - o(1),$$

matching the upper bounds in Section 5.3. (See Exercise 11.9 for other choices.) In brief, Corollary 11.3.4 shows that the error of the empirical risk minimizers (5.3.1) is asymptotically optimal.

Finally, we return to prove Lemma 11.3.3.

Proof of Lemma 11.3.3 The result is nearly immediate once we have the implicit function theorem, a version of which we state here.

Lemma 11.3.5 (Implicit function theorem)**.** *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be continuously differentiable in its coordinates $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$. Assume that $F(x_0, y_0) = 0$ and that the x-derivative matrix $\mathsf{D}_x F(x, y) \in \mathbb{R}^{n \times n}$ is invertible at (x_0, y_0). Then there is an open neighborhood U of y_0 such that a unique solution $x(y)$ to $F(x, y) = 0$ exists, and it is continuously differentiable on U with*

$$\dot{x}(y) = -(\mathsf{D}_x F(x, y))^{-1} \mathsf{D}_y F(x, y).$$

To apply Lemma 11.3.5, we compute the derivatives of $L_t(\theta) = \mathbb{E}_0[(1 + \langle t, g(Z)\rangle)\ell(\theta, Z)]$ via

$$\nabla_t L_t(\theta) = \mathbb{E}_0[g(Z)\ell(\theta, Z)], \quad \nabla_\theta L_t(\theta) = \mathbb{E}_0[(1 + \langle t, g(Z)\rangle)\nabla\ell(\theta, Z)]$$

and the second derivatives

$$\nabla^2_{t,\theta} L_t(\theta) = \mathbb{E}_0[g(Z)\nabla\ell(\theta, Z)^\top] \in \mathbb{R}^{k \times d} \text{ and } \nabla^2_\theta L_t(\theta) = \mathbb{E}_0[(1 + \langle t, g(Z)\rangle)\nabla^2\ell(\theta, Z)].$$

At $\theta_0 = \theta(P_0)$, we have by assumption that $\nabla^2_\theta L_0(\theta) = \mathbb{E}_0[\nabla^2\ell(\theta_0, Z)] \succ 0$, and $\nabla_\theta L_0(\theta_0) = 0$. Make the identifications $x \mapsto \theta$, $y \mapsto t$ in the implicit function theorem, so we also identify $F(x, y) \mapsto \nabla_\theta L_t(\theta)$ and $\mathsf{D}_x F(x, y) \mapsto \nabla^2_\theta L_t(\theta)$ and $\mathsf{D}_y F(x, y) \mapsto \nabla^2_{t,\theta} L_t(\theta)^\top$. Then we can write

$$\theta_{t+v} = \theta_t - \mathbb{E}_0\left[(\nabla^2_\theta L_t(\theta_t))^{-1}\nabla\ell(\theta_t, Z)g(Z)^\top\right] v + o(\|v\|)$$

for t in a neighborhood of 0 and small $v \in \mathbb{R}^k$. □

11.4 Super-efficiency and instance optimality

Most of our lower bound development before this chapter was for worst-case (minimax) error measures. No statistician should find these fully compelling—one wants to know how hard estimation and learning in the actual distribution P at hand is! Consider, for instance, the family $\{\mathsf{N}(\theta, \sigma^2 I_d)\}_{\theta \in \mathbb{R}^d, \sigma^2 < \infty}$ of Gaussian distributions with unknown variance. The worst-case mean-square error for estimating the mean over this class is $+\infty$. We therefore seek more nuance: an "instance-optimal" lower bound, which applies to individual probability distributions $P \in \mathcal{P}$, describing how hard an estimation problem is for the particular P at hand.

To set the stage, we highlight two desiderata that a satisfying benchmark—we avoid the terminology of lower bounds—of a problem's complexity should enjoy:

(OPT.i) The benchmark should be *instance-specific*, providing a quantity for each $P \in \mathcal{P}$

(OPT.ii) The benchmark should be *uniformly achievable*, in that there should exist a procedure achieving the benchmark performance on *each* instance $P \in \mathcal{P}$

At some level, minimax complexity guarantees satisfy the two desiderata (OPT.i)–(OPT.ii): a constant function at least provides a quantity for each $P \in \mathcal{P}$, and minimaxity means that the result is achievable. The key new requirement is that a bound should provide a converse:

(OPT.iii) The benchmark should provide a *super-efficiency* converse: no procedure outperforms the benchmark except on a negligible collection of instances.

The bounds we develop in the local Fano method in Chapter 9.4 approach the desiderata (OPT.i)–(OPT.iii): they localize to a distribution P_θ, and (frequently) the bounds are achievable. For example, Example 9.4.4 shows that for some constant $c < \infty$,

$$\sup_{\|\theta-\theta_0\|_2 \le c\sqrt{d/n}} \mathbb{E}_\theta \left[\left\| \widehat{\theta}_n - \theta \right\|_2^2 \right] \gtrsim \frac{d\sigma^2}{n} \tag{11.4.1}$$

in the d-dimensional normal location family $\{\mathsf{N}(\theta, \sigma^2 I_d)\}_{\theta \in \mathbb{R}^d}$. What is missing, however, is a super-efficiency (OPT.iii) converse: nothing in the statement of the bound (11.4.1) prevents an estimator from achieving much smaller than $\frac{d\sigma^2}{n}$ error at all except for a few worst-case points.

The variations on the Van Trees inequality in Sections 11.2 and 11.3 show that for parametric families (and even beyond), the Fisher information provides a super-efficiency guarantee for the squared error: by Corollary 11.2.9, for smooth enough prior densities π, any estimator $\widehat{\theta}_n$ satisfies

$$\int \mathbb{E}_\theta \left[\left\| \widehat{\theta}_n - \theta \right\|_2^2 \right] \pi(\theta) d\theta \ge \frac{1}{n} \int \mathrm{tr}(J(\theta)^{-1}) \pi(\theta) d\theta - O(1/n^2).$$

Because $J(\theta) = \sigma^{-2} I_d$ in the normal location family, the lower bound (11.4.1) then strengthens to $\int \mathbb{E}_\theta[\|\widehat{\theta}_n - \theta\|_2^2] \pi(\theta) d\theta \ge \frac{d\sigma^2}{n}(1 - o(1))$.

To make this discussion precise, consider Section 11.2.3, where we wish to estimate a differentiable function $\psi(\theta) \in \mathbb{R}^p$ of the parameter $\theta \in \mathbb{R}^d$. (This includes the non-parametric settings in Section 11.3.) For a procedure $\widehat{\psi}$, or more accurately, a sequence of procedures $\{\widehat{\psi}_n\}$ defined for each sample size n, define the pointwise limiting squared error at θ by

$$L_A\left(\theta, \{\widehat{\psi}_n\}\right) := \limsup_n n \cdot \mathbb{E}_\theta \left[(\widehat{\psi}_n(X_1^n) - \psi(\theta))^\top A(\theta)^{-1} (\widehat{\psi}_n(X_1^n) - \psi(\theta)) \right]$$

For technical reasons, we will require our estimators $\widehat{\psi}$ to converge uniformly for the prior π, by which we mean that there exists a function B such that

$$n \cdot \mathbb{E}_\theta \left[(\widehat{\psi}_n(X_1^n) - \psi(\theta))^\top A(\theta)^{-1} (\widehat{\psi}_n(X_1^n) - \psi(\theta)) \right] \le B(\theta) \quad \text{and} \quad \int B(\theta) \pi(\theta) d\theta < \infty. \tag{11.4.2}$$

(Any estimator failing the convergence (11.4.2) exhibits somewhat irregular behavior.)

Theorem 11.4.1. *Let the conditions of Theorem 11.2.7 hold. Then for any estimator sequence $\{\widehat{\psi}_n\}$ that converges uniformly for the prior π,*

$$\int L_A\big(\theta, \{\widehat{\psi}_n\}\big) \pi(\theta) d\theta \ge \int \mathrm{tr}\left(\dot{\psi}(\theta) J(\theta)^{-1} \dot{\psi}(\theta)^\top A(\theta)^{-1} \right) \pi(\theta) d\theta.$$

Proof Use the shorthand $L_A(\theta) = L_A(\theta, \{\widehat{\psi}_n\})$. Define the nonnegative functions

$$f_n(\theta) := n \mathbb{E}_\theta \left[(\widehat{\psi}_n(X_1^n) - \psi(\theta))^\top A(\theta)^{-1} (\widehat{\psi}_n(X_1^n) - \psi(\theta)) \right],$$

which satisfy $\limsup_n f_n(\theta) = L_A(\theta)$. Taking the matrix $C(\theta) = J(\theta)^{-1} \dot{\psi}(\theta)^\top A(\theta)^{-1}$, Corollary 11.2.8 implies

$$\mathbb{E}_\pi \left[f_n(T) \right] \ge \mathbb{E}_\pi[\mathrm{tr}(\dot{\psi}(T) J(T)^{-1} \dot{\psi}(T)^\top A(T)^{-1})] - O(1/n),$$

where the big-O hides terms that depend on the prior π. Fatou's lemma (Proposition A.1.1) and the uniform convergence assumption imply

$$0 = \int \left(L_A(\theta) - \limsup_n f_n(\theta)\right) \pi(\theta) d\theta = \int \liminf_n \left(L_A(\theta) - \limsup_n f_n(\theta)\right) \pi(\theta) d\theta$$
$$\le \liminf_n \int \left(L_A(\theta) - f_n(\theta)\right) \pi(\theta) d\theta.$$

Rearranging yields the desired result. □

Taking $A = I_d$ and $\psi(\theta) = \theta$ to be the identity mapping, Theorem 11.4.1 thus shows that for smooth enough compactly supported priors π, the limiting mean squared error of any uniformly convergent estimator satisfies

$$\int \limsup_n n\mathbb{E}_\theta \left[\left\|\widehat{\theta}_n - \theta\right\|_2^2\right] \pi(\theta) d\theta \ge \int \mathrm{tr}(J(\theta)^{-1}) \pi(\theta) d\theta. \tag{11.4.3}$$

With this in mind, define $\widehat{\theta}_n$ to be *efficient* for the squared error in a neighborhood of θ_0 if

$$n \cdot \mathbb{E}_\theta \left[\left\|\widehat{\theta}_n - \theta\right\|_2^2\right] \le (1 + o(1)) \cdot \mathrm{tr}(J(\theta)^{-1})$$

uniformly in θ near θ_0, so it (asymptotically) achieves the information lower bound near θ_0. Call $\widehat{\theta}$ *super-efficient* at a point θ if

$$\limsup_n n \cdot \mathbb{E}_\theta[\|\widehat{\theta}_n - \theta\|_2^2] < \mathrm{tr}\left(J(\theta)^{-1}\right).$$

Theorem 11.4.1 implies that the points of super-efficiency of an efficient estimator necessarily have measure 0, so long as $\int \mathrm{tr}(J(\theta)^{-1}) \pi(\theta) d\theta < \infty$. Indeed, inequality (11.4.3) shows that

$$0 \ge \int \left(\mathrm{tr}(J(\theta)^{-1}) - \limsup_n n\mathbb{E}_\theta \left[\left\|\widehat{\theta}_n - \theta\right\|_2^2\right]\right) \pi(\theta) d\theta \ge 0$$

because the integrand is nonnegative, so the integrand must be 0 for almost every θ. Summarizing, we have the following corollary:

Corollary 11.4.2. *An efficient estimator $\widehat{\theta}_n$ can be super-efficient on at most a (Lebesgue) measure 0 set of points.*

More sophisticated notions of efficiency and super-efficiency have long animated mathematical and theoretical statistics; see the bibliographic section and exercises for more on this.

11.5 Case study: a lower bound for differential privacy

The preceding sections show that score correlation—that $\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x)$ must correlate with accurate estimators—provides sharp lower bounds for squared-error estimation in regular statistical models. This insight applies beyond classical asymptotic theory, and it has become a key approach to providing lower bounds in private estimation, memorization lower bounds (that accurate estimators necessarily retain precise information about the sample P_n), and related research areas. We outline the technique here through an extended application to private estimation, extending the

constrained lower bounds of Chapter 10 to work with central differential privacy. This provides a self-contained case study in which the ingredients above—Fisher information, integration by parts, local alternatives, and correlation with the score—combine to yield lower bounds. We present the general idea in the next section and defer the technical details to Sections 11.5.2*–11.5.4*.

To study private estimators, we recall Chapter 8. A mechanism $M : \mathcal{P}_n \to \Theta$ is a (randomized) mapping from the collection $\mathcal{P}_n$ of empirical distributions P_n on n observations X_1^n. Then M is (ε, δ)-differentially private if $\mathbb{P}(M(P_n) \in A) \le e^\varepsilon \mathbb{P}(M(P_n') \in A) + \delta$ for all sets A and neighboring samples P_n, P_n'. Otherwise, we maintain the setting of this chapter, assuming a parametric family of probability distributions $\{P_\theta\}$ with densities $\{p_\theta\}$, and for which the Fisher information $J(\theta) := \mathbb{E}_\theta[\dot\ell_\theta(X)\dot\ell_\theta(X)^\top]$ is well-defined, and that $\mathbb{E}_\theta[\dot\ell_\theta(X)] = 0$ as is typical.

11.5.1 The alignment (or score attack)

To prove the lower bounds, we leverage the *alignment*, a measure of the correlation between the score $\dot\ell_\theta$ and the error $M(P_n) - \theta$,

$$\mathsf{A}_\theta(x, M(P_n)) := \langle M(P_n) - \theta, \dot\ell_\theta(x)\rangle. \tag{11.5.1}$$

Researchers also call the quantity (11.5.1) the *score attack*, with the intuition that if $\langle M(P_n) - \theta, \dot\ell_\theta(x)\rangle$ is large, then we identify that x is near an element of the sample X_i.

The alignment (11.5.1) enjoys two important conditions that lead to optimality results. First, it satisfies an *identifiability* condition, in complete analogy with the correlation (11.2.1), that guarantees that $\sum_{i=1}^n \mathbb{E}[\mathsf{A}_\theta(X_i, M(P_n))]$ is large whenever $M(P_n)$ accurately estimates θ. Second, it satisfies a *soundness* condition, in that $\mathsf{A}_\theta(X, M(P_n))$ is mean-zero whenever X is independent of P_n, so it can be large only if X belongs to P_n. While the identifiability condition is somewhat messy to write, the soundness condition is quite straightforward:

Lemma 11.5.1. *Let P_n be the empirical distribution of an i.i.d. sample $X_i \stackrel{\text{iid}}{\sim} P_\theta$ and $X' \sim P_\theta$ be independent of P_n. Then*

$$\mathbb{E}[\mathsf{A}_\theta(X', M(P_n))] = 0 \quad \textit{and} \quad \mathbb{E}\left[\left|\mathsf{A}_\theta(X', M(P_n))\right|\right] \le \sqrt{\mathbb{E}\left[\|M(P_n) - \theta\|_2^2\right]}\sqrt{\|J(\theta)\|_{\rm op}}.$$

Deferring the proof of Lemma 11.5.1 temporarily, we show how to leverage it to prove lower bounds. Assume we have the identifiability condition analogizing the correlation identity (11.2.1). Differential privacy guarantees that $\mathsf{A}_\theta(X_i, M(P_n))$ and $\mathsf{A}_\theta(X_i', M(P_n))$ should have similar distributions, where X_i' is an independent observation. Then for a correlation measure Cor we obtain

$$\begin{aligned}\mathsf{Cor} \stackrel{(\mathsf{id})}{\le} \sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] &\stackrel{(\mathsf{priv})}{\le} O(\varepsilon)\sum_{i=1}^n \mathbb{E}\left[\left|\mathsf{A}_\theta(X_i', M(P_n))\right|\right] \\ &\le O(\varepsilon n)\, \|J(\theta)\|_{\rm op}^{1/2}\, \mathbb{E}[\|M(P_n) - \theta\|_2^2]^{1/2},\end{aligned} \tag{11.5.2}$$

where inequality (id) relies on identifiability—that $\dot\ell_\theta(X_i)$ and $M(P_n)$ correlate—and inequality (priv) heuristically uses differential privacy to swap X_i and X_i'. The final inequality applies Lemma 11.5.1. Then we would have

$$\mathbb{E}\left[\|M(P_n) - \theta\|_2^2\right] \gtrsim \frac{\mathsf{Cor}^2}{\|J(\theta)\|_{\rm op}} \cdot \frac{1}{n^2\varepsilon^2},$$

where Cor stands in for the correlation (identifiability) $\mathsf{A}_\theta(X_i, M(P_n))$ captures.

Once we set terms appropriately, this recipe provides explicit lower bounds. The next theorem instantiates this, which we preface by assuming we have what we term a *regular private problem* in a neighborhood of size $a > 0$ around θ_0:

(i) the mechanism M is (ε, δ)-differentially private

(ii) the parameter $\delta \geq 0$ is small enough that the score $\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x)$ satisfies

$$16\delta \sup_x \left\|\dot{\ell}_\theta(x)\right\|_2^2 \leq (e^\varepsilon - 1)^2 \left\|J(\theta)\right\|_{\rm op} < \infty,$$

where $J(\theta) = \mathrm{Cov}(\dot{\ell}_\theta(X))$ is the Fisher information of θ, for all $\left\|\theta - \theta_0\right\|_\infty \leq a$.

Condition (ii) implies that $\dot{\ell}_\theta$ is bounded, but most private estimators require this regardless.

Theorem 11.5.2. *Let $\theta_0 \in \mathbb{R}^d$ and assume we have a regular private problem in a neighborhood of size $a > 0$ around θ_0, that is, conditions (i) and (ii). Then there exists a prior density π supported on the box $\theta_0 + [-a, a]^d$ for which the following holds: let $T \sim \pi$, and conditional on $T = \theta$, sample X_1^n i.i.d. from P_θ and let P_n be the associated empirical distribution. Then*

$$Cd \leq \left(2n(e^\varepsilon - 1)\mathbb{E}_\pi\left[\left\|J(T)\right\|_{\rm op}\right]^{1/2} + c\sqrt{d}/a\right)\mathbb{E}_\pi\left[\mathbb{E}\left[\left\|M(P_n) - T\right\|_2^2 \mid T\right]\right]^{1/2},$$

where $C > 0$ and $c < \infty$ are numerical constants independent of all problem parameters.

We prove the theorem in three steps over the remainder of the section, following the plan the inequalities (11.5.2) lay out. Section 11.5.2* demonstrates the bound (id), while Section 11.5.3* demonstrates how we obtain a bound instantiating the inequality (priv). Section 11.5.4* finalizes the proof by putting the pieces together.

A few corollaries make the theorem statement cleaner. First, we assume $\varepsilon = O(1)$, so that $e^\varepsilon - 1 \lesssim \varepsilon$, and take the neighborhood size a to scale as $a = a(n, \varepsilon, \theta) \asymp \sqrt{d}/(\left\|J(\theta)\right\|_{\rm op}^{1/2} n\varepsilon)$, where we assume the Fisher information is continuous near θ. Then Theorem 11.5.2 implies that for an appropriate prior supported on $\theta + [-a, a]^d$, the average error of $M(P_n)$ satisfies

$$\mathbb{E}_\pi\left[\mathbb{E}\left[\left\|M(P_n) - T\right\|_2^2 \mid T\right]\right] \gtrsim \frac{1}{\left\|J(\theta)\right\|_{\rm op}} \cdot \frac{d^2}{n^2\varepsilon^2}. \tag{11.5.3}$$

Comparing this to the non-private case, recall the Van Trees inequality in Theorem 11.2.3, which gives $\mathbb{E}_\pi[\left\|M(P_n) - T\right\|_2^2] \geq (1 - o(1))\frac{d^2}{n\,\mathrm{tr}(J(\theta))}$ for priors supported on slightly larger than $n^{-1/2}$-sized neighborhoods of θ. For well-conditioned Fisher information $J(\theta)$, we have $\mathrm{tr}(J(\theta)) \approx d\left\|J(\theta)\right\|_{\rm op}$, in which case the non-private lower bound yields $\mathbb{E}_\pi[\left\|M(P_n) - T\right\|_2^2] \gtrsim \frac{d}{n\left\|J(\theta)\right\|_{\rm op}}$. Combining this with inequality (11.5.3), we obtain

$$\mathbb{E}_\pi\left[\left\|M(P_n) - T\right\|_2^2\right] \gtrsim \frac{d}{n\left\|J(\theta)\right\|_{\rm op}}\left(1 + \frac{d}{n\varepsilon^2}\right),$$

So (ε, δ)-differential privacy imposes a $\frac{d}{n\varepsilon^2}$ penalty over the non-private convergence rate.

Example 11.5.3 (d-dimensional means)**:** Let $\theta \in [-1,1]^d$ parameterize signs $X \in \{-1,1\}^d$ via

$$P_\theta(X = x) = \prod_{j=1}^d \frac{1 + x_j\theta_j}{2}.$$

Then $\mathbb{E}_\theta[X] = \theta$, and a quick calculation yields $\dot\ell_\theta(x) = [\frac{x_j - \theta_j}{1-\theta_j^2}]_{j=1}^d$, and $J(\theta) = (I - \mathrm{diag}(\theta\theta^\top))^{-1}$ is diagonal with entries $1/(1-\theta_j^2)$. At $\theta = 0$, $\|J(\theta)\|_{\mathrm{op}} = 1$, so

$$\sup_{\theta\in[-1,1]^d} \mathbb{E}_\theta\left[\|M(P_n) - \theta\|_2^2\right] \gtrsim \frac{d}{n}\left(1 + \frac{d}{n\varepsilon^2}\right)$$

whenever M is (ε,δ)-differentially private and $\delta \ll \frac{1}{n}$.

The Gaussian mechanism (Example 8.2.3 from Section 8.2.1) applies in this case: the empirical mean $\theta(P_n) = \mathbb{E}_{P_n}[X]$ has global ℓ_2-sensitivity $\mathrm{GS}_2(\theta) \le \frac{2\sqrt{d}}{n}$. So the mechanism

$$M(P_n) := \mathbb{E}_{P_n}[X] + O(1)\frac{\sqrt{d}}{n}\frac{\log^{\frac{1}{2}}\frac{1}{\delta}}{\varepsilon}\mathsf{N}(0, I_d)$$

is (ε,δ)-differentially private when $\varepsilon \le 1$, and it evidently satisfies

$$\mathbb{E}\left[\|M(P_n) - \theta\|_2^2\right] \le \mathbb{E}\left[\left\|\overline{X}_n - \theta\right\|_2^2\right] + O(1)\frac{d^2\log\frac{1}{\delta}}{n^2\varepsilon^2} = O(1)\frac{d}{n}\left(1 + \frac{d\log\frac{1}{\delta}}{n\varepsilon^2}\right).$$

Excepting the logarithmic factor in $\frac{1}{\delta}$, we have an order-optimal lower bound. $\diamond$

We finally return to prove Lemma 11.5.1; the remaining subsections prove the components of Theorem 11.2.3.

Proof of Lemma 11.5.1 We have

$$\mathbb{E}_\theta\left[\mathsf{A}_\theta(X', M(P_n))\right] = \mathbb{E}_\theta\left[\langle M(P_n) - \theta, \mathbb{E}_\theta[\dot\ell_\theta(X')]\rangle\right] = 0$$

by assumption that the scores are mean zero. For the upper bound, we apply Cauchy-Schwartz:

$$\begin{aligned}
\mathbb{E}\left[\left|\mathsf{A}_\theta(X', M(P_n))\right|\right]^2 &\le \mathbb{E}\left[\langle\dot\ell_\theta(X'), M(P_n) - \theta\rangle^2\right] \\
&= \mathbb{E}\left[(M(P_n) - \theta)^\top\mathbb{E}[\dot\ell_\theta(X)\dot\ell_\theta(X)^\top](M(P_n) - \theta)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(J(\theta)(M(P_n) - \theta)(M(P_n) - \theta)^\top\right)\right]
\end{aligned}$$

because X' and P_n are independent. Note that $u^\top Au \le \|u\|_2^2\|A\|_{\mathrm{op}}$ for any matrix A. $\square$

11.5.2* The identifiability inequality

We first provide a quantity we can use to control the putative correlation (id) in the score attack lower bound (11.5.2), analogizing the non-private correlation identity (11.2.1).

Lemma 11.5.4. *Assume that for each $j \in [d]$, the score $\dot{\ell}_\theta(x)$ is continuous in θ and dominated by a function $g : \mathcal{X} \to \mathbb{R}$, i.e., $\|\dot{\ell}_\theta(x)\|_2 \le g(x)$, with $\mathbb{E}[g(X)^2] < \infty$, and $\mathbb{E}[\|M(P_n)\|_2^2] < \infty$. Then*

$$\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] = \sum_{j=1}^d \frac{\partial}{\partial \theta_j} \mathbb{E}_\theta[M_j(P_n)].$$

Proof Denote the coordinates of $M(P_n)$ by $M_j(P_n)$, let $\dot{\ell}_{\theta,j}$ be the jth component of $\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x)$, and let $\partial_j = \frac{\partial}{\partial \theta_j}$ for shorthand. Then

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] &= \sum_{j=1}^d \sum_{i=1}^n \mathbb{E}\left[\dot{\ell}_{\theta,j}(X_i)(M_j(P_n) - \theta_j)\right] \\ &= \sum_{j=1}^d \mathbb{E}\left[\dot{\ell}_{\theta,j}(P_n)(M_j(P_n) - \theta_j)\right],\end{aligned}$$

where $\dot{\ell}_{\theta,j}(P_n) = \sum_{i=1}^n \partial_j \log p_\theta(X_i)$ is the derivative of the log-likelihood of the sample P_n. By assumption on the Fisher information and scores for the family $\{P_\theta\}$, $\mathbb{E}[\dot{\ell}_{\theta,j}(P_n)] = 0$, and so

$$\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] = \sum_{j=1}^d \mathbb{E}\left[\partial_j \log p_\theta^n(X_1^n) M_j(P_n)\right].$$

Computing a single j term by integrating, abusing notation to let $M(x_1^n)$ be the application of M to the sample supported on x_1^n, and recalling that p_θ denotes the density of P_θ with respect to a base measure μ, we see that

$$\begin{aligned}\mathbb{E}\left[\partial_j \log p_\theta^n(X_1^n) M_j(P_n)\right] &= \int \frac{\partial_j p_\theta^n(x_1^n)}{p_\theta^n(x_1^n)} M_j(x_1^n) p_\theta^n(x_1^n) d\mu^n(x_1^n) \\ &= \int \partial_j p_\theta^n(x_1^n) M_j(x_1^n) d\mu^n(x_1^n) \\ &= \partial_j \int p_\theta^n(x_1^n) M_j(x_1^n) d\mu^n(x_1^n).\end{aligned}$$

We may justify the exchange of integration and differentiation by the regularity assumptions on the densities p_θ and that $\mathbb{E}[\|M(P_n) - \theta\|_2^2] < \infty$. □

In the lower bound (11.5.2), Lemma 11.5.4 suggests we ought to see $\mathsf{Cor} \gtrsim d$. To see this, recognize that an accurate estimator $M(P_n)$ of θ should satisfy $\mathbb{E}[M(P_n)] \approx \theta$, and obviously $\frac{\partial}{\partial \theta_j}\theta_j = 1$. To make this rigorous, we proceed as in the development of the Van Trees family of inequalities in Section 11.2, where integrating guarantees that $\int \frac{\partial}{\partial \theta_j} \mathbb{E}_\theta[M_j(P_n)] \pi_j(\theta_j) d\theta_j \gtrsim 1$ for an accurate estimator, that is, the hoped for inequality that $\frac{\partial}{\partial \theta_j} \mathbb{E}_\theta[M_j(P_n)] \gtrsim 1$ holds on average.

The key step is integration by parts (sometimes this result is called *Stein's lemma*, for its specific statement in the case of Gaussian random variables that $\mathbb{E}[g'(Z)] = \mathbb{E}[Zg(Z)]$ for $Z \sim \mathsf{N}(0,1)$):

Lemma 11.5.5. *Let π be a differentiable density on $[a,b]$ with $\pi(a) = \pi(b) = 0$ and score $\dot{\ell}_\pi(t) = \frac{\pi'(t)}{\pi(t)}$. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and satisfy $\mathbb{E}_\pi[|g'(T)|] < \infty$ and $\mathbb{E}_\pi[|g(T)\dot{\ell}_\pi(T)|] < \infty$. Then*

$$\mathbb{E}_\pi\left[g'(T)\right] = -\mathbb{E}_\pi\left[g(T)\dot{\ell}_\pi(T)\right].$$

Proof We perform integration by parts, using $u = \pi$ and $dv = g'$, so that

$$\int_a^b g'(t)\pi(t)dt = g(t)\pi(t)|_a^b - \int_a^b g(t)\pi'(t)dt = -\int_a^b g(t)\frac{\pi'(t)}{\pi(t)}\pi(t)dt,$$

because $\pi(a) = \pi(b) = 0$. □

We apply Lemma 11.5.5 to functions $g_j(t) := \mathbb{E}_{\theta+te_j}[M_j(P_n)]$, using the notation that $T \in \mathbb{R}^d$ is a random variable taking on the values of θ, so $\mathbb{E}[M(P_n) \mid T = \theta] = \mathbb{E}_\theta[M(P_n)]$. We take π to be a product density, that is, $\pi(\theta) = \prod_{j=1}^d \pi_j(\theta_j)$, where each coordinate prior density π_j on θ_j satisfies the conditions of Lemma 11.5.5. Define the prior scores

$$\dot\ell_{\pi,j}(\theta) = \frac{\partial}{\partial\theta_j}\log\pi(\theta) = \frac{\partial}{\partial\theta_j}\log\pi_j(\theta_j) \quad \text{and} \quad \dot\ell_\pi(\theta) = \nabla_\theta \log\pi(\theta) = \left[\dot\ell_{\pi,j}(\theta)\right]_{j=1}^d,$$

using that π is a product density. Then iterating integrals, we obtain

$$\begin{aligned}
\mathbb{E}_\pi\left[\partial_j\mathbb{E}[M(P_n) \mid T]\right] &= \int\left(\int \partial_j\mathbb{E}[M_j(P_n) \mid T = \theta]\pi_j(\theta_j)d\theta_j\right)\pi_{\setminus j}(\theta_{\setminus j})d\theta_{\setminus j} \\
&= \int\left(-\int \mathbb{E}[M_j(P_n) \mid T = \theta]\dot\ell_{\pi,j}(\theta)\pi_j(\theta_j)d\theta_j\right)\pi_{\setminus j}(\theta_{\setminus j})d\theta_{\setminus j} \\
&= -\mathbb{E}_\pi\left[\mathbb{E}\left[M_j(P_n) \mid T\right]\dot\ell_{\pi,j}(T)\right].
\end{aligned}$$

Adding and subtracting T_j gives

$$\mathbb{E}_\pi\left[\partial_j\mathbb{E}[M(P_n) \mid T]\right] = -\mathbb{E}_\pi[T_j\dot\ell_{\pi,j}(T)] + \mathbb{E}_\pi\left[(T_j - \mathbb{E}[M_j(P_n) \mid T])\dot\ell_{\pi,j}(T)\right],$$

and substituting into Lemma 11.5.4 yields

$$\begin{aligned}
\mathbb{E}_\pi\left[\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_T(X_i, M(P_n)) \mid T\right]\right] &= -\sum_{j=1}^d \mathbb{E}_\pi[T_j\dot\ell_{\pi,j}(T)] + \mathbb{E}_\pi\left[\left\langle T - \mathbb{E}[M(P_n) \mid T], \dot\ell_\pi(T)\right\rangle\right] \\
&= -\mathbb{E}_\pi\left[\left\langle T, \dot\ell_\pi(T)\right\rangle\right] + \mathbb{E}_\pi\left[\left\langle T - \mathbb{E}[M(P_n) \mid T], \dot\ell_\pi(T)\right\rangle\right].
\end{aligned}$$

Cauchy-Schwarz yields $|\mathbb{E}_\pi[\langle T - \mathbb{E}[M(P_n) \mid T], \dot\ell_\pi(T)\rangle]| \le \mathbb{E}_\pi[\|T - \mathbb{E}[M(P_n) \mid T]\|_2^2]^{1/2}\operatorname{tr}(J(\pi))^{1/2}$.

Summarizing everything to this point, we state the following bound to address the identification inequality (id) in the plan (11.5.2).

Lemma 11.5.6. *Let π be a product density, $\pi(\theta) = \prod_{j=1}^d \pi_j(\theta_j)$, where each coordinate density π_j is continuously differentiable and has support $[a_j, b_j]$ with $\pi_j(a_j) = \pi_j(b_j) = 0$. Define the prior score $\dot\ell_\pi(\theta) = \nabla_\theta\log\pi(\theta)$, and let $J(\pi) = \mathbb{E}_\pi[\dot\ell_\pi(T)\dot\ell_\pi(T)^\top]$ be the prior information. Then*

$$\mathbb{E}_\pi\left[\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_T(X_i, M(P_n)) \mid T\right]\right] \ge -\mathbb{E}_\pi\left[\left\langle T, \dot\ell_\pi(T)\right\rangle\right] - \mathbb{E}_\pi\left[\|T - \mathbb{E}[M(P_n) \mid T]\|_2^2\right]^{1/2}\operatorname{tr}(J(\pi))^{1/2}.$$

11.5.3* The privacy inequality

We address the step (priv) in the bound (11.5.2) via the next two lemmas. In both lemmas, we tacitly assume the smoothness conditions on $\dot{\ell}_\theta$ throughout the section hold.

Lemma 11.5.7. *For any $K \ge 0$,*

$$\mathbb{E}[\mathsf{A}_\theta(X_i, M(P_n))] \le (e^\varepsilon - 1)\sqrt{\mathbb{E}[\|M(P_n) - \theta\|_2^2]}\sqrt{\|J(\theta)\|_{\rm op}} + 2\delta K \\ + \int_K^\infty \left(\mathbb{P}(\mathsf{A}_\theta(X_i, M(P_n)) \ge u) + \mathbb{P}(\mathsf{A}_\theta(X_i', M(P_n)) \ge u)\right) du.$$

Proof Let P_n and P_n' be samples differing on example i, so $\mathsf{A}_\theta(X_i', M(P_n)) \stackrel{\rm dist}{=} \mathsf{A}_\theta(X_i, M(P_n'))$ for $X_i, X_i' \stackrel{\rm iid}{\sim} P_\theta$. For shorthand, let $\mathsf{A} = \mathsf{A}_\theta(X_i, M(P_n))$ and $\mathsf{A}' = \mathsf{A}_\theta(X_i', M(P_n))$, and let $\mathsf{A}_+ = [\mathsf{A}]_+$ and $\mathsf{A}_- = [-\mathsf{A}]_+$ be the positive and negative parts of A (and similarly for A'). Then

$$\mathbb{E}[\mathsf{A}] = \mathbb{E}[\mathsf{A}_+ - \mathsf{A}_-] = \int_0^\infty \mathbb{P}(\mathsf{A} \ge u)du - \int_0^\infty \mathbb{P}(\mathsf{A} \le -u)du.$$

By differential privacy, $\mathbb{P}(\mathsf{A} \ge u) \le e^\varepsilon \mathbb{P}(\mathsf{A}' \ge u) + \delta$, and so for any $K > 0$,

$$\int_0^\infty \mathbb{P}(\mathsf{A} \ge u)du \le e^\varepsilon \int_0^K \mathbb{P}(\mathsf{A}' \ge u)du + \delta K + \int_K^\infty \mathbb{P}(\mathsf{A} \ge u)du \\ \le e^\varepsilon \mathbb{E}[\mathsf{A}_+'] + \delta K + \mathbb{E}[\mathsf{A}\mathbf{1}\left\{\mathsf{A} \ge K\right\}],$$

and similarly

$$\int_0^\infty \mathbb{P}(\mathsf{A} \le -u)du \ge e^{-\varepsilon} \int_0^K \mathbb{P}(\mathsf{A}' \le -u)du - \delta K + \int_K^\infty \mathbb{P}(\mathsf{A} \le -u)du \\ = e^{-\varepsilon}\mathbb{E}[\mathsf{A}_-'] - \delta K - e^{-\varepsilon}\mathbb{E}[\mathsf{A}_-'\mathbf{1}\left\{\mathsf{A}_-' \ge K\right\}] + \int_K^\infty \mathbb{P}(\mathsf{A} \le -u)du.$$

Combining these and using $e^{-\varepsilon} \le 1$, we have

$$\begin{aligned}\mathbb{E}[\mathsf{A}] &\le e^\varepsilon \mathbb{E}[\mathsf{A}_+'] - e^{-\varepsilon}\mathbb{E}[\mathsf{A}_-'] + 2\delta K + \mathbb{E}[|\mathsf{A}|\mathbf{1}\left\{\mathsf{A} \ge K\right\}] + \mathbb{E}[|\mathsf{A}'|\mathbf{1}\left\{\mathsf{A}' \le -K\right\}] \\ &= \mathbb{E}[\mathsf{A}'] + (e^\varepsilon - 1)\mathbb{E}[\mathsf{A}_+'] + (1 - e^{-\varepsilon})\mathbb{E}[\mathsf{A}_-'] + \mathbb{E}[|\mathsf{A}|\mathbf{1}\left\{\mathsf{A} \ge K\right\}] + \mathbb{E}[|\mathsf{A}'|\mathbf{1}\left\{\mathsf{A}' \le -K\right\}] + 2\delta K \\ &\le (e^\varepsilon - 1)\mathbb{E}[|\mathsf{A}'|] + 2\delta K + \int_K^\infty \left(\mathbb{P}(\mathsf{A} \ge u) + \mathbb{P}(\mathsf{A}' \le -u)\right) du,\end{aligned}$$

where we used Lemma 11.5.1 to substitute $\mathbb{E}[\mathsf{A}'] = 0$. □

Lemma 11.5.8. *Let $0 < \gamma < 1$ and let $\dot{\ell}_\theta$ have $2\frac{1+\gamma}{1-\gamma}$ moment $\mathbb{E}[\|\dot{\ell}_\theta(X)\|_2^{2(1+\gamma)/(1-\gamma)}] \le s(\gamma)$. Then*

$$\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] \le n \cdot \mathbb{E}\left[\|M(P_n) - \theta\|_2^2\right]^{1/2} \cdot \left((e^\varepsilon - 1)\left\|J(\theta)\right\|_{\rm op}^{1/2} + \left(2 + \frac{2}{\gamma}\right)\delta^{\frac{\gamma}{1+\gamma}} s(\gamma)^{\frac{1-\gamma}{2(1+\gamma)}}\right).$$

Proof Because $\frac{1+\gamma}{2} + \frac{1-\gamma}{2} = 1$, Hölder's inequality implies that

$$\begin{aligned}\mathbb{E}\left[|\mathsf{A}_\theta(X, M(P_n))|^{1+\gamma}\right] &\le \mathbb{E}\left[\left\|\dot\ell_\theta(X)\right\|_2^{1+\gamma} \left\|M(P_n) - \theta\right\|_2^{1+\gamma}\right] \\ &\le \mathbb{E}\left[\left\|\dot\ell_\theta(X)\right\|_2^{2\frac{1+\gamma}{1-\gamma}}\right]^{\frac{1-\gamma}{2}} \mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{\frac{1+\gamma}{2}}.\end{aligned}$$

For $u \ge 0$,

$$\mathbb{P}(\mathsf{A}_\theta(X, M(P_n)) \ge u) \le \frac{s(\gamma)^{\frac{1-\gamma}{2}} \mathbb{E}[\|M(P_n) - \theta\|_2^2]^{\frac{1+\gamma}{2}}}{u^{1+\gamma}},$$

so Lemma 11.5.7 implies

$$\mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] \le (e^\varepsilon - 1)\sqrt{\mathbb{E}[\|M(P_n) - \theta\|_2^2] \left\|J(\theta)\right\|_{\rm op}} + 2\delta K + \frac{2}{\gamma} K^{-\gamma} s(\gamma)^{\frac{1-\gamma}{2}} \mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{\frac{1+\gamma}{2}}.$$

Choosing K to minimize this bound by solving

$$\delta = K^{-(1+\gamma)} s(\gamma)^{\frac{1-\gamma}{2}} \mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{\frac{1+\gamma}{2}}, \quad \text{i.e.} \quad K = \delta^{-\frac{1}{1+\gamma}} s(\gamma)^{\frac{1-\gamma}{2(1+\gamma)}} \mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{1/2},$$

then substituting and summing over n gives the lemma. □

11.5.4* Putting the bounds together and finalizing Theorem 11.5.2

To prove the lower bound in Theorem 11.5.2, we combine Lemmas 11.5.6 and 11.5.8 and make judicious choices for their parameters. First, recognize that under the assumptions of the theorem (that $\dot\ell_\theta$ is bounded), we may take the limit $\gamma \uparrow 1$ in Lemma 11.5.8, as $s(\gamma)^{\frac{1-\gamma}{2(1+\gamma)}} \to \sup_x \|\dot\ell_\theta(x)\|_2$. Then the lemma implies that for any θ,

$$\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] \le n\mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{1/2} \left((e^\varepsilon - 1)\left\|J(\theta)\right\|_{\rm op}^{1/2} + 4\sqrt{\delta} \sup_x \|\dot\ell_\theta(x)\|_2\right).$$

By the assumption that δ is small enough that $16\delta \sup_x \|\dot\ell_\theta(x)\|_2^2 \le (e^\varepsilon - 1)^2 \left\|J(\theta)\right\|_{\rm op}$, we see that

$$\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_\theta(X_i, M(P_n))\right] \le 2n(e^\varepsilon - 1)\left\|J(\theta)\right\|_{\rm op}^{1/2} \mathbb{E}\left[\left\|M(P_n) - \theta\right\|_2^2\right]^{1/2}. \tag{11.5.4}$$

We choose an appropriate prior π in Lemma 11.5.6 so that $-\mathbb{E}[\langle T, \dot\ell_\pi(T)\rangle] \gtrsim d$ while $\mathrm{tr}(J(\pi))$ remains reasonable. This is fairly straightforward: take the prior $\pi = \prod_{j=1}^d \pi_j$ to be the product of identical coordinate densities with locations shifted to be around $\theta \in \mathbb{R}^d$. Nearly any symmetric, compactly supported, $\mathcal{C}^\infty$ unimodal density f will serve here; as a particular example, take $f(t) = C\exp(-1/\left[1 - t^2\right]_+)$, where C is a numerical constant chosen so that $\int_{-1}^1 f(t)dt = 1$. Then

$$f'(t) = -2C\frac{t}{(1-t^2)^2} \exp\left(-\frac{1}{1-t^2}\right) \mathbf{1}\left\{|t| \le 1\right\}$$

is odd, $\mathcal{C}^\infty$, compactly supported, satisfies $\int_{-1}^1 f'(t)dt = 0$, and

$$-\int_{-1}^1 tf'(t)dt = 2C\int_{-1}^1 \frac{t^2}{(1-t^2)^2}\exp\left(-\frac{1}{1-t^2}\right)dt \gtrsim 1.$$

For any $a > 0$, to obtain a density supported on $[-a, a]$, we set $f_a(u) = \frac{1}{a}f(u/a)$, which satisfies $f_a'(u) = \frac{1}{a^2}f'(u/a)$. Then

$$J(f_a) := \int_{-a}^a \frac{f_a'(u)^2}{f_a(u)}du = \frac{1}{a^3}\int_{-a}^a \frac{f'(u/a)^2}{f(u/a)}du = \frac{1}{a^2}\int_{-1}^1 \frac{f'(t)^2}{f(t)}dt = \frac{1}{a^2}J(f),$$

where $J(f) = \int f'(t)^2/f(t)dt = O(1)$ denotes the Fisher-information of the one-dimensional density f. Similarly, for any fixed value t_0,

$$-\int_{-a}^a (u+t_0)f_a'(u)du = -\int_{-a}^a uf_a'(u)du = -\frac{1}{a^2}\int_{-a}^a uf'(u/a)du = -\int_{-1}^1 tf'(t)dt \gtrsim 1.$$

Setting the priors $\pi_j(t) = f_a(t - \theta_j)$, where $\theta \in \mathbb{R}^d$ is any desired point, we obtain prior Fisher information $J(\pi) = \frac{1}{a^2}J(f)I_d$, where $J(f) = O(1)$ and I_d is the $d \times d$ identity matrix. We may therefore summarize the consequences of Lemma 11.5.6 we have derived via the following result.

Lemma 11.5.9. *Let $a > 0$ and $\theta \in \mathbb{R}^d$. Then there is a coordinate independent prior $\pi = \prod_{j=1}^d \pi_j$ supported on $\theta + [-a, a]^d$, where each density π_j is $\mathcal{C}^\infty$, and*

$$\mathbb{E}_\pi\left[\sum_{i=1}^n \mathbb{E}\left[\mathsf{A}_T(X_i, M(P_n)) \mid T\right]\right] \geq C \cdot d - c\sqrt{da^{-2}\mathbb{E}_\pi\left[\mathbb{E}\left[\|M(P_n) - T\|_2^2 \mid T\right]\right]}$$

where $C > 0$ and $c < \infty$ are numerical constants independent of all problem parameters.

This lemma provides the formal statement of inequality (id) in the plan (11.5.2), while inequality (11.5.4) provides the formal statement of inequality (priv). Summarizing, we have shown the existence of a prior π on the variable T supported on the box $\theta + [-a, a]^d$ for which

$$C \cdot d \leq \left(2n(e^\varepsilon - 1)\mathbb{E}_\pi\left[\|J(T)\|_{\rm op}\right]^{1/2} + c\sqrt{d}/a\right)\mathbb{E}_\pi\left[\mathbb{E}\left[\|M(P_n) - T\|_2^2 \mid T\right]\right]^{1/2},$$

where we integrated over $T \sim \pi$ and applied Cauchy-Schwarz in inequality (11.5.4). This implies Theorem 11.5.2.

11.6 Bibliography and further reading

The material in this section ranges from the classical to the modern, as even classical lower bound techniques reappear in new guises and applications. The idea that the Fisher information (11.0.1) should determine whether an estimator is statistically efficient dates back to the earliest days of statistics, with Fisher's incorrect attempts to demonstrate the optimality of maximum likelihood estimation [187]. Cramér and Rao [119, 355] developed their eponymous bound on the variance of unbiased estimators in the 1940s, but Stein [382] demonstrated convincingly that statistical efficiency could not rely on pointwise notions of the Fisher information alone, as the sample mean (the maximum likelihood estimate of a normal location) is inadmissible in $d \geq 3$ dimensions. James

and Stein [253] followed this with the derivation of their estimator, which exhibits strictly smaller squared error (11.1.1); see also Efron [177] and Stigler [387] for elegant treatments.

Le Cam [279, 280, 281] and Hájek [226] resolved many of the issues through their work on local asymptotic normality. Essentially, Le Cam shows that if one draws $H \sim \mathsf{Uniform}(c_n \mathbb{B}_2)$ for $1 \ll c_n \ll \sqrt{n}$, and conditional on $H = h$ samples $X_1^n \stackrel{\rm iid}{\sim} P_{\theta_0 + h/\sqrt{n}}$, the posterior on H satisfies

$$\mathbb{E}\left[\left\|\pi(H \in \cdot \mid X_1^n) - \mathsf{N}(\mu(X_1^n), J(\theta_0)^{-1})\right\|_{\rm TV}\right] \to 0$$

where $\mu(X_1^n)$ is a "posterior" mean. That is, the posterior is nearly Gaussian with covariance $J(\theta_0)^{-1}$. As a consequence, for any bounded symmetric quasi-convex function φ and sequence of estimators $\widehat{\theta}_n$, if π_n is uniform on $T \in \theta_0 + \frac{c_n}{\sqrt{n}} \mathbb{B}_2$ and we draw $X_1^n \stackrel{\rm iid}{\sim} P_T$, then

$$\begin{aligned} \mathbb{E}_{\pi_n}\left[\mathbb{E}\left[\varphi(\sqrt{n}(\widehat{\theta}_n(X_1^n) - T)) \mid T\right]\right] &\geq \mathbb{E}\left[\inf_h \varphi\left(h - \mathsf{N}(\mu(X_1^n), J(\theta_0)^{-1})\right)\right] - o(1) \\ &= \mathbb{E}\left[\varphi(\mathsf{N}(0, J(\theta_0)^{-1}))\right] - o(1). \end{aligned} \tag{11.6.1}$$

The final inequality follows from Anderson's lemma [17], which guarantees that μ best estimates the random vector $Z \sim \mathsf{N}(\mu, \Sigma)$. So by integrating over neighborhoods of θ_0 of size ever so slightly larger than $1/\sqrt{n}$, the inverse Fisher information indeed becomes fundamental. The books [407, 408] treat this material, and van der Vaart [407, Ch. 25] inspires the tilting approach we take in Section 11.3; see also Duchi and Ruan [155]. The Bayesian perspective we adopt in the Van Trees inequalities follows Gill and Levit [210] quite closely, though Van Trees [410] proved the initial results.

Super-efficiency (Section 11.4) motivated much of the study of optimal estimation, including Le Cam and Hájek's work; see van der Vaart [406]. The framing in terms of instance optimality (OPT.i)–(OPT.iii) we borrow from Cai and Low [96] (see also Chapter 13). Many of these one can remember in the shorthand that the inequality (11.6.1) holds nearly pointwise. (If $\widehat{\theta}_n$ must be computable, the set at which strict inequality may hold is countable [416].)

Section 11.5, which uses the techinques of correlating the score $\dot{\ell}_\theta$ with performant estimators that the Van Trees inequality prefigures, provides a glimpse into a larger research agenda of demonstrating optimality of private estimators. In the setting of pure $(\varepsilon, 0)$-differential privacy, Hardt and Talwar [227] provide lower bounds on estimating certain sample-based linear functions. Barber and Duchi [43] extend these arguments to the estimation of population-level quantities. In the pure privacy case, one typically has the lower bound

$$\sup_{\theta \text{ near } \theta_0} \mathbb{E}_\theta\left[\|M(P_n) - \theta\|_2^2\right] \gtrsim \frac{1}{\|J(\theta_0)\|_{\rm op}} \cdot \frac{d^3}{n^2 \varepsilon^2},$$

which suffers an additional dimension (d)-dependent multiplier over the lower bounds (11.5.3).

The *score attack* (11.5.1) that forms the basis for the privacy lower bounds descends from *fingerprinting codes* in security. Bun, Steinke, Ullman, and Vadhan [94, 384, 385] pioneered their use in proving lower bounds for differential privacy and adaptive data analysis (recall Section 6.3). Many researchers have adapted their use to provide optimality results both in privacy [175, 262, 263] and for adaptive data analysis [347, 297]. The use of Stein-type identities—integration by parts—is implicit in some of this work [175], and Portella and Harvey [347] extend these ideas. Cai et al. [98, 99] make the connection between fingerprinting codes and Stein identities in statistical lower bounds in differential privacy explicit, providing an elegant recipe for proving lower bounds; our development in Section 11.5 closely follows [99]. In some cases, these approaches can give sharper

results that also incorporate $\log \frac{1}{\delta}$ penalties in (ε, δ)-privacy [385, 98]. Leveraging these score attacks has proven fruitful beyond private estimators and adaptive analysis; for example, Attias et al. [33] show how to use similar approaches to quantify that (convex) M-estimators *must* memorize some of the data used to fit them for optimal performance.

11.7 Exercises

Exercise 11.1: Let Θ be a compact convex set with non-empty interior, and let $\{P_\theta\}_{\theta \in \Theta}$ be a family of mutually absolutely continuous distributions.

(a) Show that if $\widehat{\theta}$ is an unbiased estimator of θ, then for each $\theta \in \Theta$, $P_\theta(\widehat{\theta}(X) \notin \Theta) > 0$. *Hint.* Extend Example 11.1.2 to show that if $\widehat{\theta}$ is unbiased and $\widehat{\theta} \in \Theta$, then Θ cannot have an interior. It may be useful to leverage supporting hyperplanes of convex bodies (see Appendix B.1.3).

(b) Show that if $P_\theta(\widehat{\theta}(X) \notin \Theta) > 0$, the projected estimator $\widetilde{\theta}(x) = \mathrm{Proj}_\Theta(\widehat{\theta}(x))$ satisfies

$$\mathbb{E}_\theta \left[\left\| \widetilde{\theta}(X) - \theta \right\|_2^2 \right] < \mathbb{E}_\theta \left[\left\| \widehat{\theta}(X) - \theta \right\|_2^2 \right].$$

Hint. Theorem B.1.17 in Appendix B.1.3 characterizes projection onto a convex set.

(c) Conclude that if Θ is a convex body and $\{P_\theta\}_{\theta \in \Theta}$ satisfies the assumed conditions of the exercise, then for *all* $\theta \in \Theta$, the projected estimator $\widetilde{\theta}$ satisfies

$$\mathbb{E}_\theta \left[\left\| \widetilde{\theta}(X) - \theta \right\|_2^2 \right] < \mathbb{E}_\theta \left[\left\| \widehat{\theta}(X) - \theta \right\|_2^2 \right].$$

Exercise 11.2: Let $Y = X\theta + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, I_n)$ and $X \in \mathbb{R}^{n \times d}$. For $\lambda > 0$ define $\widehat{\theta}_\lambda = \mathrm{argmin}_t \{ \|Xt - Y\|_2^2 + \lambda \|t\|_2^2 \}$. Define the risk $R(\lambda) := \mathbb{E}[\|\widehat{\theta}_\lambda - \theta\|_2^2]$ and $R(0) = \lim_{\lambda \downarrow 0} R(\lambda)$. Give

$$R'(0^+) := \lim_{\lambda \downarrow 0} \frac{R(\lambda) - R(0)}{\lambda},$$

show that $R'(0^+) < 0$, and conclude that for any θ, there exists $\lambda > 0$ such that $R(\lambda) < R(0)$.

Exercise 11.3 (Properties of compactly supported $\mathcal{C}^\infty$ functions)**:**

(a) Let $\phi(t) = \exp(-\frac{1}{1-t^2})$ for $|t| < 1$ and $\phi(t) = 0$ for $|t| \geq 1$. Show that ϕ is $\mathcal{C}^\infty$.

(b) Show that for ϕ as in part (a), $\int_{-1}^1 \phi'(t)^2 / \phi(t) dt < \infty$.

(c) Let ϕ be a symmetric, nonnegative, $\mathcal{C}^\infty$ function with support $[-1, 1]$ and $\int \phi(t) dt = 1$. For $0 < a \leq 1$ define $\pi_a(t) = \phi(t/a)/a$. Show that π_a is a density on $[-a, a]$ and give $J(\pi_a)$ as a function of $J(\pi_1)$.

Exercise 11.4: Prove Corollary 11.2.4.

Exercise 11.5: Demonstrate the claim of Example 11.2.5 that

$$t^\star := \mathop{\mathrm{argmin}}_t \inf_{\beta_0, \beta} \mathbb{E} \left[(Y - \beta_0 - tZ - \beta^\top X)^2 \right] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

Exercise 11.6: In conditions (i)–(v) for Theorem 11.2.7, replace all instances of "suitably continuous" with "continuously differentiable," assume that $J(\theta) = \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^\top]$ is continuous in θ, and assume instead of condition (v) that π is a product measure on $[a_1, b_1] \times \cdots \times [a_d, b_d]$, as in Theorem 11.2.6. Prove Theorem 11.2.7 under these conditions.

Exercise 11.7 (Proving Corollary 11.2.10)**:** Take $A(\theta) = J_\psi(\theta)^{-1}$ to be the inverse information for ψ and let $C(\theta) = J(\theta)^{-1}\dot{\psi}(\theta)^\top J_\psi(\theta) \in \mathbb{R}^{d\times p}$. Define the prior information $\widetilde{J}(\pi \mid \psi) = J(\pi \mid A, C)$ in the notation (11.2.2) for these A and C. Then use Theorem 11.2.7 to prove Corollary 11.2.10.

Exercise 11.8 (Estimation with nuisance parameters)**:** Consider probabilistic models $\{P_{\theta,\eta}\}$, $\theta \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^k$, with densities $p_{\theta,\eta}$ and (joint) Fisher information matrix

$$J(\theta, \eta) = \begin{bmatrix} J_{\theta,\theta} & J_{\theta,\eta} \\ J_{\theta,\eta}^\top & J_{\eta,\eta} \end{bmatrix},$$

where the notation indicates partitioning J to match parameters.

(a) Define the similarly partitioned block matrices

$$M := \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \quad \text{and} \quad M^{-1} = \begin{bmatrix} X & Y \\ Y^\top & Z \end{bmatrix}.$$

Assume that $M \succ 0$. Show that $X \preceq A^{-1}$, with equality if and only if $B = 0$.

(b) Show that $J_{\theta,\theta}^{-1} \preceq [J(\theta, \eta)^{-1}]_{\theta,\theta}$. When does equality hold?

(c) Use Corollary 11.2.10 to interpret the above inequality in the context of in problems in which the nuisance parameters η are known versus unknown.

Exercise 11.9: Consider M-estimation problems as in Corollary 11.3.4. Let $H = \nabla^2 L(\theta_0)$ be the Hessian of L at $\theta_0 = \operatorname{argmin}_\theta L(\theta)$, and let $\Sigma = \operatorname{Cov}(\nabla \ell(\theta_0, Z))$ be the covariance of the gradients. Give appropriate choices of the matrices $A(t)$ and tilting functions $g : \mathcal{Z} \to \mathbb{R}^d$ to show the following:

(a) For smooth enough priors π_n on T supported on $\frac{c}{\sqrt{n}}\mathbb{B}_2^d$, any estimator $\widehat{\theta}$ satisfies

$$\liminf_{c\to\infty} \liminf_n n \cdot \mathbb{E}_{\pi_n}\left[\left\|\widehat{\theta}(Z_1^n) - \theta(P_T)\right\|^2\right] \geq \operatorname{tr}(H^{-1}\Sigma H^{-1}).$$

(b) For smooth enough priors π_n on T supported on $\frac{c}{\sqrt{n}}\mathbb{B}_2^d$, any estimator $\widehat{\theta}$ satisfies

$$\liminf_c \liminf_n n \cdot \mathbb{E}_{\pi_n}\left[(\widehat{\theta}(Z_1^n) - \theta(P_T))^\top H\Sigma^{-1}H(\widehat{\theta}(Z_1^n) - \theta(P_T))^\top\right] \geq d.$$

What do these results say about the empirical risk minimizer (5.3.1)?

Exercise 11.10: Show that a stronger version of Theorem 11.5.2 is possible: under the conditions and notation of the theorem, for any invertible matrix B,

$$Cd \leq \left(2n(e^\varepsilon - 1)\mathbb{E}_\pi\left[\left\|B^{-1}J(T)B^{-1}\right\|_{\text{op}}\right]^{1/2} + c/a\right)\mathbb{E}_\pi\left[\mathbb{E}\left[\|B(M(P_n) - T)\|_2^2 \mid T\right]\right]^{1/2}.$$

Hint. For any vector u, symmetric matrix A, and invertible B, $|u^\top Au| \leq \|Bu\|_2^2\left\|B^{-1}AB^{-1}\right\|_{\text{op}}$. Replace each application of Cauchy-Schwarz in the proof of Theorem 11.5.2 with this.

Chapter 12

Beyond local minimax techniques: a miscellany

Most lower bound techniques we have studied to this point have been local, whether by looking at parameters $\theta \pm \delta v$ for some vectors $v \in \mathcal{V}$, or locally tilting the underlying distribution via density modifications $p_t(x) = (1 + tg(x))p_0(x)$. These approaches were successful in problems with sample size n tending to infinity with fixed dimension d, demonstrating that appropriately choosing the separation $\delta > 0$ yields constant probability of error in a related testing problem. Many problems instead let the dimension grow with n, are infinite-dimensional, or call for global rather than local guarantees. The same lower-bound tools we have developed also illuminate multiple testing and detection, where sharp phase transitions separate impossible and possible regimes. Finally, refined converses strengthen our constant-probability lower bounds by showing that too accurate estimation can itself be overwhelmingly unlikely. This chapter introduces some of the techniques to develop these extensions.

12.1 Nonparametric regression: minimax upper and lower bounds

To motivate global lower bound techniques and illustrate minimax optimality beyond finite-dimensional problems, we give a brief treatment of *nonparametric* problems. In non-parametric regression, we assume there is an unknown function $f : \mathbb{R} \to \mathbb{R}$, where f belongs to a pre-determined class of functions $\mathcal{F}$; usually some type of smoothness guarantee parameterizes this class. In the standard nonparametric regression problem, we obtain observations of the form

$$Y_i = f(X_i) + \varepsilon_i \tag{12.1.1}$$

where ε_i are independent, mean zero random variables with $\mathrm{Var}(\varepsilon_i) \le \sigma^2$. See Figure 12.1 for an example. For simplicity, we evenly distribute the X_i as $X_i = i/n \in [0, 1]$. Given n observations Y_i, we consider two questions: (1) how can we estimate f? and (2) what are the optimal rates at which it is possible to estimate f? We begin by answering (1), proceeding to (2) momentarily.

12.1.1 Kernel estimates of the function

The simplest strategy to estimate $f(x)$ in the model (12.1.1) involves locally averaging values of Y_i for X_i near x. To operationalize this, we place small "bumps" around the observed points, using

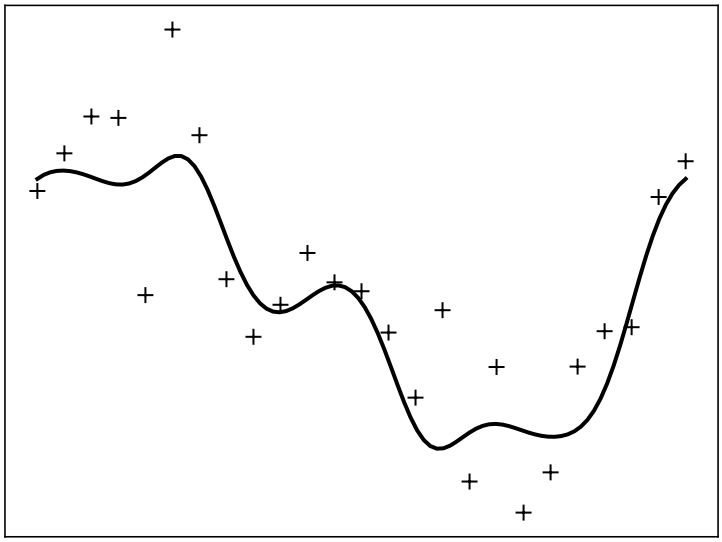


Figure 12.1: Observations in a non-parametric regression with $f(x) = \sin(2x + \cos^2(3x))$.

these to determine a weighted average. Let $K : \mathbb{R} \to \mathbb{R}_+$ be a *kernel function*, which is continuous, not identically zero, has support $\operatorname{supp} K = [-1, 1]$, and satisfies the technical condition

$$\lambda_0 \sup_x K(x) \le \inf_{|x| \le 1/2} K(x) \tag{12.1.2}$$

for some $\lambda_0 > 0$. A simple example is the "tent" function $K_{\text{tent}}(x) = [1 - |x|]_+$; we can also construct compactly supported $\mathcal{C}^\infty$ functions, such as $K(x) = \mathbf{1}\{|x| < 1\} \exp(-\frac{1}{1-x^2})$. See Figure 12.2 for these two examples.

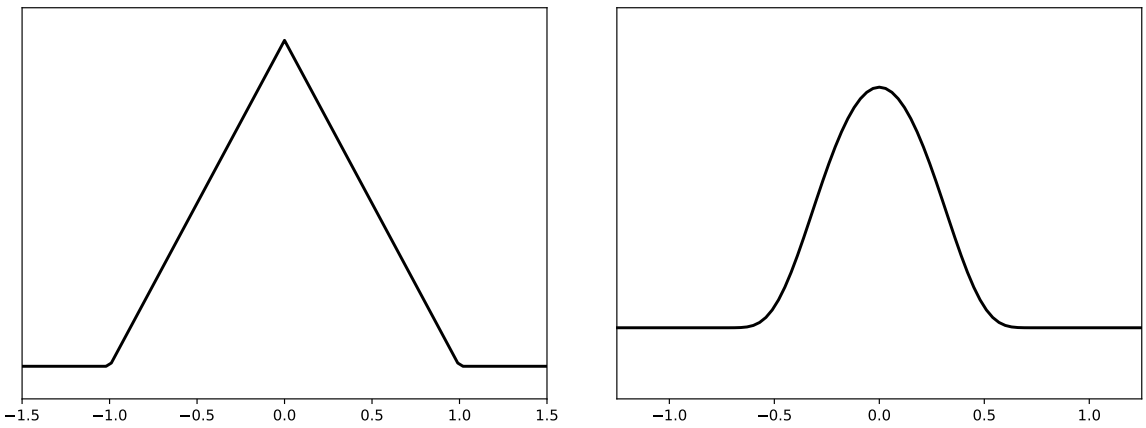


Figure 12.2: Left: "tent" kernel. Right: infinitely differentiable compactly supported kernel.

The *Nadaraya-Watson* smoothing estimator implements averaging using these kernels. Fix a

bandwidth h, which determines the width of the averaging region, and for each x define weights

$$W_{ni}(x) := \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)}.$$

Then we estimate

$$\widehat{f}_n(x) := \sum_{i=1}^n Y_i W_{ni}(x). \tag{12.1.3}$$

This regression gives points X_i near x higher weight than distant observations. The next proposition provides a guarantee on the bias and variance of the estimated $\widehat{f}_n$ at each point $x \in [0,1]$.

Proposition 12.1.1. *Let the observation model (12.1.1) hold, assume condition (12.1.2), and assume that f is L-Lipschitz. In addition assume the bandwidth $h \geq 2/n$ and that the X_i are evenly spaced on $[0,1]$. Then for each $x \in [0,1]$, the smoother (12.1.3) satisfies*

$$|\mathbb{E}[\widehat{f}_n(x)] - f(x)| \leq Lh \quad \text{and} \quad \mathrm{Var}(\widehat{f}_n(x)) \leq \frac{4\sigma^2}{\lambda_0 nh}.$$

Proof As the weights are nonnegative, sum to one, and vanish when $|X_i - x| > h$,

$$|\mathbb{E}[\widehat{f}_n(x)] - f(x)| \leq \sum_{i=1}^n |f(X_i) - f(x)| W_{ni}(x) \leq Lh.$$

Independence and the variance bound give

$$\mathrm{Var}(\widehat{f}_n(x)) = \sum_{i=1}^n W_{ni}(x)^2 \mathrm{Var}(\varepsilon_i) \leq \sigma^2 \max_i W_{ni}(x).$$

Every interval of length $h/2$ in $[0,1]$ contains at least $nh/4$ grid points when $h \geq 2/n$. Hence condition (12.1.2) yields

$$W_{ni}(x) \leq \frac{\sup_u K(u)}{\lambda_0 \sup_u K(u) |\{j : |X_j - x| \leq h/2\}|} \leq \frac{4}{\lambda_0 nh},$$

proving the claim. □

With the proposition in place, we can bound the worst-case pointwise mean squared error for estimation of a function $f \in \mathcal{F}$, the class of L-Lipschitz functions $[0,1] \to \mathbb{R}$.

Theorem 12.1.2. *Under the conditions of Proposition 12.1.1, choose $h = (\sigma^2/L^2\lambda_0)^{1/3} n^{-1/3}$. Then there exists a universal (numerical) constant $C < \infty$ such that for any $f \in \mathcal{F}$,*

$$\sup_{x \in [0,1]} \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \leq C \left(\frac{L\sigma^2}{\lambda_0}\right)^{2/3} n^{-\frac{2}{3}}.$$

Proof Using Proposition 12.1.1, we have for any $x \in [0,1]$ that

$$\mathbb{E}[(\widehat{f}_n(x) - f(x))^2] = \left(\mathbb{E}[\widehat{f}_n(x)] - f(x)\right)^2 + \mathbb{E}[(\widehat{f}_n(x) - \mathbb{E}[\widehat{f}_n(x)])^2] \leq \frac{4\sigma^2}{\lambda_0 nh} + L^2 h^2.$$

Choose h to balance this bias/variance decomposition. □

Integrating the result in Theorem 12.1.2 over the interval $[0,1]$, provides the following corollary.

Corollary 12.1.3. *Under the conditions of Theorem 12.1.2, the estimator (12.1.3) with kernel K_{tent} obtains*

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\|\widehat{f}_n - f\|_2^2] \le C \left(\frac{L\sigma^2}{n}\right)^{2/3}.$$

In Proposition 12.1.1, one can show that a more clever higher-order attains bias $\mathbb{E}[\widehat{f}_n(x)] - f(x) = O(h^\beta)$ if f has Lipschitz $(\beta - 1)$th derivative. In this case, the rate improves to

$$\sup_x \mathbb{E}[(\widehat{f}_n(x) - f(x))^2] \le C n^{-\frac{2\beta}{2\beta+1}}, \tag{12.1.4}$$

and every additional degree of smoothness gives a corresponding improvement in convergence rate. While deriving the convergence rates (12.1.4) is beyond our scope, the bibliographic section includes pointers to further reading. Such *nonparametric rates* of convergence, slower than n^{-1}, typify nonparametric problems; we must balance an effective dimension with sample size, making parametric rates impossible to achieve.

12.1.2 Minimax lower bounds on estimation with Assouad's method

We return to the question of whether the results we have given are sharp; do there exist estimators attaining a faster rate of convergencethan the kernel-based estimator (12.1.3)? Using Assouad's method, we show that these results are indeed tight. Letting $\mathcal{F}$ be the class of of 1-Lipschitz functions on $[0, 1]$, we have the following lower bound.

Theorem 12.1.4. *Let the observation points X_i be spaced evenly on $[0, 1]$ and $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ in the observation model (12.1.1). Then there exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_2^2\right] \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2}{3}}.$$

Deferring the proof of the theorem temporarily, we make a few remarks. The same technique (see Exercise 12.1) shows that if $\mathcal{F}_\beta$ denotes the class of functions f on $[0, 1]$ with $\beta - 1$ derivatives, where the $(\beta - 1)$th derivative is Lipschitz, then

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+1}}. \tag{12.1.5}$$

So while the parametric σ^2/n rate is unachievable, we can come arbitrarily close. In dimensions $d \ge 1$, the minimax rate for estimating functions f with Lipschitz $(\beta - 1)$th derivative scales as

$$\mathfrak{M}_n(\mathcal{F}_\beta, \|\cdot\|_2^2) \ge c \left(\frac{\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+d}}. \tag{12.1.6}$$

One may prove such a result using variants of Assouad's or Fano's method, and Exercise 12.2 works through the $\beta = 1$ case of this bound. (See the bibliographic section for pointers to comprehensive references.) The bound (12.1.6) presents a striking example of the curse of dimensionality. For example, take $\beta = 1$. In one dimension, we require $n \ge 90 \approx (.05)^{-3/2}$ observations to achieve squared error .05, while we require $n \ge 8000 = (.05)^{-(2+d)/2}$ even when the dimension $d = 4$, and

$n \geq 64 \cdot 10^6$ observations even in 10 dimensions, which is a relatively small problem. The sample size requirement grows exponentially with the dimension.

To establish Theorem 12.1.4, we show how to construct a family of problems—indexed by binary vectors $v \in \{-1, 1\}^k$—so that our estimation problem satisfies the Hamming separation (9.5.1), then we show that observing noisy versions of the functions we define yields little information. Choosing k to make the resulting lower bound as large as possible completes the argument.

Construction of a separated family of functions To construct our separation in Hamming metric, as required by inequality (9.5.1), fix $k \in \mathbb{N}$ to be chosen. This approach is somewhat different from our standard approach of using a fixed dimensionality and scaling the separation directly; in nonparametric problems, we scale the "dimension" itself to adjust the difficulty of the estimation problem. Define the "mother" function $g(x) = [1/2 - |x - 1/2|]_+$, which is 1-Lipschitz and 0 outside of the interval $[0, 1]$. For $v \in \{-1, 1\}^k$, we represent f_v with shifted and scaled copies of g, defining the "bump" functions

$$g_j(x) := \frac{1}{k} g\left(k\left(x - \frac{j-1}{k}\right)\right) \quad \text{and} \quad f_v(x) := \sum_{j=1}^{k} v_j g_j(x).$$

This is 1-Lipschitz; see Figure 12.3 for an illustration.

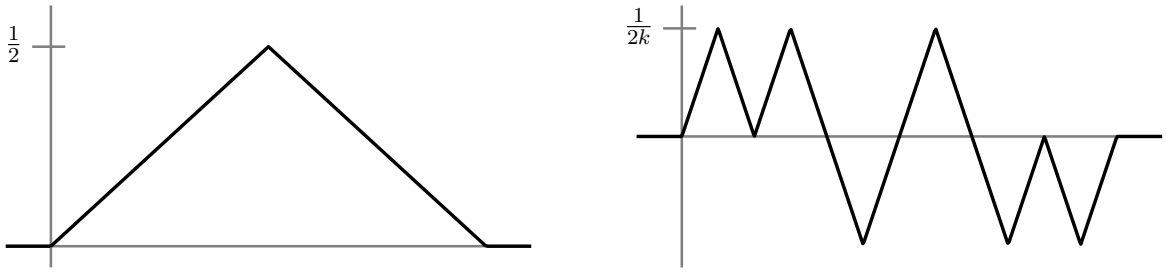


Figure 12.3. Illustration of the bump function construction. Left: the "mother" function g. Right: the shifted and scaled collection f_v, where in this case $k = 6$ and $v = (1, 1, -1, 1, -1, -1)$.

Consider a function $f : [0, 1] \to \mathbb{R}$, and let E_j be shorthand for the intervals $E_j = [(j-1)/k, j/k]$ for $j = 1, \ldots, k$. To follow the Hamming separation recipe in Assouad's method (Section 9.5), we must find a mapping identifying f with points in the hypercube $\{-1, 1\}^k$. To that end, define $\widehat{\mathsf{v}}(f) \in \{-1, 1\}^k$ by

$$\widehat{\mathsf{v}}_j(f) = \operatorname*{argmin}_{s \in \{-1,1\}} \int_{E_j} \left(f(t) - s g_j(t)\right)^2 dt.$$

We claim that

$$\left(\int_{E_j} (f(t) - f_v(t))^2 dt\right)^{\frac{1}{2}} \geq \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} f_v(t)^2 dt\right)^{\frac{1}{2}}. \tag{12.1.7}$$

Indeed, on the set E_j, we have $v_j g_j(t) = f_v(t)$, and thus $\int_{E_j} g_j(t)^2 dt = \int_{E_j} f_v(t)^2 dt$. Then by the

triangle inequality, we have

$$\begin{aligned}2 \cdot \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \left(\int_{E_j} g_j(t)^2 dt\right)^{\frac{1}{2}} &= \left(\int_{E_j} \left((\widehat{\mathsf{v}}_j(f) - v_j) g_j(t)\right)^2 dt\right)^{\frac{1}{2}} \\ &\le \left(\int_{E_j} \left(f(t) - v_j g_j(t)\right)^2 dt\right)^{\frac{1}{2}} + \left(\int_{E_j} \left(f(t) - \widehat{\mathsf{v}}_j(f) g_j(t)\right)^2 dt\right)^{\frac{1}{2}} \\ &\le 2\left(\int_{E_j} \left(f(t) - f_v(t)\right)^2 dt\right)^{\frac{1}{2}}.\end{aligned}$$

With the definition of $\widehat{\mathsf{v}}$ and inequality (12.1.7), we see that for any vector $v \in \{-1, 1\}^k$

$$\|f - f_v\|_2^2 = \sum_{j=1}^k \int_{E_j} \left(f(t) - f_v(t)\right)^2 dt \ge \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\} \int_{E_j} f_v(t)^2 dt.$$

Evaluating the final integral yields that $\int_{E_j} f_v(t)^2 dt = \frac{1}{k^3} \int_0^1 g(u)^2 du \ge \frac{c}{k^3}$ for a numerical constant $c > 0$. In particular, we have the desired Hamming separation

$$\|f - f_v\|_2^2 \ge \frac{c}{k^3} \sum_{j=1}^k \mathbf{1}\left\{\widehat{\mathsf{v}}_j(f) \neq v_j\right\}. \tag{12.1.8}$$

Bounding the binary testing error Let P_v^n denote the distribution of the n observations $Y_i = f_v(X_i) + \varepsilon_i$ when f_v is the true regression function. Then inequality (12.1.8) implies via Assouad's lemma (Thm. 9.5.2) that for $P_{\pm j}^n = \frac{2}{|\mathcal{V}|} \sum_{v : v_j = \pm 1} P_v^n$ as in the theorem,

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \ge \frac{c}{k^3} \sum_{j=1}^k \left[1 - \left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}\right]. \tag{12.1.9}$$

By convexity and Pinsker's inequality (2.2.10),

$$\left\|P_{+j}^n - P_{-j}^n\right\|_{\mathrm{TV}}^2 \le \max_v \left\|P_{v,+j}^n - P_{v,-j}^n\right\|_{\mathrm{TV}}^2 \le \max_v \frac{1}{2} D_{\mathrm{kl}}\left(P_{v,+j}^n \| P_{v,-j}^n\right).$$

To bound this KL divergence, note that under f_v the observations Y_i are independent normals with means $f_v(X_i)$. Thus

$$D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right) = \sum_{i=1}^n \frac{1}{2\sigma^2} (f_v(X_i) - f_{v'}(X_i))^2. \tag{12.1.10}$$

If v and v' differ only in coordinate j, the summands vanish unless $X_i \in E_j$, and $|f_v(X_i) - f_{v'}(X_i)| \le 1/k$. Since at most $O(n/k)$ points X_i lie in E_j and $\|f_v\|_\infty \le \frac{1}{k}$,

$$D_{\mathrm{kl}}\left(P_v^n \| P_{v'}^n\right) \le \frac{1}{2\sigma^2} \cdot \frac{O(1) n}{k} \cdot \frac{1}{k^2}.$$

Pinsker's inequality and the lower bound (12.1.9) therefore give

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_2^2) \ge \frac{c}{k^3} \sum_{j=1}^k \left[1 - C\sqrt{\frac{n}{k^3 \sigma^2}}\right] = \frac{c}{k^2}\left[1 - C\sqrt{\frac{n}{k^3\sigma^2}}\right]$$

for universal constants $0 < c, C < \infty$.

Choosing k for optimal tradeoffs Now we simply choose k to make the test error probabilities constant. Taking $k = O(1) \cdot (\frac{n}{2\sigma^2})^{1/3}$ evidently suffices, and Theorem 12.1.4 follows.

12.2 Global Fano Method

This section replaces the local construction of Fano's method in Section 9.4 with a more global construction. In particular, we show that, rather than constructing a local packing and optimizing a local perturbation scale, we can instead develop minimax lower bounds directly from packing numbers and KL metric entropies.

12.2.1 A mutual information bound based on metric entropy

To begin, we recall the classical Fano inequality in Corollary 9.4.2, which says that for any Markov chain $V \to X \to \widehat{V}$, where V is uniform on the finite set $\mathcal{V}$, we have

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V; X) + \log 2}{\log(|\mathcal{V}|)}.$$

Thus, there are two ingredients in proving lower bounds on the error in a hypothesis test: upper bounding the mutual information and lower bounding the size $|\mathcal{V}|$. The key in the global Fano method is an upper bound on the former (the information $I(V; X)$) using covering numbers.

Before stating the result, we require a bit of notation. First, we assume that V is drawn from a distribution μ, and conditional on $V = v$, the sample $X \sim P_v$. Then a standard calculation (or simply the definition of mutual information; recall equation (9.4.4)) gives that

$$I(V; X) = \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v), \quad \text{where} \quad \overline{P} = \int P_v d\mu(v).$$

We can say slightly more: letting Q be any distribution,

$$\begin{aligned}
\int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) &= \int_{\mathcal{V}} \int_{\mathcal{X}} dP_v \left[\log \frac{dP_v}{dQ} + \log \frac{dQ}{d\overline{P}}\right] d\mu(v) \\
&= \int_{\mathcal{V}} D_{\rm kl}\left(P_v \| Q\right) d\mu(v) + \int_{\mathcal{X}} \underbrace{\int_{\mathcal{V}} d\mu(v) dP_v}_{= d\overline{P}} \log \frac{dQ}{d\overline{P}} \\
&= \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) - D_{\rm kl}\left(\overline{P} \| Q\right),
\end{aligned}$$

so that

$$I(V; X) = \int D_{\rm kl}\left(P_v \| \overline{P}\right) d\mu(v) \leq \int D_{\rm kl}\left(P_v \| Q\right) d\mu(v) \tag{12.2.1}$$

for any distribution Q on the space $\mathcal{X}$. We can connect this mutual information bound (12.2.1) to a covering number of a set of distributions.

Assume that for all v, we have $P_v \in \mathcal{P}$, where $\mathcal{P}$ is a collection of distributions. In analogy with Definition 5.1, we say that the collection of distributions $\{Q_i\}_{i=1}^N$ form an ϵ-cover of $\mathcal{P}$ in KL-divergence if for all $P \in \mathcal{P}$, there exists some i such that $D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2$. With this, we may define the KL-covering number of the set $\mathcal{P}$ as

$$N_{\rm kl}\left(\epsilon, \mathcal{P}\right) := \inf\left\{N \in \mathbb{N} \mid \text{ there exist } \{Q_i\}_{i=1}^N \text{ s.t. } \sup_{P \in \mathcal{P}} \min_{i \in [N]} D_{\rm kl}\left(P \| Q_i\right) \leq \epsilon^2\right\}, \tag{12.2.2}$$

where $N_{\rm kl}(\epsilon, \mathcal{P}) = +\infty$ if no such cover exists. With definition (12.2.2) in place, we have the following proposition.

Proposition 12.2.1. *Under conditions of the preceding paragraphs,*

$$I(V; X) \le \inf_{\epsilon > 0} \left\{ \epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\}. \tag{12.2.3}$$

Proof We may carefully choose the distribution Q in the upper bound (12.2.1) to obtain the proposition. Let the distributions Q_i, $i = 1, \ldots, N$ form an ϵ^2-cover of the family $\mathcal{P}$, meaning that

$$\min_{i \in [N]} D_{\rm kl}(P \| Q_i) \le \epsilon^2 \quad \text{for all } P \in \mathcal{P}.$$

Let p_v and q_i denote the densities of P_v and Q_i with respect to some fixed base measure on $\mathcal{X}$ (the choice of base measure does not matter). Then defining the distribution $Q = (1/N) \sum_{i=1}^N Q_i$, we obtain for any v that in expectation over $X \sim P_v$,

$$\begin{aligned} D_{\rm kl}(P_v \| Q) &= \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q(X)}\right] = \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{N^{-1} \sum_{i=1}^n q_i(X)}\right] \\ &= \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\sum_{i=1}^N q_i(X)}\right] \le \log N + \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{\max_i q_i(X)}\right] \\ &\le \log N + \min_i \mathbb{E}_{P_v}\left[\log \frac{p_v(X)}{q_i(X)}\right] = \log N + \min_i D_{\rm kl}(P_v \| Q_i). \end{aligned}$$

By our assumption that the Q_i form a cover, this gives the desired result, as $\epsilon \ge 0$ was arbitrary, as was our choice of the cover. □

By a completely parallel proof, we also immediately obtain the following corollary.

Corollary 12.2.2. *Assume that $X_1, \ldots, X_n$ are drawn i.i.d. from P_v conditional on $V = v$. Let $N_{\rm kl}(\epsilon, \mathcal{P})$ denote the KL-covering number of a collection $\mathcal{P}$ containing the distributions (over a single observation) P_v for all $v \in \mathcal{V}$. Then*

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon \ge 0} \left\{ n\epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\}.$$

Corollary 12.2.2 and Proposition 12.2.1 show that the global covering numbers in KL-divergence bound the information.

12.2.2 Minimax bounds using global packings

There is now a four step process to proving minimax lower bounds using the global Fano method. Recall the Fano minimax lower bound in Proposition 9.4.3, which begins with the construction of a set of points $\{\theta(P_v)\}_{v \in \mathcal{V}}$ that form a 2δ-packing of a set Θ in some ρ-semimetric. With this construction in mind, we perform the following four steps:

(i) *Bound the packing entropy.* Give a lower bound on the packing number of the set Θ with 2δ-separation (call this lower bound $M(\delta)$).

(ii) *Bound the metric entropy.* Give an upper bound on the KL-metric entropy of the class $\mathcal{P}$ of distributions containing all the distributions P_v, that is, an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.

(iii) *Find the critical radius.* Recalling (Corollary 12.2.2) that with n i.i.d. observations, we have

$$I(V; X_1, \ldots, X_n) \le \inf_{\epsilon \ge 0} \left\{ n\epsilon^2 + \log N_{\rm kl}(\epsilon, \mathcal{P}) \right\},$$

we balance the information $I(V; X_1^n)$ and the packing entropy $\log M(\delta)$. To that end, we choose ϵ_n and $\delta > 0$ at the *critical radius* as follows: choose any ϵ_n such that

$$n\epsilon_n^2 \ge \log N_{\rm kl}(\epsilon_n, \mathcal{P}),$$

and choose the largest $\delta_n > 0$ such that

$$\log M(\delta_n) \ge 4n\epsilon_n^2 + 2\log 2,$$

which necessarily therefore satisfies

$$\log M(\delta_n) \ge 2\left(I(V; X_1^n) + \log 2\right).$$

(One could instead minimize the bound in Corollary 12.2.2; the present choice requires only an upper bound on $\log N_{\rm kl}(\epsilon, \mathcal{P})$.)

(iv) *Apply the Fano minimax bound.* Given any δ_n and ϵ_n as above, we immediately obtain that for the Markov chain $V \to X_1^n \to \widehat{V}$,

$$\mathbb{P}(V \neq \widehat{V}) \ge 1 - \frac{I(V; X_1, \ldots, X_n) + \log 2}{\log M(\delta_n)} \ge 1 - \frac{1}{2} = \frac{1}{2},$$

and thus, the Fano minimax bound in Proposition 9.4.3 implies

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \ge \frac{1}{2}\Phi(\delta_n).$$

12.2.3 Example: nonparametric regression

We flesh out the outline above to obtain a minimax lower bound for a nonparametric regression problem directly via packing and metric entropies, suppressing numerical constants. In so doing, We recover an analogue of Theorem 12.1.4 on minimax risks for estimation of 1-Lipschitz functions on $[0, 1]$.

We use a random-design Gaussian regression model: the pairs (X_i, Y_i) are i.i.d., $X_i \sim \mathsf{Uniform}[0, 1]$, and $Y_i \mid X_i = x \sim \mathsf{N}(f(x), \sigma^2)$. Letting

$$\mathcal{F} := \{f : [0, 1] \to \mathbb{R}, \ f(0) = 0, \ f \text{ is 1-Lipschitz}\}$$

be the family of 1-Lipschitz functions with $f(0) = 0$, we have

Proposition 12.2.3. *There exists a universal constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty \right] \ge c \min\left\{ 1, \left(\frac{\sigma^2}{n}\right)^{1/3} \right\},$$

where $\widehat{f}_n$ is constructed based on the n independent observations $f(X_i) + \varepsilon_i$.

The rate in Proposition 12.2.3 is sharp to within factors logarithmic in n; a more precise analysis yields

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\|\widehat{f}_n - f\|_\infty \right] \asymp \left(\frac{\sigma^2 \log n}{n} \right)^{1/3}.$$

See, for example, Tsybakov [401] for a proof.

Proof We first claim that the covering and packing numbers of $\mathcal{F}$ in the ℓ_∞ metric satisfy

$$\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log M(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta}. \tag{12.2.4}$$

To see this, fix $\delta \in (0, 1)$ and assume for simplicity that $1/\delta$ is an integer. Define the sets $E_j = [\delta(j-1), \delta j)$, and for each $v \in \{-1, 1\}^{1/\delta}$ define $h_v(x) = \sum_{j=1}^{1/\delta} v_j \mathbf{1}\{x \in E_j\}$. Then define the function $f_v(t) = \int_0^t h_v(u)du$, which increases or decreases linearly on each interval of width δ in $[0, 1]$. Then these f_v form an $O(1)\delta$-packing and an $O(1)\delta$-cover of $\mathcal{F}$, and there are $2^{1/\delta}$ such f_v. Thus the asymptotic approximation (12.2.4) holds. (Identify $1/\delta$ with k in the right figure of Figure 12.3 to see an instance of the construction.)

Abusing notation in the KL-divergence, for any $x \in [0, 1]$ and $f, g \in \mathcal{F}$, we have

$$D_{\rm kl}\left(f(x) + \varepsilon \| g(x) + \varepsilon\right) = \frac{1}{2\sigma^2}(f(x) - g(x))^2 \le \frac{\|f - g\|_\infty^2}{2\sigma^2}.$$

Letting P_f^n be the distribution of the n observations $f(X_i) + \varepsilon_i$, $i = 1, \ldots, n$, we thus have

$$D_{\rm kl}\left(P_f^n \| P_g^n\right) = \sum_{i=1}^n \frac{1}{2\sigma^2}(f(X_i) - g(X_i))^2 \le \frac{n}{2\sigma^2} \|f - g\|_\infty^2.$$

Therefore a supremum-norm cover of $\mathcal{F}$ at radius $\sigma\epsilon$ induces a KL cover at radius $O(1)\epsilon$, implying

$$\log N_{\rm kl}(\epsilon, \mathcal{P}) \lesssim \frac{1}{\sigma\epsilon} \quad \text{for} \quad \mathcal{P} = \{P_f : f \in \mathcal{F}\}.$$

Thus we have completed steps (i) and (ii) in our program above.

It remains to choose the critical radius in step (iii), but this is now relatively straightforward: by choosing $\epsilon_n \asymp (1/\sigma n)^{1/3} \wedge 1$, whence (for large n) $n\epsilon_n^2 \asymp (n/\sigma^2)^{1/3}$, we find that taking $\delta \asymp (\sigma^2/n)^{1/3}$ is sufficient to ensure that $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta^{-1} \ge 4n\epsilon_n^2 + 2\log 2$. Thus we have

$$\mathfrak{M}_n(\mathcal{F}, \|\cdot\|_\infty) \gtrsim \delta_n \cdot \frac{1}{2} \gtrsim \left(\frac{\sigma^2}{n} \right)^{1/3}$$

as desired. □

12.3 Multiple hypothesis testing and detection

Modern statistical problems frequently evaluate many hypotheses simultaneously. For example, in a genome-wide association study (GWAS), a researcher receives the expression levels of many thousands of genes, and wishes to test which few associate with a particular phenotype while limiting false discoveries. Consider another scenario: among n experiments, an idiosyncratic experimental

factor leads to successes in the fraction in which it is present; when can we (without knowing this factor) reliably detect that a small fraction of experiments were successful in spite of many null results? These scenarios lead to multiple testing and signal detection problems. We will develop the tools to provide precise asymptotics for identification or detection as the number n of hypotheses grows while the strength and number of non-null hypotheses varies in these problems.

We observe n mostly null p-values, where small p-values provide evidence against the null. Recall that variable U_i is a p-value if it is *sub-uniform*, meaning that

$$\mathbb{P}(U_i \le u) \le u \quad \text{for } u \in [0,1]$$

(so p-values are less likely to be small than a uniform variable). In the Fisherian hypothesis tests here, we wish to avoid rejecting null hypotheses (making false discoveries), and so at level $\alpha \in (0,1)$, reject i if $U_i \le \alpha$, whence $\mathbb{P}(\text{reject } i \mid H_0) = \mathbb{P}(U_i \le \alpha) \le \alpha$. The simplest global null hypothesis is that each p-value is null,

$$H_{0,i} : U_i \text{ is a } p\text{-value}, \quad i = 1, \ldots, n. \tag{12.3.1}$$

To avoid making any false discoveries, one employs some type of correction, such as the conservative Bonferroni correction (union bound): reject null $H_{0,i}$ only if $U_i \le \frac{\alpha}{n}$, so

$$\mathbb{P}(\text{any false rejection}) \le \sum_{i=1}^n \mathbb{P}(\text{reject } i \mid H_{0,i}) \le n \cdot \frac{\alpha}{n} = \alpha. \tag{12.3.2}$$

We refine this setting for the remainder of this section. To provide optimality results, we introduce alternatives, many of which capture phase transition behavior that delineates quite precisely between impossible and successful identification or detection of non-nulls.

12.3.1 The multiple testing problem

In the multiple hypothesis testing problem, we have a collection of null hypotheses $\{H_{0,i}\}_{i=1}^n$ as in setting (12.3.1), and we wish to develop procedures that falsely reject few of them while providing complementary optimality guarantees. To formalize this, we consider scenarios in which a set of these null hypotheses is truly null, while the other observed p-values are non-null, partitioning the indices into sets

$$\mathcal{I}_0 := \{i \in [n] \text{ s.t. } H_{0,i} \text{ holds}\} \quad \text{and} \quad \mathcal{I}_1 := [n] \setminus \mathcal{I}_0 = \{i \in [n] \text{ s.t. } H_{0,i} \text{ does not hold}\}$$

of true nulls and non-nulls. Then given any procedure for rejecting null hypotheses, we let

$$\mathcal{R} := \{i \in [n] \mid \text{procedure rejects } i\}$$

be the rejected null hypotheses. For any procedure, we typically wish to control one of two standard error measures. The first is the *family-wise error rate* (or FWER)

$$\mathsf{FWER} := \mathbb{P}(\mathcal{I}_0 \cap \mathcal{R} \neq \emptyset), \tag{12.3.3}$$

the probability of making a single false discovery, which can be too conservative for practical use, as we require procedures to make no mistakes. The Bonferroni correction (12.3.2) controls the FWER at level α. The more forgiving *false discovery proportion* (FDP) asks to control the ratio of false discoveries to rejected nulls,

$$\mathsf{FDP} := \frac{\operatorname{card}(\mathcal{R} \cap \mathcal{I}_0)}{\max\{1, \operatorname{card}(\mathcal{R})\}}, \tag{12.3.4}$$

so keeping FDP small means that most of the discoveries (rejections) are correct.

Benjamini and Hochberg [57] develop an alternative idea that (typically) allows rejecting more hypotheses than the naive Bonferroni correction. To motivate it, suppose the null p-values are i.i.d.:

$$H_{0,i} : U_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, 1].$$

Then the order statistics $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ are uniform on $\{u \in \mathbb{R}^d \mid 0 \le u_1 \le \cdots \le u_n \le 1\}$ (see also Exercise 12.13). Thus, somewhat heuristically, observing $U_{(i)} \le \frac{\alpha i}{n}$ evidences (many) departures from the null and motivates rejecting the corresponding hypotheses. This yields the Benjamini-Hochberg step-up procedure (see Fig. 12.4), which finds the largest value

$$\widehat{k} := \max \left\{ k \in \mathbb{N} \mid U_{(k)} \le \frac{k\alpha}{n} \right\}, \tag{12.3.5}$$

then the hypotheses associated nulls $U_{(1)}, \ldots, U_{(\widehat{k})}$.

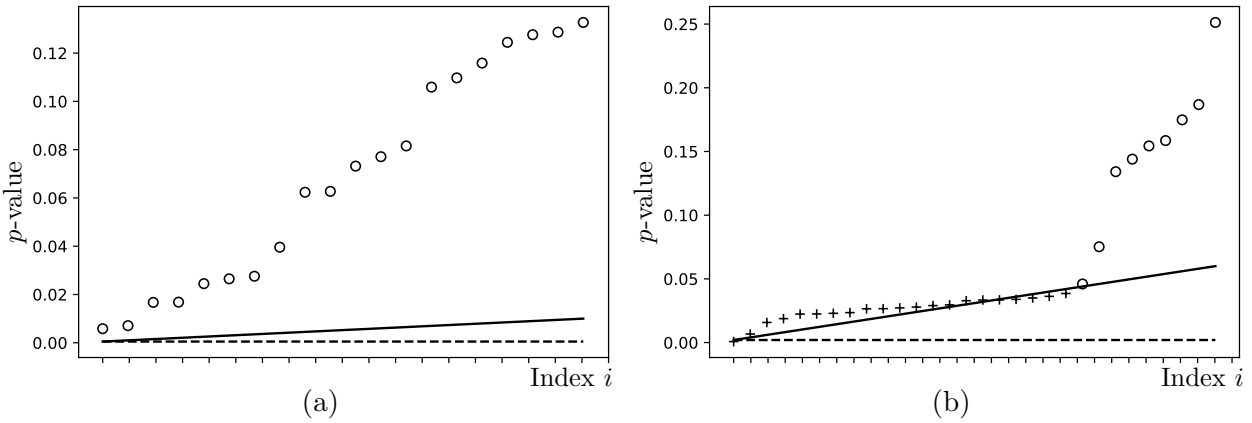


Figure 12.4. Illustration of the Benjamini-Hochberg (BH) procedure versus naive Bonferroni correction. BH rejects all hypotheses with p-values to the left of the last p-value below the solid black line. Bonferroni correction rejects hypotheses below dotted line. (a) All null p-values, where $n = 200$. Neither procedure rejects any hypothesis. (b) BH rejects all hypotheses represented with $+$ icon.

When all the null hypotheses yield independent p-values (including that they are independent of the non-nulls), the procedure (12.3.5) controls the expected proportion of false discoveries, which we term the *false discovery rate* (FDR):

$$\mathsf{FDR} := \mathbb{E}[\mathsf{FDP}].$$

Theorem 12.3.1. *Assume that for every $j \in \mathcal{I}_0$, the null p-value U_j is independent of all the other p-values, and is sub-uniform. Then the false-discovery rate of the Benjamini-Hochberg procedure (12.3.5) at level α satisfies $\mathsf{FDR} \le \frac{\mathrm{card}(\mathcal{I}_0)}{n}\alpha$.*

Proof We provide a proof that follows by iteratively replacing observed values U_i with 0, which can only increase the number of rejections in the Benjamini-Hochberg (BH) procedure (12.3.5). Let $F = \mathrm{card}(\mathcal{I}_0 \cap \mathcal{R})$ be the number of false rejections and $R = \mathrm{card}(\mathcal{R})$ be the number of total

rejected hypotheses, with $R_+ = \max\{R, 1\}$. Define $\alpha_i := \frac{i}{n}\alpha$, so that on the event $R = r$, the BH procedure rejects hypothesis j if and only if $U_j \le \alpha_r$. Then $\mathsf{FDP} = \frac{F}{R_+}$ and

$$\mathsf{FDR} = \mathbb{E}[\mathsf{FDP}] = \sum_{r=1}^{n} r^{-1}\mathbb{P}(R = r)\mathbb{E}[F \mid R = r] = \sum_{j \in \mathcal{I}_0}\sum_{r=1}^{n} \frac{1}{r}\mathbb{E}\left[\mathbf{1}\left\{U_j \le \alpha_r\right\}\mathbf{1}\left\{R = r\right\}\right]. \quad (12.3.6)$$

Fixing j and r, consider artificially setting $U_j = 0$ and performing BH (12.3.5); define R_j to be the number of rejections in this case. Then for each j, $\{U_j \le \alpha_r, R = r\} = \{U_j \le \alpha_r, R_j = r\}$ by construction, and the independence hypothesis for the nulls implies

$$\mathbb{E}\left[\mathbf{1}\left\{U_j \le \alpha_r\right\}\mathbf{1}\left\{R = r\right\}\right] = \mathbb{P}(U_j \le \alpha_r)\mathbb{P}(R_j = r) \le \frac{\alpha r}{n}\mathbb{P}(R_j = r)$$

for $j \in \mathcal{I}_0$. Substituting into the expansion (12.3.6) and using $\sum_{r=1}^{n}\mathbb{P}(R_j = r) = 1$, we have

$$\mathsf{FDR} \le \sum_{j \in \mathcal{I}_0}\sum_{r=1}^{n} \frac{\alpha r}{rn}\mathbb{P}(R_j = r) = \frac{\mathrm{card}(\mathcal{I}_0)}{n}\alpha,$$

as claimed. □

12.3.2 Optimality results in multiple testing

Many hypothesis testing procedures control the false discovery rate at level α; the trivial procedure that never rejects the null certainly does, as does the overconservative Bonferroni correction (12.3.2), as $\mathsf{FDP} \le \sum_{j \in \mathcal{I}_0}\mathbf{1}\left\{U_j \le \alpha/n\right\}$ trivially, and $\mathbb{E}[\mathsf{FDP}] \le \sum_{j \in \mathcal{I}_0}\frac{\alpha}{n} = \alpha$. So to understand the power (or optimality) of any procedure, we must evaluate it in the presence of alternatives to the nulls $H_{0,i}$, and we investigate the procedure's ability to identify true non-nulls. We formulate this as a problem of recovering a binary vector $v \in \{0,1\}^n$, where $v_j = 0$ indicates the null $H_{0,j}$ holds, while $v_j = 1$ will indicate the presence of an alternative $H_{1,j}$.

Because we focus on p-values, it is natural to consider alternatives that tend to be smaller than $\mathsf{Uniform}[0,1]$, so we will consider settings with nulls and alternatives following

$$U_j \overset{\text{ind}}{\sim} P_0 := \mathsf{Uniform}[0,1], \quad j \in \mathcal{I}_0 \quad \text{and} \quad U_j \overset{\text{ind}}{\sim} P_1, \quad j \in \mathcal{I}_1,$$

where the non-null distribution P_1 takes the form

$$P_1 := \mathsf{Uniform}[0,\tau] \quad \text{or} \quad P_1 := (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]. \quad (12.3.7)$$

Above, $\tau \in (0,1)$ indicates the strength of the alternatives, with $\tau \approx 0$ providing an easier test. The choice of alternatives (12.3.7) impacts the performance of procedures, and, for example, we might instead take nulls and alternatives

$$P_0 : Y_i \sim \mathsf{N}(0,1) \quad \text{versus} \quad P_1 : Y_i \sim \mathsf{N}(\mu, 1)$$

for some $\mu > 0$, asking for the magnitude $\mu > 0$ at which reliable discovery is possible; the setting (12.3.7) keeps things relatively simple for our exposition, and we provide further references in the bibliographic section (see also Exercise 12.14).

Continuing with the setting for our optimality results, let $V \in \{0,1\}^n$, and draw

$$V_j \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(\epsilon) \quad \text{and} \quad U_j \mid V_j = v_j \overset{\text{ind}}{\sim} P_{v_j}. \tag{12.3.8}$$

In this *canonical multiple testing problem* (12.3.8), we then ask at which settings of the pairs (ϵ, τ) we can reliably recover the zero and non-zero indices V_j, meaning we make true discoveries. We investigate problem (12.3.8) under the scaling

$$\epsilon = n^{-\beta} \quad \text{and} \quad \tau = n^{-r} \tag{12.3.9}$$

for values $\beta \in (0,1)$ and $r \in (0,\infty)$. We measure risk coordinatewise, so that for any testing procedure $\widehat{v} : [0,1]^n \to \{0,1\}^n$, we have test risk

$$R_n(\widehat{v} \mid P_0, P_1) = \sum_{j=1}^n \mathbb{P}\left(\widehat{v}_j(U_1^n) \neq V_j\right) = \sum_{j=1}^n (1-\epsilon) P_0(\widehat{v}_j \neq 0) + \epsilon P_1(\widehat{v}_j \neq 1), \tag{12.3.10}$$

where $\mathbb{P}$ denotes the joint distribution over V and U_1^n. Letting $\mathcal{R} = \{j \in [n] \mid \widehat{v}_j = 1\}$ be the procedure's rejections, the number of mistakes $M = \mathrm{card}(\mathcal{R} \cap \mathcal{I}_0) + \mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1)$ determines the risk, as

$$R_n(\widehat{v} \mid P_0, P_1) = \mathbb{E}[M].$$

The independence of the coordinates in the model (12.3.8) makes the optimal test $\widehat{v}^\star : [0,1]^n \to \{0,1\}^n$ a simple per-coordinate likelihood ratio test. Using this, a calculation yields that when the alternative $P_1 = \mathsf{Uniform}[0,\tau]$,

$$\inf_{\widehat{v}} \mathbb{P}(V_j \neq \widehat{v}_j) = \min\left\{\tau(1-\epsilon), \epsilon\right\},$$

while full-support alternative $P_1 = (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]$ gives

$$\inf_{\widehat{v}} \mathbb{P}(V_j \neq \widehat{v}_j) = \min\left\{\tau(1-\epsilon), \epsilon(1-\tau)\right\} + \epsilon\tau$$

whenever $\epsilon, \tau < \frac{1}{2}$. (See Exercise 12.11.)

In this sparse hypothesis setting, we can directly provide a lower bound on the number of mistakes of any procedure, showing that if $r \leq \beta$ in the parameter scaling (12.3.9), then no procedure can outperform the completely conservative procedure of never rejecting a null hypothesis. We provide the result for the canonical multiple hypothesis testing setting (12.3.8) with alternatives $P_1 = (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]$.

Proposition 12.3.2. *Let the conditions above hold. Assume that $r \leq \beta$ in the scaling (12.3.9) and let $\widehat{v}$ be any test with rejected nulls $\mathcal{R}$. Then the number of mistakes $M = \mathrm{card}(\mathcal{R} \cap \mathcal{I}_0) + \mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1)$ satisfies $R_n(\widehat{v} \mid P_0, P_1) = \mathbb{E}[M] \geq \mathbb{E}[\mathrm{card}(\mathcal{I}_1)] = n^{1-\beta}$, and*

$$\mathbb{E}\left[\frac{M}{\mathrm{card}(\mathcal{I}_1) + 1}\right] \geq 1 - o(1).$$

Proof The set $\mathcal{I}_1 = \{j \mid V_j = 1\}$ has cardinality $\mathrm{card}(\mathcal{I}_1) = \sum_{j=1}^n V_j$, where $V_j \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(n^{-\beta})$, so $\mathbb{E}[\mathrm{card}(\mathcal{I}_1)] = n^{1-\beta}$ and $\mathrm{Var}(V_j) \leq n^{-\beta}$. For the lower bound on $\mathbb{E}[M]$, recall the optimal

per-coordinate Bayes error $\inf_{\widehat{v}^\star} \mathbb{P}(V_j \neq \widehat{v}_j^\star) = \min\{\tau(1-\epsilon), \epsilon(1-\tau)\} + \epsilon\tau$, so

$$\begin{aligned}\mathbb{E}\left[M\right] &= \sum_{j=1}^n \mathbb{P}(\widehat{v}_j \neq V_j) \geq n \min\left\{\tau(1-\epsilon), \epsilon(1-\tau)\right\} + n\epsilon\tau \\ &= n \min\left\{n^{-r}(1-n^{-\beta}), n^{-\beta}(1-n^{-r})\right\} + n^{1-\beta-r} \\ &= n^{1-\beta} - n^{1-\beta-r} + n^{1-\beta-r} = n^{1-\beta}.\end{aligned}$$

Let $N_1 = \mathrm{card}(\mathcal{I}_1)$ and $m_n = n^{1-\beta}$. For $\eta > 0$, a Bernstein-type bound (4.1.10) gives $\mathbb{P}(N_1 > (1+\eta)m_n) \leq \exp(-c_\eta m_n)$ for some $c_\eta > 0$. Consequently,

$$\mathbb{E}\left[\frac{M}{N_1+1}\right] \geq \frac{\mathbb{E}[M\mathbf{1}\left\{N_1 \leq (1+\eta)m_n\right\}]}{(1+\eta)m_n+1} \geq \frac{\mathbb{E}[M] - n\mathbb{P}(N_1 > (1+\eta)m_n)}{(1+\eta)m_n+1} \geq \frac{1}{1+\eta} - o_n(1).$$

Letting $n \to \infty$ and then $\eta \downarrow 0$ proves the second claim. □

So for any interesting testing accuracy (or recovery of the non-null hypotheses), we necessarily must have stronger signal in the scaling (12.3.9), that is, $r > \beta$. Somewhat remarkably, the Benjamini-Hochberg procedure (12.3.5) is asymptotically optimal in this case.

Theorem 12.3.3. *Assume $r > \beta$ in the scaling (12.3.9). Let $\mathcal{R}$ denote the set of indices that the Benjamini-Hochberg procedure (12.3.5) with level $\alpha = \alpha_n$ rejects, and assume $n^{r-\beta}\alpha_n \gg 1$. Then for $\delta > 0$,*

$$\mathbb{P}\left(\mathrm{card}(\mathcal{R} \cap \mathcal{I}_1) \leq (1-\delta)\,\mathrm{card}(\mathcal{I}_1)\right) \to 0. \tag{12.3.11}$$

Moreover, whenever the level $\alpha_n \to 0$ as well,

$$\mathbb{E}\left[\frac{\mathrm{card}(\mathcal{R} \cap \mathcal{I}_0) + \mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1)}{\mathrm{card}(\mathcal{R}) + 1}\right] \to 0. \tag{12.3.12}$$

Proof We prove the convergence (12.3.11) first. Let $N_1 = \mathrm{card}(\mathcal{I}_1) = \|V\|_1$ be the (random) cardinality of the true non-nulls. For $\mathcal{R}$ to not capture at least a fraction $(1-\delta)$ of the non-nulls, it must be the case that $U_j > \frac{(1-\delta)N_1\alpha}{n}$ for at least δN_1 of the values $\{U_j\}_{j\in\mathcal{I}_1}$, as otherwise, BH (12.3.5) would have rejected them. Because $U_j \stackrel{\mathrm{iid}}{\sim} P_1 = (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]$ conditional on $j \in \mathcal{I}_1$, conditioning on the value $N_1 > 0$ and ignoring immaterial integer roundings, we observe

$$\mathbb{P}\left(\mathrm{card}(\mathcal{R} \cap \mathcal{I}_1) \leq (1-\delta)N_1 \mid N_1\right) \leq \binom{N_1}{N_1\delta}\mathbb{P}\left(Y_\tau > \frac{(1-\delta)N_1\alpha}{n}\right)^{\delta N_1},$$

where $Y_\tau \sim P_1$ is an independent draw. Observe that

$$\mathbb{P}\left(Y_\tau > \frac{(1-\delta)N_1\alpha}{n}\right) \leq (1-\tau)\mathbb{P}\left(\mathsf{Uniform}[0,1] > \frac{(1-\delta)N_1\alpha}{n\tau}\right) + \tau.$$

Let us now guarantee that N_1 is large. As in the proof of Proposition 12.3.2, by a Bernstein-type bound, there is a numerical constant C such that for all $t \geq 0$, $N_1 \geq n^{1-\beta} - C(\sqrt{n^{1-\beta}t} + t)$ with probability at least $1 - e^{-t}$. So long as $t \ll n^{1-\beta}$, then we have $N_1 \geq n^{1-\beta}(1-o(1))$, or equivalently,

$\frac{(1-\delta)N_1\alpha}{n\tau} \ge n^{r-\beta}(1-\delta)\alpha(1-o(1))$, which is at least 1 eventually as $r > \beta$ so $n^{r-\beta}\alpha \gg 1$. So for all large enough n, whenever $N_1 \ge n^{1-\beta}(1-o(1))$ we obtain

$$\mathbb{P}\left(Y_\tau > \frac{(1-\delta)N_1\alpha}{n}\right) \le \tau.$$

Substituting above, we have

$$\begin{aligned}\mathbb{P}\left(\mathrm{card}(\mathcal{R} \cap \mathcal{I}_1) \le (1-\delta)N_1 \mid N_1\right) \le \left(\frac{e}{\delta}\right)^{N_1\delta} \tau^{\delta N_1} &= \exp\left(N_1\delta \log\frac{e}{\delta} + N_1\delta\log\tau\right) \\ &= \exp\left(N_1\delta\left(\log\frac{e}{\delta} - r\log n\right)\right).\end{aligned}$$

Whenever $N_1 \ge n^{1-\beta}(1-o(1))$ this converges to zero exponentially quickly. Taking $t = 4\log n$ we therefore have $N_1 \ge n^{1-\beta}(1 - C\sqrt{\log n / n^{1-\beta}})$ with probability at least $1 - n^{-4}$, and so

$$\mathbb{P}\left(\mathrm{card}(\mathcal{R} \cap \mathcal{I}_1) \le (1-\delta)N_1\right) \le \exp\left(-cr\delta n^{1-\beta}\log n\right) + n^{-4},$$

where $c > 0$ is a numerical constant, the inequality holding as soon as $n^{r-\beta} \ge \frac{1}{(1-\delta)\alpha}$.

To see the second convergence guarantee (12.3.12), use that $\mathrm{card}(\mathcal{R}\cap\mathcal{I}_1) \ge (1-\delta)N_1$ with high probability, and on this event $\mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1) \le \delta N_1$, while $\mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1) \le N_1$ regardless. Thus

$$\mathbb{E}\left[\frac{\mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1)}{N_1 + 1}\right] \le \mathbb{E}\left[\frac{\delta N_1}{N_1}\right] + o(1) = \delta + o(1).$$

The preceding calculations demonstrate that $\mathrm{card}(\mathcal{R}) \ge (1-\delta)N_1$ with probability at least $1 - n^{-4}$, so

$$\mathbb{E}\left[\frac{\mathrm{card}(\mathcal{R}^c \cap \mathcal{I}_1)}{\mathrm{card}(\mathcal{R}) + 1}\right] \le \mathbb{E}\left[\frac{\delta N_1}{(1-\delta)N_1}\right] + n\mathbb{P}(\mathrm{card}(\mathcal{R}) < (1-\delta)N_1) + o(1) = \frac{\delta}{1-\delta} + o(1).$$

By the guarantee of Theorem 12.3.1 that the BH procedure controls the false discovery rate at level $\alpha = \alpha_n$, we trivially have $\mathbb{E}[\mathrm{card}(\mathcal{R} \cap \mathcal{I}_0)/(\mathrm{card}(\mathcal{R}) + 1)] \le \alpha_n \to 0$. Combining these guarantees and that $\delta > 0$ was arbitrary gives inequality (12.3.12). □

Summarizing the results, the Benjamini-Hochberg procedure achieves asymptotically optimal behavior under the scalings (12.3.9) whenever $r > \beta$, while if $r \le \beta$, then no procedure can succeed. We therefore see a *phase transition* behavior: accurate multiple hypothesis testing is asymptotically possible if and only if the signal strength satisfies $r > \beta$, while otherwise no procedure can achieve better risk than simply rejecting no nulls. We remark in passing—see Exercise 12.12—that the more naive Bonferroni correction achieves essentially the same risk as never rejecting a hypothesis.

12.3.3 The signal detection problem

The results in Section 12.3.2 allow us to delineate those scenarios where reliable identification of null and non-null hypotheses is possible. Essentially, we can recover (up to a vanishing error fraction) the non-null hypotheses in the canonical model (12.3.8) when the fraction ϵ of non-nulls scales as $\epsilon = n^{-\beta}$ and the strength, represented by the upper bound $\tau = n^{-r}$ on the observed p-values, satisfies $r > \beta$ in the scaling (12.3.9). As we mention in the introduction to this section, in some scenarios we would like to *detect* the presence of non-nulls, even if we cannot reliably identify

them. This detection is possible, unsurprisingly, in a wider array of settings than those in which reliable identification is possible, though we still often see a phase transition behavior delineating asymptotically perfect versus impossible detection.

Revisiting the canonical multiple testing problem (12.3.8), we model the sparse signal detection problem as testing mixtures, where for individual observations Y_i we have a null distribution P_0, a known alternative P_1, and a sequence of observations Y_i, $i = 1, \ldots, n$, following either

$$H_0 : Y_i \stackrel{\text{iid}}{\sim} P_0 \quad \text{or} \quad H_1 : Y_i \stackrel{\text{iid}}{\sim} (1-\epsilon)P_0 + \epsilon P_1. \tag{12.3.13}$$

In the alternative H_1 we observe from a mixture of P_0 and P_1, and about ϵ fraction of the time we observe data from P_1. The question in such a sparse signal detection problem is the rate at which we can take $\epsilon \downarrow 0$ while still reliably testing between H_0 and H_1.

An elementary example highlights the gap between detection and recovery:

Example 12.3.4 (Gaussian detection)**:** In the problem (12.3.13), take null $P_0 = \mathsf{N}(0, 1)$ and alternative $P_1 = \mathsf{N}(\mu, 1)$. In this case, we wish to understand the scaling of μ with the sparsity level $\epsilon = n^{-\beta}$ in the setting (12.3.9). Identification of non-nulls has logarithmic signal threshold $\mu = \sqrt{2r \log n}$, and Exercise 12.14 demonstrates that if $r \lesseqgtr \beta$ governs recovery. Detection is easier: every fixed $\mu > 0$ is detectable when $\beta < \frac{1}{2}$.

Indeed, define $B_j = \mathbf{1}\{Y_j \ge \mu/2\}$, and $\delta = 2(\mathbb{P}(\mathsf{N}(0,1) \le \mu/2) - \frac{1}{2}) = \frac{1}{\sqrt{2\pi}}\mu + O(\mu^2) > 0$. Under the hypotheses,

$$H_0 : B_j \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}\left(\frac{1-\delta}{2}\right) \quad \text{and} \quad H_1 : B_j \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}\left((1-\epsilon)\frac{1-\delta}{2} + \epsilon\frac{1+\delta}{2}\right)$$

The means of $\sum_{j=1}^n B_j$ differ by $\epsilon\delta n = n^{1-\beta}\delta \gg \sqrt{n}$, while their standard deviations are $O(1)\sqrt{n}$. Thresholding between the means gives summed error $o(1)$. $\diamond$

In the testing problem (12.3.13), we will investigate optimal error under the test risk

$$R_n(\Psi \mid H_0, H_1) := \mathbb{P}(\Psi(Y_1^n) \neq 0 \mid H_0) + \mathbb{P}(\Psi(Y_1^n) \neq 1 \mid H_1), \tag{12.3.14}$$

summing the probabilities of error under the null and alternative. Because problem (12.3.13) is a simple hypothesis test of

$$P_0^n \quad \text{versus} \quad ((1-\epsilon)P_0 + \epsilon P_1)^n,$$

the likelihood ratio (Neyman-Pearson) test is always optimal, though unless we can compute it precisely, this provides little insight. Alternatively, the equivalence between Hellinger and variation distance in Proposition 2.2.8 combines with the identity $\inf_\Psi\{P(\Psi = 1) + Q(\Psi = 0)\} = 1 - \|P - Q\|_{\text{TV}}$ to demonstrate

$$d_{\text{hel}}^2(P_0^n, ((1-\epsilon)P_0 + \epsilon P_1)^n) \to 0 \text{ if and only if } \inf_\Psi R_n(\Psi \mid H_0, H_1) \to 1$$

while

$$d_{\text{hel}}^2(P_0^n, ((1-\epsilon)P_0 + \epsilon P_1)^n) \to 1 \text{ if and only if } \inf_\Psi R_n(\Psi \mid H_0, H_1) \to 0.$$

Because $d_{\text{hel}}^2(P^n, Q^n) = 1 - (1 - d_{\text{hel}}^2(P, Q))^n$, the sparse mixture testing problem (12.3.13) satisfies

$$\inf_\Psi R_n(\Psi \mid H_0, H_1) \to \begin{cases} 1 & \text{if } d_{\text{hel}}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) \ll \frac{1}{n} \\ 0 & \text{if } d_{\text{hel}}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) \gg \frac{1}{n}. \end{cases} \tag{12.3.15}$$

While a fully general theory characterizing the limits (12.3.15) does not exist, examples can help to delineate when we might hope to detect sparse signals. As a simple one that captures some of the techniques, consider the following Bernoulli detection problem.

Example 12.3.5 (Bernoulli detection)**:** Consider a null distribution $P_0 = \mathsf{Bernoulli}(\frac{1}{2})$ and alternative P_1 a point mass at 1. Then $(1-\epsilon)P_0 + \epsilon P_1 = \mathsf{Bernoulli}(\frac{1+\epsilon}{2})$, so

$$2d_{\rm hel}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) = 2 - \sqrt{1+\epsilon} - \sqrt{1-\epsilon} = \frac{\epsilon^2}{4} + O(\epsilon^4).$$

So perfect testing is asymptotically possible whenever $\epsilon^2 \gg \frac{1}{n}$ and impossible whenever $\epsilon^2 \ll \frac{1}{n}$. A simple test achieves this: taking $S_n = \sum_{i=1}^n Y_i$, $\Psi = \mathbf{1}\{S_n > \frac{n}{2} + n\epsilon/4\}$. Then Hoeffding's inequality gives $\mathbb{P}(\Psi(S_n) \text{ incorrect}) \le \exp(-n\epsilon^2/8)$. $\diamond$

This example fails to capture the fuller complexity of signal detection, because no individual signal can be too strong—the observations lie in $\{0,1\}$ regardless. Examples 12.3.4 and 12.3.5 do, however, show that the "interesting" regime is when the signals are quite sparse, so that $\beta > \frac{1}{2}$. Indeed, let $\beta < \frac{1}{2}$ and assume that $\|P_0 - P_1\|_{\rm TV} \ge c > 0$. Then it is relatively simple to develop a test Ψ that achieves risk $R_n(\Psi \mid H_0, H_1) \to 0$ by counting observations more likely to have come from P_0 or P_1, and a slight elaboration of this procedure works for $\beta = \frac{1}{2}$ as well.

Corollary 12.3.6. *Let P_0 and P_1 be distributions satisfying $\|P_0 - P_1\|_{\rm TV} = c > 0$, and consider the testing problem (12.3.13) with $\epsilon = \epsilon_n = n^{-\beta}$. If $\beta < \frac{1}{2}$, then*

$$\inf_\Psi R_n(\Psi \mid H_0, H_1) \to 0.$$

If $P_{0,n}$ and $P_{1,n}$ are sequences of distributions satisfying $\|P_{0,n} - P_{1,n}\|_{\rm TV} \to 1$, then if $\beta \le \frac{1}{2}$,

$$\inf_\Psi R_n(\Psi \mid H_0, H_1) \to 0.$$

In each limit we consider the hypothesis test of $H_0 : P_{0,n}^n$ versus $H_1 : ((1-\epsilon_n)P_{0,n} + \epsilon_n P_{1,n})^n$.

See Exercise 12.9 for a sketch of the proof, which more or less formalizes the argument in Example 12.3.4. This motivates the next modification of our scenario.

12.3.4 The sparse signal detection problem

Corollary 12.3.6 shows that if the set of non-null hypotheses is too dense—meaning $\beta < \frac{1}{2}$ in the scaling (12.3.9), then detection is quite easy, and this is true for most non-trivial nulls P_0 and alternatives P_1. Thus, for detection in particular, we consider the setting that $\beta > \frac{1}{2}$, which we call *sparse signals*, that is, we wish to test

$$H_0 : Y_i \stackrel{\rm iid}{\sim} P_0 \quad \text{versus} \quad H_1 : Y_i \stackrel{\rm iid}{\sim} (1-\epsilon)P_0 + \epsilon P_1 \quad \text{where} \quad \epsilon = n^{-\beta}, \; \beta > \frac{1}{2}. \tag{12.3.16}$$

The main objective in the sparse detection problem (12.3.16) is to identify a signal strength parameter r for the alternative P_1 (whose precise meaning depends on the problem at hand) and a smallest possible value $r^*(\beta)$ as a function of β, where if $r > r^*(\beta)$ detection is asymptotically possible, in that the summed type-I and type-II errors go to 0, i.e., $\inf_\Psi R_n(\Psi \mid H_0, H_1) \to 0$. We will see sharp phase transitions between detectable and impossible scenarios appear in several examples. A second goal, which is mostly beyond our scope, is the development of adaptive procedures that

achieve this optimal signal strength $r^*(\beta)$ without knowledge of β. We focus on developing some of the information-theoretic tools that allow us to identify $r^*(\beta)$.

As the limit (12.3.15) dictates, we follow the general recipe to compute the Hellinger distance between P_0 and the mixture $(1-\epsilon)P_0 + \epsilon P_1$ in a neighborhood of $\epsilon = 0$. We let $L = \frac{dP_1}{dP_0}$ be the likelihood ratio between P_1 and P_0, so that

$$d_{\rm hel}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) = \frac{1}{2}\mathbb{E}_0\left[(1 - \sqrt{(1-\epsilon) + \epsilon L})^2\right] = 1 - \mathbb{E}_0\left[\sqrt{1 - \epsilon + \epsilon L}\right], \tag{12.3.17}$$

and controlling these expectations gives our results. In cases where L is typically negligible relative to $1/\epsilon$, the Taylor expansion of $\sqrt{1-\epsilon+\epsilon L} = 1 + \frac{1}{2}\epsilon(L-1) + O(\epsilon^2(L-1)^2)$ makes control easy whenever $\epsilon^2(L-1)^2 \to 0$ in some uniform sense. Otherwise, because $\sqrt{\cdot}$ has highly non-uniform Taylor expansions, these types of analyses require more analytical care.

We can now instantiate the general setting (12.3.13) with a few different variants. We begin by considering a pure p-value setting, in alignment with our treatment in Section 12.3.1, where we consider nulls $P_0 = \mathsf{Uniform}[0,1]$ and $P_1 = \mathsf{Uniform}[0,\tau]$ or $(1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]$, exactly in definition (12.3.7). In either case, under the alternative P_1, we ought to see more significant (smaller) p-values. We expect that as $\tau \downarrow 0$, detection should become easier—we ought to see too many small p-values—and the next two examples make this quantitative.

Example 12.3.7 (A p-value detection problem)**:** As in the multiple testing scaling (12.3.9), consider nulls and alternatives

$$P_0 = \mathsf{Uniform}([0,1]) \quad \text{and} \quad P_1 = \mathsf{Uniform}([0,\tau]), \quad \text{with} \quad \tau = n^{-r},$$

where r indicates the "signal strength," with larger r decreasing the threshold τ. We will show that for sparsity levels $\frac{1}{2} \le \beta < 1$, the value $r^*(\beta) = 2\beta - 1$ provides a precise phase transition between impossible and possible detection:

$$\lim_{n\to\infty} \inf_{\Psi} R_n(\Psi \mid H_0, H_1) = \begin{cases} 1 & \text{if } r < r^*(\beta) = 2\beta - 1 \\ 0 & \text{if } r > r^*(\beta) = 2\beta - 1. \end{cases} \tag{12.3.18}$$

So testing is possible if and only if the signal for the non-null p-values is suitably strong, that is, they have support $[0, n^{-r}]$ for some $r > 2\beta - 1$.

Let $P_\epsilon = (1-\epsilon)P_0 + \epsilon P_1$. Then direct calculation gives

$$d_{\rm hel}^2(P_0, P_\epsilon) = 1 - \tau\sqrt{1 - \epsilon + \epsilon/\tau} - (1-\tau)\sqrt{1-\epsilon}. \tag{12.3.19}$$

If $r < \beta$, then $\epsilon/\tau \to 0$ as $n \to \infty$, and a second-order expansion yields

$$d_{\rm hel}^2(P_0, P_\epsilon) = \frac{1 + o(1)}{8}\frac{\epsilon^2}{\tau} = n^{-2\beta + r + o(1)}.$$

If $r \ge \beta$, then $\tau \le \epsilon$ and equation (12.3.19) gives $d_{\rm hel}^2(P_0, P_\epsilon) \ge c\epsilon$ for a universal constant $c > 0$. Combining these estimates with (12.3.15) proves the transition at $r^*(\beta) = 2\beta - 1$. $\diamond$

In Example 12.3.7, the choice $P_1 = \mathsf{Uniform}[0,\tau]$ places no mass on $(\tau, 1]$, so observations in that region are definitively null. A full-support alternative makes this less brutal while still attaining the same (precise) asymptotics:

Example 12.3.8 (A gentler p-value detection problem)**:** Consider nulls and alternatives

$$P_0 = \mathsf{Uniform}[0,1] \quad \text{and} \quad P_1 = (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1],$$

where as in Example 12.3.7, we take $\epsilon = n^{-\beta}$ and $\tau = n^{-r}$. In this case, P_1 and P_0 are absolutely continuous, giving likelihood ratio $L = L(u) = \frac{1-\tau}{\tau}\mathbf{1}\left\{0 \le u \le \tau\right\} + \frac{\tau}{1-\tau}\mathbf{1}\left\{\tau \le u \le 1\right\}$. If $\beta > r$, then the expansion that $d_{\rm hel}^2(P_0, P_\epsilon) = \frac{1+o(1)}{8}\frac{\epsilon^2}{\tau}$ implicit in Example 12.3.7 still holds, while if $\beta \le r$, we can still show that $d_{\rm hel}^2(P_0, P_\epsilon) \gtrsim \epsilon = n^{-\beta} \gg 1/n$. (See Exercise 12.10.) The critical threshold $r^*(\beta) = 2\beta - 1$ thus still provides the phase transition between perfect detectability and impossibility in the limits (12.3.18). $\diamsuit$

Examples 12.3.7 and 12.3.8 show that in situations in which a small proportion of p-values in a multiple hypothesis test come from non-null distributions, so long as they concentrate near enough to 0, we can detect them: the *critical threshold* $r^*(\beta) = 2\beta - 1$ guarantees that if $r < r^*(\beta)$, no procedure can reliably test the null against the alternative, while if $r > r^*(\beta)$, then the optimal test has risk (summed error probability) $\inf_\Psi R_n(\Psi \mid H_0, H_1) \to 0$.

We briefly discuss a related phenomenon in the context of detection of departure from the null for Gaussians, as in Example 12.3.4, but now we consider the variant of the scaling (12.3.9) that

$$\epsilon = n^{-\beta} \quad \text{and} \quad \mu = \sqrt{2r\log n}. \tag{12.3.20}$$

We wish to discover (and demonstrate the existence of) another critical threshold $r^*(\beta)$ such that

$$\inf_\Psi R_n(\Psi \mid H_0, H_1) \to \begin{cases} 0 & \text{if } r > r^*(\beta) \\ 1 & \text{if } r < r^*(\beta). \end{cases}$$

The derivation of this threshold is more delicate.

Theorem 12.3.9. *Consider the detection problem* (12.3.13) *with* $P_0 = \mathsf{N}(0,1)$ *and* $P_1 = \mathsf{N}(\mu, 1)$ *under the scaling* (12.3.20)*. Then the critical threshold is*

$$r^*(\beta) = \begin{cases} \left[\beta - \frac{1}{2}\right]_+ & \text{if } 0 < \beta \le \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \beta > \frac{3}{4}. \end{cases}$$

If $\beta \le \frac{1}{2}$, Corollary 12.3.6 yields the threshold $r^*(\beta) = 0 = [\beta - \frac{1}{2}]_+$, so really Theorem 12.3.9 only says something interesting when $\beta > \frac{1}{2}$. Exercise 12.16 proves the "easy" part of this theorem, that $r^*(\beta) \ge \beta - \frac{1}{2}$, while Exercise 12.17 oulines a proof of the more sophisticated threshold.

We close with a comment on adaptive achievability in Examples 12.3.7 and 12.3.8 and in Theorem 12.3.9, though much of this is beyond the scope of this book. In each of the examples and the theorem, direct calculation with Hellinger distances identifies the critical signal threshold $r^*(\beta)$, which shows (implicitly) that the likelihood ratio test has asymptotic power either 0 or 1. We seek instead a test that can adapt to the various problem parameters. For the Gaussian model, set $U_i = 1 - \Phi(Y_i)$, which is uniform under the null. More generally, because we seek departure from the null H_0 that all observed $U_i \stackrel{\rm iid}{\sim} \mathsf{Uniform}[0,1]$, but cannot (typically) identify specific departures, a reasonable idea is to look at the bulk behavior of the error

$$P_n(U \le t) - t := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{U_i \le t\right\} - t.$$

When $U_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$, the central limit theorem shows that $\sqrt{n}(P_n(U \le t) - t) \stackrel{d}{\rightsquigarrow} \mathsf{N}(0, t(1-t))$ for each t. Tukey's *Higher Criticism* seeks the largest value of this statistic,

$$T_n^{\text{HC}} := \sup_t \left\{ \sqrt{n} \frac{P_n(U \le t) - t}{\sqrt{t(1-t)}} \mid \frac{1}{n} \le t \le 1 - \frac{1}{n} \right\}. \tag{12.3.21}$$

Then under H_0, careful empirical process arguments yield $T_n^{\text{HC}} / \sqrt{2 \log \log n} \stackrel{p}{\to} 1$ (see, e.g., [379, Theorem 16.1.2]). Donoho and Jin [147] leverage this convergence to show that there is a sequence $a_n \to 1$ and a test that rejects if $T_n^{\text{HC}} \ge a_n \sqrt{2 \log \log n}$ achieving the optimal critical testing radius $r^*(\beta)$ in the Gaussian case (Theorem 12.3.9). Exercise 12.8 adjusts their arguments to show the statistic (12.3.21) adaptively achieves the optimal testing error in Examples 12.3.7 and 12.3.8.

12.4* Strong converses and high-probability lower bounds

The results we have developed thus far provide *weak* converse results: if one attempts to estimate a parameter, there is a constant probability of error for estimating to within a certain accuracy, as our reductions from estimation to testing provide lower bounds by demonstrating a constant error probability. In the information and communication theory literature, on the other hand, more nuanced results allow *strong converses*, which for our purposes we interpret as follows: if one attempts to send too many distinct messages through a communication channel, then the probability of error in decoding them necessarily approaches 1. Recall the setting of Section 9.4, where we choose $V \in \mathcal{V}$ uniformly at random, then have the Markov chain $V \to Y \to \widehat{V}$. Then letting $\epsilon = \mathbb{P}(\widehat{V} \ne V)$ denote the probability of error, Fano's inequality (Corollary 9.4.2) states that

$$\log \operatorname{card}(\mathcal{V}) \le \frac{I(V;Y)}{1-\epsilon} + \frac{\log 2}{1-\epsilon}.$$

In a communication setting, where we wish to send a message $v \in \mathcal{V}$ over a noisy channel, we interpret this result as stating that if we wish to have vanishing error $\epsilon \to 0$, then the maximum number of messages it is possible to send satisfies $\log \operatorname{card}(\mathcal{V}) \le I(V;Y) + \log 2$.

An elegant way to derive strong converses is to provide refined versions of Fano's inequality. To develop these refinements, we typically consider somewhat more specific settings than the arbitrary Markov chain $V \to Y \to \widehat{V}$. The most common scenarios involve independent sampling, where conditional on $V = v$, we draw $Y_1^n \in \mathcal{Y}^n$ from a product distribution P_v^n. Letting ϵ be some measure of the probability of error and $g(\epsilon)$ be a function for which $g(\epsilon) < \infty$ whenever $\epsilon < 1$, a typical refinement of Fano's inequality takes the form

$$\log \operatorname{card}(\mathcal{V}) \le I(V; Y_1^n) + \widetilde{O}\left(\sqrt{n} \cdot g(\epsilon)\right), \tag{12.4.1}$$

where the big-$\widetilde{O}$ notation may hide lower order or logarithmic factors.

That inequality (12.4.1) is stronger than the standard Fano inequality may not be immediately apparent, so it is useful to consider a communication scenario using a fixed channel n times. We think of V as a message from a (large) collection $\mathcal{V}$, and upon choosing $V \in v$, we encode it into a vector $x \in \mathcal{X}^n$; the ith output of the channel Y_i is then drawn independently according to $P(\cdot \mid x_i)$. Figure 12.5 shows the setting. Most commonly, we encode bit strings $x \in \{0,1\}^n$, and the channel corrupts each bit $x_i \in \{0,1\}$ independently to $Y_i \in \{0,1\}$. By the Gilbert-Varshamov packing in Lemma 9.2.4, we expect to encode exponentially many well-separated messages, and the

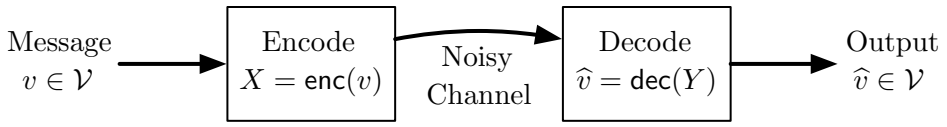


Figure 12.5: Communication over a noisy channel

information $I(V; Y_1^n)$ should grow linearly the number n of bits sent. Inequality (12.4.1) implies we *must* have $\epsilon \to 1$ whenever $\log \mathrm{card}(\mathcal{V}) - I(V; Y_1^n) \gg \sqrt{n}$, ignoring logarithmic factors.

We provide one examplar result of the form (12.4.1) for finite output spaces $\mathcal{Y}$ in the next section, then show how it implies high-probability lower bounds for different estimation problems. In particular, we will show that for many estimation problems, there is a threshold ϵ_n for which the probability of estimating to accuracy better than ϵ_n tends to zero exponentially quickly.

12.4.1 Refined Fano inequalities

We begin by revisiting Fano's inequality. Rather than repeat the proof of Proposition 2.3.3, we derive Fano's inequality from the Donsker-Varadhan variational formula (Theorem 6.1.1). This argument combines naturally with the blowing-up lemma, Corollary 7.3.2, to further refine Fano's inequality.

For the first statement, we continue to assume a Markov chain $V \to Y \to \widehat{V}$ without relying on any structure in the channel, and it is no loss of generality to assume for now that the "decoder" $\widehat{V} = \widehat{v}(Y)$ is deterministic (as we measure average probability of error).

Proposition 12.4.1. *Let V be uniform on a set of size M and $\epsilon = \mathbb{P}(\widehat{v}(Y) \neq V)$ be the average error of the estimator $\widehat{v}$. Then*

$$(1 - \epsilon) \log(M + 1) \le I(V; Y) + \log 2.$$

Proof When V is uniform, conditional on $V = v$, we draw $Y \sim P_v$. Let $Q = \frac{1}{M} \sum_v P_v$ be the marginal distribution on Y. Then by the definition of the mutual information and the Donsker-Varadhan variational inequality (Theorem 6.1.1), for any functions g_v we have

$$\begin{aligned} I(V; Y) = \frac{1}{M} \sum_v D_{\rm kl}\left(P_v \| Q\right) &\ge \frac{1}{M} \sum_v \mathbb{E}_v[g_v] - \frac{1}{M} \sum_v \log \mathbb{E}_Q[e^{g_v}] \\ &\ge \frac{1}{M} \sum_v \mathbb{E}_v[g_v] - \log \left(\frac{1}{M} \sum_v \mathbb{E}_Q[e^{g_v}] \right), \end{aligned} \tag{12.4.2}$$

where the final inequality uses the concavity of the logarithm and $\mathbb{E}_v$ denotes expectation with respect to P_v. Now we make a judicious choice of the function g_v. Let $A_v = \{y \in \mathcal{Y} \mid \widehat{v}(y) = v\}$, so that the sets A_v partition $\mathcal{Y}$, and for a $\delta > 0$ to be chosen, define

$$g_v(y) = \log(\mathbf{1}\left\{y \in A_v\right\} + \delta).$$

Use the shorthands $p_v = P_v(A_v)$ and $q_v = Q(A_v)$, so that $\sum_v q_v = 1$. Then $\mathbb{E}_v[g_v] = p_v \log(1 + \delta) + (1 - p_v) \log \delta$, while $\mathbb{E}_Q[e^{g_v}] = \delta + q_v$, and so

$$I(V; Y) \ge \frac{1}{M} \sum_v \left[p_v \log(1 + \delta) - (1 - p_v) \log \frac{1}{\delta} \right] + \log M - \log\left(1 + M\delta\right).$$

Finally, we use the definition of the average error $\epsilon = \frac{1}{M}\sum_v (1 - p_v)$, and so

$$I(V;Y) \geq (1-\epsilon)\log(1+\delta) - \epsilon \log \frac{1}{\delta} + \log M - \log(1+\delta M).$$

Choose $\delta = \frac{1}{M}$ and simplify. □

Proposition 12.4.1 recovers (with the ever-so-slightly stronger quantity $\log(M+1)$ instead of $\log M$) the initial (weak) Fano inequality in Corollary 9.4.2. The next theorem extends this result to independent sampling: conditional on $V = v$, we draw a vector $Y_1^n \sim P_v^n$, a product measure whose coordinates need not be i.i.d. As in the strong converse for hypothesis testing (Proposition 7.3.5), we provide bounds on the probabilities of "enlarged" subsets of a partition $\mathcal{Y}^n$. In this variant, we control the *maximal* error rather than the average error present in the weaker Fano inequalities, and we now require the estimator $\widehat{v}$ to be deterministic.

Theorem 12.4.2. *Let V be uniform on a set of size M, and assume conditional on $V = v$, we draw $Y_i \sim P_{v,i}$ independently on the finite alphabet $\mathcal{Y}$. Let $\epsilon = \max_v \mathbb{P}(\widehat{v}(Y_1^n) \neq v \mid V = v)$ be the maximal error of the estimator $\widehat{v}$. Then for all $t \geq 0$,*

$$(1 - e^{-t^2})\log(M+1) \leq I(V;Y_1^n) + \left[2 + \sqrt{\frac{n}{2}}\left(t + \sqrt{\log \frac{1}{1-\epsilon}}\right)\right] \log(2n \operatorname{card}(\mathcal{Y})). \tag{12.4.3}$$

Proof Let $P_v^n = P_{v,1} \otimes \cdots \otimes P_{v,n}$ for shorthand, and let $A_v = \{y \in \mathcal{Y}^n : \widehat{v}(y) = v\}$ be the acceptance regions, which partition $\mathcal{Y}^n$. Set

$$s := \sqrt{\frac{n}{2}}\left(t + \sqrt{\log \frac{1}{1-\epsilon}}\right).$$

If $s \geq n$, the result follows from $M \leq \operatorname{card}(\mathcal{Y})^n$ when $\epsilon < 1$ (and is trivial when $\epsilon = 1$). We therefore take the integer radius $\rho = \lceil s \rceil \leq n$ and define the enlargements $A_{v,\rho} = \{y : d_{\rm ham}(y, A_v) \leq \rho\}$.

Letting $Q = \frac{1}{M}\sum_v P_v^n$ be the marginal distribution of Y_1^n, define $p_v^\rho = P_v^n(A_{v,\rho})$ and $q_v^\rho = Q(A_{v,\rho})$. Define $g_v(y) = \log(\mathbf{1}\{y \in A_{v,\rho}\} + \delta)$. Then because $\mathbb{E}_v[g_v] = p_v^\rho \log(1+\delta) - (1-p_v^\rho)\log\frac{1}{\delta}$ and $\mathbb{E}_Q[e^{g_v}] = q_v^\rho + \delta$, we may apply the variational information bound (12.4.2) and the choice $\delta = 1/M$ to obtain

$$\begin{aligned} I(V;Y_1^n) &\geq \frac{1}{M}\sum_v \left[p_v^\rho \log(1+\delta) - (1-p_v^\rho)\log\frac{1}{\delta}\right] - \log\left(\frac{1}{M}\sum_v q_v^\rho + \delta\right) \\ &= (1-\bar{\epsilon}_\rho)\log(M+1) - \log\left(1 + \sum_v q_v^\rho\right), \end{aligned} \tag{12.4.4}$$

where we define $\bar{\epsilon}_\rho := \frac{1}{M}\sum_v (1 - p_v^\rho)$. A point $y \in \mathcal{Y}^n$ belongs to at most

$$B_\rho := \sum_{j=0}^{\rho} \binom{n}{j} (\operatorname{card}(\mathcal{Y}) - 1)^j \leq (2n \operatorname{card}(\mathcal{Y}))^\rho$$

enlargements $A_{v,\rho}$, because A_v partition $\mathcal{Y}^n$. So by a trivial counting argument, $\sum_v q_v^\rho \leq B_\rho$.

Finally, by assumption in the theorem, $P_v^{(n)}(A_v) \geq 1 - \epsilon$ for each v. Corollary 7.2.5 shows that each product law P_v^n satisfies a Hamming transportation (T_1) inequality, so the blowing-up

lemma, Theorem 7.3.1, therefore gives $p_v^\rho \ge 1 - e^{-t^2}$ and $\bar{\epsilon}_\rho \le e^{-t^2}$. Substituting in (12.4.4), using $1 + B_\rho \le (2n \operatorname{card}(\mathcal{Y}))^{\rho+1}$, and noting $\rho + 1 \le s + 2$ proves inequality (12.4.3). □

The cardinality of $\mathcal{Y}$ enters Theorem 12.4.2 only logarithmically. In particular, if $n \ge \max\{2, \operatorname{card}(\mathcal{Y})\}$ and $t \ge \sqrt{1/n}$, there is a universal (numerical) constant $c < \infty$ such that

$$(1 - e^{-t^2}) \log(M + 1) \le I(V; Y_1^n) + c\sqrt{n} \left(t + \sqrt{\log \frac{1}{1 - \epsilon}} \right) \log n. \tag{12.4.5}$$

Typical applications of Theorem 12.4.2 (or inequality (12.4.5)) arise in *channel coding*, where one wishes to send messages over a noisy channel, and asks for the minimal probability of error. Attempting to send too many messages through the channel invariably yields errors (with high probability); see the bibliographic section and Exercise 12.6 for some material in this direction.

We can rearrange inequality (12.4.5) to give a direct error lower bound.

Corollary 12.4.3. *Let the conditions of Theorem 12.4.2 hold, and assume $M \ge 2$ and $n \ge \max\{2, \operatorname{card}(\mathcal{Y})\}$. For the universal constant c in (12.4.5), define*

$$\gamma_n := \frac{\log(M + 1) - I(V; Y_1^n) - 1 - c \log n}{c\sqrt{n} \log n} - \sqrt{2 \log \log(M + 1)}.$$

Then the maximal probability of error ϵ satisfies

$$\epsilon := \max_v P_v(\widehat{V} \ne v) \ge 1 - \exp(-[\gamma_n]_+^2).$$

Proof We make the choice $t = \sqrt{2 \log \log M}$ in the bound (12.4.5). Then $e^{-t^2} \log(M + 1) = \frac{1}{\log(M+1)} \le 1$. Rearranging yields $\sqrt{\log(1/(1 - \epsilon))} \ge \gamma_n$, and solving for ϵ proves the claim. □

Theorem 12.4.2 relies on the product structure of the output space $\mathcal{Y}^n$, so that we can apply isoperimetric-type inequalities; roughly, the acceptance region A_v being large enough means that the enlargements A_{vr} necessarily overlap a large amount even for small $r > 0$, so that it is impossible to pack more "messages" through the channel. Extensions of the result beyond finiteness of the output space $\mathcal{Y}$ or non-i.i.d. channels exist, though their proofs rely on reverse hypercontractivity of Markov semigroups and related functional inequalities and so are beyond our scope; Liu et al. [292, Theorem 3.2] discuss these ideas.

12.4.2 High probability estimation lower bounds

Connecting the refinement of Fano's inequality in Theorem 12.4.2 to estimation problems provides new insights into the fundamental limits of estimation. As in our development of the local Fano method in Sec. 9.4, we will work with parametric problems for which $D_{\rm kl}\left(P_\theta \| P_{\theta'}\right) \lesssim \|\theta - \theta'\|_2^2$, but we will require slightly more delicate control. We consider probabilistic models for predicting a target y from a covariate vector $x \in \mathcal{X} \subset \mathbb{R}^d$, so that

$$Y \mid X = x \sim P_\theta(\cdot \mid x).$$

We say that the model has κ *quadratic information bound* if for each $x \in \mathcal{X}$, there exists a distribution $P_0(\cdot \mid x)$ on Y such that

$$D_{\rm kl}\left(P_\theta(\cdot \mid x) \| P_0(\cdot \mid x)\right) \le \kappa^2 (x^\top \theta)^2, \tag{12.4.6}$$

refining the local quadratic bound (9.4.6). Typical generalized linear models satisfy such bounds; concretely, Example 3.4.7 in Section 3.4.2 shows that for binary logistic regression of a label $y \in \{0,1\}$, the null model P_0 that $Y \sim \mathsf{Uniform}\{0,1\}$ satisfies

$$D_{\rm kl}\left(P_\theta(\cdot \mid x) \| P_0(\cdot \mid x)\right) \le \min\left\{\frac{1}{8}(x^\top\theta)^2, \log 2\right\}.$$

The key is that, for any model with a quadratic information bound (12.4.6), we can upper bound the information in estimation problems. We proceed conditionally on the covariates $X_1^n \in \mathcal{X}^n$, and take our packing set $\mathcal{V} \subset \{-1,1\}^d$.

Theorem 12.4.4. *Fix a covariate matrix $X = [x_1 \ \cdots \ x_n]^\top \in \mathbb{R}^{n\times d}$, assume Y takes values in a set $\mathcal{Y}$ with $\mathrm{card}(\mathcal{Y}) \ge 2$, and assume the model P_θ satisfies the quadratic information bound (12.4.6) for each x_i. Then there exist numerical constants $0 < c, C < \infty$ such that the following holds. For $\delta \ge 0$, define*

$$\gamma(\delta) := \frac{cd - C\kappa^2\delta^2 \left\|X\right\|_{\rm Fr}^2}{\sqrt{n}\log n} - C\sqrt{\log d}.$$

If $n \ge \mathrm{card}(\mathcal{Y})$, there exists θ with $\left\|\theta\right\|_2 \le \delta\sqrt{d}$ such that

$$\mathbb{P}_\theta\left(\left\|\widehat{\theta}(Y_1^n) - \theta\right\|_2 \ge \frac{1}{2}\delta\sqrt{d}\right) \ge 1 - \exp(-\left[\gamma(\delta)\right]_+^2).$$

Unpacking Theorem 12.4.4, let us assume the covariates are standardized to $x_i \in \{\pm 1\}^d$, so that $\left\|X\right\|_{\rm Fr}^2 = nd$. Then taking $\delta^2 = c_0/(n\kappa^2)$ for a sufficiently small universal $c_0 > 0$ gives

$$\gamma(\delta) \ge \frac{cd}{C\sqrt{n}\log n} - C\sqrt{\log d}.$$

We thus obtain the following corollary, which shows that there is essentially no probability that an estimator can achieve accuracy better than $O(1)\kappa^{-1}\sqrt{d/n}$ when the dimension scales.

Corollary 12.4.5. *In addition to the conditions of Theorem 12.4.4, assume the covariates $x_i \in [-1,1]^d$ and the dimension $d \ge \sqrt{n}\log^2 n$. Then there exists a numerical constant $c > 0$ such that*

$$\sup_{\left\|\theta\right\|_2 \le \kappa\sqrt{d/n}} \mathbb{P}_\theta\left(\left\|\widehat{\theta}_n - \theta\right\|_2 \ge \frac{c}{\kappa}\sqrt{\frac{d}{n}}\right) \ge 1 - n^{-c\log n}.$$

It is instructive to compare this bound to those available from the techniques in Chapter 9. As a particular example, let us consider binary logistic regression:

Example 12.4.6 (A strong converse in binary logistic regression)**:** Let the matrix $X = [x_1 \ \cdots \ x_n]^\top$ have rows with norm $\left\|x_i\right\|_2 = \sqrt{d}$, and consider the logistic regression model on $Y \in \{0,1\}$ that $p_\theta(y \mid x) = \exp(yx^\top\theta)/(1 + e^{x^\top\theta})$. Inequality (3.4.3) in Example 3.4.7 shows that for any parameters θ, θ', we have $D_{\rm kl}\left(P_\theta(\cdot \mid x) \| P_{\theta'}(\cdot \mid x)\right) \le \frac{1}{8}(x^\top(\theta - \theta'))^2$. Applying the local Fano method (Corollary 9.4.7) to $\mathcal{V} = \{-1,1\}^d$,

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_{\delta v}\left(\left\|\widehat{\theta}(Y_1^n) - \theta\right\|_2 \ge c\delta\sqrt{d}\right) \ge 1 - C\frac{\left\|X\right\|_{\rm Fr}^2\delta^2}{d}$$

for $\delta \le \frac{1}{2}$ and n, d large enough, where $0 < c, C < \infty$ are numerical constants. Taking $\delta^2 \asymp \frac{d}{\|X\|_{\rm Fr}^2} = \frac{1}{n}$ then gives that for some numerical constant $c > 0$,

$$\sup_{\|\theta\|_2 \le \sqrt{d/n}} P_\theta \left(\|\widehat{\theta}(Y_1^n) - \theta\|_2 \ge c\sqrt{d/n} \right) \ge \frac{1}{2}.$$

By contrast, Theorem 12.4.4 shows that in the same setting,

$$\sup_{\|\theta\|_2 \le \sqrt{d/n}} P_\theta \left(\|\widehat{\theta}(Y_1^n) - \theta\|_2 \ge c\sqrt{d/n} \right) \ge 1 - n^{-c \log n}$$

once the dimension satisfies $d \ge \sqrt{n} \log^2 n$, that is, the lower bound holds with (exceedingly) high probability. $\diamond$

12.4.3 Proof of Theorem 12.4.4

As in our discussion of the local Fano method, we will construct a packing set $\mathcal{V} \subset \{-1, 1\}^d$, and for each $v \in \mathcal{V}$ identify $\theta_v = \delta v$ for some $\delta > 0$ to be chosen.

The key for us will be to construct a packing that has *both* strong separation and for which $V \sim \mathsf{Uniform}(\mathcal{V})$ has small second moment matrix, in particular, satisfying $\mathbb{E}[VV^\top] \preceq (1 + o(1))I_d$. The probabilistic method allows this:

Lemma 12.4.7. *There exist numerical constants $0 < c$ and $C < \infty$ such that a packing $\mathcal{V} \subset \{-1, 1\}^d$ of the hypercube exists with the following properties: its cardinality satisfies* $\mathrm{card}(\mathcal{V}) \ge \exp(cd)$,

$$\left\|v - v'\right\|_2 > \sqrt{d} \text{ for } v \ne v' \in \mathcal{V}, \quad \text{and} \quad \mathbb{E}[VV^\top] \preceq \left(1 + C\sqrt{d/e^{cd}}\right) I_d$$

for $V \sim \mathsf{Uniform}(\mathcal{V})$.

We defer the proof of the lemma temporarily, as it follows by straightforward application of the concentration guarantees we have already developed.

Consider the refined Fano bound in Theorem 12.4.2. Let $\mathcal{V}$ be the packing Lemma 12.4.7 guarantees, and for a $\delta \ge 0$ to be chosen, set $\theta_v = \delta v$ for each $v \in \mathcal{V}$. Let $V \sim \mathsf{Uniform}(\mathcal{V})$, and conditional on $V = v$, draw

$$Y_i \stackrel{\rm ind}{\sim} P_{\theta_v}(\cdot \mid X = x_i).$$

Let $\widehat{\theta}(Y_1^n)$ be any estimator of θ. Because θ_v is well-separated, with $\|\theta_v - \theta_{v'}\|_2 > \delta\sqrt{d}$ we see that if $\|\widehat{\theta} - \theta_v\|_2 \le \frac{\delta}{2}\sqrt{d}$, then $\|\widehat{\theta} - \theta_{v'}\|_2 > \frac{\delta}{2}\sqrt{d}$ for all $v \ne v'$. Thus, the test

$$\Psi(Y_1^n) := \begin{cases} v & \text{if } \|\widehat{\theta} - \theta_v\|_2 \le \frac{\delta}{2}\sqrt{d} \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

satisfies

$$\mathbb{P}(\Psi(Y_1^n) \ne v \mid V = v) \le \mathbb{P}\left(\left\|\widehat{\theta} - \theta_v\right\|_2 > \frac{\delta}{2}\sqrt{d} \mid V = v \right).$$

Corollary 12.4.3 implies that the maximal probability of error $\epsilon := \max_v \mathbb{P}(\|\widehat{\theta} - \theta_v\|_2 > \frac{\delta}{2}\sqrt{d} \mid V = v)$ satisfies $\epsilon \ge 1 - \exp(-[\gamma]_+^2)$ for

$$\gamma := \frac{\log(M+1) - I(V; Y_1^n) - 1}{c\sqrt{n}\log n} - \frac{1}{\sqrt{n}} - \sqrt{2\log\log(M+1)},$$

where $2^d \ge M \ge \exp(c_0 d)$ and $0 < c_0, c < \infty$ are numerical constants.

For the final step, we upper bound the mutual information. Letting $\overline{P^n} = \frac{1}{\mathrm{card}(\mathcal{V})} \sum_{v \in \mathcal{V}} P_v^n$ and P_0^n be any other product distribution, we obtain

$$I(V; Y_1^n) = \frac{1}{\mathrm{card}(\mathcal{V})} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v^n \| \overline{P^n}\right) \le \frac{1}{\mathrm{card}(\mathcal{V})} \sum_{v \in \mathcal{V}} D_{\rm kl}\left(P_v^n \| P_0^n\right).$$

Taking P_0 to be the assumed conditional distribution providing the quadratic bound (12.4.6), we have $D_{\rm kl}\left(P_v^n \| P_0^n\right) \le \kappa^2 \sum_{i=1}^n (x_i^\top \theta_v)^2$, and so using $\theta_v = \delta v$,

$$I(V; Y_1^n) \le \kappa^2 \delta^2 \mathbb{E}\left[\sum_{i=1}^n (x_i^\top V)^2\right] \le \kappa^2 \delta^2 \left(1 + C\sqrt{d/e^{cd}}\right) \sum_{i=1}^n \|x_i\|_2^2$$

by Lemma 12.4.7. Substituting into the definition of γ proves the theorem.

Proof of Lemma 12.4.7 Let the vectors $U_i \stackrel{\rm iid}{\sim} \mathsf{Uniform}\{\pm 1\}^d$ be uniform on the binary hypercube, $i = 1, \ldots, N$, where we will choose N presently. For $i \ne j$ $\|U_i - U_j\|_2^2 = 2d - 2\langle U_i, U_j\rangle$, and $\langle U_i, U_j\rangle \stackrel{\rm dist}{=} \langle \mathbf{1}_d, U_i\rangle$ by independence. So for $t \ge 0$,

$$\mathbb{P}(\|U_i - U_j\|_2^2 \le 2d - 2t) = \mathbb{P}(\langle \mathbf{1}, U_i\rangle \ge t) \le \exp\left(-\frac{t^2}{2d}\right)$$

by standard sub-Gaussian concentration. The union bound implies

$$\mathbb{P}\left(\min_{i \ne j} \|U_i - U_j\|_2^2 \le 2d - 2t\right) \le \binom{N}{2} \exp\left(-\frac{t^2}{2d}\right) \le \exp\left(-\frac{t^2}{2d} + 2\log N\right),$$

and the choice $t = d/2$ gives $\mathbb{P}(\min_{i \ne j} \|U_i - U_j\|_2^2 \le d) \le \exp(-d/8 + 2\log N)$. Taking $N = \exp(\frac{d}{16} - \frac{1}{2})$ gives that $\|U_i - U_j\|_2 > \sqrt{d}$ for all $i \ne j$ with probability at least $1 - 1/e$.

To obtain a covariance bound, we use either Proposition 5.1.11 or Proposition 7.1.5, which for $N \ge \exp(cd)$ guarantees that $\frac{1}{N}\sum_{i=1}^N U_i U_i^\top \preceq (1 + C\sqrt{(d+t)/N}) I_d$ with probability at least $1 - e^{-t}$. □

12.5 Bibliography and further reading

Nonparametric estimation has an extensive literature. Early work often studied density estimation rather than regression, and Assouad's original paper [32] developed lower bounds in that setting. The regression arguments are sufficiently similar that we focus on them here. Èlizbar Nadaraya [179] and Watson [429] independently proposed the weighted kernel smoothing estimator (12.1.3) in 1964, and since then, smoothing methods and nonparametric estimators have enjoyed substantial attention in statistics and machine learning. Wahba [420] extended the ideas of smoothing to introduce reproducing kernel Hilbert space methods [63, 121], of which the smoothing estimators here are special cases. Tsybakov [401] gives a broad introduction to nonparametric estimation and minimax optimality. The monograph of Györfi et al. [225] treats nonparametric regression in depth; its Chapter 3 derives the dimension-dependent lower bounds (12.1.6). The book of Hastie et al. [230] includes a methodological perspective on smoothing and function estimation.

Our development of the global Fano method in Section 12.2 derives from the elegant paper of Yang and Barron [433]. The quantity (12.2.3), and its i.i.d. analogue in Corollary 12.2.2, is known

as the *index of resolvability*, and it controls estimation rates and redundancy of coding schemes for unknown distributions in a variety of scenarios; see, for example, Barron [45] and Barron and Cover [46]. It is also similar to notions of complexity in Dudley's entropy integral (cf. Dudley [164]) in empirical process theory, where the fluctuations of an empirical process are governed by a tradeoff between covering number and approximation of individual terms in the process.

Results on multiple testing, p-values, and their optimality could span several volumes alone, and we have only scratched the surface here. Statisticians have long argued for controlling multiplicity in experiments (as we also note in Chapter 6), with classical procedures for controlling the family-wise error rate (12.3.3) typically variants of the Bonferroni correction, e.g., the Holm method [239]. Benjamini and Hochberg [57], followed by Benjamini and Yekutieli [58], proposed methods for controlling the false discovery proportion and rate (12.3.4), from which a substantial literature developed. This includes Bayesian interpretations (see, for example, the book of Efron [177]), empirical process approaches [388], online and sequential control of false discoveries [197, 254], other filtering methods such as Barber and Candès's knockoffs [42], and more recent work on E-values (see the survey by Ramdas and Wang [352]). Early work on optimal sparse detection [143] takes motivation from signal processing. Ingster [250] and Donoho and Jin [147] develop important theory around sparse detection in Gaussian scenarios (e.g., those in Theorem 12.3.9), which others have extended to more general distributions [97]. More broadly, detection of the presence of combinatorial objects—even without identifying them—has inspired substantial work [7, 22]. In the context of estimation, Abramovich et al. [2] demonstrate various adaptive optimality properties of FDR-controlling procedures similar to the Benjamini-Hochberg rule (12.3.5).

Our development of strong converses follows the exposition in the paper [292]. The standard reference for these is the book by Csiszár and Körner [124], which develops information theory from first principles while providing numerous strong converse results. (See also Polyanskiy and Wu [346].) The intuition to use isoperimetric-type inequalities, especially the blowing-up lemma (Corollary 7.3.2), in the proof of strong converse results descends from work of Margulis [301], which Ahlswede et al. [12, 14] leverage to develop converses. A recent line of work has exploited these results to develop phase transition results in many "structured" estimation problems, where below a critical sample size, it is impossible to estimate almost anything about a parameter of interest, while above it, one can estimate almost perfectly. Such results are available, for example, in sparse regression, group testing, and other problems with some type of "combinatorial" core [328, 204, 116]. The types of phase transitions we develop in Sections 12.3 and 12.4* are information-theoretic, and there is a growing body of work suggesting analogous computational phase transitions; the corresponding lower bounds are generally conditional on conjectured hard problems. These information-computation gaps exhibit distinct signal-to-noise thresholds: estimation may be information-theoretically possible below the level required by known computationally efficient algorithms [e.g. 65, 87, 372].

12.6 Exercises

Exercise 12.1 (Lower bounds for estimating smooth functions)**:** Use an argument paralleling that for Theorem 12.1.4 to demonstrate the lower bound (12.1.5). *Hint.* Consider powers of the bump function $g(x) = [1/2 - |x - 1/2|]_+$ and recognize that $g^{(\beta-1)}$ is Lipschitz if and only if $\sup_x |g^{(\beta)}(x)| \le 1$.

Exercise 12.2 (The curse of dimensionality in nonparametric regression)**:** Consider the nonparametric regression problem in Section 12.1. Let $\mathbb{B}^d$ be the unit ℓ_2-ball in $\mathbb{R}^d$ and consider the

function class $\mathcal{F}$ of 1-Lipschitz functions taking values in $[-1,1]$ on $\mathbb{B}^d$, and consider the error $\|f-g\|_2^2 = \int_{\mathbb{B}^d}(f(x)-g(x))^2dx$. (Here, 1-Lipschitz means $|f(x)-f(x')| \le \|x-x'\|_2$ for any x, x'.) We show the minimax lower bound (12.1.6) for this function class using Fano's method. Fix $\delta \in [0,1]$ to be chosen and let $\{x_j\}_{j=1}^M$ be the centers of a maximal 2δ-packing of $\mathbb{B}^d$, so that $M \ge (\frac{1}{2\delta})^d$ (by Lemma 5.1.10), and define the "bump" functions

$$g_j(x) = \delta\left[1 - \|x - x_j\|_2 / \delta\right]_+,$$

whose supports have disjoint interiors. Then for a vector $v \in \{\pm 1\}^M$, define

$$f_v(x) := \sum_{j=1}^M v_j g_j(x).$$

(a) Show that $f_v \in \mathcal{F}$.

(b) Show that $\int g_j(x)^2 dx = \frac{2 \cdot \mathrm{SA}(d)}{d(d+1)(d+2)}\delta^{2+d}$, where $\mathrm{SA}(d)$ denotes the surface area of $\mathbb{S}^{d-1}$.

(c) Use Lemma 9.2.4 to show there is a collection $\mathcal{V} \subset \{\pm 1\}^M$ of cardinality $\exp(M/8)$ with $\|f_v - f_{v'}\|_2^2 \ge c_d \delta^2$ for all $v \ne v' \in \mathcal{V}$, where c_d depends only on the dimension d.

(d) Prove the minimax lower bound (12.1.6) for $\beta = 1$.

Exercise 12.3 (Compactly supported $\mathcal{C}^\infty$ functions)**:**

(a) Let $f(x) = \exp(-1/x)$ for $x > 0$, and $f(x) = 0$ otherwise. Show that f is $\mathcal{C}^\infty$.

(b) Construct a $\mathcal{C}^\infty$ function K so that $K(x) > 0$ for $x \in (-1,1)$ and $K(x) = 0$ otherwise.

Exercise 12.4 (Ellipsoid classes and covering numbers [316])**:** For non-decreasing scalars $0 < b_1 \le b_2 \le \ldots$, define the *ellipsoids*

$$\mathcal{E}_b^n := \left\{\theta \in \mathbb{R}^n \mid \sum_{j=1}^n b_j^2\theta_j^2 \le 1\right\} \text{ and } \mathcal{E}_b := \left\{\theta \in \mathbb{R}^{\mathbb{N}} \mid \sum_{j \ge 1} b_j^2\theta_j^2 \le 1\right\}.$$

We evaluate the covering and packing numbers of $\mathcal{E}_b$ with respect to the $\|\cdot\|_2$-norm.

(a) For $t \ge 0$, define $m_b(t) := \sup\{i \mid b_i \le t\}$ to be the largest index i with $b_i \le t$ (taking $m_b(t) = 0$ if no such index exists). Show that if $m > m_b(2/\epsilon)$, then for $\theta, \theta' \in \mathcal{E}_b$,

$$\sum_{j=1}^m (\theta_j - \theta_j')^2 - \epsilon^2 \le \left\|\theta - \theta'\right\|_2^2 \le \sum_{j=1}^m (\theta_j - \theta_j')^2 + \epsilon^2$$

(b) Show by a volume argument (as in Lemma 5.1.10) that the covering number $N(\epsilon) = N(\epsilon, \mathcal{E}_b^m, \|\cdot\|_2)$ of $\mathcal{E}_b^m$ in the ℓ_2-norm satisfies

$$m \log \frac{1}{\epsilon} - \sum_{i=1}^m \log b_i \le \log N(\epsilon) \le m \log \frac{1}{\epsilon} - \sum_{i=1}^m \log b_i + m \log\left(1 + \frac{b_m \epsilon}{2}\right).$$

(c) Show that

$$\sum_{j=1}^{\infty} \left[\log \frac{1}{b_j \epsilon}\right]_+ \le \log N(\epsilon, \mathcal{E}_b, \|\cdot\|_2) \le \sum_{j=1}^{\infty} \left[\log \frac{C}{b_j \epsilon}\right]_+,$$

where $C < \infty$ is a numerical constant.

(d) Using $m_b(t) = \sup\{i \mid b_i \le t\}$ from part (a), show that for $l_b(t) := \int_0^t \frac{m_b(u)}{u} du$,

$$l_b\left(\frac{c}{\epsilon}\right) \le \log N(\epsilon, \mathcal{E}_b, \|\cdot\|_2) \le l_b\left(\frac{C}{\epsilon}\right), \tag{12.6.1}$$

where $0 < c \le C < \infty$ are numerical constants.

Exercise 12.5 (Minimax rates over ellipsoidal classes [433])**:** Let $\{\phi_i\}_{i \in \mathbb{N}}$ be a uniformly bounded orthonormal collection of functions $\phi : [0,1] \to \mathbb{R}$, satisfying $\langle \phi_i, \phi_j \rangle := \int_0^1 \phi_i(x)\phi_j(x)dx = \mathbf{1}\{i = j\}$. Let b_j be non-decreasing and define the *ellipsoidal function class* parameterized by $\theta \in \mathbb{R}^{\mathbb{N}}$ by

$$\mathcal{F}_b := \left\{ f_\theta(x) = \sum_{j \ge 1} \theta_j \phi_j(x) \mid \sum_{j \ge 1} b_j^2 \theta_j^2 \le 1 \right\} \subset L^2([0,1]).$$

We evaluate minimax rates over this class for particular parameterized choices of the vectors $b \in \mathbb{R}^{\mathbb{N}}$ given observations

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1), \quad X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1].$$

(a) Let $m_b(u) = \sup\{i \mid b_i \le u\}$ as in Exercise 12.4. Show that there is a numerical constant $c > 0$ such that if $n\epsilon^2 \le c \int_0^{c/\epsilon} \frac{m_b(u)}{u} du$, then the minimax mean-squared error over $\mathcal{F}_b$ satisfies

$$\mathfrak{M}_n(\mathcal{F}_b, \|\cdot\|_2^2) \gtrsim \epsilon^2.$$

(b) Let $b_j = j^\alpha$ for some $\alpha > 0$. Show that

$$\mathfrak{M}_n(\mathcal{F}_b, \|\cdot\|_2^2) \gtrsim n^{-\frac{2\alpha}{2\alpha+1}}.$$

(c) Define the truncation estimator

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^n Y_i \phi_j(X_i) \quad \text{for } j \le k, \quad \widehat{\theta}_j = 0 \text{ otherwise.}$$

Let $\widehat{f} = f_{\widehat{\theta}}$. Give a value of k to show that the preceding minimax lower bound is sharp.

Exercise 12.6 (A strong converse in channel coding)**:** Let $Q(\cdot \mid x)$ denote a channel with finite output space $\mathcal{Y}$, and for $Y \mid X = x \sim Q(\cdot \mid x)$, define the capacity

$$C = C(Q) := \sup_X I(X;Y),$$

where the supremum is over all distributions on the input X. Abusing that we use base-e-logarithms, call a communication scheme a *rate R communication scheme with block length n*, denoted an (R, n)

communication scheme, if for $M = e^{nR}$ messages, the communicator chooses a message $v \in [M]$, encodes it via $x_1^n = x(v) \in \mathcal{X}^n$, and then sends x_1^n through the (memoryless) channel, so that $Y_i \mid x_i \sim Q(\cdot \mid x_i)$, drawn independently, are received. Use Theorem 12.4.2 or Corollary 12.4.3 to show that if $R > C$, then for all large enough n, any decoding scheme $\widehat{v} : \mathcal{Y}^n \to [M]$ must satisfy

$$\max_v \mathbb{P}_v(\widehat{v}(Y_1^n) \neq v) \geq 1 - \exp\left(-c_n \cdot n(R - C)^2\right),$$

where $c_n \gtrsim \frac{1}{\log^2 n}$.

Exercise 12.7: Take the sampling model (12.3.13), where $\epsilon = \epsilon_n = n^{-\beta}$ for some $\beta \in (\frac{1}{2}, 1)$, and $P_1 = \mathsf{N}(\mu_n, 1)$ for $\mu_n = \sqrt{2r \log n}$.

(a) Show that

$$1 + D_{\chi^2}((1-\epsilon)P_0 + \epsilon P_1 \| P_0) = 1 - \epsilon^2 + \epsilon^2 e^{\mu_n^2}.$$

(b) Show that if $r < \beta - \frac{1}{2}$, then

$$d_{\rm hel}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) \ll 1/n.$$

Hint. Use Proposition 2.2.10.

(c) Conclude that if $\mu_n \leq \sqrt{2r \log n}$ for $r < \beta - \frac{1}{2}$, it is asymptotically impossible to test between $H_0 : P_0^n$ and $H_1 : ((1-\epsilon)P_0 + \epsilon P_1)^n$.

Exercise 12.8 (Higher criticism and sparse detection [147])**:** Tukey's *Higher Criticism* statistic (12.3.21) considers the normalized process

$$Z_n(t) := \sqrt{n}\frac{P_n(U \leq t) - t}{\sqrt{t(1-t)}} \quad \text{and} \quad T_n^{\rm HC} := \sup_t\{Z_n(t) \mid 1/n \leq t \leq 1 - 1/n\}.$$

This converges in probability: $T_n^{\rm HC}/\sqrt{2 \log \log n} \xrightarrow{p} 1$ when $U_i \overset{\rm iid}{\sim} \mathsf{Uniform}([0,1])$ (see, e.g., [379, Theorem 16.1.2]). This exercise shows how to leverage $T_n^{\rm HC}$ to adaptively attain the critical testing radius $r^*(\beta) = 2\beta - 1$ in the limit (12.3.18) in Example 12.3.7.

(a) Show that for any $\alpha \in (0,1)$, there is a sequence $a_n \to 1$ such that under the null H_0,

$$\limsup_n \mathbb{P}_{H_0}\left(T_n^{\rm HC} \geq a_n\sqrt{2 \log \log n}\right) \leq \alpha.$$

(b) Show that under the alternative H_1 that $U_i \overset{\rm iid}{\sim} (1-\epsilon)\mathsf{Uniform}[0,1] + \epsilon\mathsf{Uniform}[0,\tau]$, where $\epsilon = n^{-\beta}$ and $\tau = n^{-r}$ for some $\beta \in (\frac{1}{2}, 1)$ and $r > 2\beta - 1$, we have

$$\mathbb{P}_{H_1}\left(T_n^{\rm HC} \geq a_n\sqrt{2 \log \log n}\right) \to 1.$$

So higher criticism has optimal power to detect on the interior of the region $r > r^*(\beta) = 2\beta - 1$.

Exercise 12.9 (Dense detection is easy)**:** Consider observations X_i drawn i.i.d. from

$$H_0 : X_i \overset{\rm iid}{\sim} P_0 \quad \text{or} \quad H_1 : X_i \overset{\rm iid}{\sim} (1-\epsilon)P_0 + \epsilon P_1.$$

We investigate the rate at which we may allow $\epsilon = \epsilon_n \to 0$ while still testing accurately (to prove Corollary 12.3.6). Take $\epsilon_n = n^{-\beta}$ for some $\beta \in [0,1]$, and let R_n be the test risk (12.3.14).

(a) Let $\|P_0 - P_1\|_{\rm TV} \ge c > 0$, and assume that $\beta < \frac{1}{2}$. Let A satisfy $P_0(A) + P_1(A^c) \le 1 - c$. Construct a test Ψ_n based on $S_n := \sum_{i=1}^n \mathbf{1}\{X_i \in A\}$ that achieves $\lim_n R_n(\Psi_n \mid H_0, H_1) = 0$.

(b) Assume that as $n \to \infty$, we allow P_1 to vary so that $\|P_0 - P_1\|_{\rm TV} \to 1$, and $\beta \le \frac{1}{2}$. Show that it is possible to construct a test Ψ_n such that $R_n(\Psi_n \mid H_0, H_1) \to 0$ as $n \to \infty$.

(c) Let $P_0 = \mathsf{N}(0, 1)$ and $P_1 = \mathsf{N}(\mu, 1)$ for $\mu = \sqrt{2r \log n}$, where $r > 0$. Show that

$$\|P_0 - P_1\|_{\rm TV} = 2\Phi(\mu/2) - 1 \quad \text{and so} \quad \|P_0 - P_1\|_{\rm TV} \ge 1 - \frac{O(1)}{\sqrt{\log n}} n^{-r/4}.$$

Exercise 12.10 (A less brutal multiple testing scenario)**:** Replace the testing problem in Example 12.3.7 by the gentler alternative from Example 12.3.8, with nulls and alternatives

$$P_0 = \mathsf{Uniform}([0, 1]) \quad \text{and} \quad P_1 = (1 - \tau)\mathsf{Uniform}([0, \tau]) + \tau\mathsf{Uniform}([\tau, 1]),$$

where again $\epsilon = n^{-\beta}$ while $\epsilon \ll \tau \ll 1$. Then $U \sim P_1$ has density

$$\frac{1 - \tau}{\tau} \mathbf{1}\{0 \le u \le \tau\} + \frac{\tau}{1 - \tau} \mathbf{1}\{\tau \le u \le 1\},$$

making P_1 and P_0 absolutely continuous.

(a) Show that for $\epsilon \ll \tau$, that is, $\beta > r$, the expansion $d_{\rm hel}^2(P_0, (1 - \epsilon)P_0 + \epsilon P_1) = \frac{1 + o(1)}{8} \frac{\epsilon^2}{\tau}$ holds.

(b) Mimic Example 12.3.7 to show that $d_{\rm hel}^2(P_0, (1 - \epsilon)P_0 + \epsilon P_1) \gg 1/n$ whenever $r \ge \beta$.

(c) Show that the asymptotics in the conclusion of Example 12.3.7 still hold.

Exercise 12.11: Consider the canonical multiple hypothesis testing setting (12.3.8).

(a) Let $P_1 = \mathsf{Uniform}[0, \tau]$. Show that for each coordinate j,

$$\inf_{\widehat{v}} \mathbb{P}(V_j \ne \widehat{v}_j) = \min\{\tau(1 - \epsilon), \epsilon\}.$$

(b) Let $P_1 = (1 - \tau)\mathsf{Uniform}[0, \tau] + \tau\mathsf{Uniform}[\tau, 1]$. Show that if $\epsilon, \tau < \frac{1}{2}$, then

$$\inf_{\widehat{v}} \mathbb{P}(V_j \ne \widehat{v}_j) = \min\{\tau(1 - \epsilon), \epsilon(1 - \tau)\} + \epsilon\tau.$$

Exercise 12.12 (Bonferroni and optimal recovery)**:** Consider the canonical multiple hypothesis testing problem (12.3.8) with sparsity scaling (12.3.9), and assume $1 > r > \beta$. Let $\mathcal{I}_0 = \{i \in [n] \mid V_i = 0\}$ and $\mathcal{I}_1 = [n] \setminus \mathcal{I}_0$ be the null and non-null hypotheses, respectively. Recall the Bonferroni correction (12.3.2), and let $\mathcal{R}_{\rm bc} = \{j \in [n] \mid U_j \le \frac{\alpha}{n}\}$ be the hypotheses it rejects.

(a) Show that

$$\mathbb{E}[\mathrm{card}(\mathcal{R}_{\rm bc}^c \cap \mathcal{I}_1)] \ge \epsilon n(1 - o(1)) = n^{1-\beta}(1 - o(1)).$$

(b) Show that

$$\mathbb{E}\left[\frac{\mathrm{card}(\mathcal{R}_{\rm bc}^c \cap \mathcal{I}_1)}{\mathrm{card}(\mathcal{I}_1) + 1}\right] \ge 1 - o(1).$$

(c) Comparing with Theorem 12.3.3, what can we say about the relative merits of the Bonferroni correction and the BH procedure (12.3.5)?

Exercise 12.13 (Multiple testing and the Benjamini-Hochberg procedure)**:** Consider the Benjamini-Hochberg step-up procedure (12.3.5). Here we provide a few results that suggest a failure mode of this procedure. (Though Theorem 12.3.3 shows complementary optimality properties.)

(a) Let $U_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$. For $c \in [0, 1/n]$, define the region

$$A_c := \{x \in \mathbb{R}^n \mid 0 \le x_1 \le \cdots \le x_n \le 1, x_i \ge ci \text{ for each } i\}.$$

Letting $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ be the order statistics of $U \in [0,1]^n$, show that

$$\mathbb{P}((U_{(1)}, \ldots, U_{(n)}) \in A_c) = \mathbb{P}(U_{(1)} \ge c, U_{(2)} \ge 2c, \ldots, U_{(n)} \ge nc) = 1 - nc.$$

Hint. The density of the collection of order statistics is uniform on A_0 and satisfies $p(u_1^n) = 1/n!$.

(b) Let $\widehat{k}$ be the number of hypothesis the BH procedure (12.3.5) rejects on data $U_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$. Show that $\mathbb{P}(\widehat{k} > 0) = \alpha$.

(c) Suppose that $U_1, \ldots, U_k \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0, \tau]$, where $k = n^{1-\beta}$ and $\tau = n^{-r}$, mimicking the setting of Example 12.3.7, where $2\beta - 1 < r < \beta$. Show that if we only compute order statistics for this set of variables,

$$\mathbb{P}\left(\text{any } U_{(i)} \le \frac{\alpha i}{n}\right) \to 0.$$

That is, the BH procedure rejects no hypotheses with high probability.

Exercise 12.14: Consider the canonical multiple testing problem (12.3.8), with $\epsilon = n^{-\beta}$, in which we test n nulls and alternatives

$$H_{0,j} : Y_j \sim P_0 = \mathsf{N}(0,1) \quad \text{versus} \quad H_{1,j} : Y_j \sim P_1 = \mathsf{N}(\mu, 1), \quad \text{for } \mu = \sqrt{2r \log n}.$$

Let $R_n(\widehat{v} \mid P_0, P_1) = \sum_{j=1}^n \mathbb{P}(\widehat{v}_j(Y_1^n) \ne V_j)$ be the test risk (12.3.10).

(a) Define $t = t(\mu, \epsilon) := \frac{1}{\mu} \log \frac{1-\epsilon}{\epsilon} + \frac{1}{2}\mu$. Show that the test $\widehat{v}_j = \mathbf{1}\{Y_j > t\}$ minimizes $R_n(\cdot \mid P_0, P_1)$.

(b) Show that if $r > \beta$, then there are tests $\widehat{v} : \mathbb{R}^n \to \{0,1\}^n$ satisfying

$$R_n(\widehat{v} \mid P_0, P_1) \ll n^{1-\beta}.$$

(c) Show that if $r \le \beta$,

$$\lim_{n\to\infty} \frac{1}{n^{1-\beta}} \inf_{\widehat{v}} R_n(\widehat{v} \mid P_0, P_1) = \begin{cases} \frac{1}{2} & \text{if } r = \beta \\ 1 & \text{if } r < \beta. \end{cases}$$

(d) Conclude that if $r > \beta$, we can recover all but a vanishing fraction of the non-null hypotheses, while if $r < \beta$, no procedure obtains smaller risk than one that never rejects the null.

Exercise 12.15 (Phase diagrams)**:** Consider the multiple testing scenarios of Section 12.3.1 and the sparse p-value detection problems in Section 12.3.4. Assume that we wish to test (or detect) observations $Y_j \sim P_0$ versus $Y_j \sim P_1$, where the fraction of non-nulls scales as $\epsilon = n^{-\beta}$. Consider

i. p-values: multiple testing (12.3.8) with $P_0 = \mathsf{Uniform}[0,1]$ and $P_1 = \mathsf{Uniform}[0,\tau]$ or $P_1 = (1-\tau)\mathsf{Uniform}[0,\tau] + \tau\mathsf{Uniform}[\tau,1]$, with signal strength $\tau = n^{-r}$ for some $r > 0$.

ii. Gaussians: multiple testing (12.3.8) with $P_0 = \mathsf{N}(0,1)$ and $P_1 = \mathsf{N}(\mu,1)$, with signal strength $\mu = \sqrt{2r\log n}$ for some $r > 0$.

In the former case, as in Examples 12.3.7 and 12.3.8, define the critical threshold $r^*(\beta) = [2\beta - 1]_+$, and in the latter, as in Theorem 12.3.9, define the critical threshold $r^*(\beta) = [\beta - 1/2]_+$ for $\beta \le 3/4$ and $r^*(\beta) = (1 - \sqrt{1-\beta})^2$ for $\beta > 3/4$. Use these results, Theorem 12.3.3, and the results of Exercise 12.14, to justify the phase diagram in Figure 12.6, which depicts the following:

(a) If $r > \beta$, then it is possible to asymptotically recover all but a negligible fraction of null and non-null signals; if $r < \beta$, it is impossible.

(b) If $\beta > r > r^*(\beta)$, then it is possible to detect that there are non-null signals, but not to identify them; if $r < r^*(\beta)$, then even detection is impossible.

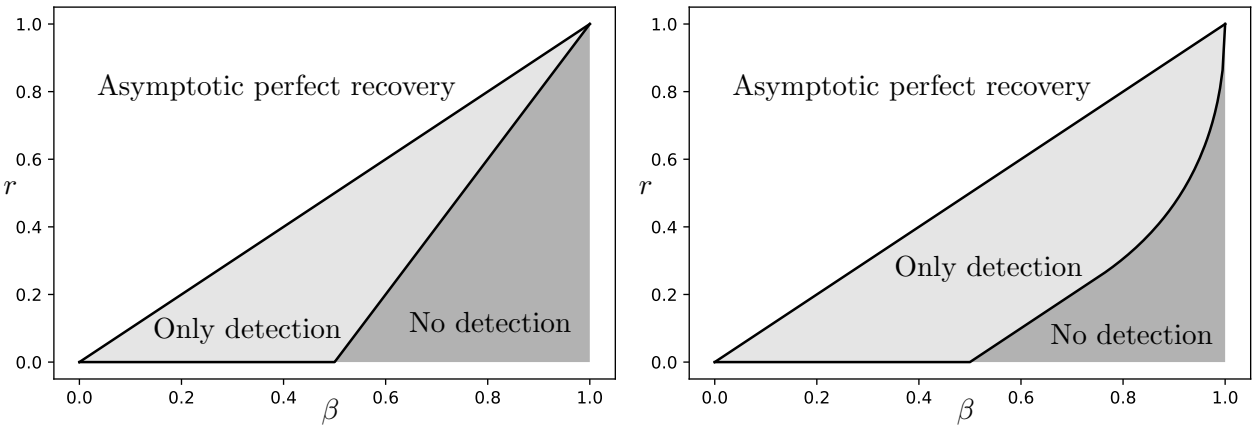


Figure 12.6. Phase diagrams for the p-value testing and detection problems (left) and the Gaussian mean testing problems (right), where the fraction of non-nulls scales as $\epsilon = n^{-\beta}$. Left: the setting $U_j \sim \mathsf{Uniform}[0,1]$ versus $U_j \sim \mathsf{Uniform}[0,\tau]$, where $\tau = n^{-r}$. Right: the setting $Y_j \sim \mathsf{N}(0,1)$ versus $Y_j \sim \mathsf{N}(\mu,1)$, where $\mu = \sqrt{2r\log n}$. The line $r = r^*(\beta)$ separates detection from non-detection, while $r = \beta$ separates detection from recovery. See Exercise 12.15.

Exercise 12.16: Follow the recipe (12.3.15) to prove $r^*(\beta) \ge \beta - \frac{1}{2}$ in Theorem 12.3.9:

(a) Explain why $-\sqrt{1+x} = -1 - \frac{x}{2} + \frac{x^2}{8(1+y)^{3/2}}$ for some y between 0 and x.

(b) Define the likelihood ratio $L = p_1/p_0$ and use the expansion (12.3.17) to obtain

$$d_{\rm hel}^2(P_0, (1-\epsilon)P_0 + \epsilon P_1) \le \frac{\epsilon^2}{8(1+o(1))}\mathbb{E}_0[(L-1)^2].$$

(c) Calculate $\mathbb{E}[(L-1)^2]$ directly to see that if $r < \beta - \frac{1}{2}$, then $\inf_\Psi R_n(\Psi \mid H_0, H_1) \to 1$.

Exercise 12.17* ([250, 147])**:** This question works through a proof of the critical threshold for sparse Gaussian detection in Theorem 12.3.9. Let $L = L(y) = \frac{p_1(y)}{p_0(y)}$ be shorthand for the likelihood ratio between $P_0 = \mathsf{N}(0,1)$ and $P_1 = \mathsf{N}(\mu,1)$, and let $P_\epsilon = (1-\epsilon)P_0 + \epsilon P_1$ be the mixture.

(a) Show that for all $x \ge -1$, $\sqrt{1+x} \le 1 + \frac{x}{2} - \frac{x^2}{8 \cdot 2^{3/2}} \mathbf{1}\left\{x \le 1\right\}$.

(b) Show that for all $x \ge -\frac{1}{2}$, $\sqrt{1+x} \ge 1 + \left(\frac{x}{2} - \frac{x^2}{4}\right) \mathbf{1}\left\{x \le 1\right\}$.

(c) Use the tighter bound $\sqrt{1+x} \le 1 + \frac{x}{2} - \frac{x^2}{12} \mathbf{1}\left\{x \le 1\right\}$ to show that

$$\frac{\epsilon^2}{12} \mathbb{E}_0[(L-1)^2 \mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}]$$
$$\le d_{\mathrm{hel}}^2(P_0, P_\epsilon) \le \frac{\epsilon}{2} \mathbb{E}_0[(L-1)\mathbf{1}\left\{L > \epsilon^{-1} + 1\right\}] + \frac{\epsilon^2}{4} \mathbb{E}_0[(L-1)^2 \mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}].$$

(d) Show that $L = L(y) \le \frac{1}{\epsilon} + 1$ if and only if $y \le C_n(\beta, r) \cdot \mu$ where $C_n(\beta, r) = \frac{1}{2} + \frac{\beta}{2r} + \frac{\log(1+n^{-\beta})}{2r \log n}$.

(e) Let $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$ be the standard normal CDF. Show that

$$\mathbb{E}_0[L\mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}] = \Phi\left((C_n(\beta, r) - 1)\mu\right) \quad \text{and}$$
$$\mathbb{E}_0[L^2\mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}] = e^{\mu^2} \Phi\left((C_n(\beta, r) - 2)\mu\right).$$

(f) Define $C(\beta, r) = \frac{1}{2} + \frac{\beta}{2r}$. Show that if $\frac{1}{2} < \beta < 1$, then for an error term $\gamma_n \to 0$,

$$\epsilon^2 \mathbb{E}_0\left[(L-1)^2 \mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}\right] = O(n^{-2\beta}) + n^{2r - 2\beta + \gamma_n} \cdot \begin{cases} n^{-r(\frac{\beta}{2r} - \frac{3}{2})^2} & \text{if } C(\beta, r) \le 2 \\ 1 & \text{if } C(\beta, r) \ge 2. \end{cases}$$

Use this to conclude that for $r^*(\beta)$ as in Theorem 12.3.9,

$$n \cdot \epsilon^2 \mathbb{E}_0\left[(L-1)^2 \mathbf{1}\left\{L \le \epsilon^{-1} + 1\right\}\right] \to \begin{cases} 0 & \text{if } r < r^*(\beta) \\ \infty & \text{if } r > r^*(\beta). \end{cases}$$

Hint. Use Mills ratio.

(g) Complete the proof of Theorem 12.3.9.

Exercise 12.18 (Detecting a sparse signal through the maximum)**:**

(a) Show that if $Z_i \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0,1)$, then $\max_{i \le n} Z_i / \sqrt{2 \log n} \stackrel{p}{\to} 1$. *Hint.* Use Mills ratio, Exercise 4.3.

(b) Show that if $Z_i \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0,1)$, then for any $\alpha \in [0,1]$, there exists a sequence $a_n = a_n(\alpha) \to 1$ for which $\mathbb{P}(\max_{i \le n} Z_i \ge a_n \sqrt{2 \log n}) \to \alpha$.

Recall the testing problem of Theorem 12.3.9 to distinguish $H_0 : Y_i \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0,1)$ from the alternative $H_1 : Y_i \stackrel{\mathrm{iid}}{\sim} (1 - \epsilon_n)\mathsf{N}(0,1) + \epsilon_n \mathsf{N}(\mu_n, 1)$, where $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$, and $\beta \in (\frac{1}{2}, 1)$.

(c) Show that if $r > (1 - \sqrt{1-\beta})^2$, then for $a_n = a_n(0) \to 1$ from part (b), the test

$$\Psi_n := \begin{cases} 0 & \text{if } \max_{i \le n} Y_i \le a_n \sqrt{2 \log n} \\ 1 & \text{if } \max_{i \le n} Y_i > a_n \sqrt{2 \log n} \end{cases}$$

satisfies $L_n(\Psi_n \mid H_0, H_1) \to 0$.

Chapter 13

Testing and functional estimation

When we wish to estimate a complete "object," such as the parameter θ in a linear regression $Y = X\theta + \varepsilon$, or a density when we observe $X_1, \ldots, X_n$ i.i.d. with a density f, the previous chapters give approaches to proving fundamental optimality results. In many cases, however, we desire to estimate *functionals* $T(P)$ of a distribution P rather than P itself. A naive estimator might construct an estimate $\widehat{P}$ of P and simply plug it in: use $\widehat{T} = T(\widehat{P})$. But this plug-in approach can be inefficient. As a trivial example, suppose we have an unknown distribution P supported on $[-1, 1]$ and desired statistic $T(P) = \mathbb{E}_P[X]$. Then the sample mean estimator

$$T_n := \overline{X}_n$$

satisfies $\mathbb{E}[(T_n - \mathbb{E}[X])^2] \le \frac{1}{n}$, but an estimator that first approximates the P via some $\widehat{P}$ and then estimates $\int x d\widehat{P}(x)$ is likely to incur substantial additional error (cf. Section 12.1).

Techniques in functional estimation are useful for testing problems as well. In these, we have a null H_0 that the observed data X follow a distribution belonging to a family $\mathcal{P}_0$, and we wish to test the closeness of P to $\mathcal{P}_0$. This reduces to estimating distance-like functionals, such as $\inf_{P_0 \in \mathcal{P}_0} \|P_0 - P\|_{\mathrm{TV}}$. In related *two-sample tests*, we are receive $X_1^n \stackrel{\mathrm{iid}}{\sim} P$ and $Y_1^m \stackrel{\mathrm{iid}}{\sim} Q$ and would like to test whether $Q = P$ or Q and P are far from one another.

Broadly, we wish to develop tools to understand the optimality of estimators and tests of scalar valued parameters of a distribution P. Such functionals could include the norm $\|\theta\|_2$ of a regression vector or the probability mass of outcomes a study fails to observe. This chapter develops tools to understand these problems.

13.1 Geometrizing rates of convergence

One-dimensional estimands often admit "geometric" characterizations of minimax error, which relate the continuity of a function to be estimated to distances between underlying probability distributions. This approach makes the intuition that estimation should be hard when the statistic of interest is sensitive to the underlying distribution quantitative. To proceed, we define the *local modulus of continuity* of a functional $\theta : \mathcal{P} \to \mathbb{R}$ on a family $\mathcal{P}$ of distributions at a fixed distribution P_0 with respect to the Hellinger distance,

$$\omega_{\mathrm{hel}}(\epsilon; \theta, P_0, \mathcal{P}) := \sup_{P_1 \in \mathcal{P}} \left\{ |\theta(P_0) - \theta(P_1)| \mid d_{\mathrm{hel}}(P_0, P_1) \le \epsilon \right\}. \tag{13.1.1}$$

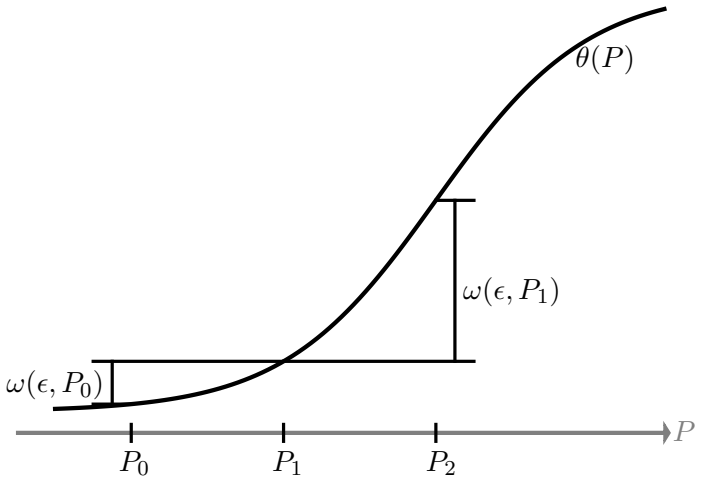


Figure 13.1. Illustration of the (Hellinger) modulus of continuity of a statistic $\theta(P)$. The modulus $\omega(\epsilon, P_0)$ at P_0 is relatively small, while the modulus $\omega(\epsilon, P_1)$ at P_1 is larger and achieved by $\theta(P_2)$.

Figure 13.1 illustrates this modulus, which we recognize as the local Lipschitz constant of the parameter θ for the Hellinger distance around a fixed distribution P_0. The *global modulus of continuity* is simply the supremum of the local modulus over all $P_0 \in \mathcal{P}$,

$$\omega_{\rm hel}(\epsilon; \theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \omega_{\rm hel}(\epsilon; \theta, P, \mathcal{P}). \tag{13.1.2}$$

These are "geometric" quantities in that each measures how much θ can vary over small neighborhoods with respect to the Hellinger distance. Using Le Cam's two point method (Chapter 9.3), we can nearly immediately see that this Hellinger modulus implies lower bounds on estimation error. In this case, we lower bound the "hardest one-dimensional sub-problem,"

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) := \sup_{P_1 \in \mathcal{P}} \inf_{\widehat{\theta}_n} \max_{P \in \{P_0, P_1\}} \mathbb{E}_{P^n}\left[\Phi(|\widehat{\theta}_n(X_1^n) - \theta(P)|)\right]. \tag{13.1.3}$$

The quantity (13.1.3) differs from the standard minimax risk in that the distribution P_0 is fixed, and nature must first choose an alternative P_1; the estimator $\widehat{\theta}_n$ then *knows* P_1 and P_0. Any lower bound on the hardest subproblem (13.1.3) immediately lower bounds the minimax risk, as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi) \geq \sup_{P_0 \in \mathcal{P}} \mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi).$$

Proposition 13.1.1. *Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be any nondecreasing loss function and $\theta : \mathcal{P} \to \mathbb{R}$. Define the sequence $\epsilon_n^2 = \frac{2-\sqrt{2}}{n}$. Then for $n \geq 2$,*

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) \geq \frac{1}{2}\left(1 - \sqrt{3/8}\right) \Phi\left(\frac{1}{2}\omega_{\rm hel}(\epsilon_n; \theta, P_0, \mathcal{P})\right).$$

Additionally, for all $c > 0$,

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) \geq \frac{1 - \sqrt{1 - e^{-c}}\sqrt{1 - e^{-2c}}}{2} \cdot \Phi\left(\frac{1}{2}\omega_{\rm hel}\left(\sqrt{\frac{c}{n}}; \theta, P_0, \mathcal{P}\right)\right) \cdot (1 - o(1)).$$

Proof This is nearly immediate from the minimax lower bound (9.3.3) in Le Cam's two point method, which implies

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) \ge \frac{1}{2} \sup_{P_1 \in \mathcal{P}} \Phi(|\theta(P_0) - \theta(P_1)|/2) \left(1 - \|P_0^n - P_1^n\|_{\rm TV}\right).$$

Now we use Proposition 2.2.8 to observe that

$$\begin{aligned}\|P_0^n - P_1^n\|_{\rm TV} &\le d_{\rm hel}(P_0^n, P_1^n)\sqrt{2 - d_{\rm hel}^2(P_0^n, P_1^n)} \\ &= \left(1 - (1 - d_{\rm hel}^2(P_0, P_1))^n\right) \sqrt{1 + (1 - d_{\rm hel}^2(P_0, P_1))^n}\end{aligned}$$

by the tensorization identity (9.2.5) for Hellinger distance. Because $(1 - c/n)^n \to e^{-c}$, the asymptotic limit in the proposition follows from the inequality

$$\limsup_n \sup_{d_{\rm hel}^2(P,Q) \le \frac{c}{n}} \|P^n - Q^n\|_{\rm TV} \le (1 - e^{-c})\sqrt{1 + e^{-c}} = \sqrt{1 - e^{-c}}\sqrt{1 - e^{-2c}}.$$

Let us perform the evaluation in finite samples to obtain explicit constants. Because $x \mapsto x\sqrt{2 - x}$ is increasing for $0 \le x \le 1$, we see that if $d_{\rm hel}^2(P_0, P_1) \le \frac{c}{n}$ then

$$\|P_0^n - P_1^n\|_{\rm TV} \le \left(1 - \left(1 - \frac{c}{n}\right)^n\right) \sqrt{1 + \left(1 - \frac{c}{n}\right)^n}.$$

If $(1 - \frac{c}{n})^n \ge \frac{1}{2}$, then $\|P_0^n - P_1^n\|_{\rm TV} \le \frac{1}{2}\sqrt{3/2} = \sqrt{3/8}$. Because $(1 - \frac{c}{n})^n \uparrow e^{-c}$ as n grows, taking $n = 2$ we solve $(1 - c/2)^2 = \frac{1}{2}$, or $c = 2 - \sqrt{2}$, to obtain that $d_{\rm hel}^2(P_0, P_1) \le \frac{2 - \sqrt{2}}{n}$ implies $\|P_0^n - P_1^n\|_{\rm TV} \le \sqrt{3/8}$ for $n \ge 2$. Substituting into the minimax lower bound gives the result. □

So long as the local modulus does not vary too wildly as $\epsilon \to 0$, it in fact *characterizes* the difficulty of the hardest one-dimensional subproblem at P_0. We say that the modulus is regular at P_0 if there exist $0 < r_1 \le r_0$ and K_0, K_1 such that

$$K_0 \epsilon^{r_0} \le \omega_{\rm hel}(\epsilon; \theta, P_0, \mathcal{P}) \le K_1 \epsilon^{r_1}$$

for all small $\epsilon > 0$. Then we have the following complement to Proposition 13.1.1, showing that the Hellinger modulus indeed is the fundamental quantity governing the risk (13.1.3) of the hardest one-dimensional subproblem.

Proposition 13.1.2. *The risk of the hardest one-dimensional subproblem satisfies*

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) \le \sup_{\epsilon \ge 0} e^{-n\epsilon^2} \Phi\left(\omega_{\rm hel}(\epsilon; \theta, P_0, \mathcal{P})\right). \tag{13.1.4}$$

If additionally $\Phi(t) = t^p$ for some $p > 0$ and the local modulus is regular at P_0, then for large enough n,

$$\mathfrak{M}_n^{\rm 1d}(\theta(\mathcal{P}), P_0, \Phi) \le e^{-\frac{1}{2} r_1 p} \cdot \Phi\left(\omega_{\rm hel}\left(\sqrt{\frac{r_0 p}{2n}}; \theta, P_0, \mathcal{P}\right)\right).$$

Proof Let $\omega(\epsilon) = \omega_{\rm hel}(\epsilon; \theta, P_0, \mathcal{P})$ for shorthand. Given P_0, P_1, let $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, and let Ψ_n be the optimal test between P_0^n and P_1^n, so that

$$P_0(\Psi_n \ne 0) + P_1(\Psi_n \ne 1) = 1 - \|P_0^n - P_1^n\|_{\rm TV} \le 1 - d_{\rm hel}^2(P_0^n, P_1^n) = (1 - d_{\rm hel}^2(P_0, P_1))^n$$

by Proposition 2.2.8 and the tensorization identity (9.2.5). Define the estimator $\widehat{\theta}_n = \theta_1$ if $\Psi_n = 1$ and $\widehat{\theta}_n = \theta_0$ otherwise. Then

$$\max_{P \in \{P_0, P_1\}} \mathbb{E}_P \left[\Phi(|\widehat{\theta}_n - \theta(P)|)\right] \le (1 - d_{\rm hel}^2(P_0, P_1))^n \Phi(|\theta_0 - \theta_1|).$$

Because $(1-c)^n \le e^{-nc}$, we obtain inequality (13.1.4).

By the regularity assumption on $\omega_{\rm hel}$, the $\epsilon_n^\star$ minimizing the upper bound (13.1.4) satisfies

$$\sqrt{\frac{r_1 p}{2n}} \le \epsilon_n^\star \le \sqrt{\frac{r_0 p}{2n}}$$

(take the ϵ values minimizing, respectively, $e^{-n\epsilon^2} K_1^p \epsilon^{r_1 p}$ and $e^{-n\epsilon^2} K_0^p \epsilon^{r_0 p}$). Substitute. □

Noting that $\frac{1}{2} - \frac{1}{2}\sqrt{1-e^{-1}}\sqrt{1-e^{-2}} > 1/8$, Proposition 13.1.1 implies that for large enough n, we have the (global) minimax lower bound

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi) \ge \frac{1}{8} \cdot \Phi\left(\frac{1}{2}\omega_{\rm hel}\left(\frac{1}{\sqrt{n}}; \theta, \mathcal{P}\right)\right). \tag{13.1.5}$$

So the modulus of continuity of the parameter θ at a radius of roughly $\frac{1}{\sqrt{n}}$ for the Hellinger provides a lower bound on estimation, reducing the problem of obtaining a minimax lower bound to one of lower bounding the modulus of continuity of θ. (The attainability of the implied bound is more involved; see the bibliographic section for some discussion of this and related issues.)

Example 13.1.3 (Estimating the value of a nonparametric function)**:** Let us revisit the nonparametric regression problems of Section 12.1. Assume we receive observations $Y_i = f(X_i) + \varepsilon_i$, where ε_i are i.i.d. $\mathsf{N}(0,1)$, $X_i \stackrel{\rm iid}{\sim} \mathsf{Uniform}([-1,1])$, and f is a 1-Lipschitz function. We wish to estimate the value of f at 0, i.e., $\theta(P) = f(0)$. Once we provide a lower bound on the Hellinger modulus in this observation model, Proposition 13.1.1 gives a minimax lower bound.

We now construct a particular function f. Let $\phi(x) = [1 - |x|]_+$, which is 1-Lipschitz, and for $t \in [0,1]$ define $f_t(x) = t\phi(x/t)$, which is 1-Lipschitz (and f_0 is identically 0). Letting P_t and P_0 denote the joint distributions of (X, Y) when $f = f_t$ or $f = f_0$, we have

$$d_{\rm hel}^2(P_t, P_0) = \frac{1}{2}\int_{-1}^1 d_{\rm hel}^2\big(\mathsf{N}(f_t(x), 1), \mathsf{N}(f_0(x), 1)\big) dx.$$

The Hellinger distance between two Gaussians (see Exercise 2.2) satisfies

$$d_{\rm hel}^2(\mathsf{N}(\mu_0, \sigma^2), \mathsf{N}(\mu_1, \sigma^2)) = 1 - \exp\left(-\frac{1}{8\sigma^2}(\mu_0 - \mu_1)^2\right) \le \frac{(\mu_0 - \mu_1)^2}{8\sigma^2},$$

where we use that $e^x \ge 1 + x$, or $1 - e^x \le -x$ for all $x \in \mathbb{R}$. We therefore obtain

$$d_{\rm hel}^2(P_t, P_0) \le \frac{1}{16}\int_{-1}^1 f_t(x)^2 dx = \frac{1}{16}\int_{-t}^t t^2 \phi^2(x/t) dx = \frac{t^3}{16}\int_{-1}^1 \phi^2(x) dx = \frac{t^3}{24}.$$

(The factor 24 is, of course, unimportant.) Observing that the separation between the parameters of interest is $\theta(P_t) - \theta(P_0) = t$, the modulus has lower bound

$$\omega_{\rm hel}(\epsilon; \theta, \mathcal{P}) \ge \sup\left\{t \mid t^3 \le 24\epsilon^2\right\} = \sqrt[3]{24} \cdot \epsilon^{2/3}.$$

Substituting in Proposition 13.1.1 with $\epsilon_n = \frac{1}{\sqrt{n}}$, we obtain minimax squared error

$$\mathfrak{M}_n\left(\theta(\mathcal{P}), (\cdot)^2\right) \gtrsim \frac{1}{n^{2/3}}$$

for the class of Lipschitz functions. Estimators (such as the kernel smoothing estimators of Section 12.1) achieve this rate, showing it is sharp. $\diamond$

13.1.1 Fisher information and divergence measures

We have seen the Fisher information matrix appear in lower bounds in previous chapters; here, we show that for parametric models, we can frequently characterize the local modulus of continuity using it, in turn implying a lower bound characterized by the Fisher information. This perspective holds beyond moduli with respect to Hellinger distance, so that for suitably regular families of distributions, f-divergences with twice differentiable f will all be locally equivalent.

To demonstrate the characterization of moduli via the Fisher information, we will provide quadratic expansions of Hellinger and other divergence measures in terms of the Fisher information. To motivate things, we begin heuristically, providing rigorous regularity conditions later. Assume that the distributions $P_\theta \in \mathcal{P}$ have densities p_θ, and recall the Fisher score $\dot{\ell}_\theta := \nabla_\theta \log p_\theta$, which is (typically) mean-zero. Let f be any twice differentiable convex function with $f(1) = 0$ and $f''(1) > 0$. Then for v small, we have $p_{\theta+v}(x)/p_\theta(x) - 1 = o_x(\|v\|)$, where $o_x(\|v\|)$ means that a term that may depend on x but satisfies $o_x(\|v\|)/\|v\| \to 0$ as $v \to 0$, and

$$f\left(\frac{p_{\theta+v}(x)}{p_\theta(x)}\right) = f(1) + f'(1)\left(\frac{p_{\theta+v}(x)}{p_\theta(x)} - 1\right) + \frac{f''(1)}{2}\left(\frac{p_{\theta+v}(x)}{p_\theta(x)} - 1\right)^2 + o_x(\|v\|^2).$$

Ignoring the remainder terms (which require some type of domination to integrate them; we provide representative statements in the appendix 13.5.1 to this chapter), we obtain

$$\begin{aligned} D_f\left(P_{\theta+v} \| P_\theta\right) &= \frac{f''(1)}{2}\int\left(\frac{p_{\theta+v}(x)}{p_\theta(x)} - 1\right)^2 p_\theta(x)dx + \int o_x(\|v\|^2)p_\theta(x)dx \\ &\stackrel{(?)}{=} \frac{f''(1)}{2} D_{\chi^2}(P_{\theta+v}\|P_\theta) + o(\|v\|^2), \end{aligned} \tag{13.1.6}$$

where the equality (?) heuristically integrates the remainder. Continuing with our heuristics, we use the expansion

$$\frac{p_{\theta+v}(x)}{p_\theta(x)} - 1 = \frac{\nabla_\theta p_\theta(x)^\top v + o_x(\|v\|)}{p_\theta(x)} = \dot{\ell}_\theta(x)^\top v + o_x(\|v\|)$$

to obtain

$$D_{\chi^2}(P_{\theta+v}\|P_\theta) \stackrel{(?)}{=} \int (\dot{\ell}_\theta(x)^\top v)^2 p_\theta(x)dx + o(\|v\|^2) = v^\top J(\theta) v + o(\|v\|^2), \tag{13.1.7}$$

where once again equality (?) is heuristic. Expressions (13.1.6) and (13.1.7) show that, at least under regularity conditions (again, see Section 13.5.1), we expect that

$$D_f\left(P_{\theta+v}\|P_\theta\right) = \frac{f''(1)}{2} v^\top J(\theta) v + o(\|v\|^2), \tag{13.1.8}$$

so that the Fisher information matrix $J(\theta)$ induces a (local) geometry on probability distributions equivalent to that for f-divergences.

Under the condition (13.1.8), it becomes straightforward to compute the modulus of continuity. Indeed, we may generalize the Hellinger modulus to define

$$\omega_f(\epsilon; T, P_\theta, \mathcal{P}) := \sup \left\{ |T(\theta') - T(\theta)| \mid D_f\left(P_{\theta'} \| P_\theta\right) \le \epsilon^2 \right\},$$

where we square ϵ to match the Hellinger case (13.1.1), which corresponds to $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$. Then the Fisher information (asymptotically) characterizes the modulus of continuity whenever we have the identifiability condition that for all $\epsilon > 0$,

$$\inf_{\theta'} \left\{ D_f\left(P_{\theta'} \| P_\theta\right) \mid \left\|\theta - \theta'\right\|_2 \ge \epsilon \right\} > 0. \tag{13.1.9}$$

Proposition 13.1.4. *Let $\{P_\theta\}$ be a family of distributions for which the f-divergence satisfies the expansion (13.1.8) at $\theta \in \mathbb{R}^d$, and let $T : \mathbb{R}^d \to \mathbb{R}$ be differentiable at θ. Then*

$$\omega_f(\epsilon; T, P_\theta, \mathcal{P}) \ge \sqrt{2/f''(1)} \cdot \epsilon \left\|J(\theta)^{-1/2} \nabla T(\theta)\right\|_2 - o(\epsilon).$$

If in addition the identifiability condition (13.1.9) holds, then

$$\omega_f(\epsilon; T, P_\theta, \mathcal{P}) \le \sqrt{2/f''(1)} \cdot \epsilon \left\|J(\theta)^{-1/2} \nabla T(\theta)\right\|_2 + o(\epsilon).$$

Deferring the proof of the proposition temporarily, in a sense, then, so long as we have enough regularity, most divergence measures and related moduli of continuity induce the same geometry via the Fisher information. The next subsection (which is purely technical) develops conditions under which these Fisher information expansions hold, but here we preview them a bit by claiming that the expansion (13.1.8) holds for the squared Hellinger distance, corresponding to $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2 = \frac{t}{2} - \sqrt{t} + 1$ for most distributions, where $f''(1) = \frac{1}{4}$. Thus, for the Hellinger distance, we have

$$\omega_{\rm hel}(\epsilon; T, P_\theta, \mathcal{P}) = \sqrt{8}(1 + o(1))\epsilon \left\|J(\theta)^{-1/2} \nabla T(\theta)\right\|_2,$$

and so applying Le Cam's method, we have the following corollary.

Corollary 13.1.5. *Let the conditions of Proposition 13.1.4 hold. Then there exists a numerical constant $c > 0$ such that*

$$\sup_{\theta_1} \inf_{\widehat{\theta}_n} \max_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left[\varphi\left(\left| \widehat{\theta}_n(X_1^n) - \theta \right| \right) \right] \ge c\varphi\left(\frac{c}{\sqrt{n}} \left\|J(\theta)^{-1/2} \nabla T(\theta)\right\|_2 \right)$$

for all large enough n.

Proof of Proposition 13.1.4 Given the expansion (13.1.8), as $\epsilon \downarrow 0$, for any $v \in \mathbb{R}^d$ we have

$$\frac{1}{\epsilon^2} D_f\left(P_{\theta + \epsilon v} \| P_\theta\right) \to \frac{f''(1)}{2} v^\top J(\theta) v.$$

Then because $T(\theta + v) = T(\theta) + \langle \nabla T(\theta), v \rangle + o(\|v\|)$, we have

$$\liminf_{\epsilon \to 0} \frac{1}{\epsilon} \omega(\epsilon) \ge \sup \left\{ \langle \nabla T(\theta), v \rangle \mid v^\top J(\theta) v \le \frac{2}{f''(1)} \right\} = \sqrt{2/f''(1)} \left\|J(\theta)^{-1/2} \nabla T(\theta)\right\|_2.$$

For the upper bound, by the identifiability assumption (13.1.9), there exists a function $\delta(\epsilon) \to 0$ as $\epsilon \downarrow 0$ for which $D_f\left(P_{\theta'} \| P_\theta\right) \le \epsilon^2$ implies that $\|\theta' - \theta\|_2 \le \delta(\epsilon)$. The identity (13.1.8) shows that for small enough $\epsilon > 0$, we have $D_f\left(P_{\theta'} \| P_\theta\right) = \frac{f''(1)}{2}(\theta' - \theta)^\top J(\theta)(\theta' - \theta) + o(\|\theta' - \theta\|^2)$ whenever $\|\theta' - \theta\| \le \delta(\epsilon)$, so that in fact it must be the case that $D_f\left(P_{\theta'} \| P_\theta\right) \le \epsilon^2$ implies that $\theta' = \theta + v$ for some v satisfying $v^\top J(\theta) v \le \frac{2}{f''(1)}\epsilon^2(1 + o(1))$ as $\epsilon \to 0$. So

$$\begin{aligned}
\omega(\epsilon) & \le \sup\left\{|T(\theta + v) - T(\theta)| \mid v^\top J(\theta) v \le (2/f''(1))\epsilon^2(1 + o(1))\right\} \\
& = \sup\left\{|\langle \nabla T(\theta), v\rangle| + o(\|v\|) \mid v^\top J(\theta) v \le (2/f''(1))\epsilon^2(1 + o(1))\right\} \\
& \le \sup\left\{\left\|J(\theta)^{-1/2}\nabla T(\theta)\right\|_2 \left\|J(\theta)^{1/2} v\right\|_2 + o(\|v\|) \mid \left\|J(\theta)^{1/2} v\right\|_2^2 \le (2/f''(1))\epsilon^2(1 + o(1))\right\}
\end{aligned}$$

by Cauchy-Schwarz, giving the result. □

13.2 Le Cam's convex hull method

While the geometric perspective of two-point lower bounds is elegant, it can be misleading and suggest wrong rates of convergence. (We will see examples where this occurs.) We can generalize the two-point approach by instead comparing families of distributions that all induce separations between statistics of one another, and then computing the distance between the convex hulls of the families, which is at most the distance between the pair. This leads to Le Cam's *convex hull method*, which we state abstractly and specialize later to different scenarios. Let $\mathcal{P}$ be a collection of distributions on an underlying space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \mathbb{R}^d$ be a parameter. We say that two subsets $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1$ are *δ-separated* in $\|\cdot\|$ if

$$\|\theta(P_0) - \theta(P_1)\| \ge \delta \quad \text{for all } P_0 \in \mathcal{P}_0 \text{ and } P_1 \in \mathcal{P}_1. \tag{13.2.1}$$

Note that the definition (13.2.1) requires only pairwise separation.

Let $\mathrm{Conv}(\mathcal{P})$ be the collection of mixtures of elements of $\mathcal{P}$, that is,

$$\mathrm{Conv}(\mathcal{P}) = \left\{\sum_{i=1}^m \lambda_i P_i \mid m \in \mathbb{N},\ \lambda \succeq 0,\ \langle \lambda, \mathbf{1}\rangle = 1,\ P_i \in \mathcal{P}\right\}.$$

Defining the minimax risk

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\|\widehat{\theta} - \theta(P)\|\right]$$

(note the temporary lack of sample size n), we have the following generalization of inequality (9.3.3).

Theorem 13.2.1 (Le Cam's Convex Hull Lower Bound)**.** *Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be δ-separated in $\|\cdot\|$, with both $\mathcal{P}_0, \mathcal{P}_1 \subset \mathcal{P}$. Then*

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) \ge \frac{\delta}{2} \sup\left\{\left[1 - \left\|\overline{P}_0 - \overline{P}_1\right\|_{\mathrm{TV}}\right] \mid \overline{P}_0 \in \mathrm{Conv}(\mathcal{P}_0), \overline{P}_1 \in \mathrm{Conv}(\mathcal{P}_1)\right\}$$

Proof For any parameter θ, the separation $\|\theta(P_0) - \theta(P_1)\| \ge \delta$ and the triangle inequality guarantee that for each pair $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_1$, at least one of $\|\theta - \theta(P_0)\| \ge \delta/2$ or $\|\theta - \theta(P_1)\| \ge \delta/2$

holds. Let $\overline{P}_0 = \sum_{j=1}^m \alpha_j P_j$ and $\overline{P}_1 = \sum_{j=1}^m \beta_j Q_j$ for $P_j \in \mathcal{P}_0$ and $Q_j \in \mathcal{P}_1$, respectively, where α, β are convex combinations. Then by Markov's inequality,

$$\begin{aligned}
\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) &\geq \frac{1}{2}\sum_{j=1}^m \alpha_j \mathbb{E}_{P_j}\left[\|\widehat{\theta} - \theta(P_j)\|\right] + \frac{1}{2}\sum_{j=1}^m \beta_j \mathbb{E}_{Q_j}\left[\|\widehat{\theta} - \theta(Q_j)\|\right] \\
&\geq \frac{\delta}{2}\left[\sum_{j=1}^m \alpha_j \mathbb{E}_{P_j}\left[\mathbf{1}\{\|\widehat{\theta} - \theta(P_j)\| \geq \delta/2\}\right] + \sum_{j=1}^m \beta_j \mathbb{E}_{Q_j}\left[\mathbf{1}\{\|\widehat{\theta} - \theta(Q_j)\| \geq \delta/2\}\right]\right] \\
&\geq \frac{\delta}{2}\sum_{j=1}^m \left(\alpha_j \mathbb{E}_{P_j}\left[\inf_{P_0 \in \mathcal{P}_0} \mathbf{1}\{\|\widehat{\theta} - \theta(P_0)\| \geq \delta/2\}\right] + \beta_j \mathbb{E}_{Q_j}\left[\inf_{P_1 \in \mathcal{P}_1} \mathbf{1}\{\|\widehat{\theta} - \theta(P_1)\| \geq \delta/2\}\right]\right) \\
&= \frac{\delta}{2}\left(\mathbb{E}_{\overline{P}_0}\left[\inf_{P_0 \in \mathcal{P}_0} \mathbf{1}\{\|\widehat{\theta} - \theta(P_0)\| \geq \delta/2\}\right] + \mathbb{E}_{\overline{P}_1}\left[\inf_{P_1 \in \mathcal{P}_1} \mathbf{1}\{\|\widehat{\theta} - \theta(P_1)\| \geq \delta/2\}\right]\right).
\end{aligned}$$

If we define $f_v(x) = \inf_{P \in \mathcal{P}_v} \mathbf{1}\{\|\widehat{\theta}(x) - \theta(P)\| \geq \delta/2\}$ for $v = 0, 1$, then $f_0 + f_1 \geq 1$. For such f_0, f_1, we claim the following lemma, which extends Le Cam's lemma (Proposition 2.3.1):

Lemma 13.2.2. *For any two distributions P_0 and P_1,*

$$\inf_{f_0 + f_1 \geq 1} \mathbb{E}_{P_0}[f_0] + \mathbb{E}_{P_1}[f_1] \geq 1 - \|P_0 - P_1\|_{\mathrm{TV}}.$$

We leave this form of total variation distance as an exercise (see Exercise 2.1). Substituting it into the display above, we find that for any $\overline{P}_v \in \mathrm{Conv}(\mathcal{P}_v)$, $v \in \{0, 1\}$, we have

$$\mathfrak{M}(\theta(\mathcal{P}), \|\cdot\|) \geq \frac{\delta}{2}\left[1 - \left\|\overline{P}_0 - \overline{P}_1\right\|_{\mathrm{TV}}\right].$$

Taking a supremum over $\overline{P}_0$ and $\overline{P}_1$ gives the theorem. □

13.2.1 The χ^2-mixture bound

Theorem 13.2.1 provides a powerful tool for developing lower bounds between collections of well-separated distributions. The most typical approach is to take the class $\mathcal{P}_0$ to consist of a single "base" distribution P_0, and then let $\mathcal{P}_1$ vary around P_0 in some prescribed way, so that for an index set $\mathcal{V}$, we let $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} P_v$. Even so, when we have a sample of size n from one of the distributions, this results in a total variation quantity of the form

$$\left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}} \quad \text{where} \quad \overline{P^n} = \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} P_v^n,$$

yielding a mixture of product distributions—something frequently quite challenging to control.

The key technique here leverages the inequalities relating divergences from Chapter 2, which allows us to replace the variation distance with something more convenient. In previous chapters, this was the KL-divergence; now, we use the χ^2-divergence, as it interacts more nicely with the mixture product structure. Essentially, we replace an expectation over $X \sim P$ with two expectations: one over $X \sim P$ and another over independent samples $V, V' \sim \mathsf{Uniform}(\mathcal{V})$.

To obtain the bound, first note that

$$2\left\|P_0 - \overline{P}\right\|_{\mathrm{TV}}^2 \leq D_{\mathrm{kl}}\left(\overline{P}\|P_0\right) \leq \log(1 + D_{\chi^2}\left(\overline{P}\|P_0\right)) \leq D_{\chi^2}\left(\overline{P}\|P_0\right)$$

by Propositions 2.2.9 and 2.2.10. We then have the following technical lemma.

Lemma 13.2.3. *Let $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v$ and P_v and P_0 have densities p_v, p_0 with respect to some base measure μ on a set $\mathcal{X}$. Then*

$$D_{\chi^2}\left(\overline{P}\|P_0\right) = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}}\int\frac{p_v(x)p_{v'}(x)}{p_0(x)}d\mu(x) - 1 = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}}\mathbb{E}_0\left[\frac{p_v(X)p_{v'}(X)}{p_0^2(X)}\right] - 1,$$

where the expectation is taken with respect to $X \sim P_0$.

Proof For any two distributions P and Q we have $D_{\chi^2}\left(P\|Q\right) = \int(dP/dQ - 1)^2 dQ = \int\frac{dP^2}{dQ} - 2\int\frac{dP}{dQ}dQ + \int dQ = \int\frac{dP^2}{dQ} - 1$. Then we proceed by recognizing that $(\frac{1}{N}\sum_{i=1}^N x_i)^2 = \frac{1}{N^2}\sum_{i,j} x_i x_j$ for any sequence x_i, and so

$$D_{\chi^2}\left(\overline{P}\|P_0\right) + 1 = \int\frac{((1/|\mathcal{V}|)\sum_{v\in\mathcal{V}} dP_v)^2}{dP_0} = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}}\int\frac{dP_v dP_{v'}}{dP_0}$$

as desired. □

When we apply Lemma 13.2.3 for product distributions, we can obtain tensorization-type inequalities; the next result is an immediate consequence of Lemma 13.2.3.

Lemma 13.2.4. *Let $\overline{P^n} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v^n$, where P_v and P_0 have densities as in Lemma 13.2.3. Then*

$$1 + D_{\chi^2}\left(\overline{P^n}\|P_0^n\right) = \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}}\mathbb{E}_0\left[\frac{p_v(X)p_{v'}(X)}{p_0^2(X)}\right]^n.$$

Going through a few examples will best show how to leverage these lemmas in a few of their many applications. Roughly, our typical approach is the following: we identify $\mathcal{V}$ with $\{\pm 1\}^d$ or some other suitable collection of vectors. We then choose distributions P_v and P_0 with densities such that the ratios p_v/p_0 "act" like exponentials involving inner products of $v \in \mathcal{V}$ with some other quantity. Because v is uniform in $\mathcal{V}$ in Lemma 13.2.3, we can leverage all the tools we have developed to control moment generating functions and concentration inequalities in Chapter 4 to bound the χ^2-divergence and then apply Theorem 13.2.1.

Let us give one example of this approach to bound the distance between a $\mathsf{N}(0, I_d)$ vector and a convex combination of normals with means away from 0; the technique we use to prove the bounds will arise frequently in the sequel. Let $P_0 = \mathsf{N}(0, \sigma^2 I_d)$ be the (scaled) standard normal distribution on $\mathbb{R}^d$, and for $\mathcal{V} = \{-1, 1\}^d$ and some $\delta \geq 0$ to be chosen, let $P_v = \mathsf{N}(\delta v, \sigma^2 I_d)$. Then we can show that while $D_{\rm kl}\left(P_v\|P_0\right) = \frac{d\delta^2}{2\sigma^2}$ for each individual P_v, the divergence for the average can be quadratically smaller in the ratio δ^2/σ^2.

Lemma 13.2.5. *Let P_0 and P_v be Gaussian distributions as above, and define the mixture $\overline{P} = \frac{1}{2^d}\sum_{v\in\{\pm1\}^d} P_v$. Then*

$$2\left\|P_0 - \overline{P}\right\|_{\rm TV}^2 \leq \log(1 + D_{\chi^2}\left(\overline{P}\|P_0\right)) \leq \frac{d\delta^4}{2\sigma^4}.$$

Proof The first inequality combines Pinsker's inequality (Proposition 2.2.9) with the bound $D_{\rm kl}\left(P\|Q\right) \leq \log(1 + D_{\chi^2}(P\|Q))$ in Proposition 2.2.10. Now we expand the χ^2-divergence, yielding

$$1 + D_{\chi^2}\left(\overline{P}\|P_0\right) = \mathbb{E}\left[\exp\left(-\frac{1}{2\sigma^2}\left\|Y - \delta V\right\|_2^2 - \frac{1}{2\sigma^2}\left\|Y - \delta V'\right\|_2^2 + \frac{1}{\sigma^2}\left\|Y\right\|_2^2\right)\right],$$

where the expectation is over $Y \sim \mathsf{N}(0, \sigma^2 I_n)$ and $V, V' \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V})$. Taking the expectation over Y first, before averaging over the packing elements, allows more careful control. Indeed, expanding the squares and recognizing that $\|v\|_2^2 = d$ for each $v \in \{\pm 1\}^d$, we have

$$\begin{aligned} 1 + D_{\chi^2}\left(\overline{P} \| P_0\right) = \mathbb{E}\left[\exp\left(\frac{\delta}{\sigma^2}\langle Y, V + V'\rangle - \frac{d\delta^2}{\sigma^2}\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\delta^2}{2\sigma^2}\left\|V + V'\right\|_2^2 - \frac{d\delta^2}{\sigma^2}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\delta^2}{\sigma^2}\langle V, V'\rangle\right)\right] \\ &\le \exp\left(\frac{d\delta^4}{2\sigma^4}\right), \end{aligned}$$

where the final inequality follows because an individual $U \sim \mathsf{Uniform}(\{\pm 1\})$ is 1-sub-Gaussian, so $\langle V, V'\rangle$ is d-sub-Gaussian. □

As we typically wish to control distances between product distributions arising from i.i.d. sampling, we also seek "tensorized" bounds. The next result exhibits this for Gaussians.

Lemma 13.2.6. *Let P_0 and P_v be Gaussian distributions as in Lemma 13.2.4, and define the mixture $\overline{P^n} = \frac{1}{2^d}\sum_{v \in \{\pm 1\}^d} P_v^n$. Then*

$$2\left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}}^2 \le \log\left(1 + D_{\chi^2}\left(\overline{P^n} \| P_0^n\right)\right) \le \frac{dn^2\delta^4}{2\sigma^4}.$$

Proof Tracing the proof of Lemma 13.2.5 and using Lemma 13.2.4, we have

$$\begin{aligned} 1 + D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) &= \frac{1}{4^d}\sum_{v, v' \in \{\pm 1\}^d} \mathbb{E}_0\left[\exp\left(\frac{\delta}{\sigma^2}\langle Y, v + v'\rangle - \frac{d\delta^2}{\sigma^2}\right)\right]^n \\ &= \frac{1}{4^d}\sum_{v, v' \in \{\pm 1\}^d} \exp\left(\frac{n\delta^2}{2\sigma^2}\left\|v + v'\right\|_2^2 - \frac{nd\delta^2}{\sigma^2}\right). \end{aligned}$$

Recognizing that $\|v + v'\|_2^2 = 2d + 2\langle v, v'\rangle$, we have $1 + D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) = \mathbb{E}[e^{\frac{\delta^2 n}{\sigma^2}\langle V, V'\rangle}] \le \exp(\frac{d\delta^4 n^2}{2\sigma^4})$ for $V, V' \stackrel{\text{ind}}{\sim} \mathsf{Uniform}(\{\pm 1\}^d)$. □

13.2.2 Estimating the norm of a Gaussian vector and multiple hypothesis testing

We revisit the signal detection problems from Section 12.3 in a stylized example that also shows why two-point lower bounds may fail. We assume a *Gaussian sequence model*

$$Y = \theta^\star + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2 I_n) \tag{13.2.2}$$

where $\theta^\star \in \mathbb{R}^n$ satisfies $\|\theta^\star\|_2 \le r$, and we wish to estimate the statistic

$$T(P) := \|\theta^\star\|_2^2.$$

As motivation, consider a global-null testing problem with n independent p-values $U_j \stackrel{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$. Under the null, $Y_j = \Phi^{-1}(U_j) \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$, where Φ denotes the standard normal CDF, while alternatives produce $Y \sim \mathsf{N}(\theta, I_n)$ for some $\theta \neq 0$. We detect departures from the null by testing whether $\|\theta\|_2$ is far from zero.

In the sequence model (13.2.2), the identity $\mathbb{E}[\|Y\|_2^2] = \|\theta^\star\|_2^2 + n\sigma^2$ makes the debiased quantity

$$T_n := \|Y\|_2^2 - n\sigma^2$$

a natural estimator. Using that $\mathrm{Var}((\theta_j^\star + \sigma\varepsilon_j)^2) = 2\sigma^4 + 4\sigma^2(\theta_j^\star)^2$ for $\varepsilon_j \sim \mathsf{N}(0,1)$,

$$\mathbb{E}\left[\left|T_n - \|\theta^\star\|_2^2\right|^2\right] = \sum_{j=1}^n \mathrm{Var}\left((\theta_j^\star + \sigma\varepsilon_j)^2\right) = 2n\sigma^4 + 4\|\theta^\star\|_2^2\sigma^2 \le 2n\sigma^4 + 4r^2\sigma^2.$$

Thus the family $\mathcal{P}_{\sigma,r}$ of Gaussian sequence models (13.2.2) with variance σ^2 and $\|\theta^\star\|_2 \le r$ has minimax absolute estimation risk at most

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \le \sqrt{2n\sigma^4 + 4r^2\sigma^2} \le \sqrt{2n}\sigma^2 + 2r\sigma. \tag{13.2.3}$$

The upper bound (13.2.3) is sharp. To demonstrate this, we might be tempted to use the approach via modulus of continuity for a 1-dimensional sub-problem in Section 13.1. This cannot work, as the global Hellinger modulus (13.1.2) scales as $\min\{r\sigma, \sigma^2\}$ (see Section 13.5.4 for a proof):

Lemma 13.2.7. *Let $\mathcal{P}_{\sigma,r}$ denote the Gaussian sequence models (13.2.2) satisfying $\|\theta\|_2 \le r$ and $0 < \epsilon < 1$ be arbitrary. Then for $T(\theta) := \|\theta\|_2^2$, the global Hellinger modulus satisfies*

$$\min\left\{\sqrt{-8\log(1-\epsilon^2)} \cdot r\sigma, r^2\right\} \le \omega_{\mathrm{hel}}(\epsilon; T, \mathcal{P}_{\sigma,r}) \le \min\left\{2\sqrt{-8\log(1-\epsilon^2)} \cdot r\sigma, r^2\right\}.$$

Taking the constant $\epsilon = \sqrt{1 - e^{-1/8}} \approx .34$ yields $\min\{r\sigma, r^2\} \le \omega_{\mathrm{hel}}(\epsilon; T, \mathcal{P}_{\sigma,r}) \le \min\{2r\sigma, r^2\}$. As a consequence (recall Propositions 13.1.1 and 13.1.2), the strongest lower bound that the two-point lower bound strategy can provide on the minimax risk scales as

$$\mathfrak{M}_n^{\mathrm{1d}}(T(\mathcal{P}_{\sigma,r}), |\cdot|) \asymp \min\{r\sigma, r^2\}, \tag{13.2.4}$$

which can be smaller than the bound (13.2.3).

We therefore turn to the mixture approach. Let $P_0 = \mathsf{N}(0, \sigma^2 I_n)$, and for $\mathcal{V} = \{\pm 1\}^n$ define $P_v = \mathsf{N}(\delta v, \sigma^2 I_n)$. Then $T(P_0) = 0$ while $T(P_v) = \delta^2 n$, so Theorem 13.2.1 gives

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{2}\left\{1 - \sqrt{\frac{1}{2}\log(1 + D_{\chi^2}(\overline{P}\|P_0))}\right\}$$

for $\overline{P} = \frac{1}{2^n}\sum_{v\in\mathcal{V}} P_v$. Substituting the result of Lemma 13.2.5 into the lower bound yields

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{2}\left(1 - \sqrt{\frac{n\delta^4}{4\sigma^4}}\right).$$

We choose $\delta = \min\{\frac{\sigma}{n^{1/4}}, \frac{r}{\sqrt{n}}\}$, to achieve the minimax lower bound

$$\mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \ge \frac{\delta^2 n}{4} = \frac{1}{4}\min\left\{\sigma^2\sqrt{n}, r^2\right\}. \tag{13.2.5}$$

Combining the result (13.2.3) with the lower bounds (13.2.4) and (13.2.5) characterizes the risk:

Corollary 13.2.8. *Let $\mathcal{P}_{\sigma,r}$ be the Gaussian sequence model family $\{\mathsf{N}(\theta, \sigma^2 I_n) \mid \|\theta\|_2 \le r\}$, and $T(\theta) = \|\theta\|_2^2$. Then there is a numerical constant $c > 0$ such that*

$$c \cdot \min\left\{\sigma^2\sqrt{n} + r\sigma, r^2\right\} \le \mathfrak{M}_n(T(\mathcal{P}_{\sigma,r}), |\cdot|) \le \min\left\{\sqrt{2n\sigma^4 + 4r^2\sigma^2}, r^2\right\}.$$

Proof The only thing to recognize is that $r\sigma \ge \sigma^2\sqrt{n}$ whenever $r \ge \sigma\sqrt{n}$, in which case $\min\{r^2, \sigma r\} = \sigma r$ in the bound (13.2.4). □

13.2.3* Lower bounds on estimating integral functionals

In problems such as nonparametric regression or nonparametric density estimation, a frequent concern is the degree of smoothness of the underlying functional, which can then impact the type of regularization one employs. For example, if the underlying function is sufficiently differentiable, cubic smoothing splines [e.g. 230, Chapter 5.4] estimate regression functions by solving

$$\widehat{f} = \operatorname*{argmin}_f \left\{\frac{1}{2}\sum_{i=1}^n (Y_i - f(x_i))^2 + \frac{\lambda}{2}\int (f''(x))^2 dx\right\},$$

and the regularization λ is chosen to enforce sufficient smoothness of $\widehat{f}$.

It is therefore interesting to estimate various functionals based on integration to help understand the gross character of the function being estimated. Here we show how to provide lower bounds on estimation error for functionals of the form

$$T_k(f) := \int (f^{(k)}(x))^2 dx,$$

the L^2-norm of the kth derivative of an (unknown) function f, using the convex hull methodology. We consider estimation of such functionals over classes of functions with bounded higher-order derivatives, defining

$$\mathcal{F}_s := \left\{f : [0,1] \to \mathbb{R} \mid f(0) = 0, f \in \mathcal{C}^s \text{ and } \|f^{(s)}\|_\infty \le 1\right\},$$

functions that are s-times continuously differentiable with uniformly bounded sth derivative, where we choose $f(0) = 0$ for normalization. We adopt the nonparametric regression observation model in Example 13.1.3, so that $Y_i = f(X_i) + \varepsilon_i$, where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$ and $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}([-1,1])$.

Proposition 13.2.9. *Fix $k \in \mathbb{N}$. Then for any $s \ge 0$,*

$$\mathfrak{M}_n\left(T_k(\mathcal{F}_{k+s}), |\cdot|\right) = \inf_{\widehat{T}} \sup_{f \in \mathcal{F}_{k+s}} \mathbb{E}_f\left[|\widehat{T}(Y_1^n, X_1^n) - T_k(f)|\right] \gtrsim n^{-\frac{4s}{4(k+s)+1}}.$$

Proof The key is to construct "bump-like" functions that induce separation in the functional T_k, as we have done in the construction of nonparametric lower bounds in Chapter 12.1. Pick any "bump" function $\phi : [0,1] \to \mathbb{R}$ with $\phi(0) = \phi(1) = 0$ for which $\|\phi^{(k+s)}\|_\infty \le 1$ and $\int_0^1 \phi^{(k)}(x)^2 dx > 0$. For a value $m \in \mathbb{N}$ to be chosen, define the rescaled function

$$g(x) := \frac{1}{m^{k+s}}\phi(mx),$$

which satisfies $g^{(k)}(x) = \frac{1}{m^s}\phi^{(k)}(mx)$ and $g^{(k+s)}(x) = \phi^{(k+s)}(mx)$ so $\|g^{(k+s)}\|_\infty \le 1$. Then for $v \in \{-1, 0, 1\}^m$, define the (repeating bump) functions

$$f_v(x) := \sum_{j=1}^m v_j g\left(x - \frac{j-1}{m}\right) = \frac{1}{m^{k+s}} \sum_{j=1}^m v_j \phi\left(mx - (j-1)\right).$$

To apply the convex hull method in Theorem 13.2.1, we exhibit a separation between the functions f_v and $f_{\mathbf{0}}$, then bound the divergence between the observations from the null $f_{\mathbf{0}}$ and the convex hull of the alternatives f_v. The separation is straightforward: we have $T_k(f_{\mathbf{0}}) = 0$, while

$$\begin{aligned} T_k(f_v) = \int_0^1 f_v^{(k)}(x)^2 dx = m \int_0^{1/m} (g^{(k)}(x))^2 dx = \frac{m}{m^{2s}} \int_0^{1/m} (\phi^{(k)}(mx))^2 dx \\ = \frac{1}{m^{2s}} \int_0^1 (\phi^{(k)}(u))^2 du \gtrsim \frac{1}{m^{2s}}. \end{aligned} \tag{13.2.6}$$

The divergence bounds for the convex hull are more intricate: we relate the χ^2 divergence to exponential moments of (independent) random packing vectors, as in Lemma 13.2.5, and then use sub-Gaussianity of random signs (Example 4.1.5) to bound this quantity. Define P_v^n to be the joint distribution of $(X_i, Y_i)_{i=1}^n$ when $Y_i = f_v(X_i) + \varepsilon_i$, and let $\overline{P^n} = \frac{1}{2^m} \sum_{v \in \{\pm 1\}^m} P_v^n$. Let $p_v(y \mid x)$ denote the density of $Y \sim \mathsf{N}(f_v(x), 1)$. By Lemma 13.2.4, we have

$$\begin{aligned} 1 + D_{\chi^2}\left(\overline{P^n} \| P_{\mathbf{0}}^n\right) &= \frac{1}{2^{2m}} \sum_{v, v' \in \{\pm 1\}^m} \mathbb{E}_{\mathbf{0}}\left[\frac{p_v(Y \mid X) p_{v'}(Y \mid X)}{p_{\mathbf{0}}(Y \mid X)^2}\right]^n \\ &= \frac{1}{2^{2m}} \sum_{v, v' \in \{\pm 1\}^m} \mathbb{E}_{\mathbf{0}}\left[\exp\left(Y(f_v(X) + f_{v'}(X)) - \frac{1}{2} f_v(X)^2 - \frac{1}{2} f_{v'}(X)^2\right)\right]^n, \end{aligned}$$

where we have used that $X \sim \mathsf{Uniform}([0,1])$ and that $Y \sim \mathsf{N}(0,1)$ under $P_{\mathbf{0}}$. To bound the expectations, define the coordinate functions $g_j(x) = m^{-(k+s)}\phi(mx - (j-1))$ and vector $\vec{g}(x) = [g_j(x)]_{j=1}^m$, so $f_v(x) = \langle v, \vec{g}(x)\rangle$ and $\vec{g}(x)$ has at most one non-zero element. Then

$$\begin{aligned} \mathbb{E}_0\left[\frac{p_v(Y \mid X) p_{v'}(Y \mid X)}{p_0(Y \mid X)^2} \mid X = x\right] &= \exp\left(\frac{1}{2}\langle v + v', \vec{g}(x)\rangle^2 - \frac{1}{2}\langle v, \vec{g}(x)\rangle^2 - \frac{1}{2}\langle v', \vec{g}(x)\rangle^2\right) \\ &= \exp\left(\langle v, \vec{g}(x)\rangle\langle v', \vec{g}(x)\rangle\right) = \exp\left(v^\top \operatorname{diag}(\vec{g}(x))^2 v'\right) \end{aligned}$$

because of the disjoint support of the elements of $\vec{g}$. Using that $e^t \le 1 + t + t^2$ for $|t| \le 1$,

$$\begin{aligned} \mathbb{E}_0\left[\frac{p_v(Y \mid X) p_{v'}(Y \mid X)}{p_0(Y \mid X)^2}\right] &\le 1 + v^\top \mathbb{E}_0[\operatorname{diag}(\vec{g}(X))^2] v' + \int_0^1 \left(v^\top \operatorname{diag}(\vec{g}(x))^2 v'\right)^2 dx \\ &= 1 + \int_0^{1/m} g^2(x) dx \cdot \langle v, v'\rangle + m \int_0^{1/m} g^4(x) dx \\ &= 1 + \frac{1}{m^{2(k+s)+1}} \int_0^1 \phi^2(u) du \cdot \langle v, v'\rangle + \frac{1}{m^{4(k+s)}} \int_0^1 \phi^4(u) du. \end{aligned}$$

So for the numerical constants $c = \int_0^1 \phi^2(u) du$ and $c' = \int_0^1 \phi^4(u) du$, we use the inequality $1 + t \le e^t$

to obtain that for $V, V' \stackrel{\rm ind}{\sim} \mathsf{Uniform}(\{\pm 1\}^m)$,

$$1 + D_{\chi^2}\left(\overline{P^n} \| P_{\mathbf{0}}^n\right) \le \mathbb{E}\left[\exp\left(\frac{nc}{m^{2(k+s)+1}}\langle V, V'\rangle + \frac{nc'}{m^{4(k+s)}}\right)\right] \le \exp\left(\frac{c^2 n^2 m}{2m^{4(k+s)+2}} + \frac{nc'}{m^{4(k+s)}}\right). \tag{13.2.7}$$

In particular, we can choose m scaling as $n^{\frac{2}{4(k+s)+1}}$, which (with an appropriate constant) yields $1 + D_{\chi^2}\left(\overline{P^n} \| P_{\mathbf{0}}^n\right) \le 2$, so that $\left\|P_{\mathbf{0}}^n - \overline{P^n}\right\|_{\rm TV}^2 \le \log(1 + D_{\chi^2}\left(\overline{P^n} \| P_{\mathbf{0}}^n\right)) \le \log 2 < 1$. Substituting this choice of m into the separation bound (13.2.6) gives the result. □

Proposition 13.2.9 gives a few consequences. First, if we wish to estimate the integral of the kth derivative of a function f, if all we know is that f has bounded and continuous kth derivative then the minimax risk is constant—estimation is impossible. Typical choices are that $k = 1$ and the number of additional degrees of smoothness $s = 1$ in Proposition 13.2.9, which implies a lower bound on the minimax risk of

$$\mathfrak{M}_n(T_1(\mathcal{F}_2), |\cdot|) \gtrsim n^{-\frac{1}{2+1/4}} = n^{-\frac{4}{9}}.$$

More generally, if the number of additional degrees of smoothness $s < k + \frac{1}{4}$, then standard parametric rates of $n^{-1/2}$ are unachievable. Additionally, it is possible (but tedious) to provide a result completely analogous to Lemma 13.2.7, which shows that the two-point approach cannot give sharp lower bounds for this problem. Exercise 13.15 asks you to prove that even estimating $T_0(f) = \int f^2(x)dx$ is challenging in higher dimensions, exhibiting a curse-of-dimensionality phenomenon.

13.3 Instance-optimal bounds, super-efficiency, and risk transfer

Super-efficiency converses allow us to state that a putative benchmark for estimation performance is sharp, so that outperforming the benchmark on more than a handful of problems is impossible. Recalling Section 11.4, we seek instance specific (point (OPT.i)) and uniformly achievable (point (OPT.ii)) benchmarks. To show that a benchmark provides the "right" lower bound, Section 11.4 considers super-efficiency converses that guarantee no procedure can outperform the lower bound except on a negligible collection of instances (point (OPT.iii)). In classical estimation, Theorem 11.4.1 highlights how Fisher information addresses this. In one-dimensional estimation, however, we can frequently demonstrate that achieving small risk on one instance "transfers" that risk onto other instances:

(OPT.iii') There is a *risk-transfer* converse: if a procedure outperforms the benchmark on a population P, there exist other populations where the procedure underperforms the benchmark.

Proposition 13.1.1 shows that the local modulus of continuity provides a potential benchmark lower bound, as it gives an instance-specific guarantee (desideratum (OPT.i)). The question of achieving the local modulus risk (desideratum (OPT.ii)) typically depends on the problem, and no general characterization exists (but see the bibliographic discussion). Nonetheless, by considering the modulus of continuity of a parameter θ with respect to the χ^2-divergence rather than the Hellinger distance, we will obtain a benchmark for estimation complexity that, essentially, always satisfies the desiderata (OPT.i) and (OPT.iii'). Some literatures term (OPT.iii') a super-efficiency

converse, and others call the technique a "constrained risk inequality," but we use the risk transfer nomenclature to distinguish the super-efficiency converses in Section 11.4 and to avoid confusion with our optimality results on constrained estimators in Chapter 10.

13.3.1 Risk transfer inequalities

By using local moduli of continuity of the parameter to be estimated, we can provide pointwise guarantees of the form (OPT.iii'): outperforming a benchmark at a single instance implies quantitatively worse performance at others. We focus on real-valued parameters for simplicity. Let $\theta(P) \in \mathbb{R}$ be a parameter, and for a loss $\Phi : \mathbb{R}_+ \to \mathbb{R}$, let L_P denote the risk, or expected loss,

$$L_P(\widehat{\theta}) := \mathbb{E}_P \left[\Phi \left(\left| \widehat{\theta}(X) - \theta(P) \right| \right) \right],$$

where $X \sim P$. We shall use the χ^2-*affinity*

$$\rho_{\chi^2}(P_1 \| P_0) := D_{\chi^2}(P_1 \| P_0) + 1 = \int \frac{dP_1^2}{dP_0} = \mathbb{E}_0 \left[\frac{dP_1^2}{dP_0^2} \right] = \mathbb{E}_1 \left[\frac{dP_1}{dP_0} \right], \tag{13.3.1}$$

where $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectation under P_0 and P_1, respectively, to show that if $L_{P_0}(\widehat{\theta})$ is small under P_0, then it must be large under alternative distributions P_1. The following theorem provides a lower bound on the risk of an estimator $\widehat{\theta}$ under P_1 given an upper bound on its risk under P_0:

Theorem 13.3.1. *Assume the loss $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is convex. Let $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, and define the separation $\Delta := 2\Phi(\frac{1}{2}|\theta_0 - \theta_1|)$. If the estimator $\widehat{\theta}$ satisfies $L_{P_0}(\widehat{\theta}) \le \gamma$, then*

$$L_{P_1}(\widehat{\theta}) \ge \left[\sqrt{\Delta} - \sqrt{\gamma \rho_{\chi^2}(P_1 \| P_0)} \right]_+^2.$$

The theorem immediately extends to product distributions, as $\rho_{\chi^2}(P_1^n \| P_0^n) = \rho_{\chi^2}(P_1 \| P_0)^n$, so that if $\widehat{\theta}_n$ satisfies $\mathbb{E}_0[\Phi(|\widehat{\theta}_n(X_1^n) - \theta_0|)] \le \gamma$, then

$$\mathbb{E}_{P_1} \left[\Phi \left(\left| \widehat{\theta}_n(X_1^n) - \theta_1 \right| \right) \right] \ge \left[\sqrt{2\Phi(|\theta_0 - \theta_1|/2)} - \sqrt{\gamma \rho_{\chi^2}(P_1 \| P_0)^n} \right]_+^2.$$

Proof We assume without loss of generality that $\widehat{\theta}(x) \in [\theta_0, \theta_1]$, as otherwise, projecting it onto the interval $[\theta_0, \theta_1]$ decreases the risk. For $\theta \in [\theta_0, \theta_1]$, we have $\theta = t\theta_0 + (1-t)\theta_1$ for a $t \in [0,1]$, so

$$\begin{aligned} \sqrt{\Phi(|\theta - \theta_0|)} + \sqrt{\Phi(|\theta - \theta_1|)} &= \sqrt{\Phi((1-t)|\theta_1 - \theta_0|)} + \sqrt{\Phi(t|\theta_1 - \theta_0|)} \\ &\ge \sqrt{\Phi((1-t)|\theta_1 - \theta_0|) + \Phi(t|\theta_1 - \theta_0|)} \ge \sqrt{2\Phi\left(\frac{1}{2}|\theta_1 - \theta_0|\right)}, \end{aligned} \tag{13.3.2}$$

because $t = \frac{1}{2}$ minimizes $\Phi(ta) + \Phi((1-t)a)$. Using the majorization (13.3.2), we thus obtain

$$\mathbb{E}_1 \left[\Phi \left(\left| \widehat{\theta} - \theta_0 \right| \right)^{1/2} \right] + \mathbb{E}_1 \left[\Phi \left(\left| \widehat{\theta} - \theta_1 \right| \right)^{1/2} \right] \ge \sqrt{2\Phi(|\theta_1 - \theta_0|/2)} = \sqrt{\Delta}.$$

Rearranging, we have by Cauchy-Schwarz that

$$L_{P_1}(\widehat{\theta}) \ge \mathbb{E}_1 \left[\Phi \left(\left| \widehat{\theta} - \theta_1 \right| \right)^{1/2} \right]^2 \ge \left[\sqrt{\Delta} - \mathbb{E}_1 \left[\Phi \left(\left| \widehat{\theta} - \theta_0 \right| \right)^{1/2} \right] \right]_+^2.$$

Applying Cauchy-Schwarz once more to see that

$$\mathbb{E}_1\left[\sqrt{\Phi(|\widehat{\theta}-\theta_0|)}\right] = \mathbb{E}_0\left[\frac{dP_1}{dP_0}\sqrt{\Phi(|\widehat{\theta}-\theta_0|)}\right] \le \mathbb{E}_0\left[\frac{dP_1^2}{dP_0^2}\right]^{1/2}\sqrt{\mathbb{E}_0\left[\Phi\left(|\widehat{\theta}-\theta_0|\right)\right]}$$

gives the theorem. □

Using Theorem 13.3.1 involves a two step process: we assume an estimator achieves small risk, then show that there exist distributions close in χ^2-distance whose parameters differ substantially.

Example 13.3.2 (Risk transfer in normal mean estimation)**:** Let $\mathcal{P} = \{\mathsf{N}(\theta, 1)\}_{\theta\in\mathbb{R}}$, and assume we observe $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Let $\widehat{\theta}_n$ be an estimator satisfying

$$\mathbb{E}_{\theta_0}[|\widehat{\theta}_n - \theta_0|] \le \frac{\gamma}{\sqrt{n}}$$

at some $\theta_0 \in \mathbb{R}$. For an alternative distribution $P_1 = \mathsf{N}(\theta_1, 1)$, we observe that

$$\rho_{\chi^2}\left(P_1\|P_0\right) = \exp\left(|\theta_1-\theta_0|^2\right) \quad \text{so} \quad \rho_{\chi^2}\left(P_1^n\|P_0^n\right) = \exp\left(n|\theta_1-\theta_0|^2\right).$$

Then by Theorem 13.3.1, we have

$$\mathbb{E}_{P_1}\left[\left|\widehat{\theta}_n - \theta_1\right|\right] \ge \left[\sqrt{|\theta_0-\theta_1|} - \frac{1}{n^{1/4}}\exp\left(\frac{n}{2}|\theta_1-\theta_0|^2 + \frac{1}{2}\log\gamma\right)\right]_+^2.$$

Taking $\theta_1 = \theta_0 + \sqrt{\frac{1}{n}\log\frac{1}{\gamma}}$ then yields

$$\mathbb{E}_{P_1}\left[\left|\widehat{\theta}_n - \theta_1\right|\right] \ge \frac{1}{\sqrt{n}}\left[\sqrt[4]{\log\frac{1}{\gamma}} - 1\right]_+^2 \gtrsim \sqrt{\frac{\log\frac{1}{\gamma}}{n}}.$$

So if $\gamma \ll 1$, meaning the estimator exhibits convergence faster than $1/\sqrt{n}$ at θ_0, it must pay a substantial penalty at points roughly $\sqrt{\frac{1}{n}\log\frac{1}{\gamma}}$ away. ◇

Example 13.3.2 provides one fairly precise description of the failures of "super-efficient" estimators in one-dimensional problems, long a bugaboo in classical statistics. Exercise 13.14 works through the Hodges super-efficient estimator, a variant of this phenomenon where an estimator of a normal location parameter "better" than the sample mean seems to exist.

These results are examples of a broader class of risk transfer results that arise from considering the χ^2 modulus of continuity instead of the familiar Hellinger modulus, with a slight modification which does not change the asymptotic but makes computation simpler. With that in mind, recall the χ^2-affinity (13.3.1) and define the local χ^2-modulus of continuity by

$$\omega_{\chi^2}(\epsilon; \theta, P_0, \mathcal{P}) := \sup_{P_1\in\mathcal{P}}\left\{|\theta(P_1) - \theta(P_0)| \mid \log\rho_{\chi^2}\left(P_1\|P_0\right) \le 2\epsilon^2\right\}. \tag{13.3.3}$$

By Proposition 2.2.10, $2d_{\text{hel}}^2(P_0, P_1) \le \log(1 + D_{\chi^2}(P_1\|P_0)) = \log\rho_{\chi^2}\left(P_1\|P_0\right)$, so the Hellinger modulus dominates the χ^2-modulus:

$$\omega_{\chi^2}(\epsilon; \theta, P_0, \mathcal{P}) \le \omega_{\text{hel}}(\epsilon; \theta, P_0, \mathcal{P}).$$

Thus $\omega_{\chi^2}(1/\sqrt{n}; \theta, P_0, \mathcal{P})$ always lower bounds the 1-dimensional minimax risk (13.1.3). In most "common" cases, such as finite-dimensional parametric problems, divergences between distributions are locally quadratic (13.1.8) (as in Examples 13.5.2 and 13.5.3), and the moduli are equivalent to within constant factors. Then $\omega_{\chi^2}(1/\sqrt{n})$ also upper bounds the 1-dimensional risk (Proposition 13.1.2), providing an equivalent benchmark to the Hellinger modulus (13.1.1). But outperforming this χ^2-modulus comes with a cost:

Corollary 13.3.3. *Let the conditions of Theorem 13.3.1 hold, and define the shorthand* $\omega(\epsilon) = \omega_{\chi^2}(\epsilon; \theta, P_0, \mathcal{P})$. *Let* $\widehat{\theta}_n$ *be any estimator satisfying*

$$\mathbb{E}_{P_0}\left[\Phi(|\widehat{\theta}_n(X_1^n) - \theta(P_0)|)\right] \le \gamma \cdot \Phi\left(\omega(1/\sqrt{n})\right)$$

for some $\gamma < 1$. *Then for all distributions* P_1 *such that* $\log(1 + D_{\chi^2}(P_1 \| P_0)) \le \frac{1}{n} \log \frac{1}{\gamma}$,

$$\mathbb{E}_{P_1}\left[\Phi\left(|\widehat{\theta}_n(X_1^n) - \theta(P_1)|\right)\right] \ge \left[\sqrt{2\Phi(|\theta(P_1) - \theta(P_0)|/2)} - \sqrt{\Phi(\omega(1/\sqrt{n}))}\right]_+^2.$$

In particular,

$$\sup_{\log \rho_{\chi^2}(P_1 \| P_0) \le \frac{1}{n} \log \frac{1}{\gamma}} \mathbb{E}_{P_1}\left[\Phi\left(|\widehat{\theta}_n(X_1^n) - \theta(P_1)|\right)\right] \ge \left[\sqrt{2\Phi\left(\frac{1}{2}\omega\left(\left(\frac{1}{n}\log\frac{1}{\gamma}\right)^{1/2}\right)\right)} - \sqrt{\Phi(\omega(1/\sqrt{n}))}\right]_+^2.$$

Proof Let $\Phi_n = \Phi(\omega_{\chi^2}(\frac{1}{\sqrt{n}}; \theta, P_0, \mathcal{P}))$ for shorthand. By Theorem 13.3.1, for any P_1 and associated parameters θ_0, θ_1, we have

$$L_{P_1}(\widehat{\theta}) \ge \left[\sqrt{2\Phi(|\theta_1 - \theta_0|/2)} - \sqrt{\gamma \Phi_n \cdot \rho_{\chi^2}\left(P_1 \| P_0\right)^n}\right]_+^2.$$

Take P_1 to be any distribution satisfying $\log \rho_{\chi^2}\left(P_1 \| P_0\right) \le \frac{1}{n} \log \frac{1}{\gamma}$, so that $\gamma \Phi_n \cdot \rho_{\chi^2}\left(P_1 \| P_0\right)^n \le \gamma \Phi_n \exp(\log \frac{1}{\gamma}) = \Phi_n$. Take a supremum over such P_1. □

If an estimator outperforms the risk the χ^2-modulus of continuity predicts at a distribution P_0, then its loss on nearby distributions P_1 must be larger than that the modulus predicts. A few subtleties are worth noting: the moduli in Corollary 13.3.3 are local to P_0, not to the alternative P_1, meaning that we have not fully satisfied the converse (OPT.iii'). If the modulus (13.3.3) itself is continuous enough in the argument P_0, we obtain the desired risk transfoer.

Example 13.3.4 (Risk transfer in parametric families)**:** Let $\{P_\theta\}_{\theta \in \mathbb{R}^d}$ satisfy the regularity condition that $D_{\chi^2}(P_{\theta+v} \| P_\theta) = v^\top J(\theta) v + o(\|v\|^2)$, where $J(\theta) := \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top]$ is the Fisher information for the parameter θ. Let T be continuously differentiable and $\widehat{T}_n$ satisfy

$$\mathbb{E}_{\theta_0}\left[\left(\widehat{T}_n(X_1^n) - T(\theta_0)\right)^2\right] \le \gamma \frac{\nabla T(\theta_0)^\top J(\theta_0)^{-1} \nabla T(\theta_0)}{n}$$

at some θ_0. We recognize $\nabla T(\theta)^\top J(\theta)^{-1} \nabla T(\theta)$ as the factor characterizing the modulus of continuity for problems with Fisher information, as in Proposition 13.1.4. Applying Corollary 13.3.3 with the squared error, we obtain that for all θ satisfying $(\theta - \theta_0)^\top J(\theta_0)(\theta - \theta_0) \lesssim \frac{1}{n} \log \frac{1}{\gamma}$ that

$$\mathbb{E}_\theta\left[\left(\widehat{T}_n(X_1^n) - T(\theta)\right)^2\right] \ge \left[|T(\theta) - T(\theta_0)| - O(1)\frac{\|J(\theta_0)^{-1/2} \nabla T(\theta_0)\|_2}{\sqrt{n}}\right]_+^2.$$

By differentiability, we have $|T(\theta) - T(\theta_0)| = |\langle \nabla T(\theta_0), \theta - \theta_0 \rangle| + o(\|\theta - \theta_0\|)$, so choosing

$$\theta = c\sqrt{\frac{1}{n}\log\frac{1}{\gamma}}\frac{J(\theta_0)^{-1/2}\nabla T(\theta_0)}{\left\|J(\theta_0)^{-1/2}\nabla T(\theta_0)\right\|_2}$$

for some numerical constant $c > 0$, we see that

$$\mathbb{E}_\theta\left[\left(\widehat{T}_n(X_1^n) - T(\theta)\right)^2\right] \gtrsim \frac{\nabla T(\theta_0)^\top J(\theta_0)^{-1}\nabla T(\theta_0)}{n}\log\frac{1}{\gamma} \gtrsim \frac{\nabla T(\theta)^\top J(\theta)^{-1}\nabla T(\theta)}{n}\log\frac{1}{\gamma},$$

where we use the continuity of J and $\nabla T(\theta)$. $\diamond$

13.3.2 A general risk transfer bound

In some cases we wish to lower bound non-convex losses, such as the probability of error, which corresponds to the indicator of an estimate belonging to a set or not, so Theorem 13.3.1 fails to apply. We therefore extend it to general losses.

Corollary 13.3.5. *Let the loss $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be non-decreasing. Let $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, and define the separation $\Delta := \Phi(\frac{1}{2}|\theta_0 - \theta_1|)$. If the estimator $\widehat{\theta}$ satisfies $L_{P_0}(\widehat{\theta}) \le \gamma$, then*

$$L_{P_1}(\widehat{\theta}) \ge \left[\sqrt{\Delta} - \sqrt{\gamma\rho_{\chi^2}\left(P_1\|P_0\right)}\right]_+^2.$$

Proof The only modification we need to make to the proof of Theorem 13.3.1 is to replace the majorization inequality (13.3.2) with a slightly smaller lower bound. To that end, let $\theta \in [\theta_0, \theta_1]$ so that $\theta = t\theta_0 + (1-t)\theta_1$ for some $t \in [0,1]$. Then at least one of $|\theta - \theta_0| \ge \frac{1}{2}|\theta_1 - \theta_0|$ or $|\theta - \theta_1| \ge \frac{1}{2}|\theta_1 - \theta_0|$, and so

$$\sqrt{\Phi(|\theta - \theta_0|)} + \sqrt{\Phi(|\theta - \theta_1|)} \ge \sqrt{\Phi(|\theta_1 - \theta_0|/2)}.$$

The rest of the proof is identical. $\square$

Applying Corollary 13.3.5 follows much the same technique as our applications of Theorem 13.3.1 have. To highlight this, we revisit normal mean estimation, Example 13.3.2, in a testing scenario.

Example 13.3.6 (Super-efficient testing of a normal mean)**:** Let $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2)\}_{\theta\in\mathbb{R}}$ be the collection of normal distributions, and consider a zero-one loss indicating closeness of the estimated mean to the true mean, $\Phi(t) = \mathbf{1}\{|t| \ge \sigma/\sqrt{n}\}$, so that

$$L_{P_\theta^n}(\widehat{\theta}) = P_\theta^n\left(|\widehat{\theta}(X_1^n) - \theta| \ge \frac{\sigma}{\sqrt{n}}\right).$$

Minimax considerations dictate that the risk of most estimators should be roughly of constant order. Assume, however, that the estimator $\widehat{\theta}_n$ satisfies $P_0^n(|\widehat{\theta}_n| \ge \frac{\sigma}{\sqrt{n}}) \le \delta_n \to 0$. Fix $0 < c < 1$ and define the sequence of alternative parameter spaces

$$\Theta_n := \left\{\theta \in \mathbb{R} \mid 2\frac{\sigma}{\sqrt{n}} \le |\theta| \le \frac{\sigma}{\sqrt{n}}\sqrt{c\log\frac{1}{\delta_n}}\right\}.$$

Then we claim that

$$\liminf_n \inf_{\theta \in \Theta_n} P_\theta^n \left(|\widehat{\theta}_n(X_1^n) - \theta| \geq \frac{\sigma}{\sqrt{n}} \right) = 1, \tag{13.3.4}$$

that is, for θ in a shell around 0, the estimator is *never* within $\sigma/\sqrt{n}$ of the true mean.

To see the limit (13.3.4), assume that n is large enough that $c \log \frac{1}{\delta_n} \geq 2$, so that Θ_n is non-empty. Then calculating the χ^2-affinity for $\theta \in \Theta_n$, we obtain

$$\rho_{\chi^2}\left(P_\theta \| P_0\right)^n = \exp\left(\frac{n\theta^2}{\sigma^2}\right) \leq \exp\left(\frac{c\sigma^2 n \log \frac{1}{\delta_n}}{\sigma^2 n}\right) = \delta_n^{-c}.$$

For any $\theta \in \Theta_n$, we have $\Phi(\frac{1}{2}|\theta|) = \mathbf{1}\left\{|\theta| \geq \sigma/\sqrt{n}\right\} = 1$, so Corollary 13.3.5 gives

$$L_{P_\theta^n}(\widehat{\theta}_n) \geq \left[1 - \delta_n^{(1-c)/2}\right]_+^2 \to 1,$$

implying inequality (13.3.4). $\diamond$

13.3.3* Risk transfer with mixtures

Section 13.2 develops lower bound techniques via Le Cam's convex hull method, which can provide stronger minimax lower bounds than two-point methods on hardest 1-dimensional subproblems. An extended local modulus of continuity with respect to the χ^2-divergence over convex hulls of distributions admits risk transfer inequalities in these settings, making it possible to extend the pointwise converses of Section 13.3.1. For simplicity we focus on absolute losses $\Phi(t) = |t|$.

To that end, we define the *mixture modulus*, which extends the local modulus (13.3.3), by

$$\overline{\omega}_{\chi^2}(\epsilon; \theta, P_0, \mathcal{P}) := \tag{13.3.5}$$
$$\sup_{m \in \mathbb{N}} \sup_{P_i \in \mathcal{P}} \left\{ \min_{i \leq m} \{|\theta(P_0) - \theta(P_i)|\} \mid \log \rho_{\chi^2}\left(\overline{P} \| P_0\right) \leq 2\epsilon^2 \text{ for some } \overline{P} \in \operatorname{Conv}\{P_i\}_{i=1}^m \right\}.$$

The quantity (13.3.5) measures how much it is possible to perturb the parameter θ while making sure that some element of the convex hull $\operatorname{Conv}\{P_i\}_{i=1}^m$ is close to the base distribution P_0. We abuse notation slightly to let $\mathcal{P}^n = \{P^n\}_{P \in \mathcal{P}}$ be the collection of product distributions and

$$\overline{\omega}_{\chi^2}(\epsilon; \theta, P_0^n, \mathcal{P}^n) :=$$
$$\sup_{m \in \mathbb{N}} \sup_{P_i \in \mathcal{P}} \left\{ \min_{i \leq m} \{|\theta(P_0) - \theta(P_i)|\} \mid \log \rho_{\chi^2}\left(\overline{P^n} \| P_0^n\right) \leq 2\epsilon^2 \text{ for some } \overline{P^n} \in \operatorname{Conv}\{P_i^n\}_{i=1}^m \right\}.$$

The mixture modulus will provide a benchmark sufficient to guarantee pointwise super-efficiency converses. First, however, we note that as a corollary of the convex hull method in Theorem 13.2.1, that the mixture modulus always lower bounds the minimax risk.

Corollary 13.3.7. *Let the conditions of Theorem 13.2.1 hold. Then for each n and $P_0 \in \mathcal{P}$,*

$$\mathfrak{M}_n(\theta(\mathcal{P}), |\cdot|) \geq \frac{1}{4}\overline{\omega}_{\chi^2}\left(\frac{1}{2}; \theta, P_0^n, \mathcal{P}^n\right).$$

Proof Let $P_0 \in \mathcal{P}$ and $\mathcal{P}_1 \subset \mathcal{P}$ be any δ-separated collection, that is, satisfying $|\theta(P_0) - \theta(P)| \geq \delta$ for all $P \in \mathcal{P}_1$. Then Theorem 13.2.1 implies that

$$\mathfrak{M}_n(\theta(\mathcal{P}), |\cdot|) \geq \frac{\delta}{2}\left(1 - \left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}}\right)$$

for any $\overline{P^n} \in \mathrm{Conv}\{\mathcal{P}_1^n\}$. By Pinsker's inequality (Proposition 2.2.9) and Proposition 2.2.10,

$$2\left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}}^2 \le \log \rho_{\chi^2}\left(\overline{P^n} \| P_0^n\right),$$

so if $\log \rho_{\chi^2}\left(\overline{P^n}\|P_0^n\right) \le \frac{1}{2}$ then $\left\|P_0^n - \overline{P^n}\right\|_{\mathrm{TV}} \le \frac{1}{2}$. Take a supremum over the convex hull. □

We turn to risk transfer inequalities, where we show in analogy with Theorem 13.3.1 that achieving small risk at some distribution P_0 implies achieving larger risk at others.

Theorem 13.3.8. *Let $P_v \in \mathcal{P}$ for $v = 0, 1, \ldots, M$, let $\theta_v = \theta(P_v)$ for each v, and define the separation $\Delta = \min_{v \ge 1} |\theta_0 - \theta_v|$. If $\mathbb{E}_{P_0}[|\widehat{\theta} - \theta_0|] \le \gamma$, then for any $\lambda \in \mathbb{R}_+^M$ with $\langle \lambda, \mathbf{1}\rangle = 1$ and $P_\lambda := \sum_{v=1}^M \lambda_v P_v$,*

$$\sum_{v=1}^M \lambda_v \mathbb{E}_{P_v}\left[|\widehat{\theta} - \theta_v|\right] \ge \left[\sqrt{\Delta} - \sqrt{\gamma \rho_{\chi^2}\left(P_\lambda \| P_0\right)}\right]_+^2.$$

Proof The proof mirrors that of Theorem 13.3.1. Fix $v \in [M]$. For any $\theta \in \mathbb{R}$, we have as in inequality (13.3.2) that

$$\sqrt{|\theta - \theta_0|} + \sqrt{|\theta - \theta_v|} \ge \sqrt{|\theta_0 - \theta_v|} \ge \sqrt{\Delta}.$$

So for any v,

$$\mathbb{E}_v\left[|\widehat{\theta} - \theta_v|^{1/2} + |\widehat{\theta} - \theta_0|^{1/2}\right] \ge \sqrt{\Delta},$$

implying by Cauchy-Schwarz that

$$\mathbb{E}_v[|\widehat{\theta} - \theta_v|] \ge \mathbb{E}_v\left[|\widehat{\theta} - \theta_v|^{1/2}\right]^2 \ge \left[\sqrt{\Delta} - \mathbb{E}_v\left[|\widehat{\theta} - \theta_0|^{1/2}\right]\right]_+^2.$$

As λ forms a convex combination, the convexity of $t \mapsto [t]_+^2$ yields by Jensen's inequality that

$$\sum_v \lambda_v \mathbb{E}_v\left[|\widehat{\theta} - \theta_v|\right] \ge \left[\sqrt{\Delta} - \sum_v \lambda_v \mathbb{E}_v\left[|\widehat{\theta} - \theta_0|^{1/2}\right]\right]_+^2.$$

Applying the Cauchy-Schwarz inequality once again, we obtain

$$\begin{aligned}\sum_v \lambda_v \mathbb{E}_v\left[|\widehat{\theta} - \theta_0|^{1/2}\right] = \mathbb{E}_0\left[\frac{dP_\lambda}{dP_0}|\widehat{\theta} - \theta_0|^{1/2}\right] &\le \mathbb{E}_0\left[\frac{dP_\lambda^2}{dP_0^2}\right]^{1/2} \mathbb{E}_0\left[|\widehat{\theta} - \theta_0|\right]^{1/2} \\ &= \rho_{\chi^2}\left(P_\lambda \| P_0\right)^{1/2} \mathbb{E}_0\left[|\widehat{\theta} - \theta_0|\right]^{1/2}.\end{aligned}$$

The theorem follows. □

We can leverage Theorem 13.3.8 to show how the convex hull modulus (13.3.5) implies super-efficiency converses. For shorthand in the corollary, we define

$$\overline{\omega}_n(\epsilon) := \overline{\omega}_{\chi^2}(\epsilon; \theta, P_0^n, \mathcal{P}^n).$$

Corollary 13.3.9. *Let $P_0 \in \mathcal{P}$, and let $\widehat{\theta}_n$ satisfy $\mathbb{E}_{P_0}[|\widehat{\theta}_n(X_1^n) - \theta(P_0)|] \le \gamma\overline{\omega}_n(1)$. Then there exists a collection of distributions $\{P_v\}_{v=1}^M \subset \mathcal{P}$ and a vector $\lambda \in \mathbb{R}_+^M$, $\langle \mathbf{1}, \lambda\rangle = 1$, such that*

$$\sum_{v=1}^M \lambda_v \mathbb{E}_{P_v^n}\left[|\widehat{\theta}_n(X_1^n) - \theta_v|\right] \ge \left[\sqrt{\overline{\omega}_n(1/\sqrt{\gamma})} - \sqrt{\overline{\omega}_n(1)}\right]_+^2.$$

Proof By Theorem 13.3.8, for any collection $P_1, \ldots, P_M \in \mathcal{P}$ with associated parameters $\theta_v = \theta(P_v)$, if we let $\Delta = \min_v |\theta_v - \theta_0|$ then

$$\sum_{v=1}^M \lambda_v \mathbb{E}_{P_v^n}\left[|\widehat{\theta}(X_1^n) - \theta_v|\right] \ge \left[\sqrt{\Delta} - \sqrt{\gamma\overline{\omega}_n(1)\rho_{\chi^2}\left(P_\lambda^n \| P_0^n\right)}\right]_+^2$$

for any vector $\lambda \in \mathbb{R}_+^M$ with $\langle \mathbf{1}, \lambda\rangle = 1$, where $P_\lambda^n = \sum_v \lambda_v P_v^n$. So long as $\log \rho_{\chi^2}\left(P_\lambda^n \| P_0^n\right) \le \frac{1}{\gamma}$, we therefore have

$$\sum_{v=1}^M \lambda_v \mathbb{E}_{P_v^n}\left[|\widehat{\theta}(X_1^n) - \theta_v|\right] \ge \left[\sqrt{\Delta} - \sqrt{\overline{\omega}_n(1)}\right]_+^2.$$

Take a supremum over all such convex combinations of $\mathcal{P}^n = \{P^n\}_{P\in\mathcal{P}}$ and let $\Delta = \overline{\omega}_n(1/\sqrt{\gamma})$. □

Corollary 13.3.9 shows how the convex-hull modulus (13.3.5) around P_0 provides an unbeatable bound: if an estimator achieves a faster convergence rate by a factor $\gamma < 1$ than $\overline{\omega}_n(1)$, then on average over some collection of alternative distributions, its risk must scale at least as the modulus across a larger neighborhood.

Example 13.3.10 (An integral functional)**:** We revisit estimation of $T_k(f) := \int_0^1 (f^{(k)}(x))^2 dx$ given observations $Y_i = f(X_i) + \varepsilon_i$, as in Section 13.2.3*. Take $k = s = 1$, so that f has bounded first derivative and we wish to estimate $T_1(f) = \int f'(x)^2 dx$. Suppose for the identically 0 function $f_0(x) = 0$, the estimator $\widehat{T}_n$ has risk

$$\mathbb{E}_{f_0}[|\widehat{T}_n|] \le \gamma_n n^{-\frac{4}{9}}, \quad \text{where} \quad \gamma_n \to 0,$$

which is (asymptotically) faster than the lower bound in Proposition 13.2.9. Corollary 13.3.9 implies that for the class $\mathcal{F}_2 = \{f : [0,1] \to \mathbb{R} \mid \|f''\|_\infty \le 1, f(0) = 0\}$, we have

$$\frac{\sup_{f\in\mathcal{F}_2} \mathbb{E}_f[|\widehat{T}(X_1^n, Y_1^n) - T_1(f)|]}{n^{-4/9}} \to \infty.$$

To see this limit, we revisit the construction in the proof of Proposition 13.2.9. Let P_f be the induced distribution of observations under $Y_i = f(X_i) + \varepsilon_i$, where P_0 corresponds to $f = 0$. Equations (13.2.6) and (13.2.7) imply there exist positive numerical constants c_0, c_1, c_2 such that for each $m, n \in \mathbb{N}$, there is a collection of functions $\mathcal{G}$ such that

$$T_1(f) \ge \frac{c_0}{m^2} \text{ for } f \in \mathcal{G} \quad \text{and} \quad \rho_{\chi^2}\left(\overline{P^n} \| P_0\right) \le \exp\left(\frac{c_1 n^2}{m^9} + \frac{n c_2}{m^8}\right),$$

where $\overline{P^n} = \frac{1}{|\mathcal{G}|}\sum_{f\in\mathcal{G}} P_f^n$. In particular, this implies that

$$\overline{\omega}_{\chi^2}(\epsilon; T_1, P_0^n, \mathcal{P}^n) \gtrsim \sup_{m\in\mathbb{N}}\left\{\frac{1}{m^2} \mid \frac{n^2}{m^9} + \frac{n}{m^8} \le \epsilon^2\right\} \gtrsim \left(\frac{\epsilon}{n}\right)^{4/9},$$

where we chose $m = (n/\epsilon)^{2/9}$. By Corollary 13.3.9, we thus have

$$\sup_{f \in \mathcal{F}_2} \mathbb{E}_{P_f^n}\left[|\widehat{T}_n - T_1(f)|\right] \gtrsim \left(\frac{1}{n\gamma_n}\right)^{4/9} \gg \frac{1}{n^{4/9}},$$

a quantitative lower bound on the risk of $\widehat{T}_n$. $\diamond$

Example 13.3.10 precludes some types of adaptivity: the all-zeros function belongs to the smoothness class $\mathcal{F}_s = \{f \mid \|f^{(s)}\|_\infty \le 1\}$ for all $s \in \mathbb{N}$. See the bibliographic section for commentary.

13.4 Minimax hypothesis testing

Section 12.3 discusses multiple hypothesis testing and detection when we observe many signals; we investigate a related signal detection problem here, where we observe $X \sim P$ from an unknown distribution P, and wish to detect if there is some "signal" present in P. We formulate a null model and a set of alternatives for which there is *some* signal, though we only care to test its existence. In distinction from Sections 12.3.3 and 12.3.4, we formulate this via a general composite hypothesis testing problem, where we have a family of potential distributions $\mathcal{P}$ and receive a sample $X \sim P$ for some $P \in \mathcal{P}$. We wish to distinguish hypotheses H_0 and H_1,

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1, \tag{13.4.1}$$

where the collections $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1 \subset \mathcal{P}$ are disjoint. When $\mathcal{P}_0$ and $\mathcal{P}_1$ are singletons, Le Cam's inequalities (Section 2.3.1) tell us precisely how difficult the test is, characterizing error probabilities via the variation distance. The composite case requires more care.

For a test $\Psi : \mathcal{X} \to \{0, 1\}$, we define the *test risk*

$$R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) := \sup_{P \in \mathcal{P}_0} P(\Psi \neq 0) + \sup_{P \in \mathcal{P}_1} P(\Psi \neq 1),$$

that is, the sum of the worst-case probabilities that the test is incorrect. Abstractly, we assume a metric ρ on the family of distributions $\mathcal{P}$, and instead of the general hypothesis test (13.4.1), we will consider testing whether $P \in \mathcal{P}_0$ or $\rho(P, P_0) \ge \epsilon$ for all $P_0 \in \mathcal{P}$, giving the variant

$$\begin{aligned} H_0 : &\quad P \in \mathcal{P}_0 \\ H_1 : &\quad P \in \mathcal{P}_1(\epsilon) := \{P \in \mathcal{P} \text{ s.t. } \rho(P, P_0) \ge \epsilon \text{ all } P_0 \in \mathcal{P}_0\} \end{aligned} \tag{13.4.2}$$

In this case, we can define the *risk at distance* ϵ for a sample of size n by

$$R_n(\Psi, \epsilon) := \sup_{P \in \mathcal{P}_0} P(\Psi(X_1^n) \neq 0) + \sup_{P \in \mathcal{P}_1(\epsilon)} P(\Psi(X_1^n) \neq 1), \tag{13.4.3}$$

leaving $\mathcal{P}_0$ and $\mathcal{P}$ implicit in the definition, and where we let $X_1^n \stackrel{\text{iid}}{\sim} P$. This leads to the *minimax test risk* at separation ϵ,

$$\inf_{\Psi} R_n(\Psi, \epsilon). \tag{13.4.4}$$

We use this quantity to identify the separation ϵ at which testing is possible (or impossible). Thus, while the definition allows some ambiguity, we say a sequence ϵ_n provides rate-optimal *critical testing radius* for problem (13.4.2) if there exist constants $0 < c \le C < \infty$ such that

$$\limsup_n \inf_{\Psi} R_n(\Psi, C\epsilon_n) \le \frac{1}{3} \quad \text{and} \quad \liminf_n \inf_{\Psi} R_n(\Psi, c\epsilon_n) \ge \frac{2}{3}. \tag{13.4.5}$$

For separation $c\epsilon_n$, no method can accurately test whether $P \in \mathcal{P}_0$ or $\inf_{P_0 \in \mathcal{P}_0} \rho(P, P_0) \geq c\epsilon_n$, but it is possible to test whether $P \in \mathcal{P}_0$ or $\inf_{P_0 \in \mathcal{P}_0} \rho(P, P_0) \geq C\epsilon_n$. We can make the probability of error exponentially small by increasing the sample size by a constant factor (see Exercise 13.3).

An analogous form to Le Cam's convex hull lower bound in Theorem 13.2.1 provides the same fundamental quantity (the variation distance between convex hulls of $\mathcal{P}_0$ and $\mathcal{P}_1$) for lower bounds in testing, so we may apply the techniques comparing different divergences we have developed. In particular, we have the following testing analogue of Le Cam's convex hull lower bound.

Proposition 13.4.1 (Convex hull lower bounds in testing)**.** *For any classes $\mathcal{P}_0$ and $\mathcal{P}_1$, the minimax test risk satisfies*

$$\inf_{\Psi} R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \geq 1 - \inf \left\{ \left\| \overline{P}_0 - \overline{P}_1 \right\|_{\mathrm{TV}} \mid \overline{P}_0 \in \mathrm{Conv}(\mathcal{P}_0), \overline{P}_1 \in \mathrm{Conv}(\mathcal{P}_1) \right\}.$$

Proof Let $\overline{P}_0 \in \mathrm{Conv}(\mathcal{P}_0)$ and $\overline{P}_1 \in \mathrm{Conv}(\mathcal{P}_1)$. Then for any test Ψ,

$$R(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \geq \overline{P}_0(\Psi \neq 0) + \overline{P}_1(\Psi \neq 1)$$

because suprema are always at least as large as averages. Note that $A = \{x \mid \Psi(x) = 0\}$ satisfies

$$\overline{P}_0(\Psi \neq 0) + \overline{P}_1(\Psi \neq 1) = \overline{P}_0(A^c) + \overline{P}_1(A) = 1 - (\overline{P}_0(A) - \overline{P}_1(A)),$$

and take an infimum over regions A. □

Equality typically holds, but this requires (infinite dimensional) minimax theory (cf. [282] or [251, Theorem 2.1]).

13.4.1 Detecting a difference in populations

With the worst-case hypothesis testing setup in place, we can give a general recipe for developing tests. We specialize this recipe to provide examples in signal detection in a Gaussian model and two-sample (or goodness of fit) tests for multinomials, highlighting applications of the mixture approach. In each problem, to demonstrate achievability, we develop an estimator T_n of (a function of) the distance $\rho(P_0, P_1)$ then construct a test Ψ by thresholding this estimator. For the converse results that no test can distinguish the families $\mathcal{P}_0$ and $\mathcal{P}_1$ at a particular distance, we use the mixture χ^2 approaches we have outlined.

Let us give the general recipe. A statistic T separating $\mathcal{P}_0$ and $\mathcal{P}_1$ should assign large values for samples $X \sim P_1$ for $P_1 \in \mathcal{P}_1$ and small values for samples $X \sim P_0$. To provide guarantees, we require the separation $\mathbb{E}_1[T] - \mathbb{E}_0[T]$ to be commensurate with the variance of T. To that end, we say that the statistic T *robustly C-separates* $\mathcal{P}_0$ and $\mathcal{P}_1$ if

$$\mathbb{E}_{P_1}[T] - \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] \geq C \left(\sup_{P_0 \in \mathcal{P}_0} \sqrt{\mathrm{Var}_{P_0}(T)} + \sqrt{\mathrm{Var}_{P_1}(T)} \right). \tag{13.4.6}$$

for each $P_1 \in \mathcal{P}_1$. Typically, we choose T so that $\mathbb{E}_{P_0}[T] = 0$ for each P_0 in the null $\mathcal{P}_0$. A separating T immediately gives a low error test:

Proposition 13.4.2. *Let the statistic $T : \mathcal{X} \to \mathbb{R}$ robustly C-separate P_0 from $\mathcal{P}_1$. Define the threshold $\tau = \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] + C \sup_{P_0 \in \mathcal{P}_0} \sqrt{\mathrm{Var}_{P_0}(T)}$. Then the test*

$$\Psi(X) := \mathbf{1}\{T \geq \tau\}$$

satisfies

$$R(\Psi \mid \{P_0\}, \mathcal{P}_1) \leq \frac{2}{C^2}.$$

Proof Without loss of generality we assume $\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[T] = 0$, as the test is invariant to shifts, so $\tau = C \sup_{P_0 \in \mathcal{P}_0} \sqrt{\mathrm{Var}_{P_0}(T)}$. Under any null P_0, Chebyshev's inequality gives

$$P_0(\Psi \neq 0) = P_0(T \geq \tau) \leq \frac{\mathrm{Var}_{P_0}(T)}{\tau^2} \leq \frac{1}{C^2}.$$

For alternatives $P_1 \in \mathcal{P}_1$,

$$P_1(\Psi \neq 1) = P_1(T \leq \tau) = P_1(T - \mathbb{E}_{P_1}[T] \leq \tau - \mathbb{E}_{P_1}[T]) \leq \frac{\mathrm{Var}_{P_1}(T)}{[\mathbb{E}_{P_1}[T] - \tau]_+^2}.$$

By the robust C-separation,

$$\mathbb{E}_1[T] - \tau \geq C\sqrt{\mathrm{Var}_{P_1}(T)} + C \sup_{P_0} \sqrt{\mathrm{Var}_{P_0}(T)} - C \sup_{P_0} \sqrt{\mathrm{Var}_{P_0}}(T) = C\sqrt{\mathrm{Var}_{P_1}(T)}.$$

So $P_1(\Psi \neq 1) \leq \frac{1}{C^2}$. □

13.4.2 Signal detection and testing a Gaussian mean

Returning to the motivating signal detection problem—whether there exists a signal at all—we give two examples.

Example 13.4.3 (Dense Gaussian signal detection)**:** Consider testing

$$\begin{aligned} H_0 : &\quad P = P_0 = \mathsf{N}(0, I_d) \quad \text{against} \\ H_1 : &\quad P \in \mathcal{P}_1(r) := \{\mathsf{N}(\theta, I_d) \mid \|\theta\|_2 \geq r\}. \end{aligned} \tag{13.4.7}$$

That is, we are interested in whether $X \sim P$ has a mean θ separated (in ℓ_2) by at least r from the all-zeros vector. The problem is to find the critical radius r at which testing between $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1$ becomes feasible (or infeasible). ◇

For the next example, we recall Section 13.2.2, reminding ourselves that a p-value for a hypothesis H_0 is any sub-uniform random variable Y, meaning that $P(Y \leq u) \leq u$ for $u \in [0, 1]$.

Example 13.4.4 (A global null in multiple hypothesis testing)**:** Consider the problem of testing d distinct null hypotheses $H_{0,j}$, $j = 1, \ldots, d$, where for each we have a p-value Y_j and reject $H_{0,j}$ if $Y_j \leq \tau$ for a threshold τ. If we assume the Y_j are exact p-values, that is, $P(Y_j \leq u) = u$ for $u \in [0, 1]$, then testing the global independent null

$$H_0 := \bigcap_{j=1}^d H_{0,j} = \left\{\text{each } Y_j \overset{\text{iid}}{\sim} \mathsf{Uniform}[0, 1]\right\}$$

is equivalent to Gaussian signal detection, as in Sec. 13.2.2, with the transformation $Z_j = \Phi^{-1}(Y_j)$, where Φ denotes the standard Gaussian CDF. Under the global null H_0,

$$Z \sim \mathsf{N}(0, I_d).$$

The question of the alternative class $\mathcal{P}_1$ depends on the application; Example 13.4.3 corresponds to something like dense alternatives. ◇

With these as motivation, let us find the critical radius r at which minimax testing becomes feasible in Example 13.4.3. While standard techniques for estimation give minimax mean squared error

$$\mathfrak{M}_n(\theta(\mathcal{P}), \|\cdot\|_2^2) = \frac{d\sigma^2}{n},$$

for estimating θ in a normal location family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2 I_d)\}_{\theta \in \mathbb{R}^d}$, we can *test* whether the mean of a Gaussian is zero at a smaller sample size. Effectively, we will see that we can save a dimension-dependent factor $\sqrt{d}$. The next two examples—one addressing achievability and one the fundamental limit—show that in the dense Gaussian signal detection problem of Example 13.4.3, the critical test radius (13.4.5) at which testing is feasible or infeasible scales as

$$r_n := \frac{d^{1/4}}{\sqrt{n}}.$$

We can achieve (asymptotically) accurate testing in the dense signal detection problem (13.4.7) if and only if $\sqrt{d}/n \to 0$ as $n \to \infty$.

We first leverage Proposition 13.4.2 to demonstrate achievability in Example 13.4.3.

Example 13.4.5 (Achievability in Gaussian mean testing)**:** We find an estimator of $\|\theta\|_2^2$, and then threshold it. The discussion preceding Corollary 13.2.8 (equation (13.2.3)) shows that the estimator $T_n = \|\overline{X}_n\|_2^2 - d/n$ is unbiased for $\|\theta\|_2^2$ and satisfies

$$\mathbb{E}_\theta\left[(T_n - \|\theta\|_2^2)^2\right] = \mathrm{Var}_\theta(T_n) \le \frac{2d}{n^2} + \frac{4\|\theta\|_2^2}{n}. \tag{13.4.8}$$

As $\mathbb{E}_0[T_n] = 0$ and $\mathbb{E}_\theta[T_n] - \mathbb{E}_0[T_n] = \|\theta\|_2^2$, fixing any $K \ge 1$, the statistic T_n robustly K-separates P_0 from $\mathcal{P}_1(r)$ (recall definition (13.4.6)) whenever

$$\|\theta\|_2^2 \ge K\left(\frac{\sqrt{2d}}{n} + \sqrt{\frac{2d}{n^2} + \frac{4}{n}\|\theta\|_2^2}\right)$$

for θ with $\|\theta\|_2 \ge r$. Taking radius $r^2 = C\frac{\sqrt{d}}{n}$ for a large enough constant C ensures this separation with $K = 3$. Applying Proposition 13.4.2, the test $\Psi(X_1^n) = \mathbf{1}\{T_n \ge 3\sqrt{2d}/n\}$ satisfies

$$R_n(\Psi, Cr_n) \le \frac{1}{3} \quad \text{for} \quad r_n = \frac{d^{1/4}}{\sqrt{n}},$$

which gives the achievability required for the critical test radius (13.4.5). $\diamond$

We can also provide the converse to show that the rate $r_n = \frac{d^{1/4}}{\sqrt{n}}$ in Example 13.4.5 is sharp.

Example 13.4.6 (Lower bounds in Gaussian mean testing)**:** Let $\mathcal{P}_1(r) = \{\mathsf{N}(\theta, I_d) \mid \|\theta\|_2 \ge r\}$ be a collection of Gaussians with means r-separated from the origin. We seek the radius r below which distinguishing $P_0 = \mathsf{N}(0, I_d)$ and $P_1 \in \mathcal{P}_1(r)$ given an i.i.d. sample X_1^n is impossible. Lemma 13.2.6 and Proposition 13.4.1 combine (set $\delta^2 = \frac{r^2}{d}$ in Lemma 13.2.6) to give

$$\inf_\Psi R_n(\Psi \mid \{P_0\}, \mathcal{P}_1(r)) \ge 1 - \sqrt{\frac{n^2 r^4}{4d}}.$$

The threshold $r^2 = \sqrt{d}/n$ yields constant test error probability $R_n \ge \frac{1}{2}$. $\diamond$

Combining the estimation guarantee with this lower bound shows that $r^2 = \frac{\sqrt{d}}{n}$ identifies the critical radius (13.4.5) for testing $H_0 : \mathsf{N}(0, I_d)$ against the alternatives $H_1 : \mathsf{N}(\theta, I_d)$ with $\|\theta\|_2^2 \ge r^2$.

13.4.3 Goodness of fit and two-sample tests for multinomials

In goodness of fit testing (called "property testing" in the theoretical computer science literature), given a sample $X_1, \ldots, X_n \stackrel{\iid}{\sim} P$, we wish to test whether $P = P_0$ for a prescribed base distribution P_0 or P is far from P_0. The related two-sample testing problem generalizes this, where we assume samples $X_1, \ldots, X_n \stackrel{\iid}{\sim} P$ and $Y_1, \ldots, Y_m \stackrel{\iid}{\sim} Q$, and wish to test whether $P = Q$. Each of these falls into the class of hypothesis tests (13.4.2), where the choice of the metric ρ can change the character of convergence rates dramatically. Methods for developing goodness of fit and two-sample tests typically take the broad approach in the introductory Section 13.4.1, defining a statistic T that separates the distribution P_0 (or the joint that X_i and Y_j have the same distribution) from the alternatives about which we are curious, then thresholding that statistic.

Even in what might appear to be a particularly simple case—multinomial distributions, where P corresponds to a probability mass function (p.m.f.) $p \in \Delta_d = \{q \in \mathbb{R}_+^d \mid \langle \mathbf{1}, q \rangle = 1\}$—surprising complexity arises. To give a taste for this, we work through example calculations for both the ℓ_1- (total variation) and ℓ_2-metrics on the p.m.f.s. In the uniformity testing case, we let $p_0 = \frac{1}{d}\mathbf{1}$ be the uniform distribution on $[d]$, and we seek the critical threshold ϵ at which testing

$$\|p - p_0\| = 0 \quad \text{versus} \quad \|p - p_0\| \geq \epsilon$$

from n i.i.d. observations $X_i \stackrel{\iid}{\sim} p$ becomes feasible or infeasible.

It is simpler to analyze a variant of this problem using the *Poissonization* trick. To motivate the idea, identify the observations X_i with the basis vectors (so that observing item $j \in \{1, \ldots, d\}$ corresponds to $X_i = e_j$). Then the sample mean $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ is unbiased, but its coordinates exhibit the nontrivial dependence that $\langle \mathbf{1}, \widehat{p} \rangle = 1$. Thus, we consider an alternative approach, where we assume a two-stage sampling procedure: we first draw $N \sim \mathsf{Poi}(n)$, and then conditional on $N = m$, draw $X_1^m \stackrel{\iid}{\sim} p$. As $\mathbb{E}[N] = n$ and N concentrates around its mean, this is nearly equivalent to observing $X_i \stackrel{\iid}{\sim} p$ for $i = 1, \ldots, n$, and a standard calculation shows that X_1^N conditional on $N = m$ and $X_1^m \stackrel{\iid}{\sim} p$ have identical distribution.

We can make this near equivalence quantitative: the minimax risk for estimation in the Poissonized sampling scheme tracks that for estimation in the original multinomial setting. Indeed, suppose that we wish to estimate an abstract statistic $T(p)$ of $p \in \Delta_d$, and assume for simplicity that $T(p) \in [-r, r]$ for some fixed r. Define the minimax and Poissonized minimax risks

$$\mathfrak{M}_n := \inf_{T_n} \sup_{p \in \Delta_d} \mathbb{E}_p\left[(T_n(X_1^n) - T(p))^2\right]$$

and

$$\mathfrak{M}_{\mathsf{Poi}(n)} := \inf_{\{T_m\}} \sup_{p \in \Delta_d} \mathbb{E}_p\left[(T_N(X_1^N) - T(p))^2\right],$$

where the latter expectation is taken over the sample size $N \sim \mathsf{Poi}(n)$, and $\{T_m\}$ denotes a sequence of estimators (defined for all sample sizes m). Then Poissonized risk and $\mathfrak{M}_n$ are similar:

Proposition 13.4.7. *Assume the statistic $T(p) \in [-r, r]$ for all $p \in \Delta_d$. There exist numerical constants $0 < c, C < \infty$ such that*

$$\mathfrak{M}_{\mathsf{Poi}(2n)} - Cr^2 \exp(-cn) \leq \mathfrak{M}_n \leq 2 \cdot \mathfrak{M}_{\mathsf{Poi}(n/2)}. \tag{13.4.9}$$

Exercises 13.4 and 13.5 work through a proof of this result.

We leverage these ideas to construct an estimator for the ℓ_2-distance between multinomial distributions $p, q \in \Delta_d$, identifying the critical testing radius along the way. Identifying $X \sim p$ by coordinate basis vectors, so $X \in \{e_j\}_{j=1}^d$, we have $N \sim \mathsf{Poi}(n)$ and $X_1^N \stackrel{\text{iid}}{\sim} p$ and $Y_1^N \stackrel{\text{iid}}{\sim} q$ conditional on N, and we define

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^N X_i, \quad \widehat{q} = \frac{1}{n}\sum_{i=1}^N Y_i. \tag{13.4.10}$$

This is equivalent to sampling $n\widehat{p}_j \stackrel{\text{ind}}{\sim} \mathsf{Poi}(np_j)$ and $n\widehat{q}_j \stackrel{\text{ind}}{\sim} \mathsf{Poi}(nq_j)$, $j = 1, \ldots, d$, and so we use the quantities (13.4.10) to define an estimator of $\|p - q\|_2^2$, which we can threshold and apply via Proposition 13.4.2. We begin with unbiased estimation of $\|p - q\|_2^2$.

Example 13.4.8 (The ℓ_2-distance between multinomials)**:** We claim that for any $C \ge 1$, if multinomials p and q satisfy

$$\|p - q\|_2^2 \ge C\left(\sqrt{\frac{8}{n^2}} + \sqrt{\frac{16\|p - q\|_4^2}{n} + \frac{8}{n^2}}\right), \tag{13.4.11}$$

then there exists a test $\Psi = \Psi_n$ of $H_0 : p = q$ against the alternative H_1 that p and q satisfy inequality (13.4.11) with risk $P_0(\Psi \neq 0) + P_1(\Psi \neq 1) \le \frac{2}{C^2}$. As $\|p - q\|_2 \ge \|p - q\|_4$, a necessary and sufficient condition for inequality (13.4.11) to hold is that $\|p - q\|_2 \gtrsim 1/\sqrt{n}$.

To see this, we construct an unbiased estimator of $\|p - q\|_2^2$ and control its variance. For the estimators (13.4.10), define

$$Z_j := (n\widehat{p}_j - n\widehat{q}_j)^2 - n\widehat{p}_j - n\widehat{q}_j.$$

If $W \sim \mathsf{Poi}(\lambda)$ then $\mathbb{E}[W] = \mathrm{Var}(W) = \lambda$, so $\mathbb{E}[n\widehat{p}_j] = \mathrm{Var}(n\widehat{p}_j) = np_j$, and

$$\begin{aligned}\mathbb{E}[Z_j] &= \mathbb{E}[(n\widehat{p}_j)^2] + \mathbb{E}[(n\widehat{q}_j)^2] - 2n^2p_jq_j - np_j - nq_j \\ &= \mathrm{Var}(n\widehat{p}_j) + \mathrm{Var}(n\widehat{q}_j) + (np_j)^2 + (nq_j)^2 - 2n^2p_jq_j - np_j - nq_j = n^2(p_j - q_j)^2.\end{aligned}$$

In particular,

$$T_n := \frac{1}{n^2}\langle \mathbf{1}, Z\rangle \quad \text{satisfies} \quad \mathbb{E}[T_n] = \|p - q\|_2^2.$$

To test whether p and q are identical using Proposition 13.4.2, we compute $\mathrm{Var}(\langle \mathbf{1}, Z\rangle) = \sum_{j=1}^d \mathrm{Var}(Z_j)$. Leveraging (via tedious calculation) that $W \sim \mathsf{Poi}(\lambda)$ has moments

$$\mathbb{E}[W] = \lambda, \ \mathbb{E}[W^2] = \lambda(1 + \lambda), \ \mathbb{E}[W^3] = \lambda + 3\lambda^2 + \lambda^3, \ \mathbb{E}[W^4] = \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4,$$

we obtain (see Exercise 13.10)

$$\mathrm{Var}(Z_j) = 4n^3(p_j - q_j)^2(p_j + q_j) + 2(p_j + q_j)^2n^2 \tag{13.4.12}$$

and so by Cauchy-Schwarz,

$$\mathrm{Var}(\langle \mathbf{1}, Z\rangle) \le 4n^3\|p - q\|_4^2\|p + q\|_2 + 2n^2\|p + q\|_2^2.$$

Under the null $H_0 : p = q$, $\mathrm{Var}(\langle \mathbf{1}, Z\rangle) = 2n^2\|p + q\|_2^2 \le 8n^2$, as $\sup_{p,q}\|p + q\|_2 = 2$. Proposition 13.4.2 thus shows that the test

$$\Psi := \mathbf{1}\left\{T_n \ge C\sqrt{8}/n\right\}$$

satisfies $P_0(\Psi \neq 0) + P_1(\Psi \neq 1) \le \frac{2}{C^2}$, where P_0 is any distribution with $p = q$ and P_1 is any distribution with $\|p - q\|_2$ satisfying the separation (13.4.11). $\diamond$

Summarizing, we see that if we wish to test whether two multinomials are identical or separated in ℓ_2, the critical threshold for the hypothesis test

$$H_0 : p = q \quad \text{versus} \quad H_1 : \|p - q\|_2 \geq \delta \tag{13.4.13}$$

satisfies $\delta \leq \frac{1}{\sqrt{n}}$: we can test between H_0 and H_1 at separations that are essentially "independent" of the dimension or number of categories d. This is in fact sharp, as a relatively straightforward argument with Le Cam's two-point lemma demonstrates (see Exercise 13.12). However, if we change the norm $\|\cdot\|_2$ into the ℓ_1-norm $\|\cdot\|_1$, the story changes significantly.

Let us change the hypothesis test (13.4.13) to a goodness of fit test. Let P_0 be the uniform distribution on $\{1, \ldots, d\}$, with p.m.f. $p_0 = \frac{1}{d}\mathbf{1}$. We consider the distance to uniformity test

$$H_0 : p = p_0 = \frac{1}{d}\mathbf{1} \quad \text{versus} \quad H_1 : p \in \mathcal{P}_1(\delta) := \{p \in \Delta_d \mid \|p - p_0\|_1 \geq \delta\}. \tag{13.4.14}$$

Developing a test that distinguishes these hypotheses at the optimal rate follows a similar argument to that for the ℓ_2-case, and we outline an approach to it in Exercise 13.6. To develop the correct threshold δ for which no test can reliably distinguish H_0 from H_1, we use the mixture of χ^2-distributions approach from Lemma 13.2.3.

Proposition 13.4.9. *In the testing problem* (13.4.14)*, whenever* $\delta \leq d^{1/4}/\sqrt{n}$*,*

$$\inf_{\Psi} R_n(\Psi \mid \{P_0\}, \mathcal{P}_1(\delta)) \geq 1 - \frac{1}{\sqrt{2}}.$$

Proof We construct a packing of the probability simplex $\Delta_d \subset \mathbb{R}_+^d$ that guarantees that the divergence between elements of H_0 and H_1 in the test (13.4.14) is small. For simplicity, we assume d is even. For the base distribution P_0 take p.m.f. $p_0 = \frac{1}{d}\mathbf{1}$ as required by the problem (13.4.14). To construct the alternatives, let $\mathcal{V} \subset \{\pm 1\}^d$ be the collection of $2^{d/2}$ vectors of the form $v = (v', -v')$, where $v' \in \{\pm 1\}^{d/2}$, so that $\langle \mathbf{1}, v\rangle = 0$ for each $v \in \mathcal{V}$. Then for $\delta \geq 0$ to be chosen, define the p.m.f.s $p_v = \frac{1+\delta v}{d}$. Identifying samples $X \in \{e_1, \ldots, e_d\}$, for any $x \in \{e_j\}$, we have $P_v(X = x) = \frac{1}{d}(1 + \delta\langle v, x\rangle)$, and so for any pair v, v' we have

$$\frac{P_v(X = x)P_{v'}(X = x)}{P_0(X = x)^2} = (1 + \delta\langle v, x\rangle)(1 + \delta\langle v', x\rangle).$$

From this key equality, we see that if $V, V' \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V})$, then $\overline{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v^n$ satisfies

$$\begin{aligned} 1 + D_{\chi^2}\left(\overline{P}\|P_0^n\right) &= \mathbb{E}_0\left[\prod_{i=1}^n (1 + \delta\langle V, X_i\rangle)(1 + \delta\langle V', X_i\rangle)\right] \\ &= \mathbb{E}\left[\mathbb{E}_0[(1 + \delta\langle V, X\rangle)(1 + \delta\langle V', X\rangle) \mid V, V']^n\right] = \mathbb{E}\left[\left(1 + \frac{\delta^2}{d}\langle V, V'\rangle\right)^n\right], \end{aligned}$$

where the final equality follows because $\mathbb{E}_0[\langle v, X\rangle] = \frac{1}{d}\langle v, \mathbf{1}\rangle = 0$ for each $v \in \mathcal{V}$. Now we use that $1 + t \leq e^t$ for all t to obtain

$$1 + D_{\chi^2}\left(\overline{P}\|P_0^n\right) \leq \mathbb{E}\left[\exp\left(\frac{n\delta^2}{d}\langle V, V'\rangle\right)\right] = \mathbb{E}\left[\exp\left(\frac{2n\delta^2}{d}\sum_{j=1}^{d/2} U_j\right)\right]$$

for $U_j \stackrel{\text{iid}}{\sim} \mathsf{Uniform}\{\pm 1\}$. These U_j are 1-sub-Gaussian, so

$$1 + D_{\chi^2}\left(\overline{P} \| P_0^n\right) \le \exp\left(\frac{n^2\delta^4}{d}\right).$$

Propositions 2.2.9 and 2.2.10 (Pinsker's inequality) give $2\left\|P_0^n - \overline{P}\right\|_{\text{TV}}^2 \le \frac{n^2\delta^4}{d}$. Choose $\delta^4 = \frac{d}{n^2}$. $\square$

Combining these results with those of Exercise 13.6, we see the promised phenomena: testing uniformity of $p \in \Delta_d$ in the ℓ_2-distance requires a sample scaling as $n \asymp 1/\delta^2$, where δ is the separation, while in the ℓ_1 distance requires a sample of size $n \asymp \sqrt{d}/\delta^2$.

13.5* Deferred and technical proofs

13.5.1 Valid asymptotic information expansions of divergences

We collect a few representative regularity conditions to make the expansions of f-divergences in terms of the Fisher information matrix, as in the heuristic equality (13.1.8), rigorous. The Hellinger distance admits a fairly general result allowing even non-differentiable densities.

Lemma 13.5.1. *Let $\{P_\theta\}_{\theta\in\Theta}$ be a parametric family where each P_θ has a density p_θ w.r.t. a base measure μ. Assume that $\sqrt{p_\theta(x)}$ is continuously differentiable near θ_0 for μ-almost all x (though the neighborhood may depend on x). Assume that the Fisher information matrix $J(\theta) = \mathbb{E}_\theta[\dot\ell_\theta \dot\ell_\theta^\top]$ is continuous in θ at θ_0. Then*

$$d_{\text{hel}}^2(P_{\theta_0+v}, P_{\theta_0}) = \frac{1}{8} v^\top J(\theta_0) v + o(\|v\|^2) \quad \textit{uniformly in } v \textit{ near } 0.$$

We defer the proof to Section 13.5.2.

The Hellinger distance admits particularly unrestrictive conditions on the family $\{P_\theta\}$, because the generating functional $f(t) = \frac{1}{2}(\sqrt{t}-1)^2$ grows only linearly as $t \to \infty$. When the divergences can grow more quickly, we will use stronger regularity conditions on the model family.

Definition 13.1. *Let $\{P_\theta\}_{\theta\in\Theta}$, where $\Theta \subset \mathbb{R}^d$, be a family of distributions. The family is* sufficiently regular for the χ^2-divergence *at the point $\theta_0 \in \operatorname{int}\Theta$ if the following hold:*

i. There exists a base measure μ for which P_θ has density p_θ for all θ in a neighborhood of θ_0.

ii. The Fisher information $J_{\theta_0} := \mathbb{E}_{\theta_0}[\dot\ell_{\theta_0}\dot\ell_{\theta_0}^\top]$ exists.

iii. There is $h : \mathcal{X} \to [0,\infty]$ with $\mathbb{E}_{\theta_0}[h^2] < \infty$ for which $|\frac{p_{\theta_0+v}(x)}{p_{\theta_0}(x)} - 1 - \dot\ell_{\theta_0}(x)^\top v| \le h(x)\|v\|$ for all v near 0.

Many common distribution families are sufficiently regular.

Example 13.5.2: The collection $\{\mathsf{N}(\theta, 1)\}$ of normal distributions with densities $p_\theta(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-\theta)^2)$ has score function $\dot\ell_\theta(x) = (x-\theta)$, and taking $\theta = 0$, we have

$$\frac{p_v(x)}{p_0(x)} - 1 - xv = e^{-\frac{1}{2}v^2}e^{xv} - 1 - xv = (e^{-\frac{1}{2}v^2} - 1)(1+xv) + e^{-\frac{1}{2}v^2}\sum_{k=2}^\infty \frac{(xv)^k}{k!}.$$

As $|e^{-\frac{1}{2}v^2} - 1| \le v^2$ and

$$\sum_{k=2}^{\infty} \frac{(xv)^k}{k!} \le |xv|^2 \sum_{k=0}^{\infty} \frac{|xv|^k}{(k+2)!} \le |xv|^2 \exp(|xv|/2) \le |xv|^2 \exp\left(\frac{x^2}{4} + \frac{v^2}{4}\right),$$

we may take $h(x) = |x| + 1 + x^2 e^{x^2/4}$ in Definition 13.1, which is integrable against $e^{-\frac{1}{2}x^2}$. $\diamond$

Tilted distributions, as in Section 11.3, also satisfy Definition 13.1.

Example 13.5.3 (Tilted densities and regularity)**:** Let $g : \mathcal{X} \to \mathbb{R}^d$ be a bounded function, mean zero for a distribution P_0 on $\mathcal{X}$, and for $t \in \mathbb{R}^d$ define the tilted density

$$p_t(x) := (1 + \langle t, g(x)\rangle)p_0(x)$$

as in definition (11.3.1), where p_0 is the density of P_0. This family has score $\dot{\ell}_t(x) = \nabla_t \log(1 + tg(x)) = \frac{g(x)}{1+\langle t, g(x)\rangle}$, which satisfies $\dot{\ell}_0(x) = g(x)$, so $\frac{p_t(x)}{p_0(x)} - 1 - \langle \dot{\ell}_0(x), t\rangle = 0$, and the boundedness of g ensures that p_t is nonnegative for t near 0. The family satisfies Definition 13.1. $\diamond$

For non-pathological f, we typically have the second-order expansion (13.1.8) of the f-divergence in terms of the Fisher information. We restrict to the class of f-divergences that have global quadratic approximations: we assume there is some $K < \infty$ such that

$$\left|f(1+t) - f(1) - f'(1)t - \frac{f''(1)}{2}t^2\right| \le Kt^2 \quad \text{for all } t \ge -1. \tag{13.5.1}$$

The divergences we typically encounter satisfy this condition (see Exercise 13.2).

Example 13.5.4: For the Hellinger distance with $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$ or KL-divergence with $f(t) = t \log t - t + 1$, inequality (13.5.1) holds with $K = \frac{1}{2}$. For the χ^2 divergence with $f(t) = (t-1)^2$, inequality (13.5.1) holds with $K = 0$. $\diamond$

Definition 13.1 and inequality (13.5.1) are enough to guarantee the local information approximation (13.1.8), meaning that for "regular enough" parametric families, all locally quadratic divergences are locally equivalent to the metric induced by the Fisher information matrix.

Lemma 13.5.5. *Let $\{P_\theta\}$ be sufficiently regular for the χ^2-divergence (Definition 13.1) at θ and $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$ satisfying the second-order Taylor bound (13.5.1). Then for v near 0, the f-divergence satisfies*

$$D_f\left(P_{\theta+v} \| P_\theta\right) = \frac{f''(1)}{2} D_{\chi^2}(P_{\theta+v} \| P_\theta) + o(\|v\|^2)$$

and

$$D_{\chi^2}(P_{\theta_0+v} \| P_{\theta_0}) = v^\top J(\theta_0) v + o(\|v\|^2).$$

See Section 13.5.3 for a proof of this result, which as a consequence demonstrates that the χ^2-divergence is indeed (locally) finite on $\{P_\theta\}$.

13.5.2 Proof of Lemma 13.5.1

We mostly reduce to the 1-dimensional case by considering vectors $v_s \to v \in \mathbb{S}^{d-1}$ as $s \to 0$, as the claim in the lemma is equivalent to the claim that

$$\limsup_{v \to 0} \frac{1}{\|v\|^2} \left(d_{\rm hel}^2(P_{\theta_0+v}, P_{\theta_0}) - \frac{1}{8} v^\top J(\theta_0) v \right) = 0.$$

For $s \in \mathbb{R}$ define $f_s(x) := p_{\theta+sv_s}(x)$, so that $\dot{f}_s(x) = \langle \dot{p}_{\theta+sv_s}(x), v \rangle$ and the score $\dot{\ell}_s(x) := \frac{\dot{f}_s(x)}{f_s(x)} = \langle v_s, \dot{p}_{\theta+sv_s}(x) \rangle / p_{\theta+sv_s}(x)$. Now for $t > 0$,

$$\sqrt{f_t(x)} = \sqrt{f_0(x)} + \int_0^t \frac{\dot{f}_s(x)}{2\sqrt{f_s(x)}} ds,$$

so we have

$$\frac{1}{t^2} d_{\rm hel}^2(P_{\theta+tv_t}, P_\theta) = \frac{1}{8} \int \left(\frac{1}{t} \int_0^t \frac{\dot{f}_s(x)}{\sqrt{f_s(x)}} ds \right)^2 d\mu(x).$$

By Jensen's inequality, we have the domination condition

$$\left(\frac{1}{t} \int_0^t \frac{\dot{f}_s(x)}{\sqrt{f_s(x)}} ds \right)^2 \le \frac{1}{t} \int_0^t \frac{\dot{f}_s(x)^2}{f_s(x)} ds = \frac{1}{t} \int_0^t \dot{\ell}_s^2(x) f_s(x) ds,$$

and by Fubini's theorem this quantity is integrable:

$$\int \left(\frac{1}{t} \int_0^t \frac{\dot{f}_s(x)}{\sqrt{f_s(x)}} ds \right)^2 d\mu(x) \le \frac{1}{t} \int \int_0^t \dot{\ell}_s^2(x) f_s(x) ds d\mu(x) = \frac{1}{t} \int_0^t v_s^\top J(\theta_0 + sv_s) v_s ds < \infty$$

for small enough t, as J is assumed continuous at θ_0. The assumption that for μ-almost all x there is a neighborhood of θ_0 on which $\sqrt{p_\theta}$ is continuously differentiable then allows us to apply a variant of dominated convergence (see Proposition A.1.4 in Appendix A.1), because

$$\left(\frac{1}{t} \int_0^t \frac{\langle \dot{p}_{\theta_0+sv_s}(x), v_s \rangle}{\sqrt{p_{\theta_0+sv_s}(x)}} ds \right)^2 \to \frac{\langle v, \dot{p}_{\theta_0}(x) \rangle^2}{p_{\theta_0}(x)}$$

for μ-almost all x. Then $\frac{1}{t^2} d_{\rm hel}^2(P_{\theta_0+tv_t}, P_{\theta_0}) \to \frac{1}{8} \int \langle \dot{\ell}_{\theta_0}(x), v \rangle^2 p_{\theta_0}(x) d\mu(x) = \frac{1}{8} v^\top J(\theta_0) v$. Because we allow $v_s \to v$ to vary in s, this implies the desired uniformity.

13.5.3 Proof of Lemma 13.5.5

We prove the expansions related to the χ^2-divergence first, showing that

$$\int \left(\frac{p_{\theta+v}(x)}{p_\theta(x)} - 1 - \dot{\ell}_\theta(x)^\top v \right)^2 p_\theta(x) d\mu(x) = o(\|v\|^2). \tag{13.5.2}$$

Assuming the bound (13.5.2), standard L^p-convergence give the expansion $D_{\chi^2}(P_{\theta+v} \| P_\theta) = v^\top J(\theta) v + o(\|v\|^2)$, as $\mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top] = J(\theta)$. As in the proof of Lemma 13.5.1, we essentially reduce to the 1-dimensional case. Let $v_t \to v \in \mathbb{S}^{d-1}$ as $t \to 0$. Then

$$\frac{1}{t^2} \int \left(\frac{dP_{\theta+tv_t}}{dP_\theta} - 1 - t\dot{\ell}_\theta^\top v_t \right)^2 dP_\theta \le \int h^2(x) \|v_t\|^2 dP_\theta(x) \lesssim \int h^2(x) dP_\theta(x) < \infty,$$

and using that for μ-almost all x

$$\frac{1}{t}\left(\frac{p_{\theta+tv_t}(x)}{p_\theta(x)} - 1 - \dot\ell_\theta(x)^\top v_t\right) \to 0$$

as $t \to 0$, dominated convergence implies $\int (\frac{dP_{\theta+tv_t}}{dP_\theta} - 1 - t\dot\ell_\theta^\top v_t)^2 dP_\theta = o(t^2)$, giving the claim (13.5.2).

We demonstrate the expansions of the f-divergences. Without loss of generality, we assume that $f'(1) = 0$, because we may always replace f with $\widetilde{f}(t) := f(t) - tf'(1) + f'(1)$, which gives the same divergence. Once again let $v_t \to v \in \mathbb{S}^{d-1}$ as $t \to 0$, and assume w.l.o.g. that $\|v_t\|_2 \le 2$ for all t. Define the densities $q_t(x) = p_{\theta+tv_t}(x)$, so that by assumption

$$\left|\frac{q_t(x)}{q_0(x)} - 1 - t\dot\ell_\theta(x)^\top v_t\right| \le 2th(x).$$

Define the gap

$$g_t(x) := \frac{1}{t^2}\left|f\left(\frac{q_t(x)}{q_0(x)}\right) - \frac{f''(1)}{2}\left(\frac{q_t(x)}{q_0(x)} - 1\right)^2\right|.$$

Then

$$\begin{aligned} g_t(x) \le \frac{C}{t^2}\left(\frac{q_t(x)}{q_0(x)} - 1\right)^2 &\le \frac{2C}{t^2}\left(\frac{q_t(x)}{q_0(x)} - 1 - t\dot\ell_\theta(x)^\top v_t\right)^2 + \frac{2C}{t^2}t^2\left\|\dot\ell_\theta(x)\right\|_2^2\|v_t\|_2^2 \\ &\le 2C\left(h(x)^2 + \left\|\dot\ell_\theta(x)\right\|_2^2\right)\|v_t\|_2^2 \le 4C\left(h(x)^2 + \left\|\dot\ell_\theta(x)\right\|_2^2\right), \end{aligned}$$

which is P_θ-integrable. Because $f(1+s) = f(1) + f'(1)s + \frac{f''(1)}{2}s^2 + o(s^2)$ and $f(1) = f'(1) = 0$ by assumption,

$$\begin{aligned} g_t(x) &= \frac{1}{t^2}\left|f\left(1 + \frac{q_t(x)}{q_0(x)} - 1\right) - \frac{f''(1)}{2}\left(\frac{q_t(x)}{q_0(x)} - 1\right)^2\right| \\ &= o\left(\frac{1}{t^2}\left(\frac{q_t(x)}{q_0(x)} - 1\right)^2\right) = o\left(t^{-2}\left(t\dot\ell_\theta(x)^\top v_t + o(t)\right)^2\right) \to 0 \end{aligned}$$

for all x. Dominated convergence then implies that $\lim_{t\to 0}\int g_t(x)q_0(x)d\mu(x) = 0$, and so

$$\frac{1}{t^2}\left(D_f\left(Q_t\|Q_0\right) - \frac{f''(1)}{2}D_{\chi^2}(Q_t\|Q_0)\right) \to 0$$

as $t \to 0$. Because we took $q_t(x) = p_{\theta+tv_t}$ along a path, this gives the result.

13.5.4 Proof of Lemma 13.2.7

A quick integral calculation yields the Hellinger distance

$$d_{\rm hel}^2\left(\mathsf{N}(\theta_0, \sigma^2 I), \mathsf{N}(\theta_1, \sigma^2 I)\right) = 1 - \exp\left(-\frac{1}{8\sigma^2}\left\|\theta_0 - \theta_1\right\|_2^2\right).$$

Given θ_0 and letting $c^2 = -8\log(1-\epsilon^2)$ for shorthand, the modulus of $T(\theta) = \|\theta\|_2^2$ is

$$\omega_{\rm hel}(\epsilon) = \sup_\theta\left\{\|\theta_0\|_2^2 - \|\theta\|_2^2 \mid \|\theta_0 - \theta\|_2^2 \le c^2\sigma^2\right\}.$$

The optimizer is collinear with θ_0: for fixed $\|\theta\|_2$, choosing θ in the direction of θ_0 minimizes $\|\theta_0 - \theta\|_2$. Thus, writing $\theta = (1-t)\theta_0$ with $t \in [0,1]$ and $r = \|\theta_0\|_2$, we solve

$$\sup_t \left\{(1-(1-t)^2)r^2 \mid t^2r^2 \le c^2\sigma^2\right\}. \tag{13.5.3}$$

The maximizer $t = \min\{1, c\sigma/r\}$. Thus, if $r \le c\sigma$, $t = 1$ maximizes the problem (13.5.3), yielding value r^2; if $r > c\sigma$, we obtain value

$$\left(1-(1-c\sigma/r)^2\right)r^2 = \left(\frac{2c\sigma}{r} - \frac{c^2\sigma^2}{r^2}\right)r^2 = 2c\sigma r - c^2\sigma^2 \in [c\sigma r, 2c\sigma r].$$

Substituting for $c = \sqrt{-8\log(1-\epsilon^2)}$ gives the lemma.

13.6 Bibliography and further reading

Optimal functional estimation has long animated nonparametric statistics, with Stein [383] providing the initial impetus by introducing the hardest 1-dimensional subproblems upon which Section 13.1 reposes. As van der Vaart and Wellner [409] write, this paper is "the seminal paper for semiparametric statistics, making the transition from the asymptotic theory for regular parametric models to the estimation of smooth functionals on infinite-dimensional models." Donoho and Liu [148, 149, 150], describe and extend the set of techniques we develop in Section 13.1. They never published the first paper [148], as Birgé's results essentially subsume it [71]. They show that the global Hellinger modulus (13.1.2) provides an attainable quantity in certain functional estimation problems in which the statistic $\theta(P)$ is linear in P and the class $\mathcal{P}$ is convex [149, 150].

Le Cam's convex hull method (Section 13.2) traces back to [281], and Yu [435] surveys this. The particular technique for mixtures, especially the χ^2-bounds in Section 13.2.1, we learned from a series of unpublished lecture notes by David Pollard at Yale; Ingster and Suslina [251] provide a deep investigation of the approach. The mixture approaches and "bump" constructions, such as those in Proposition 13.2.9, more or less appear in Birgé and Massart [72]. Several textbooks cover functional estimation in more detail; there is the classic text [68], and Giné and Nickl [211] develop optimality and adaptivity theory for nonparametric estimation. The paper [163] provides applications of these ideas in the context of distributionally robust estimation, showing that statistically robust optimization problems suffer nonparametric rates of convergence even in finite dimensions, and Acharya et al. [3, 4] employ them in the context of information-limited estimation.

Our results on instance-optimal lower bounds, and the identification of the desideratum (OPT.iii') (and the earlier desiderata (OPT.i) and (OPT.ii)) derive from Cai and Low [96]. Duchi and Ruan [154] give the risk transfer inequalities we present in Section 13.3.1, and Tsybakov [400, Appendix A1] provides similar results that apply for higher-dimensional optimization problems.

While questions around estimating functionals of an underlying density are classical in statistics [68, 145, 211], Theoretical computer science has considered "property testing" and estimation for discrete distributions. Here, one attempts to estimate functionals $T(p)$ of a discrete distribution with sample complexity sublinear in the dimension of p. Concrete motivations for this come from language modeling or species estimation, where one wishes to estimate the mass missing from an observed sample, famously addressed by Good and Turing to crack Enigma codes [216], and which Fisher and collaborators also considered [189]. More recent work evidences a flurry of interest, including sublinear estimators for testing ℓ_1-distance [53, 341], optimality theory for Good-Turing estimators [335, 334], entropy and support size estimators [339, 340, 403, 404, 405, 431], and other

functional estimation problems [95]. The treatment of minimax testing in Section 13.4 follows Balakrishnan and Wasserman [38], as well as Balakrishnan's lecture notes [37].

13.7 Exercises

Exercise 13.1: Recall the Hellinger distance between distributions P and Q with densities p, q is $d_{\rm hel}(P,Q)^2 = \frac{1}{2}\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$. Let P be $\mathsf{N}(\mu_0, \Sigma)$ and Q be $\mathsf{N}(\mu_1, \Sigma)$. Show that

$$\frac{1}{2}d_{\rm hel}(P,Q)^2 = 1 - \exp\left(-\frac{1}{8}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)\right).$$

Exercise 13.2: Demonstrate the claims in Example 13.5.4.

Exercise 13.3: Suppose that Ψ has test risk $R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) \le \frac{1}{3}$ for testing between $\mathcal{P}_0$ and $\mathcal{P}_1$. Let $k \in \mathbb{N}$. Given a sample of size kn, show how to develop a test $\Psi^\star$ with

$$R_{kn}(\Psi^\star \mid \mathcal{P}_0, \mathcal{P}_1) \le 2\exp\left(-c \cdot k\right)$$

for a constant $c > 0$. *Hint.* Split the sample into k samples of size n, and then apply Ψ to each.

Exercise 13.4 (Poissonization: lower bounds [431])**:** Prove the lower bound in Proposition 13.4.7, inequality (13.4.9), that is, that for numerical constants C, c,

$$\mathfrak{M}_{\mathsf{Poi}(2n)} - Cr^2\exp(-cn) \le \mathfrak{M}_n.$$

Hint. Bound $\mathfrak{M}_{\mathsf{Poi}(2n)}$ with a weighted sum of $\mathfrak{M}_m$. Use the MGF calculation that for $X \sim \mathsf{Poi}(\lambda)$, $\mathbb{E}[e^{tX}] = \exp(\lambda(e^t - 1))$ to show that $N \sim \mathsf{Poi}(2n)$ is concentrated above n.

Exercise 13.5 (Poissonization: upper bounds [431])**:** Assume the minimax result that

$$\mathfrak{M}_n = \sup_\pi \inf_{T_n} \mathbb{E}\left[(T_n(X_1^n) - T(p))^2\right],$$

where the supremum is over priors π on $p \in \Delta_k$, and the expectation is over the random draw of p and $X_i \stackrel{\rm iid}{\sim} p$. (This is a standard infinite-dimensional minimax theorem [183, 380].)

Let $\{T_m\}$ be a sequence of estimators and define the sequence of averaged risks

$$r_m := \mathbb{E}[(T_m(X_1^m) - T(p))^2].$$

Define the modified risks $\widetilde{r}_m = \min\{r_1, \ldots, r_m\} = \min\{\widetilde{r}_{m-1}, r_m\}$, and the "corrected" estimators

$$\widetilde{T}_m(x_1^m) := \begin{cases} \widetilde{T}_{m-1}(x_1^{m-1}) & \text{if } r_m \ge \widetilde{r}_{m-1} \\ T_m(x_1^m) & \text{if } r_m < \widetilde{r}_{m-1}. \end{cases}$$

(a) Show that $\mathbb{E}[(\widetilde{T}_m(X_1^m) - T(p))^2] \le \mathbb{E}[(T_m(X_1^m) - T(p))^2]$.

(b) Show that for $N \sim \mathsf{Poi}(n/2)$ and $p \sim \pi$, then X_i drawn i.i.d. conditionally on p,

$$\inf_{T_n} \mathbb{E}[(T_n(X_1^n) - T(p))^2] \le 2\mathbb{E}[(T_N(X_1^N) - T(p))^2]$$

(c) Finalize the proof of the upper bound in inequality (13.4.9).

Exercise 13.6 (Goodness of fit test in ℓ_1-distance [38]): In the notation of Section 13.4.3, let $p_0 = \frac{1}{d}\mathbf{1}$ be the uniform distribution on $[d]$. Consider testing $\|p - p_0\|_1 = 0$ against the alternative $\|p - p_0\|_1 \ge \epsilon_n$. Use Poissonization to let $N_j \sim \mathsf{Poi}(np_j)$, and define the statistic

$$T_n := d \sum_{j=1}^{d} \left((N_j - n/d)^2 - N_j \right).$$

(a) Show that under the null $H_0 : p_0$ (uniformity), $\mathbb{E}_0[T_n] = 0$.

(b) Show that generally, $\mathbb{E}_p[T_n] = n^2 d \|p - p_0\|_2^2$, and so if $\epsilon_n = \|p - p_0\|_1$, then $\mathbb{E}_p[T_n] \ge n^2 \epsilon_n^2$.

(c) Compute the variance of the individual terms defining T_n to show that

$$\mathrm{Var}_p((N_j - n/d)^2 - N_j) = 2np_j(np_j + 2n^2(p_j - 1/d)^2).$$

(d) For $\alpha \in (0, 1)$, define the test

$$\Psi_n = \mathbf{1}\left\{ T_n > \sqrt{2dn^2/\alpha} \right\}.$$

Show that under the null, $P_0(\Psi_n = 1) \le \alpha$.

(e) Show that there exists a numerical constant $C < \infty$ such that if

$$\epsilon_n^2 \ge \frac{C\sqrt{d}}{n} \left[\frac{1}{\sqrt{\alpha}} + \frac{1}{\beta} \right],$$

then $\mathbb{P}_p(\Psi_n = 0) \le \beta$ for all $\|p - p_0\|_1 \ge \epsilon_n$.

Exercise 13.7: Consider the hypothesis testing problem of testing whether a collection of independent Bernoulli random variables $X_1, \ldots, X_n$ is fair (H_0, so that $\mathbb{P}(X_i = 1) = \frac{1}{2}$ for each i) or that there are unfair subcollections. That is, for a set $C \subset [-1, 1]^n$ we wish to test

$$H_0 : X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}\left(\frac{1}{2}\right) \quad \text{versus} \quad H_1 : X_i \stackrel{\text{ind}}{\sim} \mathsf{Bernoulli}\left(\frac{1+\theta_i}{2}\right), \ \theta \in C.$$

Show that if the set C is orthosymmetric, meaning that $\theta \in C$ implies $S\theta \in C$ for any diagonal sign matrix S, then no test can reliably distinguish H_0 from H_1 in a minimax sense. *Hint.* Let $v \in \mathcal{V} := \{\pm 1\}^n$ index coordinate signs and define $\theta_v = Dv$ for some diagonal D, where $Dv \in C$. Let P_v be the product distribution with $X_i \sim \mathsf{Bernoulli}(\frac{1+D_i v_i}{2})$. What is $\frac{1}{2^n}\sum_{v \in \mathcal{V}} P_v$?

Exercise 13.8: Define the collection $\{P_v\}$ by tiltings of an underlying base distribution P_0, where each tilting is indexed by a function $g_v : \mathcal{X} \to [-1, \infty)$ with $\mathbb{E}_{P_0}[g_v(X)] = 0$, and where

$$dP_v(x) = (1 + g_v(x))dP_0(x).$$

Let P_v^n be the distribution of n observations $X_i \stackrel{\text{iid}}{\sim} P_v$ and $\overline{P^n} = \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} P_v^n$. Show that for the inner product $\langle f, g \rangle_P = \int f(x)g(x)dP(x)$,

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 \le \mathbb{E}\left[\exp(n\langle g_V, g_{V'} \rangle_{P_0})\right] \quad \text{for } V, V' \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathcal{V}).$$

Exercise 13.9: Consider testing whether a collection of Bernoulli random variables has an "upward trend" over time, by which we mean that if $X_i \sim \mathsf{Bernoulli}(p_i)$ independently, then

$$\overline{p}_{\rm end} := \frac{1}{n/4} \sum_{i=\frac{3n}{4}+1}^{n} p_i > \overline{p}_{\rm beg} := \frac{1}{n/4} \sum_{i=1}^{n/4} p_i.$$

In a more quantitative version of this problem, we wish to test

$$H_0 : X_i \stackrel{\rm iid}{\sim} \mathsf{Bernoulli}(1/2) \quad \text{versus} \quad H_1 : X_i \stackrel{\rm ind}{\sim} \mathsf{Bernoulli}(p_i),\ \overline{p}_{\rm end} - \overline{p}_{\rm beg} \geq \delta.$$

(a) Use Le Cam's two-point method to show that there exists a numerical constant $c > 0$ such that for $\delta \leq \frac{c}{\sqrt{n}}$, no test can reliably distinguish H_0 from H_1.

(b) Use Proposition 13.4.2 and the statistic

$$T_n := \frac{1}{n/4} \sum_{i=3n/4+1}^{n} X_i - \frac{1}{n/4} \sum_{i=1}^{n/4} X_i$$

to develop a test Ψ that achieves risk $R_n(\Psi \mid H_0, H_1) \leq \frac{1}{4}$ whenever $\delta \geq O(1) \cdot n^{-1/2}$.

Exercise 13.10: Prove the identity (13.4.12).

Exercise 13.11 (Unbiased distance estimators for multinomials)**:** Let $X_i \stackrel{\rm iid}{\sim} p$, $i = 1, \ldots, n$, and $Y_i \stackrel{\rm iid}{\sim} q$, $i = 1, \ldots, m$, so X_1^n and Y_1^m are multinomial samples for $p, q \in \Delta_d$. Define the empirical estimators $\widehat{p}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\left\{X_i = j\right\}$ and $\widehat{q}_j = \frac{1}{m} \sum_{i=1}^m \mathbf{1}\left\{Y_i = j\right\}$.

(a) Give $\mathbb{E}[\|\widehat{p}\|_2^2]$.

(b) Show that $T_n := \|\widehat{p} - \widehat{q}\|_2^2$ satisfies

$$\mathbb{E}[T_n] = \|p - q\|_2^2 + \frac{1}{n} + \frac{1}{m} - \frac{1}{n} \|p\|_2^2 - \frac{1}{m} \|q\|_2^2 .$$

(c) Modify T_n into a new statistic $T_n^{\rm unb}$ so that $\mathbb{E}[T_n^{\rm unb}] = \|p - q\|_2^2$.

Exercise 13.12: Show that in the hypothesis testing problem (13.4.13), there is a numerical constant $c > 0$ such that $\delta \leq c/\sqrt{n}$ implies that no test can reliably distinguish H_0 from H_1.

Exercise 13.13: Consider the linear regression model

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, I_n)$$

where $X \in \mathbb{R}^{n \times d}$ is fixed (and known), and $Y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^d$. Consider testing

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \|\theta\|_2 \geq r.$$

Use the convex hull method for lower bounds (Proposition 13.4.1) to show that if

$$r \leq \frac{d^{1/4}}{\sqrt{n} \left\|n^{-1/2} X\right\|_{\rm op}}$$

then any test of H_0 against H_1 has minimax test risk at least $\frac{1}{2}$.

Exercise 13.14 (The Hodges super-efficient estimator)**:** Consider a normal mean estimation problem, $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$ where $\theta \in \mathbb{R}$ is unknown, and define the *Hodges estimator*

$$\widehat{\theta}_{\mathrm{H}}(X_1^n) := \begin{cases} 0 & \text{if } |\overline{X}_n| \le n^{-1/4} \\ \overline{X}_n & \text{otherwise.} \end{cases}$$

(a) Show that the limiting distribution of $\widehat{\theta}_{\mathrm{H}}$ satisfies

$$\sqrt{n}(\widehat{\theta}_{\mathrm{H}}(X_1^n) - \theta) \stackrel{d}{\rightsquigarrow} \begin{cases} \mathbf{1}_0 & \text{if } \theta = 0 \\ \mathsf{N}(0,1) & \text{otherwise.} \end{cases}$$

(b) Show that for any $c < 1$,

$$\lim_{n\to\infty} \inf_{\theta} \left\{ P_\theta \left(|\widehat{\theta}_{\mathrm{H}}(X_1^n) - \theta| \ge 1/\sqrt{n} \right) \mid \frac{2}{\sqrt{n}} \le |\theta| \le \frac{c}{\sqrt{2}} \frac{1}{n^{1/4}} \right\} = 1.$$

What does this say about the performance of $\widehat{\theta}_{\mathrm{H}}$?

Exercise 13.15: Let $\mathcal{F}_s^d$ denote the collection of functions $f : \mathbb{R}^d \to [-1, 1]$ with compact support in $\mathbb{B}_2^d$ and for which the sth derivative is bounded continuous, that is, $\left\|\nabla^s f(x)\right\|_{\mathrm{op}} \le 1$ for all x. Consider the functional $T(f) := \int f^2(x)dx$. Exhibit the curse-of-dimensionality for estimating $T(f)$ with the standard regression model $Y_i = f(X_i) + \mathsf{N}(0, 1)$, where $X_i \stackrel{\text{iid}}{\sim} \mathsf{Uniform}(\mathbb{B}_2^d)$:

$$\mathfrak{M}_n(T(\mathcal{F}_s^d), |\cdot|) := \inf_{\widehat{T}_n} \sup_{f \in \mathcal{F}_s^d} \mathbb{E}_f \left[\left| \widehat{T}_n(X_1^n, Y_1^n) - T(f) \right| \right] \ge C_d \cdot n^{-\frac{4s}{4s+d}},$$

where C_d is a constant depending only on the dimension.

Part III

Entropy, predictions, divergences, and information

Précis

In typical information-theoretic settings, operational interpretations of entropy and information center around communication, such as the number of bits necessary lossless encoding, as in Section 2.4. In this part of the book, we develop a complementary statistical interpretation of information and entropy as they relate to minimizing losses in prediction problems. We shall see that mutual information becomes the gap between best possible prediction of a target value Y from a signal X, where losses are measured via the log loss, and entropy becomes a measure of achievable predictive performance. There is no reason to restrict ourselves to the logarithmic loss generating the classical Shannon entropy and information, and accordingly, by considering general loss functions we can generalize entropy and information to a decision-theoretic view of the "information in an experiment." By connecting the techniques to results coming from convex analysis and optimization theory, we will also be able to develop a theory of *surrogate risk consistency*, where solving a prediction problem with one loss φ is equivalent—or at least implies—that we have solved the same problem according to a (potentially harder to optimize) loss ℓ.

Chapter 14

Predictions, loss functions, and entropies

In prediction problems, we observe a random variable X and wish to predict a target Y, seeking a function that accurately predicts the value of Y given X. Several perspectives address such problems. We roughly divide these into three approaches, though there is considerable overlap between the techniques and goals of the three:

(1) Point prediction, where we wish to find a prediction function f so that $f(X)$ most accurately predicts Y itself.

(2) Probabilistic prediction, or forecasting, where we seek a distribution $\widehat{P}$ approximating the coniditional distribution of Y given X, so that $\mathbb{P}(Y = y \mid X = x) \approx \widehat{P}(Y = y \mid X = x)$. *Calibration*, the subject of the next chapter, relaxes this to ask that $\mathbb{P}(Y = y \mid \widehat{P}) \approx \widehat{P}(Y = y)$, meaning the distribution of Y given a predicted distribution $\widehat{P}$ is accurate.

(3) Predictive inference, where for a given level $\alpha \in (0, 1)$, we seek a prediction set $C(X)$ such that $\mathbb{P}(Y \in C(X)) \approx 1 - \alpha$.

We focus mostly on the former two.

In this first chapter of the sequence, we focus on the probabilistic prediction problem. Our main goal will be to elucidate and identify *proper* loss functions for choosing probabilistic predictions, meaning that the true distribution of Y minimizes the loss, and *strictly proper* losses, meaning that the true distribution of Y uniquely minimizes the loss. As part of this, we will develop mappings between losses and entropy-type functionals; these will repose on convex analytic techniques. Moreover, we highlight how any proper loss (which we will define) corresponds to a particular measure of entropy on the distribution P, and how these connect with *Bregman divergences* in convex optimization. For the deepest understanding of this chapter, it will be useful to review the convexity concepts (e.g., convex sets, functions, subgradients) in Appendix B, as well as results on solutions to convex optimization problems in Appendix C. We summarize the required facts in Section 14.1.1.

14.1 Proper losses, scoring rules, and generalized entropies

As motivation, consider a weather forecasting problem: a meteorologist wishes to predict a rain Y_t on days $t = 1, 2, \ldots$, where $Y_t = 1$ indicates rain and $Y_t = 0$ indicates no rain. At time t, using

covariates X_t, the forecaster predicts a probability $p_t \in [0,1]$ for the binary outcome Y_t, and a loss function $\ell(p,y)$ that scores a prediction p for an outcome y. The forecaster should have an incentive to make predictions as accurately as possible, so the distribution minimizing the expected loss should coincide with the true distribution of Y.

This leads to proper losses. In our treatment, we allow infinite values, so we work with the upper and lower extended real lines, recalling that $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ and $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$.

Definition 14.1. *Let $\mathcal{P}$ be a collection of distributions on $\mathcal{Y}$. A loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* proper *if, whenever $Y \sim P \in \mathcal{P}$,*

$$\mathbb{E}[\ell(P,Y)] \le \mathbb{E}[\ell(Q,Y)] \quad \text{for all } Q \in \mathcal{P}.$$

The loss is strictly proper *if the preceding inequality is strict whenever $Q \neq P$.*

In much of the literature on prediction, one instead considers *proper scoring rules*, which are simply negative proper losses, that is, functions $S : \mathcal{P} \times \mathcal{Y} \to \underline{\mathbb{R}}$ satisfying $S(P,y) = -\ell(P,y)$ for a (strictly) proper loss. In addition, we frequently work with discrete distributions, so that Y has a probability mass function (p.m.f.), in which case we will use $p \in \Delta_k := \{p \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, p \rangle = 1\}$ to identify the distribution and $\ell(p,y)$ instead of $\ell(P,y)$. The *relative interior* $\operatorname{relint} \Delta_k = \{p \mid p \succ 0, \langle \mathbf{1}, p \rangle = 1\}$ is the interior of Δ_k relative to the affine space containing it.

Perhaps the two most famous proper losses are the log loss and the squared loss (often termed *Brier scoring*). For simplicity let us assume that $\mathcal{Y} = \{1, 2, \ldots, k\}$, and identify distributions P on $\mathcal{Y}$ with vectors $p \in \Delta_k$. The squared loss is then

$$\ell_{\rm sq}(p,y) = (p_y - 1)^2 + \sum_{i \neq y} p_i^2 = \|p - e_y\|_2^2,$$

where e_y is the yth standard basis vector, while the log loss (the negative log-likelihood) is

$$\ell_{\log}(p,y) = -\log p_y.$$

Both of these are strictly proper. To see this propriety, let Y have p.m.f. $p \in \Delta_k$, so that $\mathbb{P}(Y = y) = p_y$. Then for the squared loss and any $q \in \Delta_k$, we have

$$\mathbb{E}[\ell_{\rm sq}(q,Y)] - \mathbb{E}[\ell_{\rm sq}(p,Y)] = \mathbb{E}[\|q - e_Y\|_2^2] - \mathbb{E}[\|p - e_Y\|_2^2] = \|q\|_2^2 - 2\langle q, p \rangle + \langle p, p \rangle = \|q - p\|_2^2.$$

For the log loss, we have

$$\mathbb{E}[\ell_{\log}(q,Y)] - \mathbb{E}[\ell_{\log}(p,Y)] = -\sum_{y=1}^k p_y \log q_y + \sum_{y=1}^k p_y \log p_y = \sum_{y=1}^k p_y \log \frac{p_y}{q_y} = D_{\rm kl}\left(p \| q\right).$$

It is immediate that $q = p$ uniquely minimizes each loss.

That the gap between the expected losses at q and p reduces to a particular divergence-like measure—the squared ℓ_2-distance in the case of the squared loss and the KL-divergence in the case of the log loss—is no accident. In fact, for proper losses, we will show that this divergence representation necessarily holds.

The key underlying our development is a particular construction, which we present in Section 14.1.2, that transforms a loss into a generalized notion of entropy. Before highlighting it, we take a brief detour through a few of the concepts in convexity we require. (Chapters B and C in the appendices contain proofs and illustrations of the results we require.)

14.1.1 A convexity primer

A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex if for $\lambda \in [0, 1]$ and $x, y \in \operatorname{dom} f$, we have

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda)f(y),$$

where for $x \notin \operatorname{dom} f$ we define $f(x) = +\infty$, and strictly convex if the inequality is strict for $x \neq y$ and $\lambda \in (0, 1)$. We consider only proper convex functions, so $f(x) > -\infty$ for each x. A convex f is *closed* if its epigraph $\operatorname{epi} f = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \le t\} \subset \mathbb{R}^{d+1}$ is closed; equivalently, f is lower semicontinuous: $\liminf_{y \to x} f(y) \ge f(x)$. A concave function f is one for which $-f$ is convex.

Three main concepts form the basis for our development. The first is the *subgradient* (see Appendix B.2). For a $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, the *subgradient set* (or the subdifferential) at the point x is

$$\partial f(x) := \left\{ s \in \mathbb{R}^d \mid f(y) \ge f(x) + \langle s, y - x \rangle \text{ for all } y \in \mathbb{R}^d \right\}. \tag{14.1.1}$$

If f is convex, $\partial f(x)$ is nonempty at all points x in the relative interior of $\operatorname{dom} f$ (Theorem B.2.4). At points $x \in \operatorname{relint} \operatorname{dom} f$, the set $\partial f(x)$ is a singleton if and only f is differentiable at x, in which case $\partial f(x) = \{\nabla f(x)\}$.. If f is differentiable at x, and f is strictly convex at x if and only if $f(y) > f(x) + \langle s, y - x \rangle$ whenever $y \neq x$ (Corollary C.1.8 in Appendix C.1.2). When losses may take infinite values, in some cases we allow subgradients to take values in the extended reals $\overline{\mathbb{R}}^k$ and $\underline{\mathbb{R}}^k$, which are nonstandard elements of the convex subdifferential (14.1.1).

The second concept is a calculus of suprema of collections of convex functions. If for some set $\mathcal{A}$, f_α is convex for each index $\alpha \in \mathcal{A}$, then

$$f(x) := \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

is convex, and f is closed if f_α is closed for each α. (See Appendix B.2.5.) If the supremum is attained at $\mathcal{A}(x) = \{\alpha \in \mathcal{A} \mid f_\alpha(x) = f(x)\}$, then (Lemma B.2.39)

$$\partial f(x) \supset \operatorname{Conv} \left\{ s_\alpha \in \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x) \right\}.$$

Lastly, we recall a special duality relationship that all closed convex functions f enjoy (see Appendix C.2). The *Fenchel-Legendre conjugate* or *convex conjugate* of a function f is

$$f^*(s) := \sup_x \left\{ \langle s, x \rangle - f(x) \right\}. \tag{14.1.2}$$

The function f^* is always convex, as it is the supremum of affine functions of s. A special function attached to the conjugate f^* is the *biconjugate* $f^{**}(x) = \sup_s \{\langle s, x \rangle - f^*(s)\}$, or $f^{**} = (f^*)^*$, and f^{**} is equal to the supremum of *all* affine functions below f:

Lemma 14.1.1 (Theorem C.2.1)**.** *Let f be a closed convex function. Then $f^{**}(x) = f(x)$.*

We immediately have the *Fenchel-Young inequality* that

$$f^*(s) + f(x) \ge \langle s, x \rangle \quad \text{for all} \quad s, x,$$

and Proposition C.2.3 shows that, for closed convex f, equality holds if and only if

$$s \in \partial f(x) \quad \text{if and only if} \quad x \in \partial f^*(s), \tag{14.1.3}$$

which are equivalent. Thus we obtain the identities $\partial f^* = (\partial f)^{-1}$ and $\partial f = (\partial f^*)^{-1}$ and the characterization $\partial f^*(s) = \operatorname{argmin}_x \{-\langle s, x \rangle + f(x)\} = \operatorname{argmax}_x \{\langle s, x \rangle - f(x)\}$.

14.1.2 From a proper loss to an entropy

The key construction underlying all of our proper losses is the optimal value of the expected loss. To any loss ℓ acting on a family $\mathcal{P}$ of distributions, we construct the *generalized entropy* associated with the loss ℓ by

$$H_\ell(Y) := \inf_{Q \in \mathcal{P}} \mathbb{E}[\ell(Q, Y)], \tag{14.1.4}$$

where we have paralleled the typical notation $H(Y)$ for the Shannon entropy. It is often more convenient to write this entropy as a function of the distribution P of Y,

$$H_\ell(P) = \inf_{Q \in \mathcal{P}} \mathbb{E}_P[\ell(Q, Y)], \tag{14.1.5}$$

where Y follows the distribution P; we will use whichever is more convenient. As the notation (14.1.5) makes clear, $H_\ell(P)$ is the infimum of a collection of linear functions of the form $P \mapsto \mathbb{E}_P[\ell(Q, Y)]$, so $H_\ell(P)$ is concave in P. The quantity (14.1.5) informs the properties of the loss ℓ and how we may always view any concave function H on a family of distributions $\mathcal{P}$ as a generalized entropy.

Section 14.2 gives a converse construction, recovering proper losses from entropies. For now, we consider examples where Y has finite support $\{1, \ldots, k\}$, with $\mathcal{P} = \Delta_k = \{p \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, p \rangle = 1\}$.

Example 14.1.2 (Entropy of the log loss)**:** The log loss $\ell_{\log}(p, y) = -\log p_y$ satisfies

$$H_{\ell_{\log}}(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[-\log q_Y] = \inf_{q \in \Delta_k} \left\{ -\sum_{y=1}^k p_y \log \frac{q_y}{p_y} - \sum_{y=1}^k p_y \log p_y \right\} = -\sum_{y=1}^k p_y \log p_y,$$

the classical Shannon entropy. $\diamond$

This highlights a new operational interpretation of entropy: the (Shannon) entropy is the minimal expected loss of a player in a prediction game, where the player chooses a distribution Q on Y, nature draws $Y \sim P$, and upon observing $Y = y$, the player suffers loss $-\log Q(Y = y)$.

Example 14.1.3 (Entropy of the 0-1 error)**:** Fix a deterministic tie-breaking rule and let $\widehat{y}(p) \in \operatorname{argmax}_j p_j$. For the 0-1 loss $\ell_{0\text{-}1}(p, y) = \mathbf{1}\{y \neq \widehat{y}(p)\}$,

$$H_{\ell_{0\text{-}1}}(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[\ell_{0\text{-}1}(q, Y)] = 1 - \max_y p_y.$$

So $H_{\ell_{0\text{-}1}}(e_y) = 0$ for any standard basis vector (point mass), and $H_{\ell_{0\text{-}1}}(p) > 0$ otherwise. The uniform vector $p = \mathbf{1}/k$ maximizes $H_{\ell_{0\text{-}1}}(p)$, with $H_{\ell_{0\text{-}1}}(\mathbf{1}/k) = \frac{k-1}{k}$. $\diamond$

Example 14.1.4 (Brier scoring and squared error)**:** For the squared error (Brier scoring) loss $\ell_{\text{sq}}(p, y) = \|p - e_y\|_2^2$, let Y have p.m.f. $p \in \Delta_k$. Then

$$H_{\ell_{\text{sq}}}(p) = \mathbb{E}[\ell_{\text{sq}}(p, Y)] = \|p\|_2^2 - 2\|p\|_2^2 + 1 = 1 - \|p\|_2^2.$$

So $H_{\ell_{\text{sq}}}(p) \geq 0$, with $H_{\ell_{\text{sq}}}(p) = 0$ if and only if $p \in \{e_y\}_{y=1}^k$ is a point mass, and the uniform distribution maximizes the entropy, with $H_{\ell_{\text{sq}}}(\mathbf{1}/k) = 1 - 1/k$. $\diamond$

14.1.3 The information in an experiment

In classical information theory, the mutual (Shannon) information between random variables X and Y is the gap between the entropy of Y and the remaining entropy given X, that is,

$$I(X;Y) = H(Y) - H(Y \mid X).$$

In complete analogy with our development in Chapter 2, we can define the information between variables X and Y relative to a particular loss function ℓ. Thus, we define the *ℓ-conditional entropy*

$$H_\ell(Y \mid X = x) := \inf_{Q \in \mathcal{P}} \mathbb{E}\left[\ell(Q, Y) \mid X = x\right]$$

and, in analogy to the definitions in Section 2.1.1, the conditional entropy of Y given X is

$$H_\ell(Y \mid X) := \mathbb{E}\left[\inf_{Q \in \mathcal{P}} \mathbb{E}[\ell(Q, Y) \mid X]\right] = \int_{\mathcal{X}} H_\ell(Y \mid X = x) dP_X(x),$$

the average minimal expected loss when one observes X.

This definition yields the *information in an experiment*. By an experiment, think of Y as a hypothesis to be tested or a value to be predicted and observe X, and we ask how much observing X helps this prediction. Then we define the information

$$I_\ell(X;Y) := H_\ell(Y) - H_\ell(Y \mid X) \tag{14.1.6}$$

as the gap between the prior entropy of Y and its posterior entropy conditional on the observation X. In view of our definition of entropy as expected minimal loss in a prediction game, the information $I_\ell(X;Y)$ is the reduction in optimal predictive loss an observation X offers a player in predicting Y, and so $I_\ell(X;Y) \geq 0$. The information (14.1.6) is typically asymmetric in X and Y, so we are careful about the ordering.

The next three examples show different information quantities, where in each we let $\mathcal{Y}$ have finite cardinality k and identify $\mathcal{P}$ with the simplex $\Delta_k = \{p \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, p \rangle = 1\}$.

Example 14.1.5 (Shannon information)**:** Taking the log loss $\ell(p, y) = -\log p_y$, we have

$$I_\ell(X;Y) = H_\ell(Y) - H_\ell(Y \mid X) = H(Y) - H(Y \mid X) = I(X;Y),$$

the classical Shannon information. $\diamond$

Example 14.1.6 (0-1 error)**:** Fix the same deterministic tie-breaking rule as in Example 14.1.3 and lett $y^\star \in \operatorname{argmax}_y \mathbb{P}(Y = y)$ and $y^\star(x) \in \operatorname{argmax}_y \mathbb{P}(Y = y \mid X = x)$. Then

$$I_{\ell_{0\text{-}1}}(X;Y) = \mathbb{P}(Y \neq y^\star) - \mathbb{E}\left[\mathbb{P}(Y \neq y^\star(X) \mid X)\right] = \mathbb{P}(Y \neq y^\star) - \mathbb{P}(Y \neq y^\star(X)),$$

the gap between the prior and posterior probabilities of making a mistake guessing Y. $\diamond$

Example 14.1.7 (Squared error)**:** For the Brier score with squared error $\ell_{\rm sq}(p, y) = \|p - e_y\|_2^2$, we have $H_{\ell_{\rm sq}}(p) = 1 - \|p\|_2^2$, and so

$$I_{\ell_{\rm sq}}(X;Y) = \sum_{j=1}^k \mathbb{E}\left[\mathbb{P}(Y = j \mid X)^2\right] - \sum_{j=1}^k \mathbb{P}(Y = j)^2 = \sum_{j=1}^k \operatorname{Var}(\mathbb{P}(Y = j \mid X)),$$

the sum of the variances of the random variables $\mathbb{P}(Y = j \mid X)$. The higher the variance of these quantities, the more information X carries about Y. $\diamond$

14.2 Characterizing proper losses and Bregman divergences

With the definition (14.1.4) of the generalized entropy, we can now characterize all proper losses. We do this in three settings: in the first (Section 14.2.1), we give a representation for proper losses when $\mathcal{Y}$ is finite and discrete, so we can identify it with $\mathcal{Y} = \{1, \ldots, k\}$ and distributions P on $\mathcal{Y}$ with probability mass functions $p \in \Delta_k$. We then demonstrate a full characterization of propriety (Section 14.2.2), which requires measure-theoretic tools and may be skipped on a first reading. Finally, we modify the results for finite $\mathcal{Y}$ to consider cases in which Y is vector-valued and $\mathcal{Y} \subset \mathbb{R}^k$ is contained in a compact set. This case transparently generalizes the finite representations of Section 14.2.1 and will form the basis of our development going forward, as it permits direct use of convex-analytic tools.

14.2.1 Characterizing proper losses for Y taking finitely many vales

We present the *Gneiting-Raftery-Savage representation* of proper losses, which characterizes all proper losses using the entropies (14.1.4) or, equivalently, (14.1.5). To avoid pathological cases, we work with regular losses, which always assign a finite value to the correct predicted distribution; we assume regularity without further comment.

Definition 14.2. *Let $\mathcal{P}$ be a family of distributions on $\mathcal{Y}$. The loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* regular for the family $\mathcal{P}$ *if $\mathbb{E}_P[\ell(P, Y)]$ is real valued for all $P \in \mathcal{P}$.*

We allow losses to attain infinite values, as in the case of the logarithmic loss. The following theorem then provides the promised representation of proper losses through their generalized entropies. In the theorem, we use $\Omega'(p)$ to denote a fixed element of the subdifferential $\partial\Omega(p)$; we will allow $\Omega'(p)$ to take extended values in some cases (and will not modify notation for this), but we will be clear when this is necessary.

Theorem 14.2.1 (Proper scoring rules: the finite case)**.** *Let $\mathcal{Y} = \{1, \ldots, k\}$ and $\mathcal{P} \subset \Delta_k$ be convex.*

(i) If the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ satisfies the representation

$$\ell(p, y) = -\Omega(p) - \langle \Omega'(p), e_y - p \rangle \tag{14.2.1}$$

for a subdifferentiable closed convex $\Omega : \mathcal{P} \to \mathbb{R}$, where $\Omega'(p) \in \partial\Omega(p)$, then ℓ is proper.

(ii) Conversely, if ℓ is proper, then the negative generalized entropy

$$\Omega_\ell(p) := -H_\ell(p) = \sup_q \left\{-\mathbb{E}_p[\ell(q, Y)] \mid q \in \mathcal{P}\right\}$$

satisfies equality (14.2.1), and Ω is closed convex.

If ℓ is real valued, then $\Omega'(p) \in \mathbb{R}^k$ in the representation (14.2.1). If ℓ can take value $+\infty$, we allow $\Omega'(p) \in \underline{\mathbb{R}}^k$ when $p \notin \operatorname{relint} \Delta_k$. The loss is strictly proper if and only if Ω is strictly convex on $\mathcal{P}$.

Proof If ℓ has the given representation and $\mathbb{P}(Y = y) = p_y$, then we have

$$\mathbb{E}[\ell(q, Y)] = -\Omega(q) - \langle \Omega'(q), p - q \rangle \geq -\Omega(p) = \mathbb{E}[\ell(p, Y)]$$

by the first-order convexity property of convex functions (definition (14.1.1) of a subdifferential).

Conversely, suppose that the loss is proper, and let $\Omega(p) = \Omega_\ell(p)$. Clearly Ω is convex, as it is the supremum of linear functionals of p. Moreover, propriety of ℓ guarantees that

$$\Omega(p) \geq -\mathbb{E}[\ell(q, Y)] = \Omega(q) + \sum_{y=1}^k -\ell(q, y)(p_y - q_y)$$

That is, for each $q \in \mathcal{P}$ the vector $[-\ell(q, y)]_{y=1}^k \in \partial\Omega(q)$, so Ω is subdifferentiable. Choosing the vector $\Omega'(p) = [-\ell(p, y)]_{y=1}^k$ gives

$$\ell(p, y) = -\Omega(p) + \ell(p, y) + \Omega(p) = -\Omega(p) - \sum_{i=1}^k p_i \ell(p, i) + \ell(p, y) = -\Omega(p) - \langle \Omega'(p), e_y - p\rangle.$$

We may have $\ell(p, i) = +\infty$ only if $p_i = 0$, in which case we use $\sum_i p_i \ell(p, i) < \infty$ to treat $0 \cdot \infty = 0$. Then our definition $\Omega'(p) = [-\ell(p, y)]_{y=1}^k$ remains sensible.

For the strict inequalities and strict propriety, trace the argument replacing inequalities with strict inequalities for $q \neq p$ and use Corollary B.2.3 or C.1.8 in the appendices. □

The negative generalized entropy Ω in Theorem 14.2.1 is essentially unique and marks an important duality between proper losses and convex functions: to each loss, we can assign a generalized entropy, and from this generalized entropy, we can reconstruct the loss. Exercise 14.2 explores this connection and uniqueness. We can also give a few examples that recover standard losses from entropies. For each, we begin with a convex function Ω, then exhibit the associated proper or strictly proper scoring rule. A weakness of the representation (14.2.1) is that, typically, we do *not* achieve a loss function convex in p. In Section 14.3 (and Section 16.4 to come), however, we show how to convert proper losses into *surrogates* convex in their arguments and which, after a particular transformation based on convex duality, are proper and yield the correct distributional predictions.

Example 14.2.2 (Logarithmic losses)**:** Consider the negative entropy $\Omega(p) = \sum_{y=1}^k p_y \log p_y$. We have $\frac{\partial}{\partial p_y}\Omega(p) = 1 + \log p_y \in [-\infty, 1]$, and

$$\ell_{\log}(p, y) = -\sum_{j=1}^k p_j \log p_j + \sum_{j=1}^k p_j(1 + \log p_j) - (1 + \log p_y) = -\log p_y,$$

yielding the log loss. For this case, we require that the gradients $\nabla\Omega(p)$ take values in the (downward) extended reals $\underline{\mathbb{R}}^k$. ◇

Example 14.2.3 (Brier scores and squared error)**:** For the squared error $\ell_{\rm sq}(p, y) = \|p - e_y\|_2^2$, we can directly check that $\Omega(p) = \|p\|_2^2$ gives the loss. Indeed,

$$-\|p\|_2^2 - 2\langle p, e_y - p\rangle = \|p\|_2^2 - 2\langle p, e_y\rangle + 1 - 1 = \|p - e_y\|_2^2 - 1.$$

So aside from an additive constant, we have the desired result. ◇

More esoteric examples exist in the literature, such as the spherical score arising from $\Omega(p) = \|p\|_2$ (note the lack of a square).

Example 14.2.4 (Spherical scores)**:** Let $\Omega(p) = \|p\|_2$, which is strictly convex on Δ_k. Then

$$\nabla\Omega(p) = p/\|p\|_2$$

and $\ell(p, y) = -\|p\|_2 - \frac{1}{\|p\|_2}\langle p, e_y - p\rangle = -p_y/\|p\|_2$, which is strictly proper but non-convex. ◇

Bregman divergences

A key aspect of the Savage representation (14.2.1) is that associated to any proper loss is a *Bregman divergence*, which measures the difference between a convex function and its first-order approximation. Extending Chapter 3 to allow non-differentiable Ω, we tacitly fix a subgradient selection $\Omega'(u) \in \partial\Omega(u)$, and define

$$D_\Omega(u, v) := \Omega(u) - \Omega(v) - \langle \Omega'(v), u - v \rangle. \tag{14.2.2}$$

When Ω is nondifferentiable, D_Ω depends on the selected subgradient, though in the present setting the loss representation (14.2.1) fixes this selection. In particular, $D_\Omega(u, v) \geq 0$ for all u and v, and moreover, if Ω is strictly convex

$$D_\Omega(u, v) > 0 \quad \text{whenever} \quad u \neq v.$$

(See, e.g., Corollaries B.2.3 and C.1.8 in the appendices.)

Familiar examples include the squared Euclidean norm $\Omega(u) = \frac{1}{2}\|u\|_2^2$, giving

$$D_\Omega(u, v) = \frac{1}{2}\|u - v\|_2^2,$$

and the negative entropy $\Omega(u) = \sum_{j=1}^k u_j \log u_j$. This gives

$$D_\Omega(u, v) = \sum_{j=1}^k u_j \log u_j - \sum_{j=1}^k v_j \log v_j - \sum_{j=1}^k (1 + \log v_j)(u_j - v_j) = \sum_{j=1}^k u_j \log \frac{u_j}{v_j} + \mathbf{1}^\top (u - v).$$

If $u, v \in \Delta_k$, then $D_\Omega(u, v) = D_{\rm kl}(u\|v)$ because $\mathbf{1}^\top u = \mathbf{1}^\top v = 1$.

Continuing this identification of distributions on $\mathcal{Y}$ with elements $p \in \Delta_k$, Theorem 14.2.1 provides a representation that immediately connects proper losses with Bregman divergences. Let $\ell(p, y) = -\Omega(p) - \langle \Omega'(p), e_y - p \rangle$ be proper. Then an algebraic manipulation shows that propriety is equivalent to nonnegativity of a Bregman risk gap:

Corollary 14.2.5. *Let the conditions of Theorem 14.2.1 hold. Then ℓ is (strictly) proper if and only if there exists a (strictly) convex $\Omega : \Delta_k \to \overline{\mathbb{R}}$ for which for all $p, q \in \Delta_k$,*

$$\mathbb{E}_p[\ell(q, Y)] - \mathbb{E}_p[\ell(p, Y)] = \Omega(p) - \Omega(q) - \langle \Omega'(q), p - q \rangle = D_\Omega(p, q).$$

14.2.2 General proper losses

Moving beyond the finite discrete case, we can consider predicting distributions P on general sets $\mathcal{Y}$. For example, recalling our meteorological forecasting motivation, suppose we wish to predict a distribution of the (real-valued) amount Y of rainfall on a given day. Many predictions place a point mass at $Y = 0$, with a decaying tail for higher amounts of rainfall. Then it is natural to predict a cumulative distribution function $F : \mathbb{R} \to [0, 1]$ and measure error relative to actual rainfall. Several losses are common; one is the *continuous ranked probability score*.

Example 14.2.6 (Continuous ranked probability score (CRPS))**:** The CRPS loss for a CDF F at y is

$$\ell_{\rm crps}(F, y) = \int (F(t) - \mathbf{1}\{y \leq t\})^2 dt. \tag{14.2.3}$$

This is strictly proper: let G be any cumulative distribution function, and Y have CDF F. Then

$$\begin{aligned}\mathbb{E}[\ell_{\rm crps}(G, Y)] - \mathbb{E}[\ell_{\rm crps}(F, Y)] &= \int \left(G(t)^2 - F(t)^2 - 2(G(t) - F(t))\mathbb{E}[\mathbf{1}\left\{Y \le t\right\}]\right) dt \\ &= \int \left(G(t) - F(t)\right)^2 dt\end{aligned}$$

because $\mathbb{E}[\mathbf{1}\left\{Y \le t\right\}] = F(t)$. This is a variant of the (squared) Cramér-von-Mises distance between F and G, and which is positive unless $F = G$ by the right-continuity of CDFs. (We have ignored integrability issues; bounded supports of F and G suffice.) $\diamond$

The difficulty of computing the continuous ranked probability score motivates considering other losses on probability distributions, where we evaluate selected quantiles rather than the entire CDF. Thus, for a probability distribution P on Y, let

$$\mathsf{Quant}_\alpha(Y) = \mathsf{Quant}_\alpha(P) := \inf\left\{t \mid P(Y \le t) \ge \alpha\right\}$$

to be the α-quantile of the distribution P. (When Y has cumulative distribution F, this is the inverse CDF mapping $F^{-1}(\alpha) = \inf\{t \mid F(t) \ge \alpha\}$.) Defining the quantile penalty

$$\rho_\alpha(t) = \alpha\left[t\right]_+ + (1 - \alpha)\left[-t\right]_+,$$

for a finite collection $\mathcal{A} \subset (0, 1)$, the *quantile loss* is

$$\ell_{{\rm quant},\mathcal{A}}(P, y) := \sum_{\alpha \in \mathcal{A}} \rho_\alpha(y - \mathsf{Quant}_\alpha(P)). \tag{14.2.4}$$

The quantile loss is proper but not strictly so:

Example 14.2.7 (Quantile loss)**:** To see that the quantile loss (14.2.4) is proper, consider the single quantile penalty ρ_α: let $g(t) = \mathbb{E}[\rho_\alpha(Y - t)] = \alpha\mathbb{E}[\left[Y - t\right]_+] + (1 - \alpha)\mathbb{E}[\left[t - Y\right]_+]$, which $\mathsf{Quant}_\alpha(Y)$ minimizes. Indeed, g is convex and has left and right derivatives

$$\begin{aligned}\mathsf{D}_- g(t) &:= \lim_{s \uparrow t} \frac{g(s) - g(t)}{s - t} = -\alpha\mathbb{P}(Y \ge t) + (1 - \alpha)\mathbb{P}(Y < t) = \mathbb{P}(Y < t) - \alpha \quad \text{and} \\ \mathsf{D}_+ g(t) &:= \lim_{s \downarrow t} \frac{g(s) - g(t)}{s - t} = -\alpha\mathbb{P}(Y > t) + (1 - \alpha)\mathbb{P}(Y \le t) = \mathbb{P}(Y \le t) - \alpha.\end{aligned}$$

By definition of the lower quantile, $t = \mathsf{Quant}_\alpha(Y)$ satisfies $\mathbb{P}(Y < t) - \alpha \le 0 \le \mathbb{P}(Y \le t) - \alpha$ and so $\mathsf{D}_- g(t) \le 0 \le \mathsf{D}_+ g(t)$. So convexity yields

$$\mathbb{E}[\rho_\alpha(Y - \mathsf{Quant}_\alpha(Y))] \le \mathbb{E}[\rho_\alpha(Y - t)]$$

for all t. Applying this argument for each $\alpha \in \mathcal{A}$, we thus have for any Q that

$$\mathbb{E}[\ell_{{\rm quant},\mathcal{A}}(Q, Y)] \ge \mathbb{E}[\ell_{{\rm quant},\mathcal{A}}(P, Y)]$$

whenever $Y \sim P$, and equality holds if Q and P have identical α quantile for each $\alpha \in \mathcal{A}$. $\diamond$

A generalized version of Theorem 14.2.1 allows us to address such scenarios, though requires measure theoretic notation. Happily, the proof remains essentially the same as the finite case. For a convex function $\Omega : \mathcal{P} \to \overline{\mathbb{R}}$, we say $\Omega'(P; \cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ is a subderivative of Ω at $P \in \mathcal{P}$ if

$$\begin{aligned} \Omega(Q) &\geq \Omega(P) + \int_{\mathcal{Y}} \Omega'(P; y)(dQ(y) - dP(y)) \\ &= \Omega(P) + \mathbb{E}_Q[\Omega'(P; Y)] - \mathbb{E}_P[\Omega'(P; Y)] \end{aligned} \quad \text{for all } Q \in \mathcal{P} \tag{14.2.5}$$

and each quantity is well-defined. When $\mathcal{Y}$ is discrete and we can identify $\mathcal{P}$ with the simplex Δ_k, the inequality (14.2.5) is the subgradient inequality (14.1.1) that $\Omega(q) \geq \Omega(p) + \langle \Omega'(p), q - p \rangle$ for $p, q \in \Delta_k$, where $\Omega'(p) \in \partial\Omega(p)$. We may thus generalize Theorem 14.2.1 (where we tacitly assume that each expectation or integral is well-defined).

Theorem 14.2.8. *Let $\mathcal{P}$ be a convex collection of distributions on $\mathcal{Y}$ and $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be regular.*

(i) If the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ has the representation

$$\ell(P, y_0) = -\Omega(P) - \Omega'(P; y_0) + \int \Omega'(P; y) dP(y), \quad \textit{for all } y_0 \in \mathcal{Y}, \tag{14.2.6}$$

where $\Omega'(P; \cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ is a subderivative of Ω at $P \in \mathcal{P}$, then ℓ is proper.

(ii) Conversely, if ℓ is proper, then the negative generalized entropy

$$\Omega(P) = \Omega_\ell(P) = -H_\ell(P) = \sup\{-\mathbb{E}_P[\ell(Q, Y)] \mid Q \in \mathcal{P}\}$$

with $\Omega'(P; y) = -\ell(P, y)$ satisfies equality (14.2.6).

The loss is strictly proper if and only if the negative entropy Ω is strictly convex on $\mathcal{P}$.

Proof If ℓ has the representation (14.2.6), then

$$-\mathbb{E}_P[\ell(P, Y)] = \Omega(P) \geq \Omega(Q) + \int \Omega'(Q; y)(dP(y) - dQ(y)) = -\mathbb{E}_P[\ell(Q, Y)]$$

for any $Q \in \mathcal{P}$ by the definition (14.2.5) of a subderivative. So ℓ is proper.

Conversely, if ℓ is proper, then as in the proof of Theorem 14.2.1 we define

$$\Omega(P) := \sup_{Q \in \mathcal{P}} -\mathbb{E}_P[\ell(Q, Y)] = -\mathbb{E}_P[\ell(P, Y)],$$

which is convex. Defining $\Omega'(P; y) = -\ell(P, y) \in \underline{\mathbb{R}}$ for $P \in \mathcal{P}$,

$$\Omega(Q) \geq -\mathbb{E}_Q[\ell(P, Y)] = \Omega(P) + \mathbb{E}_P[\ell(P, Y)] - \mathbb{E}_Q[\ell(P, Y)] = \Omega(P) + \int \Omega'(P; y)(dQ(y) - dP(y))$$

by propriety, so that evidently $\Omega'(P; y)$ is a subderivative of Ω at $P \in \mathcal{P}$. The equality $\ell(P, y_0) = -\Omega(P) - \Omega'(P; y_0) + \int \Omega'(P; y) dP(y)$ is immediate.

The arguments for strict propriety and convexity are similar, replacing non-strict inequalities with strict inequalities. □

The obvious corollary to Theorem 14.2.8, paralleling Corollary 14.2.5, follows.

Corollary 14.2.9. *Let $\mathcal{P}$ be a convex collection of probability distributions on $\mathcal{Y}$, and for convex $\Omega : \mathcal{P} \to \overline{\mathbb{R}}$ with subderivatives $\Omega'(P; \cdot) : \mathcal{Y} \to \underline{\mathbb{R}}$ define the generalized Bregman divergence*

$$D_\Omega(P, Q) := \Omega(P) - \Omega(Q) - \int \Omega'(Q; y)(dP(y) - dQ(y)).$$

Then the loss $\ell : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is (strictly) proper if and only if there exists a (strictly) convex and subdifferentiable $\Omega : \mathcal{P} \to \overline{\mathbb{R}}$ for which

$$\mathbb{E}_P[\ell(Q, Y)] - \mathbb{E}_P[\ell(P, Y)] = D_\Omega(P, Q).$$

The subdifferentials and differentiability in this potentially infinite dimensional case can make writing the particular representation (14.2.6) challenging; for example, the representation of the quantile loss in Example 14.2.7 is quite complex. In the case of predictions involving the cumulative distribution function F, one can obtain the subderivative by taking directional (Gateaux) derivatives in directions $G - F$ for cumulative distributions G. In this case, for the point cumulative distribution G_y with $G_y(t) = \mathbf{1}\{y \le t\}$, we define

$$\Omega'(F; y) = \lim_{\epsilon \downarrow 0} \frac{\Omega(F + \epsilon(G_y - F)) - \Omega(F)}{\epsilon}.$$

The continuous ranked probability score (Example 14.2.6) admits this expansion (again, we ignore integrability issues).

Example 14.2.10 (CRPS continued)**:** Under appropriate integrability conditions, the strict propriety of the CRPS loss (14.2.3) gives generalized negative entropy

$$\Omega(F) = \sup_G -\mathbb{E}[\ell(G, Y)] = -\mathbb{E}[\ell_{\text{crps}}(F, Y)] = \int (F(t) - 1)F(t)dt.$$

Expanding $\Omega(F + \epsilon(G - F))$ for small ϵ as in the recipe above, we have

$$\Omega(F + \epsilon(G - F)) = \Omega(F) - \epsilon \int (G(t) - F(t))dt + 2\epsilon \int F(t)(G(t) - F(t))dt + O(\epsilon^2),$$

yielding directional derivative

$$\Omega'(F; G - F) = \int (2F(t) - 1)(G(t) - F(t))dt.$$

Taking $G_y(t) = \mathbf{1}\{y \le t\}$ identifies $\Omega'(F; y) = \Omega'(F; G_y - F)$. By inspection, if Y has cumulative distribution function F, then $\mathbb{E}[G_Y(t)] = F(t)$ and so $\mathbb{E}[\Omega'(F; Y)] = 0$. Substituting in the representation (14.2.6) gives

$$-\Omega(F) - \Omega'(F; y) = \int \left(F(t) - \mathbf{1}\{y \le t\}\right)^2 dt = \ell_{\text{crps}}(F, y),$$

as desired. $\diamond$

We can also consider the log loss for general distributions, though the necessity of defining a base measure makes this subtle. For discrete cases—when $\mathcal{Y}$ is finite or countable—we can always use the probability mass function.

Example 14.2.11 (Log loss for general distributions)**:** For general distributions P, we fix a base measure ν on $\mathcal{Y}$, and let $\mathcal{P}$ be the collection of distributions absolutely continuous with respect to ν. For $Q \in \mathcal{P}$ with density q, we define $\ell(Q, y) = -\log q(y)$, while $\ell(Q, y) = +\infty$ for $Q \not\ll \nu$. For $P \in \mathcal{P}$,

$$\mathbb{E}_P[\ell(Q, Y)] = \mathbb{E}_P[-\log q(Y)] = \mathbb{E}_P\left[\log \frac{p(Y)}{q(Y)}\right] - \mathbb{E}_P[\log p(Y)] = D_{\rm kl}\left(P\|Q\right) + H_\nu(P),$$

where $H_\nu(P) := -\int p(y) \log p(y) d\nu(y)$ denotes the Shannon entropy for the base measure ν.

The negative entropy $\Omega(P) = -H_\nu(P)$ when $P \ll \nu$, $+\infty$ otherwise. To obtain the subderivatives in the representation (14.2.6), we heuristically compute $\frac{\partial}{\partial p(y)}\Omega(P)$ to guess that

$$\Omega'(P; y) = 1 + \log p(y).$$

We can directly check that this is indeed a subgradient for Ω: we have

$$\Omega(P) + \mathbb{E}_Q[\Omega'(P; Y)] - \mathbb{E}_P[\Omega'(P; Y)] = \mathbb{E}_Q\left[\log \frac{p(Y)}{q(Y)}\right] + \mathbb{E}_Q[\log q(Y)] = \Omega(Q) - D_{\rm kl}\left(Q\|P\right),$$

assuming $Q \ll \nu$. (Otherwise $-\mathbb{E}_Q[\log p(Y)] = D_{\rm kl}\left(Q\|P\right) = +\infty$ regardless.) Then we have

$$-\Omega(Q) - \Omega'(Q; y) + \mathbb{E}_Q[\Omega'(Q; Y)] = H_\nu(Q) - H_\nu(Q) - 1 + 1 - \log q(y) = -\log q(y),$$

verifying the loss representation (14.2.6). $\diamond$

14.2.3 Proper losses for vector-valued Y

The final variant of propriety we consider generalizes the case that $\mathcal{Y}$ is finite in Section 14.2.1. We now assume that Y is vector-valued, with $\mathcal{Y} \subset \mathbb{R}^k$, and let

$$\mathrm{Conv}(\mathcal{Y}) = \left\{\mathbb{E}_P[Y] \mid P \text{ is a distribution on } \mathcal{Y}\right\},$$

which is compact. (See Corollary B.1.24.) A few examples may be clarifying.

Example 14.2.12 (Multinomial Y as vectors)**:** If Y is a multinomial taking values in a discrete set of size k, we can instead identify Y with the first k standard basis vectors $e_1, \ldots, e_k$. Then $p = \mathbb{E}[Y] \in \Delta_k$ is the p.m.f. of Y, and $\mathrm{Conv}(\mathcal{Y}) = \Delta_k$. $\diamond$

Example 14.2.13 (Binary Y as a scalar)**:** When $Y \in \{0, 1\}$ is a Bernoulli random variable, we identify Y with itself, so that $p = \mathbb{E}[Y] = \mathbb{P}(Y = 1) \in [0, 1]$ and $\mathrm{Conv}(\mathcal{Y}) = [0, 1]$. $\diamond$

Example 14.2.14 (Ordinal Y as a scalar)**:** Consider a rating problem to predict an outcome of 1 to 5 stars. In this case, Y takes values $\{1, \ldots, 5\} \subset \mathbb{R}$, but the ordering between the elements is important. More generally, Y may take values in $\{y_1, \ldots, y_k\} \subset \mathbb{R}$, where $y_1 < \cdots < y_k$. Identify Y with its scalar value, $\mathbb{E}[Y] \in [y_1, y_k]$ and $\mathrm{Conv}(\mathcal{Y}) = [y_1, y_k]$. $\diamond$

In this vector-valued Y case, instead of distributions P, the goal is to predict the *mean mapping*

$$\mu(P) := \mathbb{E}_P[Y] \in \mathrm{Conv}(\mathcal{Y}),$$

so that $\mu : \mathcal{P} \to \mathbb{R}^k$ for the collection $\mathcal{P}$ of distributions on Y. We wish to preference predictions of the correct expectation, leading to the following definition.

Definition 14.3. *Let $\mathcal{Y}$ be finite and $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$ be the collection of mean parameters for $Y \in \mathcal{Y}$. Then $\ell : \mathcal{M} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is* proper *if*

$$\mathbb{E}_P[\ell(\mu, Y)] \geq \mathbb{E}_P[\ell(\mathbb{E}_P[Y], Y)] \quad \text{for all } \mu \in \mathcal{M},$$

and strictly proper *if the inequality is strict whenever $\mu \neq \mathbb{E}_P[Y]$.*

Definition 14.3 generalizes Definition 14.2 in the multinomial case, as Example 14.2.12 makes clear.

We can extend Theorem 14.2.1 to a more general case, where as usual we tacitly assume ℓ is *regular*, with $\mathbb{E}_P[\ell(\mathbb{E}_P[Y], Y)] = \mathbb{E}_P[\ell(\mu(P), Y)] < \infty$ for all distributions P on $\mathcal{Y}$. For $\mu \in \mathcal{M}$, define the supporting set

$$\mathcal{S}(\mu) := \{y \in \mathcal{Y} \mid \text{there exists } P \text{ with } P(Y = y) > 0 \text{ and } \mathbb{E}_P[Y] = \mu\}.$$

Lemma B.1.16 shows that $\operatorname{relint} \mathcal{M} = \{\mathbb{E}_P[Y] \mid \min_y P(Y = y) > 0\}$, so $\mathcal{S}(\mu) = \mathcal{Y}$ for $\mu \in \operatorname{relint} \mathcal{M}$.

Theorem 14.2.15. *Let $\mathcal{Y} \subset \mathbb{R}^k$ and $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$.*

(i) Let $a : \mathcal{Y} \to \mathbb{R}$ and the remainder $r(\mu, y) \geq 0$ satisfy $r(\mu, y) = 0$ if $y \in \mathcal{S}(\mu)$. Assume the loss $\ell : \mathcal{M} \times \mathcal{Y} \to \overline{\mathbb{R}}$ has representation

$$\ell(\mu, y) = -\Omega(\mu) - \langle \Omega'(\mu), y - \mu \rangle + a(y) + r(\mu, y) \tag{14.2.7}$$

for a convex $\Omega : \mathcal{M} \to \mathbb{R}$ with selection $\Omega'(\mu) \in \partial\Omega(\mu)$. Then ℓ is proper. (We allow extended values $\langle \Omega'(\mu), y \rangle \in \underline{\mathbb{R}}$.)

(ii) Conversely, if $\ell : \mathcal{M} \times \mathcal{Y} \to \mathbb{R}$ is proper, then it has representation (14.2.7), where Ω is closed and subdifferentiable on $\mathcal{M}$ and $\Omega'(\mu) \in \mathbb{R}^k$ is finite.

The loss is strictly proper if and only if Ω is strictly convex. Additionally, if $\ell : \mathcal{M} \times \mathcal{Y} \to \overline{\mathbb{R}}$ is proper, the representation (14.2.7) holds on $\operatorname{relint} \mathcal{M}$ and $\Omega'(\mu) \in \mathbb{R}^k$.

The outcome only term $a(y)$ in the representation (14.2.7) does not affect propriety, and the remainder r is necessary only at (relative) boundary means μ, which preclude some outcomes $y \in \mathcal{Y}$. Because the proof relies on much of the convex analysis in Appendix B, we therefore defer it to Section 14.5.1. An example on ordinal regression illustrates the scenarios.

Example 14.2.16 (Ordinal regression, Example 14.2.14 continued)**:** Let $Y \in \{0, 1, \ldots, k\}$ be a value to be predicted, where the ordering on Y is important, as in ratings of items. In this case, the set $\mathcal{M} = \mathrm{Conv}(\mathcal{Y}) = [0, k]$, and any strictly convex Ω with domain $[0, k]$ gives rise to a proper loss via the construction $\ell_\Omega(\mu, y) = -\Omega(\mu) - \Omega'(\mu)(y - \mu)$. First, we take $\Omega(\mu) = \frac{1}{2}\mu^2$. This gives rise to a (modified) squared error

$$\ell_\Omega(\mu, y) = \frac{1}{2}(\mu - y)^2 - \frac{1}{2}y^2,$$

which is strictly convex in μ and strictly proper.

Another natural choice for Ω is a variant of the negative binary entropy, and we define

$$\Omega(\mu) = (k - \mu) \log(k - \mu) + \mu \log \mu,$$

which is strictly convex in $\mu \in [0, k]$, with $\Omega(\mu) = +\infty$ for $\mu > k$ or $\mu < 0$, while $\Omega(0) = \Omega(k) = k \log k$. We have $\Omega'(\mu) = \log \frac{\mu}{k - \mu}$, and so

$$\ell_\Omega(\mu, y) = -y \log \mu + (y - k) \log(k - \mu),$$

for $y \in \{0, \ldots, k\}$. Note the importance of allowing infinite values in ℓ when $\mu \to \{0, k\}$. $\diamond$

14.3 From entropies to convex losses, arbitrary predictions, and link functions

Frequently, when we fit models, it is inconvenient to directly model or predict probabilities, that is, to minimize over probabilistic predictions. Instead, we wish to fit some real-valued prediction and then transform it into a probabilistic prediction. This is perhaps most familiar from binary and multiclass logistic regression, where a *link function* transforms real-valued predictions into probabilistic predictions. For the binary logistic regression case with $Y \in \{-1, 1\}$, we assume we predict a score $s \in \mathbb{R}$, where $s > 0$ indicates a prediction that Y is more likely to be 1 and $s < 0$ that it is more likely negative. The implied (modeled) probability that $Y = y$ is then

$$p(y \mid s) = \frac{1}{1 + \exp(-ys)} \quad \text{for } y \in \{-1, 1\}.$$

For k-class classification using logistic regression, we predict a score vector $s \in \mathbb{R}^k$, where s_y indicates a score associated to one of the k potential class labels y; this implies the probabilites

$$p(y \mid s) = \frac{\exp(s_y)}{\sum_{i=1}^k \exp(s_i)} = \frac{1}{1 + \sum_{i \neq y} \exp(s_i - s_y)}.$$

In each, instead of directly minimizing log probabilities of error over the probability simplex, instead we use surrogate logistic losses whose arguments can range over $\mathbb{R}$ or $\mathbb{R}^k$. In the case of binary logistic regression with $y \in \{-1, 1\}$, this is

$$\varphi(s, y) = \log(1 + \exp(-sy)),$$

while in the multiclass case we use the multiclass logistic loss

$$\varphi(s, y) = -s_y + \log\left(\sum_{i=1}^k \exp(s_i))\right) = \log\left(1 + \sum_{i \neq y} \exp(s_i - s_y)\right).$$

Each of these provides a direct relationship between the probabilistic predictions and derivatives of φ. In the binary logistic regression case, we have

$$p(y \mid s) = 1 + \frac{\partial}{\partial s}\varphi(s, y) = 1 - \frac{1}{1 + \exp(ys)} = \frac{1}{1 + \exp(-ys)},$$

while in the multiclass case we similarly have

$$p(y \mid s) = 1 + \frac{\partial}{\partial s_y}\varphi(s, y) = \frac{\exp(s_y)}{\sum_{i=1}^k \exp(s_i)}.$$

14.3.1 Convex conjugate linkages

These dualities hold in substantially more generality, and they are the key to transforming proper losses (as applied on probabilities) into proper surrogate losses that apply to real-valued scores and are convex in their arguments, allowing us to bring the tools of convex optimization to bear on fitting predictive models. We work in the general setting of Section 14.2.3 of losses for vector-valued y where $\mathcal{Y} \subset \mathbb{R}^k$, so that instead of predicting probability distributions on Y itself we predict

elements μ of the means $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$, and let ℓ be a strictly proper loss. Ignoring boundary issues, Theorems 14.2.1 and 14.2.15 demonstrate that if the loss ℓ is proper, there exists a closed convex negative generalized entropy Ω for which (up to normalization)

$$\ell(\mu, y) = -\Omega(\mu) - \langle \Omega'(\mu), y - \mu \rangle. \tag{14.3.1}$$

Let us suppose temporarily that we have a negative entropy $\Omega : \mathcal{M} \to \overline{\mathbb{R}}$ in this representation. Recalling the convex conjugate (14.1.2), the conjugate $\Omega^*(s) = \sup\{\langle s, \mu \rangle - \Omega(\mu)\}$ satisfies $\Omega^{**}(\mu) = \Omega(\mu)$. In particular, the Fenchel-Young *surrogate loss*

$$\varphi(s, y) := \Omega^*(s) - \langle s, y \rangle,$$

which is defined for all $s \in \mathbb{R}^k$ (instead of $\mathrm{Conv}(\mathcal{Y})$), satisfies

$$\mathbb{E}_P[\varphi(s, Y)] = \Omega^*(s) - \langle s, \mathbb{E}_P[Y] \rangle = \Omega^*(s) - \langle s, \mu(P) \rangle$$

for the mean mapping $\mu(P) = \mathbb{E}_P[Y]$. Moreover,

$$\inf_s \mathbb{E}_P[\varphi(s, Y)] = \inf_s \left\{\Omega^*(s) - \langle s, \mu(P) \rangle\right\} = -\Omega^{**}(\mu(P)) = -\Omega(\mu(P)),$$

and so it generates the same negative entropy as the original loss ℓ, as

$$\inf_\mu \mathbb{E}_P[\ell(\mu, Y)] = \inf_\mu \{-\Omega(\mu) - \langle \Omega'(\mu), \mu(P) - \mu \rangle\} = -\Omega(\mu(P)).$$

This identification of (generalized) entropies will underpin much of our development of loss consistency in Chapter 16. For now, we content ourselves with addressing propriety of the surrogate loss φ and how to transform predictions $s \in \mathbb{R}^k$ into probabilistic predictions μ.

The key will be to consider what we term *convex-conjugate-linkages*, or *conjugate linkages* for short. Recall the duality relationships (14.1.3) from the Fenchel-Young inequality. We first give a somewhat heuristic presentation. When the maximizer is unique, we may define the *conjugate link*

$$\mathsf{pred}_\Omega(s) := \operatorname*{argmax}_\mu \{\langle s, \mu \rangle - \Omega(\mu)\}. \tag{14.3.2}$$

If $\mu = \mathsf{pred}_\Omega(s)$ and $s = \Omega'(\mu) \in \partial\Omega(\mu)$ is the subgradient used to define the proper loss (14.3.1), the Fenchel-Young equalities (14.1.3) and assumed uniqueness of the maximizer give

$$\mu = \mathsf{pred}_\Omega(s) = \nabla\Omega^*(s) = (\partial\Omega)^{-1}(s) \;\; \text{if and only if} \;\; \Omega^*(s) + \Omega(\mu) = \langle s, \mu \rangle.$$

Substituting into the representation (14.3.1) thus gives the pointwise identity

$$\ell(\mathsf{pred}_\Omega(s), y) = \Omega^*(s) - \langle s, y \rangle.$$

That is, we have recovered the surrogate

$$\varphi(s, y) = \Omega^*(s) - \langle s, y \rangle. \tag{14.3.3}$$

The surrogate loss (14.3.3) constructed from the negative entropy Ω is the key transformation of the loss ℓ into a convex loss, and (no matter the properties of ℓ) is always convex.

All that remains is to give conditions under which the prediction (14.3.2) is unique and exists for all possible score vectors $s \in \mathbb{R}^k$. We make the following definition.

Definition 14.4. *A closed convex $\Omega : \mathbb{R}^k \to \overline{\mathbb{R}}$ is a* Legendre negative entropy *if* $\operatorname{int}\operatorname{dom}\Omega \neq \varnothing$, *it is strictly convex and continuously differentiable on* $\operatorname{int}\operatorname{dom}\Omega$, *and*

$$\|\nabla\Omega(\mu)\| \to \infty \quad \text{as either } \mu \to \operatorname{bd}\operatorname{dom}\Omega \quad \text{or} \quad \|\mu\| \to \infty. \tag{14.3.4}$$

This is precisely the condition we require to make each step in the development of the surrogate (14.3.3) airtight; as a corollary to Theorem C.3.7 in Appendix C.3.3, we have the following.

Corollary 14.3.1. *Let Ω be a Legendre negative entropy. Then the conjugate link prediction (14.3.2) is unique and exists for all $s \in \mathbb{R}^k$. In particular, the conjugate Ω^* is strictly convex, continuously differentiable, satisfies $\operatorname{dom}\Omega^* = \mathbb{R}^k$, and $\nabla\Omega^* = (\nabla\Omega)^{-1}$.*

The corollary allows us to show the equivalence of the strictly proper loss ℓ and its surrogate:

Theorem 14.3.2. *Let $\ell : \mathcal{M} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be the strictly proper loss associated with the Legendre negative entropy Ω. Then*

$$\ell(\mathsf{pred}_\Omega(s), y) = \varphi(s, y) := \Omega^*(s) - \langle s, y\rangle.$$

Moreover, the convex surrogate φ satisfies the consistency that if

$$\mathbb{E}[\varphi(s_n, Y)] \to \inf_s \mathbb{E}[\varphi(s, Y)]$$

then $\mu_n = \mathsf{pred}_\Omega(s_n)$ satisfies

$$\mathbb{E}[\ell(\mu_n, Y)] \to \inf_\mu \mathbb{E}[\ell(\mu, Y)].$$

Proof The first equality we have already demonstrated. For the minimization claim, we note that if $\mu = \mathbb{E}[Y]$, then $\mathbb{E}[\varphi(s, Y)] = \Omega^*(s) - \langle\mu, s\rangle$ and $\inf_s\{\Omega^*(s) - \langle\mu, s\rangle\} = -\Omega(\mu)$. Strict propriety of ℓ then gives $\inf_{\mu'} \mathbb{E}[\ell(\mu', Y)] = -\Omega(\mu)$. □

Said differently, the surrogate φ is consistent with the loss ℓ and (strictly) proper, in that if s minimizes $\mathbb{E}[\varphi(s, Y)]$, then $\mathsf{pred}_\Omega(s)$ minimizes $\mathbb{E}[\ell(\mu, Y)]$. The statement in terms of limits is necessary, as simple examples show, because some link functions cannot achieve extreme points of $\mathcal{M} = \operatorname{Conv}(\mathcal{Y})$. We provide two examples, the first showing necessity of limits, the second (multiclass classification) highlighting a the need to extend Theorem 14.3.2 to allow constraints.

Example 14.3.3 (Binary logistic regression)**:** For a label $Y \in \{0, 1\}$ and predictions $p \in [0, 1]$, take the generalized negative entropy

$$\Omega(p) = p \log p + (1 - p) \log(1 - p).$$

By inspection, $\operatorname{dom}\Omega = [0, 1]$, and $\Omega'(p) = \log \frac{p}{1-p}$ satisfies $|\Omega'(p)| \to \infty$ as $p \to \{0, 1\}$. For $s \in \mathbb{R}$, the conjugate

$$\Omega^*(s) = \sup_p\{sp - p \log p - (1 - p) \log(1 - p)\} = \log(1 + e^s),$$

where the supremum is achieved by $p = \mathsf{pred}_\Omega(s) = \frac{e^s}{1+e^s}$. Then

$$\varphi(s, y) = \log(1 + e^s) - sy = -\log p(y \mid s),$$

where $p(y \mid s) = \frac{e^{ys}}{1+e^s}$ is the binary logistic probability of the label $y \in \{0, 1\}$.

For the induced log loss $\ell(p, y) = -y \log p - (1 - y) \log(1 - p)$, if $\mathbb{P}(Y = 1) = 1$, then $p = 1$ minimizes $\mathbb{E}[\ell(p, Y)]$. Similarly, if $\mathbb{P}(Y = 0) = 1$, then $p = 0$ minimizes $\mathbb{E}[\ell(p, Y)]$. Each of these is achieved as $s \to \pm\infty$ in $p(y \mid s) = \frac{e^{ys}}{1+e^s}$. ◇

Sometimes, we need to elaborate the setting of Theorem 14.3.2 to address constraints:

Example 14.3.4 (Multiclass logistic regression)**:** Identify $\mathcal{Y} = \{e_1, \ldots, e_k\}$ with the k standard basis vectors, and for $p \in \Delta_k = \{p \in \mathbb{R}_+^k \mid \mathbf{1}^\top p = 1\}$, consider the negative entropy

$$\Omega(p) = \sum_{y=1}^k p_y \log p_y.$$

This function is strictly convex and of Legendre type for the positive orthant $\mathbb{R}_+^k$ but not Δ_k. Shortly, we shall allow linear constraints on the predictions to address this shortcoming.

As an alternative, take $\mathcal{Y} = \{0, e_1, \ldots, e_{k-1}\}$, so $\mathcal{M} = \mathrm{Conv}(\mathcal{Y}) = \{p \in \mathbb{R}_+^{k-1} \mid \mathbf{1}^\top p \le 1\}$ has an interior and more easily admits a conjugate duality relationship. In this case, the negative entropy

$$\Omega(p) = \sum_{y=1}^{k-1} p_y \log p_y + (1 - \mathbf{1}^\top p) \log(1 - \mathbf{1}^\top p) \tag{14.3.5}$$

is of Legendre type on $\mathcal{M}$. A calculation for $s \in \mathbb{R}^{k-1}$ yields

$$\Omega^*(s) = \log\left(1 + \sum_{y=1}^{k-1} e^{s_y}\right),$$

with

$$\mathsf{pred}_\Omega(s) = \left(\frac{e^{s_1}}{1 + \sum_{j=1}^{k-1} e^{s_j}}, \ldots, \frac{e^{s_{k-1}}}{1 + \sum_{j=1}^{k-1} e^{s_j}}\right).$$

Letting p denote the entries of this vector, we can then assign a probability to class k via $p_k = 1 - \sum_{j=1}^{k-1} p_j = 1/(1 + \sum_{j=1}^{k-1} e^{s_j})$. $\diamond$

14.3.2 Convex conjugate linkages with affine constraints

As Example 14.3.4 shows, in some cases a "natural" formulation fails to satisfy the desiderata of our link functions. Accordingly, we modify Legendre type (14.3.4) negative entropy Ω to allow for affine constraints. Continuing to work in the scenario in which $\mathcal{Y} \subset \mathbb{R}^k$, suppose now that the affine hull

$$\mathcal{A} = \mathrm{aff}(\mathcal{Y}) := \left\{\sum_{j=1}^m \alpha_j y_j \mid y_j \in \mathcal{Y}, \alpha^\top \mathbf{1} = 1, m \in \mathbb{N}\right\}$$

is a proper subspace of $\mathbb{R}^k$. The key motivating example here is the "failure" case of Example 14.3.4 on multiclass logistic regression, where $\mathcal{Y} = \{e_1, \ldots, e_k\}$, with affine hull $\{p \in \mathbb{R}^k \mid \langle p, \mathbf{1}\rangle = 1\}$. In this case we wish to predict probabilities, and so given a score vector $s \in \mathbb{R}^k$ and using the negative entropy $\Omega(p) = \sum_{y=1}^k p_y \log p_y$, we let

$$\mathsf{pred}(s) = \operatorname*{argmin}_p \left\{\Omega(p) - \langle s, p\rangle \mid \mathbf{1}^\top p = 1\right\} = \left[e^{s_y} / \sum_{j=1}^k e^{s_j}\right]_{y=1}^k.$$

Generalizing this approach to arbitrary regularizers Ω, we modify the prediction (14.3.2) to be

$$\mathsf{pred}_{\Omega,\mathcal{A}}(s) = \operatorname*{argmax}_{\mu \in \mathcal{A}} \left\{\langle s, \mu\rangle - \Omega(\mu)\right\}.$$

Then for the loss $\ell(\mu, y) = -\Omega(\mu) - \langle \nabla\Omega(\mu), y - \mu\rangle$ associated with the negative entropy Ω, we define the surrogate

$$\varphi(s, y) := \ell(\mathsf{pred}_{\Omega,\mathcal{A}}(s), y).$$

This construction still yields a well-defined convex loss with the same consistency properties as those in Theorem 14.3.2. Indeed, defining

$$\Omega_{\mathcal{A}}(\mu) = \Omega(\mu) + \mathbf{I}_{\mathcal{A}}(\mu)$$

and the associated conjugate $\Omega^*_{\mathcal{A}}(s) = \sup\{\langle s, \mu\rangle - \Omega(\mu) \mid \mu \in \mathcal{A}\}$, we have the following theorem.

Theorem 14.3.5. *Let $\ell : \mathcal{M} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be the strictly proper loss associated with the Legendre negative entropy Ω and $\mathcal{A} = \mathrm{aff}(\mathcal{Y})$ be the affine hull of $\mathcal{Y}$. Then*

$$\varphi(s, y) := \ell(\mathsf{pred}_{\Omega,\mathcal{A}}(s), y) = \Omega^*_{\mathcal{A}}(s) - \langle s, y\rangle.$$

Moreover, the convex surrogate φ satisfies the consistency that if

$$\mathbb{E}[\varphi(s_n, Y)] \to \inf_s \mathbb{E}[\varphi(s, Y)]$$

then $\mu_n = \mathsf{pred}_{\Omega,\mathcal{A}}(s_n)$ satisfies

$$\mathbb{E}[\ell(\mu_n, Y)] \to \inf_\mu \mathbb{E}[\ell(\mu, Y)].$$

We return to proving the theorem presently, first showing how it applies to Example 14.3.4.

Example 14.3.6 (Multiclass logistic regression)**:** Consider Example 14.3.4, where $\mathcal{Y} = \{e_1, \ldots, e_k\} \subset \mathbb{R}^k$ has affine hull $\mathcal{A} = \{p \in \mathbb{R}^k \mid \langle \mathbf{1}, p\rangle = 1\}$. Then taking $\Omega(p) = \sum_{y=1}^k p_k \log p_k$, a calculation with a Lagrangian shows that

$$\mathsf{pred}_{\Omega,\mathcal{A}}(s) = \operatorname*{argmin}_{p \in \Delta_k} \{-\langle s, p\rangle + \Omega(p)\} = \left[e^{s_y} / \sum_{j=1}^k e^{s_j}\right]_{y \in [k]}.$$

In turn, this gives surrogate logistic loss

$$\varphi(s, y) = \log\left(\sum_{j=1}^k e^{s_j - s_y}\right).$$

The logistic loss is *not* strictly convex, as $\varphi(s + t\mathbf{1}, y) = \varphi(s, y)$ for $t \in \mathbb{R}$. If Y is a multinomial random variable with $\mathbb{P}(Y = e_y) = p_y$, then another calculation shows the vector

$$s^\star_y = \log p_y$$

minimizes $\mathbb{E}[\varphi(s, Y)]$, which in turn gives $\mathsf{pred}_{\Omega,\mathcal{A}}(s^\star) = p$, maintaining propriety. $\diamond$

Proof of Theorem 14.3.5 Before proving the theorem proper, we show how the key identity that $s = \nabla\Omega(\mu)$ we use to develop equality (14.3.3) generalizes in the presence of the affine constraint. The function $h_{\mathcal{A}}$ is strictly convex on its domain $\mathrm{dom}\, h \cap \mathcal{A}$, and moreover, $\nabla h^*_{\mathcal{A}}$ exists and is continuous. The following lemma (a consequence of Corollary C.3.10 in Appendix C.2) extends Corollary 14.3.1 and allows us to address equality (14.3.3).

Lemma 14.3.7. *The conjugate $\Omega^*_{\mathcal{A}}$ is continuously differentiable with $\operatorname{dom}\Omega^*_{\mathcal{A}} = \mathbb{R}^k$, and if $\mu = \nabla\Omega^*_{\mathcal{A}}(s)$, then $\mu \in \operatorname{int}\operatorname{dom}\Omega$ and*

$$\nabla\Omega(\mu) = s + v$$

for some vector v normal to $\mathcal{A}$, that is, a vector $v \in \mathbb{R}^k$ satisfying $\langle v, \mu_0 - \mu_1\rangle = 0$ for all $\mu_0, \mu_1 \in \mathcal{A}$.

We return to prove the theorem. Take $s \in \mathbb{R}^k$. Then because $\mathsf{pred}_{\Omega,\mathcal{A}}(s) = \nabla\Omega^*_{\mathcal{A}}(s)$,

$$\varphi(s, y) = \ell(\mathsf{pred}_{\Omega,\mathcal{A}}(s), y) = -\Omega(\nabla\Omega^*_{\mathcal{A}}(s)) - \langle\nabla\Omega(\nabla\Omega^*_{\mathcal{A}}(s)), y - \nabla\Omega^*_{\mathcal{A}}(s)\rangle.$$

As $\nabla\Omega^*_{\mathcal{A}}(s) \in \mathcal{A}$ and using the shorthand $\mu = \nabla\Omega^*_{\mathcal{A}}(s) \in \mathcal{A}$, we have $\nabla\Omega(\mu) = s + v$ for some v normal to $\mathcal{A}$. Moreover, $\Omega(\mu) = \Omega_{\mathcal{A}}(\mu)$, and so the Fenchel-Young inequality (14.1.3) guarantees $-\Omega_{\mathcal{A}}(\mu) = \Omega^*_{\mathcal{A}}(s) - \langle s, \mu\rangle$. Substituting in the expression for φ, we obtain

$$\begin{aligned}\varphi(s, y) &= \Omega^*_{\mathcal{A}}(s) - \langle s, \mu\rangle - \langle s + v, y - \mu\rangle \\ &= \Omega^*_{\mathcal{A}}(s) - \langle s, \mu\rangle + \langle s, \mu - y\rangle = \Omega^*_{\mathcal{A}}(s) - \langle s, y\rangle\end{aligned}$$

where the second equality follows because $v \perp \mu - y$.

For the consistency argument, let $\mu_n = \mathsf{pred}_{\Omega,\mathcal{A}}(s_n)$. Then $\mathbb{E}[\ell(\mu_n, Y)] = \mathbb{E}[\varphi(s_n, Y)]$ and if $\mu = \mathbb{E}[Y]$, then $\mathbb{E}[\varphi(s, Y)] = \Omega^*_{\mathcal{A}}(s) - \langle\mu, s\rangle$ and $\inf_s \mathbb{E}[\varphi(s, Y)] = -\Omega_{\mathcal{A}}(\mu) = -\Omega(\mu)$. Strict propriety of ℓ gives $\inf_{\mu'} \mathbb{E}[\ell(\mu', Y)] = -\Omega(\mu)$. □

14.4 Exponential families, maximum entropy, and log loss

Chapter 3 introduced exponential families and their likelihood theory. Here, we revisit and investigate them in the framework we have developed, which allows us to connect proper losses, convex conjugacy, and maximum entropy. Recall that a regular exponential family $\{P_\theta\}$ has density

$$p_\theta(x) = \exp(\langle\theta, \phi(x)\rangle - A(\theta)), \quad A(\theta) = \log \int e^{\langle\theta, \phi(x)\rangle} d\nu(x)$$

with respect to a base measure ν, where the domain $\Theta := \operatorname{dom} A$ is open. The log loss $-\log p_\theta(x)$ then corresponds to the Fenchel-Young loss (14.3.3)

$$\varphi(\theta, x) := -\log p_\theta(x) = A(\theta) - \langle\theta, \phi(x)\rangle.$$

Proposition 3.2.1 guarantees φ is convex in θ, and it is $\mathcal{C}^\infty$ (Proposition 3.2.2).

The mean mapping $\mu(P) := \mathbb{E}_P[\phi(X)] \in \mathbb{R}^d$ will be central to the duality between maximum likelihood and entropy we explore here. Accordingly, throughout this section we let

$$\mathcal{P} := \{\text{distributions } P \ll \nu\} = \{\text{distributions } P \text{ with a density } p \text{ w.r.t. } \nu\}$$

be the collection of distributions with densities with respect to ν, and we define the set of potential *mean parameters*, also called the *marginal polytope*,

$$\mathcal{M} := \{\mu(P) = \mathbb{E}_P[\phi(X)] \mid P \in \mathcal{P}\}. \tag{14.4.1}$$

Any distribution $P \in \mathcal{P}$ with mean vector $\mu = \mu(P)$ has associated generalized negative entropy

$$\Omega(\mu) := \sup_\theta \{-\mathbb{E}_P[\varphi(\theta, X)]\} = \sup_\theta \{\langle\theta, \mu(P)\rangle - A(\theta)\} = A^*(\mu),$$

the convex conjugate of A. The gradients ∇A and ∇A^* encode the duality between the natural-parameter space Θ and the mean-parameter space $\mathcal{M}$, and we elucidate a few of the main properties. Because $\nabla A(\theta) = \mathbb{E}_\theta[\phi(X)]$ in the exponential family, we immediately see that

$$\nabla A(\Theta) := \{\nabla A(\theta)\}_{\theta \in \Theta} \subset \mathcal{M}.$$

Recalling the duality relationship (14.1.3) that

$$\theta \in \partial A^*(\mu) \text{ if and only if } \nabla A(\theta) = \mu,$$

we can say much more.

Proposition 14.4.1. *Let $\mathcal{M}^o = \operatorname{relint} \mathcal{M}$. Then $\nabla A(\Theta) = \mathcal{M}^o$. Additionally:*

(i) If the family is minimal, then $\mathcal{M}$ has nonempty interior and Ω is continuously differentiable on $\mathcal{M}^o$, with $\theta = \nabla\Omega(\mu)$ if and only if $\nabla A(\theta) = \mu$.

(ii) If the family is non-minimal, then Ω is continuously differentiable relative to $\operatorname{aff}(\mathcal{M})$, meaning that there exists a continuous mapping $\nabla\Omega(\mu) \in \Theta$ such that for all $\mu \in \mathcal{M}^o$,

$$\partial\Omega(\mu) = \left\{\nabla\Omega(\mu) + \operatorname{span}(\mathcal{M} - \mathcal{M})^\perp\right\}.$$

Moreover, $\Theta = \Theta + \operatorname{span}(\mathcal{M} - \mathcal{M})^\perp$.

The proof of the proposition relies on the more sophisticated duality theory we develop in Appendices B and C, so we defer it to Section 14.5.3.

Proposition 14.4.1 has a direct statistical interpretation. Consider minimizing

$$\mathbb{E}_P[\varphi(\theta, X)] = \mathbb{E}_P[-\log p_\theta(X)] = A(\theta) - \langle \mu(P), \theta\rangle.$$

If $\mu(P) = \mathbb{E}_P[\phi(X)] \in \operatorname{relint} \mathcal{M}$, there exists a parameter $\theta(P)$, unique to translation in the subspace $\operatorname{span}(\mathcal{M} - \mathcal{M})^\perp$, for which

$$\theta(P) \in \operatorname*{argmin}_\theta \mathbb{E}_P[\varphi(\theta, X)] = \operatorname*{argmin}_\theta\{A(\theta) - \langle \mu(P), \theta\rangle\}.$$

Moreover, this parameter satisfies the mean matching condition

$$\nabla A(\theta(P)) = \mu(P),$$

the familiar moment-matching condition to minimize expected log loss.

We give a last result on the duality between the negative generalized entropy $\Omega(\mu)$ and log partition A, which captures a few nuances of minimization of the surrogate log loss $\varphi(\theta, x) = -\log p_\theta(x)$ when we encounter distributions P for which the mean mapping $\mu(P)$ is on the boundary of $\mathcal{M}$ or even outside it.

Proposition 14.4.2. *Let $\{P_\theta\}$ be a regular exponential family with log partition A with domain Θ, and let $\mathcal{M}$ be the associated mean parameter space with relative interior $\mathcal{M}^o = \operatorname{relint} \mathcal{M}$. Let $\Omega(\mu) = A^*(\mu)$ be the associated negative generalized entropy. Then*

(i) $A(\theta) = \Omega^*(\theta) = A^{**}(\theta)$ for all θ.

(ii) *If $\mu \in \mathcal{M}^o$, there exists $\theta(\mu) \in \Theta$ such that the negative entropy satisfies $\Omega(\mu) = A^*(\mu) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) < \infty$. If $\mu \notin \operatorname{cl} \mathcal{M}$, then $\Omega(\mu) = +\infty$.*

(iii) *If $\mu \in \operatorname{rel\,bd} \mathcal{M} = \operatorname{cl} \mathcal{M} \setminus \mathcal{M}^o$, then for any $\mu_0 \in \mathcal{M}^o$, $\Omega(\mu) = \lim_{t \downarrow 0} \Omega(t\mu_0 + (1-t)\mu)$, and there exist $\theta_t \in \Theta$ with*

$$\nabla A(\theta_t) = t\mu_0 + (1-t)\mu \quad \text{and} \quad \lim_{t \downarrow 0} \{A(\theta_t) - \langle \mu, \theta_t \rangle\} = \inf_\theta \{A(\theta) - \langle \mu, \theta \rangle\}.$$

In particular, there exist sequences of dual pairs (μ_n, θ_n) with $\mu_n \in \mathcal{M}^o$ and $\theta_n \in \Theta$ satisfying $\mu_n = \nabla A(\theta_n)$, $\mu_n \to \mu$, $\Omega(\mu_n) \to \Omega(\mu)$, and $A(\theta_n) - \langle \mu, \theta_n \rangle \to \inf_\theta \{A(\theta) - \langle \mu, \theta \rangle\}$.

See Section 14.5.4 for the deferred proof.

While the statement of Proposition 14.4.2 is complex, Gaussian estimation provides one avenue for understanding the result.

Example 14.4.3 (Gaussian estimation)**:** We fit a distribution assuming $X \sim \mathsf{N}(\mu, \Sigma)$, where μ and $\Sigma \succ 0$ are both to be estimated. Recalling Example 3.1.4, we have exponential family form with precision $K = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, so

$$p_{\theta,K}(x) = \exp\left(\langle \theta, x \rangle - \frac{1}{2}\langle xx^\top, K \rangle - A(\theta, K)\right) \quad \text{for} \quad A(\theta, K) = \frac{1}{2}\theta^\top K^{-1}\theta - \frac{1}{2}\log\det(K/2\pi).$$

The log partition function has gradients

$$\nabla_\theta A(\theta, K) = K^{-1}\theta \quad \text{and} \quad \nabla_K A(\theta, K) = -\frac{1}{2}K^{-1}\theta\theta^\top K^{-1} - \frac{1}{2}K^{-1}.$$

When the covariance $\operatorname{Cov}(X) \succ 0$, we match moments with $M = \mathbb{E}[XX^\top] \succ 0$ to obtain

$$\mathbb{E}[X] = K^{-1}\theta \quad \text{and} \quad M = K^{-1}\theta\theta^\top K^{-1} + K^{-1}.$$

Setting $\theta = K\mathbb{E}[X]$ and setting $K^{-1} = \operatorname{Cov}(X)$ solves this.

When $\operatorname{Cov}(X) \not\succ 0$, the solution $K = \operatorname{Cov}(X)^{-1}$ does not exist, so we rely on Proposition 14.4.2. Let $\Sigma = \operatorname{Cov}(X) \in \mathbb{R}^{d \times d}$ have rank $k < d$ and $\Pi \in \mathbb{R}^{d \times d}$ be the orthogonal projector onto $\operatorname{null}(\Sigma) = \{v \mid \Sigma v = 0\}$. For $t > 0$, define $K_t = (\Sigma + t\Pi)^{-1} = \Sigma^\dagger + \frac{1}{t}\Pi$, where $\Sigma^\dagger$ is the pseudo-inverse of Σ, and let $\theta_t = K_t\mathbb{E}[X]$. Then $\nabla_\theta A(\theta_t, K_t) = \mathbb{E}[X]$ for all $t > 0$, and $\nabla_K A(\theta_t, K_t) = -\frac{1}{2}\mathbb{E}[XX^\top] - \frac{t}{2}\Pi \prec 0$. In particular, $\nabla A(\theta_t, K_t) \in \mathcal{M}^o = \operatorname{int} \mathcal{M}$. We obtain the asymptotic moment matching

$$\lim_{t \to 0} \nabla A(\theta_t, K_t) = \left(\mathbb{E}[X], -\frac{1}{2}\mathbb{E}[XX^\top]\right) \in \operatorname{bd} \mathcal{M}$$

and limit

$$A(\theta_t, K_t) - \left\langle \begin{pmatrix} \mathbb{E}[X] \\ -\frac{1}{2}\mathbb{E}[XX^\top] \end{pmatrix}, \begin{pmatrix} \theta_t \\ K_t \end{pmatrix} \right\rangle = -\frac{1}{2}\log\det(K_t) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\langle \Sigma, K_t \rangle \to -\infty$$

as $t \downarrow 0$ because $\langle \Sigma, \Pi \rangle = 0$, verifying the conditions of part (iii) of Proposition 14.4.2. $\diamond$

In general, if P is a distribution such that $\mu(P) \in \mathcal{M}^o$, then there exists a parameter $\theta(P)$ minimizing $\mathbb{E}_P[\varphi(\theta, X)]$. If $\mu(P) \in \operatorname{rel\,bd} \mathcal{M}$, there is a sequence θ_n such that

$$\mathbb{E}_P[\varphi(\theta_n, X)] \to \inf_\theta \mathbb{E}_P[\varphi(\theta, X)] = -\Omega(\mu(P)), \quad \text{and} \quad \mu(P_{\theta_n}) \to \mu(P),$$

so that they asymptotically satisfy the mean identity. Finally, if $\mu(P) \notin \operatorname{cl} \mathcal{M}$, then $\inf_\theta \mathbb{E}_P[\varphi(\theta, X)] = -\infty$, making the choice of exponential family model poor, as it cannot capture the mean parameters.

14.4.1 Maximizing entropy

Our notion of generalized entropies as the minimal values of expected losses can recapture the classical entropy $H(P) = -\sum_x p(x) \log p(x)$ when P has a probability mass function p. For exponential family models, this connection goes much further, and the negative generalized entropy $\Omega(\mu)$ for $\mu \in \mathcal{M}$ coincides with the more general *Shannon entropy*:

Definition 14.5. *Let ν be a base measure on $\mathcal{X}$ and assume P has density p with respect to ν. Then the* Shannon entropy *of P (relative to ν) is*

$$H(P) = -\int p(x) \log p(x) d\nu(x).$$

For a distribution P with probability mass function p, ν is counting measure, yielding $H(P) = -\sum_x p(x) \log p(x)$, while for a distribution P with density p (for Lebesgue measure $d\nu(x) = dx$) we recover the differential entropy $H(P) = -\int p(x) \log p(x) dx$.

Example 14.4.4: Let P be the uniform distribution on $[0, a]$. Then the differential entropy $H(P) = -\log(1/a) = \log a$. $\diamond$

Example 14.4.5: Let P be the normal $\mathsf{N}(\mu, \Sigma)$ and ν be Lebesgue measure. Then

$$H(P) = \frac{1}{2} \log(\det(2\pi\Sigma)) + \frac{1}{2}\mathbb{E}[(X-\mu)^\top \Sigma^{-1}(X-\mu)] = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det(\Sigma)$$

because $p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$. $\diamond$

For exponential families, the log partition determines the Shannon entropy directly, highlighting that $-\Omega$ is indeed a familiar entropy-like object.

Proposition 14.4.6. *Let $\{P_\theta\}$ be a regular exponential family with respect to the base measure ν. Then for any $\theta \in \Theta$,*

$$H(P_\theta) = -\Omega(\mu(P_\theta)) = A(\theta) - \langle \mu(P_\theta), \theta \rangle,$$

where $\Omega(\mu) = \sup\{\langle \mu, \theta \rangle - A(\theta)\} = A^*(\mu)$.

Proof Using $\log p_\theta(x) = \langle \theta, \phi(x) \rangle - A(\theta)$, we obtain $H(P_\theta) = -\mathbb{E}_\theta[\langle \theta, \phi(X) \rangle - A(\theta)] = A(\theta) - \langle \mu(P_\theta), \theta \rangle$, where as usual $\mu(P) = \mathbb{E}_P[\phi(X)]$. As θ and $\mu(P_\theta)$ have the duality relationship $\nabla A(\theta) = \mu(P_\theta)$, we obtain $A(\theta) - \langle \mu(P_\theta), \theta \rangle = -\Omega(\mu(P_\theta))$ as desired. $\square$

The *maximum entropy principal*, which Jaynes [255] first elucidated in the 1950s, connects entropy in statistical mechanics and information theory. The principle follows: given constraints (prior information) about a distribution P, we consider all probability distributions satisfying said constraints. Then to encode our prior information while being as "agnostic" (i.e., uncertain) as possible, we should choose the distribution P satisfying the constraints to maximize the Shannon entropy. This principal naturally gives rise to exponential family models, and connects to Bayesian and minimax procedures (see Chapter 19). One caveat is that the base measure ν determines the distributions P we consider.

Suppose for a $\phi : \mathcal{X} \to \mathbb{R}^d$ and mean vector $\mu \in \mathbb{R}^d$, and we wish to solve

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \mu \tag{14.4.2}$$

over distributions $P \ll \nu$ having densities with respect to the base measure ν. Let

$$\mathcal{P}_\mu^{\text{lin}} := \{P \ll \nu \mid \mathbb{E}_P[\phi(X)] = \mu\}$$

be distributions with densities w.r.t. ν satisfying the expectation (linear) constraint $\mathbb{E}[\phi(X)] = \mu$. We then obtain the following theorem.

Theorem 14.4.7. *For $\theta \in \Theta$, let P_θ have density*

$$p_\theta(x) = \exp(\langle \theta, \phi(x)\rangle - A(\theta)), \quad A(\theta) = \log \int \exp(\langle \theta, \phi(x)\rangle) d\nu(x),$$

with respect to the measure ν. If $\mathbb{E}_{P_\theta}[\phi(X)] = \mu$, then P_θ uniquely maximizes $H(P)$ over those $P \in \mathcal{P}_\mu^{\text{lin}}$ for which the displayed entropy and KL decomposition are well-defined (though the parameter θ need not be unique).

Proof Consider any distribution $P \in \mathcal{P}_\mu^{\text{lin}}$, and θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)] = \mu$. We expand the entropy $H(P)$ as

$$\begin{aligned}
H(P) &= -\int p \log p d\nu = -\int p \log \frac{p}{p_\theta} d\nu - \int p \log p_\theta d\nu \\
&= -D_{\text{kl}}\left(P \| P_\theta\right) - \int p(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) \\
&\stackrel{(\star)}{=} -D_{\text{kl}}\left(P \| P_\theta\right) - \int p_\theta(x)[\langle \theta, \phi(x)\rangle - A(\theta)] d\nu(x) \\
&= -D_{\text{kl}}\left(P \| P_\theta\right) + H(P_\theta),
\end{aligned}$$

where step $(\star)$ uses $\int p(x)\phi(x)d\nu(x) = \int p_\theta(x)\phi(x)d\nu(x) = \mu$. As $D_{\text{kl}}\left(P \| P_\theta\right) > 0$ unless $P = P_\theta$, we have shown that P_θ uniquely maximizes the entropy. □

We give three examples of maximum entropy, showing how the choice of the base measure ν affects the maximum entropy distribution. For each, we assume that the space $\mathcal{X} = \mathbb{R}$ is the real line. We consider maximizing the entropy over all distributions P satisfying

$$\mathbb{E}_P[X^2] = 1.$$

Example 14.4.8: Let ν be counting measure on $\{-1, 1\}$, so that $\nu(\{-1\}) = \nu(\{1\}) = 1$. Then the maximum entropy distribution is $P(X = x) = \frac{1}{2}$ for $x \in \{-1, 1\}$. ◇

Example 14.4.9: Let ν be Lebesgue measure on $\mathcal{X} = \mathbb{R}$, so that $\nu([a, b]) = b - a$ for $b \geq a$. Then by Theorem 14.4.7, the maximum entropy distribution has the form $p_\theta(x) \propto \exp(-\theta x^2)$; recognizing the normal, we see that the optimal distribution is simply $\mathsf{N}(0, 1)$. ◇

Example 14.4.10: Let ν be counting measure on the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, \ldots\}$. Then Theorem 14.4.7 shows that the optimal distribution is a discrete version of the normal: we have $p_\theta(x) \propto \exp(-\theta x^2)$ for $x \in \mathbb{Z}$. ◇

Exercises 14.10 and 14.11 extend these results to inequality constraints in the family $\mathcal{P}^{\text{lin}}$.

Lastly, we consider the empirical variant of minimizing the log loss, where we maximize the likelihood of a given sample $X_1, \ldots, X_n$. Consider solving

$$\maximize_\theta \prod_{i=1}^n p_\theta(X_i) \equiv \minimize -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i), \tag{14.4.3}$$

for the exponential family model $p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta))$. We have the following result.

Proposition 14.4.11. *Let $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \phi(X_i)$. Then any θ solving $\mathbb{E}_{P_\theta}[\phi(X)] = \widehat{\mu}_n$ is a maximum likelihood solution, which exists if and only if $\widehat{\mu}_n \in \relint \mathcal{M}$. If the sample is drawn $X_i \stackrel{\text{iid}}{\sim} P$ where $P \ll \nu$ and $\mu(P) \in \relint \mathcal{M}$, then with probability 1, $\widehat{\mu}_n \in \relint \mathcal{M}$ eventually.*

Proof Define the empirical negative log likelihood

$$\widehat{L}_n(\theta) := -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) = -\langle \widehat{\mu}_n, \theta \rangle + A(\theta),$$

which is convex. Taking derivatives and using that $\Theta = \dom A$ is open, the parameter θ is a minimizer if and only if $\nabla \widehat{L}_n(\theta) = \widehat{\mu}_n - \nabla A(\theta) = 0$ if and only if $\nabla A(\theta) = \widehat{\mu}_n$. Apply Proposition 14.4.1.

For the final statement, note that $\widehat{\mu} \in \aff(\mathcal{M})$ with probability 1. Then because $\mu(P) \in \relint \mathcal{M}$ and $\widehat{\mu}_n \to \mu(P)$ with probability 1, we see that for any $\epsilon > 0$ there is some (random, but finite) N such that $n \ge N$ implies $\|\widehat{\mu}_n - \mu(P)\| \le \epsilon$ and $\widehat{\mu}_n \in \aff(\mathcal{M})$, so that $\widehat{\mu}_n \in \relint \mathcal{M}$. □

When $\widehat{\mu}_n \in \relint \mathcal{M}$, the fitted exponential-family distribution therefore has three equivalent characterizations: it minimizes the empirical logarithmic loss, maximizes likelihood, and maximizes entropy among distributions matching the empirical moment $\widehat{\mu}_n$.

14.5* Technical and deferred proofs

14.5.1 Proof of Theorem 14.2.15

One direction is straightforward. Let ℓ have the given representation. Then for any distribution P, we have $\mathbb{E}_P[r(\mu, Y)] \ge 0$, while $r(\mu(P), y) = 0$ for all y such that $P(Y = y) > 0$, so that $\mathbb{E}_P[r(\mu(P), Y)] = 0$. Then for $\mu(P) = \mathbb{E}_P[Y]$,

$$\begin{aligned}\mathbb{E}_P[\ell(\mu, Y)] &= -\Omega(\mu) - \langle \Omega'(\mu), \mu(P) - \mu \rangle + \mathbb{E}_P[a(Y)] + \mathbb{E}_P[r(\mu, Y)] \\ &\ge -\Omega(\mu(P)) + \mathbb{E}_P[a(Y)] + \mathbb{E}_P[r(\mu(P), Y)] = \mathbb{E}_P[\ell(\mu(P), Y)],\end{aligned}$$

and the inequality is strict if Ω is strictly convex and $\mu \neq \mu(P)$.

The converse direction (from a proper loss to function Ω) is more subtle. Let $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \mathbb{R}^k$, and assume w.l.o.g. that that the matrix

$$A = [y_1 \ \cdots \ y_m] \in \mathbb{R}^{k \times m}$$

with columns y_j has rank $k \le m$ (otherwise, we simply work in a subspace). Identify $\mathcal{P}$ with the probability simplex Δ_m, so the mean mapping $\mu(p) = \sum_{i=1}^m p_i y_i = Ap$ for $p \in \mathbb{R}^m$ is surjective. Finally, we write the shorthand $\vec{\ell}(\mu) = (\ell(\mu, y_1), \ldots, \ell(\mu, y_m)) \in \mathbb{R}^m$. Then for $p \in \Delta_m$, we define the entropy on the full probability distribution by

$$H(p) := \inf_{\mu \in \mathcal{M}} \langle p, \vec{\ell}(\mu) \rangle \stackrel{(\star)}{=} \langle p, \vec{\ell}(Ap) \rangle,$$

where equality $(\star)$ follows because ℓ is proper, and H finite, concave, and upper semicontinuous. Now we observe that for each $\mu \in \mathcal{M}$, H is affine on the set $A^{-1}(\mu) = \{p \in \Delta_m \mid Ap = \mu\}$. We claim this means we can write

$$H(p) = \langle v, p \rangle - \Omega(Ap) \tag{14.5.1}$$

for a closed convex function $\Omega : \mathcal{M} \to \mathbb{R}$ and vector $v \in \mathbb{R}^m$. This follows immediately from the next technical lemma, whose proof we defer to Section 14.5.2.

Lemma 14.5.1. *Let $f : C \to \mathbb{R}$ be closed convex, where $C \subset \mathbb{R}^n$ is a closed convex set, and $A \in \mathbb{R}^{m \times n}$ with $m \le n$ have rank m. Assume that for each $u \in \mathbb{R}^m$, f is affine on $A^{-1}(u) \cap C$. Then there is a closed convex function h and vector v such that $f(x) = \langle v, x \rangle + h(Ax)$.*

We now argue that we have representation $\ell(\mu, y) = -\Omega(\mu) - \langle \Omega'(\mu), y - \mu \rangle + a(y) + r(\mu, y)$ as in the statement (14.2.7) of the theorem. By definition of $H(p)$ in terms of the infimum, we have $H(p) \le \langle p, \vec{\ell}(\mu) \rangle$ for all p, and so

$$\langle v, p \rangle - \Omega(Ap) \le \langle p, \vec{\ell}(\mu) \rangle \;\; \text{i.e.} \;\; \Omega(Ap) \ge \langle p, v - \vec{\ell}(\mu) \rangle.$$

In particular, if for each μ we define the convex function by partial minimization (see Proposition B.2.23)

$$\nu \mapsto \psi_\mu(\nu) = \inf_p \left\{ \langle p, \vec{\ell}(\mu) - v \rangle \mid Ap = \nu, p \in \Delta_m \right\},$$

then $\Omega(\nu) \ge -\psi_\mu(\nu)$ for all $\mu, \nu \in \mathcal{M}$. When $\mu = \nu$, propriety gives equalities, and so we see that $\nu \mapsto \Omega(\nu) + \psi_\mu(\nu)$ is convex, nonnegative, and minimized at μ with $\Omega(\mu) + \psi_\mu(\mu) = 0$. We use this to demonstrate that Ω is subdifferentiable and find elements $\Omega'(\mu) \in \partial\Omega(\mu)$.

Because $\Omega(\cdot) + \psi_\mu(\cdot)$ is minimized by μ, we immediately see that $0 \in \partial(\Omega(\nu) + \psi_\mu(\nu))|_{\nu=\mu}$ by definition of a zero subgradient. Because $\operatorname{dom} \psi_\mu = \operatorname{dom} \Omega = \mathcal{M}$, the subdifferential sum rule (Proposition B.2.37) guarantees that

$$\partial \left(\Omega(\nu) + \psi_\mu(\nu)\right)|_{\nu=\mu} = \partial\Omega(\mu) + \partial\psi_\mu(\nu)|_{\nu=\mu},$$

and so the two sets on the right cannot be empty. In particular, there is some $s_\mu \in \partial\psi_\mu(\nu)|_{\nu=\mu}$ with $-s_\mu \in \partial\Omega(\mu)$.

Fix the choice $-\Omega'(\mu) = s_\mu$. Because the point mass e_i at y_i is feasible for $\psi_\mu(y_i)$,

$$\ell(\mu, y_i) - v_i \ge \psi_\mu(y_i) \ge \psi_\mu(\mu) + \langle s_\mu, y_i - \mu \rangle = -\Omega(\mu) + \langle s_\mu, y_i - \mu \rangle.$$

Now define the remainder

$$r(\mu, y_i) := \ell(\mu, y_i) - v_i + \Omega(\mu) + \langle s_\mu, \mu - y_i \rangle \ge 0.$$

Evidently, this choice yields

$$\ell(\mu, y_i) = -\Omega(\mu) + \langle s_\mu, y_i - \mu \rangle + v_i + r(\mu, y_i) = -\Omega(\mu) - \langle \Omega'(\mu), y_i - \mu \rangle + v_i + r(\mu, y_i),$$

which is the representation. It remains to show that $r(\mu, y) = 0$ for each $y \in \mathcal{S}(\mu)$. For this, if $p \in \Delta_m$ satisfies $Ap = \mu$, we have $\langle p, \vec{\ell}(\mu) \rangle = H(p)$ by propriety, and

$$\langle p, r(\mu, \cdot) \rangle = \langle p, \vec{\ell}(\mu) - v \rangle + \Omega(\mu) + \langle s_\mu, \mu - \mu \rangle = H(p) - \langle v, p \rangle + \Omega(\mu) = 0$$

by equality (14.5.1). In particular, we must have $r(\mu, y) = 0$ whenever $y \in \mathcal{S}(\mu)$.

Finally, when ℓ is strictly proper, we demonstrate that Ω is strictly convex. If $\mathbb{E}_p[Y] = \mu$, then we use the representation (14.2.7) to obtain

$$\begin{aligned}\mathbb{E}_p[\ell(\mu', Y)] - \mathbb{E}_p[\ell(\mu, Y)] &= -\Omega(\mu') - \langle \nabla\Omega(\mu'), \mu - \mu' \rangle + \Omega(\mu) + \mathbb{E}_p[r(\mu', Y)] - \mathbb{E}_p[r(\mu, Y)] \\ &\geq D_\Omega(\mu, \mu'),\end{aligned}$$

because $r(\mu', Y) \geq 0$ while the support condition for the remainder term guarantees $\mathbb{E}_p[r(\mu, Y)] = 0$. This is positive whenever $\mu \neq \mu'$ if and only if Ω is strictly convex.

14.5.2 Proof of Lemma 14.5.1

By working on the affine hull of C and relative interior as necessary, we can assume w.l.o.g. that C has interior, and so $\mathbb{R}^n = \mathrm{aff}(C)$. Letting $\mathrm{null}(A) = \{x \mid Ax = 0\}$ be the dimension $n - m$ null space of A, we have $\mathbb{R}^n = \mathrm{null}(A)^\perp \oplus \mathrm{null}(A)$, and we may work in the corresponding coordinates to write $x \in \mathrm{null}(A)^\perp$ and $y \in \mathrm{null}(A)$, so that y parameterizes points in $A^{-1}(x)$; by hypothesis we can thus write

$$f(x, y) = b(x) + \langle v(x), y \rangle$$

for some functions b and v. It is then enough to show that b is closed convex and v is constant.

Because C has interior, around any interior point $(x, y) \in \mathrm{int}\, C$ there is a product neighborhood, and we may construct $m+1$ affinely independent points $y_0, \ldots, y_m$ near y with $(x, y_i) \in \mathrm{int}\, C$. Using $f(x, y_i) = b(x) + v(x)^\top y_i$, we find that

$$v(x) = \begin{bmatrix} (y_1 - y_0)^\top \\ \vdots \\ (y_m - y_0)^\top \end{bmatrix}^{-1} \begin{bmatrix} f(x, y_1) - f(x, y_0) \\ \vdots \\ f(x, y_m) - f(x, y_0) \end{bmatrix}.$$

As f is locally Lipschitz (it is convex and finite), so then are v and b. Rademacher's theorem implies b and v are almost everywhere (Fréchet) differentiable, and so we may choose x so that $\nabla b(x)$ and $\mathsf{D}v(x)$ exist, and $\nabla f(x, y) = (\nabla b(x) + \mathsf{D}v(x)y, v(x))$. By convexity, for any u and y_i, y_j,

$$\begin{aligned}f(x + u, y_i) &\geq f(x, y_j) + (\nabla b(x) + \mathsf{D}v(x)y_j)^\top u + v(x)^\top (y_i - y_j) \\ &= f(x, y_i) + (\nabla b(x) + \mathsf{D}v(x)y_j)^\top u,\end{aligned}$$

while differentiability implies $f(x + u, y_i) = f(x, y_i) + (\nabla b(x) + \mathsf{D}v(x)y_i)^\top u + o(\|u\|)$ for u near 0. Combining these yields

$$f(x, y_i) + (\nabla b(x) + \mathsf{D}v(x)y_i)^\top u + o(\|u\|) \geq f(x, y_i) + (\nabla b(x) + \mathsf{D}v(x)y_j)^\top u,$$

and rearranging gives $(\mathsf{D}v(x)(y_i - y_j))^\top u + o(\|u\|) \geq 0$ as $u \to 0$. Divide by $\|u\|_2$ to conclude that for all $i, j \in \{0, \ldots, m\}$, we have $\|\mathsf{D}v(x)(y_i - y_j)\|_2 = 0$. As the points y_i are affinely independent, we obtain $\mathsf{D}v(x) = 0$ at all points for which it is differentiable; as v is locally Lipschitz on a connected convex set, it is necessarily constant on $\mathrm{int}\, C$, and we let $v_0 = v(x)$.

Thus, on the interior of C, we have $f(x, y) - v_0^\top y = b(x)$, and so b is continuous and convex over $\mathrm{int}\{x \mid (x, y) \in C\}$. We extend this to all of C. As $f(x, y)$ is affine in y, by writing $f(x, y) - f(x, y') = v(x)^\top (y - y')$ we immediately extend the linear term and observe that $f(x, y) = b(x) + v_0^\top y$ for all $(x, y) \in C$. Finally, $(x, y) \mapsto f(x, y) - v^\top y = b(x)$ remains closed on C.

14.5.3 Proof of Proposition 14.4.1

We first give the proof in the case that $\{P_\theta\}$ is a minimal exponential family, meaning that $\langle u, \phi(x)\rangle$ is non-constant in x for each $u \neq 0$, addressing the non-minimal case at the end. Then A is strictly convex (Proposition 3.2.3). As part of this proof, we will show that $\mathcal{M}^o$ is open in this case. We show both inclusions $\mathcal{M}^o \subset \nabla A(\Theta)$ and that $\nabla A(\Theta) \subset \mathcal{M}^o$.

Showing that $\nabla A(\Theta) \subset \mathcal{M}^o$. Because $\nabla^2 A(\theta) = \mathrm{Cov}_\theta(\phi(X)) \succ 0$ for each $\theta \in \Theta$, the inverse function theorem implies that ∇A is an open map. Thus $\nabla A(\Theta) \subset \operatorname{int} \mathcal{M}$.

Showing that $\mathcal{M}^o \subset \nabla A(\Theta)$. Let $\mu \in \mathcal{M}^o$, so that there exists an $\epsilon > 0$ such that $\mu + \epsilon\mathbb{B} \subset \mathcal{M}^o$. It is enough to show that $A(\theta) - \langle \mu, \theta\rangle$ is coercive in θ, as then there necessarily exists a (unique) minimizer $\theta(\mu)$ of $A(\theta) - \langle \mu, \theta\rangle$, and this minimizer satisfies $\nabla A(\theta(\mu)) = \mu$, so $\mu \in \nabla A(\Theta)$. We show that for any non-zero vector v the *recession function* of the tilted version $f(\theta) := A(\theta) - \langle \mu, \theta\rangle$ of A,

$$f'_\infty(v) := \lim_{t\to\infty} \frac{A(\theta + tv) - \langle \mu, \theta + tv\rangle - (A(\theta) - \langle \mu, \theta\rangle)}{t}$$

where $\theta \in \Theta$ is otherwise arbitrary, satisfies $f'_\infty(v) > 0$ for all $v \neq 0$, which guarantees that $A(\cdot) - \langle \mu, \cdot\rangle$ has a minimizer. (See Proposition C.3.3 and Corollary C.3.6 in Appendix C.2.1).

For vectors $v \in \mathbb{R}^d$, define the essential supremum of $\phi(x)$ in the direction v by

$$\nu^\star(v) := \operatorname*{ess\,sup}_x \langle \phi(x), v\rangle = \inf_t \left\{t \in \mathbb{R} \mid \nu(\{x \in \mathcal{X} \mid \langle v, \phi(x)\rangle \geq t\}) = 0\right\}.$$

As $\mu \in \mathcal{M}^o$, for any $v \neq 0$ we have $\langle v, \mu\rangle < \nu^\star(v)$. Let $\epsilon > 0$ satisfy $\langle v, \mu\rangle < \nu^\star(v) - \epsilon$ be otherwise arbitrary, fix $\theta \in \Theta$, and let $\mathcal{X}_\epsilon = \{x \mid \langle v, \phi(x)\rangle \geq \nu^\star(v) - \epsilon\}$, which satisfies $\nu(\mathcal{X}_\epsilon) > 0$. Then

$$\begin{aligned}
A(\theta + tv) - \langle \mu, \theta + tv\rangle &= \log \int \exp(\langle \phi(x), \theta + tv\rangle) d\nu(x) - \langle \mu, \theta + tv\rangle \\
&\geq \log \int_{\mathcal{X}_\epsilon} \exp(\langle \phi(x), \theta\rangle) e^{t(\nu^\star(v) - \epsilon)} d\nu(x) - \langle \mu, \theta\rangle - t\langle \mu, v\rangle \\
&= t(\nu^\star(v) - \epsilon) + \log \nu(\mathcal{X}_\epsilon) - t\langle \mu, v\rangle + \log \int_{\mathcal{X}_\epsilon} e^{\langle \phi(x), \theta\rangle} d\nu(x) - \langle \mu, \theta\rangle.
\end{aligned}$$

If $\nu(\mathcal{X}_\epsilon) = +\infty$, this shows $A(\theta + tv) = +\infty$ and so $A'_\infty(v) > 0$ certainly. If $\nu(\mathcal{X}_\epsilon) < \infty$, then note that $\nu^\star(v) - \epsilon - \langle \mu, v\rangle > 0$, and so

$$A(\theta + tv) - \langle \mu, \theta + tv\rangle \geq t(\nu^\star(v) - \epsilon - \langle \mu, v\rangle) - \log \nu(\mathcal{X}_\epsilon) + \log \int_{\mathcal{X}_\epsilon} e^{\langle \phi(x), \theta\rangle} d\nu(x) - \langle \mu, \theta\rangle$$

and thus

$$\frac{A(\theta + tv) - \langle \mu, \theta + tv\rangle - (A(\theta) - \langle \mu, v\rangle)}{t} \geq \nu^\star(v) - \epsilon - \langle \mu, v\rangle + o(1) > 0 \tag{14.5.2}$$

as $t \to \infty$.

Extending to the non-minimal case. If the exponential family is not minimal, we let

$$N = \{u \in \mathbb{R}^d \mid \langle u, \phi(x)\rangle \text{ is constant } \nu\text{-a.e.}\} = \operatorname{span}(\mathcal{M} - \mathcal{M})^\perp,$$

and choose matrices $U \in \mathbb{R}^{d\times k}$ and $U_\perp \in \mathbb{R}^{d\times(d-k)}$ whose columns are orthonormal bases for N and $N^\perp$, respectively. Then there is a vector $c \in \mathbb{R}^k$ such that $c \in \mathbb{R}^k$ such that $c = U^\top \phi(x)$ for

ν-almost all x, and we see that $A(\theta + Uv) = A(\theta) + \langle c, v\rangle$ and $U^\top \mu = c$ for all $\mu \in \mathcal{M}$. We show both inclusions as above. For any $\mu \in \mathcal{M}$, we have $\mathrm{aff}(\mathcal{M}) = \mu + \mathrm{span}(U_\perp)$.

Showing that $\nabla A(\Theta) \subset \mathcal{M}^o$. The reduced statistic $U_\perp^\top \phi(x)$ defines a minimal exponential family, and so for fixed θ_0, the function $w \mapsto A(\theta_0 + U_\perp w)$ is strictly convex in $w \in \mathbb{R}^{d-k}$, so the inverse function theorem shows that $w \mapsto \nabla A(\theta_0 + U_\perp w)$ is relatively open in $\mathcal{M}$.

Showing that $\mathcal{M}^o \subset \nabla A(\Theta)$. We again follow the logic of the minimal representation case. Let $\mu \in \mathcal{M}^o = \mathrm{relint}\, \mathcal{M}$, and recall $\nu^\star(U_\perp v) = \mathrm{ess\,sup}_x \langle \phi(x), U_\perp v\rangle$. Then there exists $\epsilon > 0$ such that $\mu + u \in \mathcal{M}$ for each $u \in \mathrm{span}(U_\perp)$ with $\|u\| \le \epsilon$, so that

$$\langle \mu, U_\perp v\rangle < \sup_{\|u\|_2 \le \epsilon, u \in \mathrm{span}(U_\perp)} \langle \mu + u, U_\perp v\rangle \le \nu^\star(\phi, U_\perp v).$$

Define $g(v) = A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$. Then because $A(\theta + Uw + U_\perp v) - \langle \mu, U_\perp v + Uw\rangle = A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$ for all $w \in \mathbb{R}^k, v \in \mathbb{R}^{d-k}$, it is enough to show that $g'_\infty(v) > 0$ for all $v \neq 0$. Following the same argument, *mutatis mutandis*, as that leading to inequality (14.5.2) yields that $g'_\infty(v) > 0$ for all $v \neq 0$. That is, $v \mapsto A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$ has a minimizer $v(\mu)$ (Corollary C.3.6), which is unique by the strict convexity of $v \mapsto A(\theta + U_\perp v)$, and which necessarily satisfies $U_\perp^\top \nabla A(\theta + U_\perp v(\mu)) = U_\perp^\top \mu$. As $U^\top \nabla A(\theta) = c$ for all θ and $U^\top \mu = c$ for all $\mu \in \mathcal{M}$, this shows that there exists $\theta(\mu)$ such that $\nabla A(\theta(\mu)) = \mu$ as desired. Moreover, fixing an arbitrary θ and letting $v(\mu)$ be the unique minimizer of $A(\theta + U_\perp v) - \langle \mu, U_\perp v\rangle$, the set of all minimizers

$$\Theta^\star(\mu) = \operatorname*{argmin}_\theta \{A(\theta) - \langle \mu, \theta\rangle\} = \left\{\theta + U_\perp v(\mu) + Uw \mid w \in \mathbb{R}^k\right\}.$$

This gives Proposition 14.4.1.

14.5.4 Proof of Proposition 14.4.2

For part (i), that A is closed follows from Fatou's Lemma (Proposition A.1.1), as if $\theta_n \to \theta$ then $\liminf_n \int e^{\langle \theta_n, \phi(x)\rangle} d\nu(x) \ge \int e^{\langle \theta, \phi(x)\rangle} d\nu(x)$. Then Theorem C.2.1 guarantees $A^{**} = A$.

For part (ii), note that if $\mu \in \mathcal{M}^o$, there exists $\theta(\mu) \in \Theta$ such that $\nabla A(\theta(\mu)) = \mu$ by Proposition 14.4.1. This $\theta(\mu)$ maximizes $\langle \theta, \mu\rangle - A(\theta)$ over all θ, and so $\Omega(\mu) = \langle \theta(\mu), \mu\rangle - A(\theta(\mu)) < \infty$. By Corollary C.2.5 in Appendix C.2.1 and Proposition 14.4.1, $\mathrm{dom}\, \partial\Omega = \mathcal{M}^o$, and as Ω is subdifferentiable on the relative interior of its domain, we have $\mathrm{dom}\, \Omega \subset \mathrm{cl}\, \mathcal{M}^o = \mathrm{cl}\, \mathcal{M}$. As Ω is closed convex, any point μ outside its domain necessarily satisfies $\Omega(\mu) = +\infty$.

Finally, for part (iii), we note that the function $g(t) = \Omega(t\mu_0 + (1-t)\mu)$ is a one-dimensional closed convex function. One-dimensional closed convex functions are continuous on their domains (Observation B.2.8 in Appendix B.2.2), and so g is necessarily continuous. Thus $\lim_{t \downarrow 0} g(t) = g(0)$. The existence of θ_t follows from Proposition 14.4.1.

14.6 Bibliography and further reading

The development of proper losses and scoring rules traces back to foundational axioms of statistics, where researchers wish to elicit individual probabilities (Bayesian or subjective probabilities) and to understand the value of information in different experimental settings. Some of the important early works here come from the literature on forecasting [88] and utility theory [200], and foundational work on information and losses by McCarthy [310] and Savage [371], who first demonstrated the representation in Theorem 14.2.1; a major motivation in this work was to develop measurements

that encouraged forecasters to be "honest" in their assessments of likelihood. Matheson and Winkler [305] introduced the continuous ranked probability score (CRPS), among others, to extend proper losses beyond discrete distributions. Gneiting and Raftery [214] give a wide-ranging generalization of these approaches, providing the simplified proof of Theorem 14.2.8 that we present. The statistics and economics literatures provide a broader theory of propriety beyond proper losses to *property elicitation*, which asks for those properties of probability distributions that can be discovered by optimizing a loss; see, for example, Osband's PhD thesis [336] and the papers [272, 213, 201, 190]. Waghmare and Ziegel [419] provide a recent survey on proper losses and forecasting, noting that there remain challenges for precisely how one ought to use different proper losses to assess predictive performance.

DeGroot [136, 137] developed the information in an experiment (Section 14.1.3) as a gap between uncertainty measures, although his work does not use our entropy terminology. The duality between losses, predictions, and convex loss functions here is new, though it has precursors. Special cases in exponential family models are essentially classical [311, 229], while Duchi et al. [162] develop related results on consistency for different convex surrogates of potentially non-convex losses. Blondel et al. [77], building on Reid and Williamson [359] and Williamson et al. [430], study *Fenchel-Young* losses, which define predictions via $\widehat{\mu} \in \operatorname{argmax}_{\mu \in \mathcal{M}} \{\langle s, \mu \rangle - \Omega(\mu)\}$ as in the conjugate link (14.3.2), and develop related consistency results.

Wainwright and Jordan [423] provide an excellent treatise that connects exponential family models, convex duality, entropy, and the algorithmic consequences of the results. (As we note in Chapter 3, Brown [90] gives many of their analytic properties as well.) The terminology and development in Section 14.4, especially mean mappings and their duality with the log-partition function, follow this framework. Our treatment of duality at the boundary of the mean-parameter space $\mathcal{M}$ relies on some of the more subtle aspects of convex analysis, as Rockafellar [363] and Hiriart-Urruty and Lemaréchal [235, 237] explore. (See also Appendices B and C.)

14.7 Exercises

Exercise 14.1 (Strict propriety of the log loss)**:** Let $\Delta_k = \{p \in \mathbb{R}^k_+ \mid \mathbf{1}^\top p = 1\}$ be the probability simplex. Show that if $\ell(q, y) = -\log q_y$ and $\mathbb{P}(Y = y) = p_y$, then

$$\operatorname*{argmin}_{q \in \Delta_k} \mathbb{E}[\ell(q, Y)] = p,$$

using the convention $0 \log 0 = 0$.

Exercise 14.2 (Uniqueness of generalized entropies)**:** For a concave function $f : \Delta_k \to \mathbb{R}$, define the perspective-type transform $f_{\rm per}(p) = \langle \mathbf{1}, p \rangle f(p / \langle \mathbf{1}, p \rangle)$, which maps $f_{\rm per} : \mathbb{R}^k_+ \to \underline{\mathbb{R}}$.

(a) Let $\ell : \Delta_k \times [k] \to \overline{\mathbb{R}}$ be strictly proper and Y have p.m.f. p. Show that $H(p) = \inf_{q \in \Delta_k} \mathbb{E}[\ell(q, Y)]$ is strictly concave, and that $H_{\rm per}$ is concave and continuously differentiable on $\mathbb{R}^k_{++}$.

(b) Show the converse that if $H : \Delta_k \to \mathbb{R}$ is strictly concave and its perspective $H_{\rm per}$ is differentiable on $\mathbb{R}^k_{++}$, then there exists a proper loss ℓ satisfying

$$H(p) = \inf_{q \in \Delta_k} \mathbb{E}_p[\ell(q, Y)]$$

and that $\ell(q, y) = [\nabla H_{\rm per}(q)]_y$ for all $q \in \operatorname{dom} \nabla H_{\rm per}$.

Exercise 14.3: Give the details in the computations for Example 14.3.4.

Exercise 14.4: Let $y \in \{0, 1\}$ and take negative entropy $\Omega(p) = -\log p - \log(1-p)$.

(a) Verify that the negative entropy Ω is of Legendre type (Definition 14.4).

(b) Give the associated loss ℓ and surrogate loss φ in the sense of Section 14.3.

(c) Plot the surrogate $\varphi(s, y) + \log 8$ and the logistic regression surrogate $\log(1 + e^s) - sy$ for $y \in \{0, 1\}$, each as function of s. (The shift by $\log 8$ guarantees the losses coincide at $s = 0$.)

(d) Give $\mathsf{pred}_\Omega(s)$ for $s \in \mathbb{R}$, verifying that $\mathsf{pred}_\Omega(s) \in [0, 1]$.

Exercise 14.5: For $\Omega(p) = -\log p - \log(1-p)$ as in Exercise 14.4, show that Ω is *self-concordant*, meaning that $|\Omega'''(p)| \le 2(\Omega''(p))^{3/2}$ for all $p \in (0, 1)$. (Such functions are important in optimization; the conjugate Ω^* is then also guaranteed to be self-concordant.)

Exercise 14.6 (Surrogates for regression)**:** Define $\Omega(\mu) = \frac{1}{4}\mu^4$.

(a) Give the conjugate $\Omega^*(s)$ to Ω.

(b) Show directly that the surrogate loss $\varphi(s, y) = \Omega^*(s) - sy$ satisfies that if $\hat{s} = \operatorname{argmin}_s \mathbb{E}[\varphi(s, Y)]$, then $\mathsf{pred}_\Omega(\hat{s}) = \mathbb{E}[Y]$.

Exercise 14.7: Let P be a predicted distribution and for $\alpha \in [0, \frac{1}{2}]$, define the lower and upper quantiles $l_\alpha = \mathsf{Quant}_\alpha(P)$ and $u_\alpha = \mathsf{Quant}_{1-\alpha}(P)$. Given these quantiles, for a finite set $\mathcal{A} \subset (0, \frac{1}{2})$, define the weighted interval loss

$$W(P, y) := \sum_{\alpha \in \mathcal{A}} \left[\alpha(u_\alpha - l_\alpha) + \operatorname{dist}(y, [l_\alpha, u_\alpha])\right],$$

which penalizes P using both the size $(u_\alpha - l_\alpha)$ of the quantile intervals and the distance of the outcome y from the predicted quantiles. Define the symmetrized set $\mathcal{A}_{\mathrm{s}} = \mathcal{A} \cup \{1 - \alpha \mid \alpha \in \mathcal{A}\}$. Show that

$$W(P, y) = \ell_{\mathrm{quant}, \mathcal{A}_{\mathrm{s}}}(P, y),$$

where ℓ_{quant} is the quantile loss (14.2.4).

Exercise 14.8: We explore a particularization of the results in Section 14.4. Let $Y \sim \mathsf{Poi}(e^\theta)$, so that Y has p.m.f. $p_\theta(y) = \exp(\theta y - e^\theta)/y!$ for $y \in \mathbb{N}$. Let $A(\theta) = e^\theta$ be the log-partition function. Define the "surrogate" loss $\varphi(\theta, y) = -\log p_\theta(y)$.

(a) Give the associated negative generalized entropy $\Omega(\mu)$ for $\mu \in (0, \infty)$.

(b) Give the associated loss $\ell(\mu, y)$ in the proper representation of Theorem 14.2.15. Directly verify that it is strictly proper, in that $\operatorname{argmin}_\mu \mathbb{E}[\ell(\mu, Y)] = \mathbb{E}[Y]$ for any Y supported on $\mathbb{R}_+$.

Exercise 14.9: We explore a particularization of Example 14.4.3. Let $X \sim \mathsf{N}(0, \Sigma)$ for a covariance $\Sigma \succ 0$, and let $K = \Sigma^{-1}$ be the associated precision matrix. Then X has density $p_K(x) = \exp(-\frac{1}{2}\langle xx^\top, K\rangle + \frac{1}{2}\log\det(K))$ with respect to (a scaled) Lebesgue measure, and log partition $A(K) = -\frac{1}{2}\log\det(K)$, which has domain the positive definite matrices $K \succ 0$ (and is $+\infty$ elsewhere).

(a) Give the associated negative generalized entropy $\Omega(M)$ for symmetric matrices M. Specify the domain of Ω.

(b) Give the associated loss $\ell(M, x)$ in the proper representation of Theorem 14.2.15. Directly verify that it is strictly proper, in that if the second moment matrix $C := \mathbb{E}[XX^\top]$ of X satisfies $C \succ 0$, then $\operatorname{argmin}_M \mathbb{E}[\ell(M, X)] = C$.

Exercise 14.10: This exercise generalizes Theorem 14.4.7 to apply to general (finite-dimensional) convex cone constraints. A set $\mathcal{C}$ is a *convex cone* if it is convex and closed under positive scaling: $x \in \mathcal{C}$ implies that $tx \in \mathcal{C}$ for all $t \geq 0$. Standard examples include

i. *The orthant.* Take $\mathcal{C} = \mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_j \geq 0, j = 1, \ldots, d\}$.

ii. *The semidefinite cone.* Take $\mathcal{C} = \{X \in \mathbb{R}^{d \times d} : X = X^\top, X \succeq 0\}$, where a matrix $X \succeq 0$ means that $a^\top X a \geq 0$ for all vectors a.

Given a convex cone $\mathcal{C}$, define the induced order by $x \succeq y$ if $x - y \in \mathcal{C}$. For an inner product $\langle \cdot, \cdot \rangle$, define the dual cone

$$\mathcal{C}^* := \{y : \langle y, x \rangle \geq 0 \text{ for all } x \in \mathcal{C}\}.$$

For the Euclidean inner product, the positive orthant and semidefinite cone are each self-dual. Consider the following maximum entropy problem, where the ordering $\preceq$ derives from a cone $\mathcal{C}$:

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \mu, \quad \mathbb{E}_P[\psi(X)] \preceq \beta, \tag{14.7.1}$$

where the base measure ν is implicit. Let $\mathcal{P}^{\text{lin}}_{\mu,\beta}$ be the collection of distributions $P \ll \nu$ satisfying the constraints in problem (14.7.1). Prove the following theorem:

Theorem 14.7.1. *For $\theta \in \mathbb{R}^d$ and $K \in \mathcal{C}^*$, the dual cone to $\mathcal{C}$, let $P_{\theta,K}$ have density*

$$p_{\theta,K}(x) = \exp\left(\langle \theta, \phi(x) \rangle - \langle K, \psi(x) \rangle - A(\theta, K)\right), \quad A(\theta, K) = \log \int \exp(\langle \theta, \phi(x) \rangle - \langle K, \psi(x) \rangle) d\nu(x),$$

with respect to the measure ν. If $\mathbb{E}_{P_{\theta,K}}[\phi(X)] = \mu$ and $\mathbb{E}_{P_{\theta,K}}[\psi(X)] = \beta$, then $P_{\theta,K}$ uniquely maximizes $H(P)$ over $\mathcal{P}^{\text{lin}}_{\mu,\beta}$.

Exercise 14.11 (An application of Theorem 14.7.1)**:** Let the cone $\mathcal{C}$ be the positive semidefinite cone in $\mathbb{R}^{d \times d}$, ν be the Lebesgue measure $d\nu(x) = dx$ and define $\psi(x) = \frac{1}{2} x x^\top \in \mathbb{R}^{d \times d}$. Let $\Sigma \succ 0$. Give the density solving

$$\text{maximize } -\int p(x) \log p(x) dx \text{ subject to } \mathbb{E}_P[XX^\top] \preceq \Sigma.$$

Exercise 14.12: Prove that the log determinant function is concave over the positive semidefinite matrices. That is, show that for $X, Y \in \mathbb{R}^{d \times d}$ satisfying $X \succeq 0$ and $Y \succeq 0$, we have

$$\log \det(\lambda X + (1 - \lambda) Y) \geq \lambda \log \det(X) + (1 - \lambda) \log \det(Y)$$

for any $\lambda \in [0, 1]$. *Hint.* Consider log-partition functions.

Exercise 14.13: Let $\rho_\alpha(t) = \alpha [t]_+ + (1 - \alpha) [-t]_+$. Show that the continuous ranked probability score (CRPS, Example 14.2.6) and quantile loss (Example 14.2.7) have the connection

$$\ell_{\text{crps}}(F, y) = 2 \int_0^1 \rho_t(y - F^{-1}(t)) dt$$

whenever the CDF F admits a density f with $\int |t| f(t) dt < \infty$.

Chapter 15

Calibration and proper losses

Chapter 14 develops *proper losses* for predicting distributions or, more generally, means of outcomes Y. We observe covariate/outcome pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, and we adopt the view of Section 14.2.3, where the target $Y \in \mathcal{Y} \subset \mathbb{R}^k$ is vector-valued, and we seek a predictor $f : \mathcal{X} \to \mathcal{M} = \mathrm{Conv}(\mathcal{Y})$. The risk of f under loss ℓ is then

$$L(f) := \mathbb{E}[\ell(f(X), Y)].$$

When ℓ is proper, $f(X) = \mathbb{E}[Y \mid X]$ minimizes $L(f)$, but exact recovery of this f is unrealistic.

We therefore consider a somewhat simpler goal that a predictor should be accurate conditional on its own predictions. For example, in a binary forecasting problem with outcomes $Y \in \{0, 1\}$, we ask that when the forecaster makes a prediction $p = f(X) \in [0, 1]$, the outcome Y should be 1 with frequency roughly p. We therefore define *calibration*, which is that

$$f(X) = \mathbb{E}[Y \mid f(X)]. \tag{15.0.1}$$

That is, given that the forecaster makes a prediction with value $\mu = f(X)$, we have

$$\mathbb{E}[Y \mid f(X) = \mu] = \mu.$$

While it is challenging to achieve this perfect calibration, this chapter investigates variants of the desideratum (15.0.1) that allow for statistical and information-theoretic approaches.

This chapter provides an overview of calibration, taking the particular focus that dovetails with our treatment of proper losses. We first relate calibration to proper losses and excess risk. We then develop several calibration measures that maintain statistical meaning, connecting these measures to improvement of a predictor's losses by post-processing. We show how to implement such post-processing ideas via empirical quantile averaging, which allows us to construct prediction sets and recalibrated predictors. The final sections introduce and compare several continuous calibration measures, establishing quantitative equivalences between them via convex duality.

15.1 Proper losses and calibration error

When a proper loss measures the error $\ell(f(x), y)$ in making the prediction $f(x)$ for the value y, we will see we can *always* improve the losses by modifying f to be a calibrated version of itsef. To make this apparent, assume f makes predictions in $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$, and that the loss has the representation

$$\ell(\mu, y) = -\Omega(\mu) - \langle \Omega'(\mu), y - \mu \rangle \tag{15.1.1}$$

where Ω is convex and $\Omega'(\mu) \in \partial\Omega(\mu)$ is a tacit (measurable) selection. (Theorems 14.2.1 and 14.2.15 make clear this is essentially no loss of generality, and we can omit the outcome normalization $a(y)$, as it will cancel in all comparisons below.) The associated Bregman divergence

$$D_\Omega(\mu, \nu) = \Omega(\mu) - \Omega(\nu) - \langle \Omega'(\nu), \mu - \nu \rangle,$$

is nonnegative for convex Ω (and strictly positive when Ω is strictly convex and $\mu \neq \nu$). Then using Corollary 14.2.5, for $f : \mathcal{X} \to \mathcal{M}$, if we condition on the prediction $M = f(X)$, then

$$\mathbb{E}[\ell(M, Y) \mid M] = \mathbb{E}[\ell(\mathbb{E}[Y \mid M], Y) \mid M] + \mathbb{E}[D_\Omega(\mathbb{E}[Y \mid M], M) \mid M],$$

where we use the linearity $\mathbb{E}[\ell(\mu, Y)] = \ell(\mu, \mathbb{E}[Y])$. We record this as a theorem.

Theorem 15.1.1. *Let ℓ be a proper loss with representation (15.1.1). Then for any $f : \mathcal{X} \to \mathcal{M}$,*

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}[D_\Omega(\mathbb{E}[Y \mid f(X)], f(X))].$$

In particular, the predictor

$$g(\mu) := \mathbb{E}[Y \mid f(X) = \mu]$$

is calibrated and satisfies

$$\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] \leq \mathbb{E}[\ell(f(X), Y)],$$

and the inequality is strict whenever f is not calibrated and ℓ is strictly proper.

Proof The first statement we have already proved. For the second, note that

$$g(\mu) = \mathbb{E}[Y \mid f(X) = \mu] \;\; \text{whence} \;\; \mathbb{E}[Y \mid g(f(X))] = \mathbb{E}[\mathbb{E}[Y \mid f(X)] \mid g(f(X))] = g(f(X))$$

by construction of g, so that $\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)]$. The inequality and its strictness are immediate because Ω is strictly convex if and only if ℓ is strictly proper. □

So calibrating a predictor f by post-processing can only improve its risk when ℓ is proper. We can give an alternative version of Theorem 15.1.1, where we instead consider the conjugate linkages in Section 14.3 and convex surrogate (14.3.3). Assume that Ω is a Legendre convex function (Definition 14.4), and recall the surrogate loss (14.3.3)

$$\varphi(s, y) = \Omega^*(s) - \langle s, y \rangle = \ell(\mathsf{pred}_\Omega(s), y),$$

where $\ell(\mu, y) = -\Omega(\mu) - \langle \nabla\Omega(\mu), y - \mu \rangle$ and

$$\mathsf{pred}_\Omega(s) = \operatorname*{argmin}_\mu \{-\langle \mu, s \rangle + \Omega(\mu)\} = \nabla\Omega^*(s).$$

Then the population surrogate loss decomposes into a prediction loss and calibration error:

Theorem 15.1.2. *Let Ω be Legendre and φ be the surrogate loss defined above. Then any $f : \mathcal{X} \to \mathbb{R}^k$ satisfies*

$$\mathbb{E}[\varphi(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_\Omega(\mathbb{E}[Y \mid f(X)], \mathsf{pred}_\Omega(f(X)))\right].$$

Proof Set $S = f(X)$. Conditional on S,

$$\mathbb{E}[\varphi(S, Y) \mid S] = \Omega^*(S) - \langle S, \mathbb{E}[Y \mid S]\rangle.$$

Let $M = \mathsf{pred}_\Omega(S) = \nabla\Omega^*(S)$. Then the Fenchel-Young equality (14.1.3) gives $\Omega^*(S) = \langle S, M\rangle - \Omega(M)$ and $S = \nabla\Omega(M)$, whence

$$\Omega^*(S) - \langle S, \mathbb{E}[Y \mid S]\rangle = -\Omega(\mathbb{E}[Y \mid S]) + D_\Omega(\mathbb{E}[Y \mid S], M).$$

Properness and the normalization (15.1.1) imply $\mathbb{E}[\ell(\mathbb{E}[Y \mid S], Y) \mid S] = -\Omega(\mathbb{E}[Y \mid S])$. Taking expectations proves the claim. □

As in Theorem 15.1.1, Theorem 15.1.2 shows that calibrating a predictor f can only improve the surrogate loss associated with Ω. Any predictor $f : \mathcal{X} \to \mathbb{R}^k$ has the "unnecessary" error

$$\mathbb{E}\left[D_\Omega\left(\mathbb{E}[Y \mid f(X)], \mathsf{pred}_\Omega(f(X))\right)\right]$$

arising from the divergence of predictions from calibration. We see that any proper (or derived proper) loss decomposes into a *sharpness* term relating to the typical error in predicting Y from $\mathbb{E}[Y \mid f(X)]$ and the calibration error. Replacing $f(X)$ with the expectation of Y given $f(X)$ (or a transformation thereof) does not increase this first term, but improves the calibration error. The squared error makes the decomposition clear:

Example 15.1.3 (Squared error and calibration)**:** When $\Omega(p) = \frac{1}{2}\|p\|_2^2$, we have $\Omega^* = \Omega$ and $\nabla\Omega = \nabla\Omega^*$ is the identity. Then Theorems 15.1.1 and 15.1.2 reduce to the statement that

$$\mathbb{E}[\|Y - f(X)\|_2^2] = \mathbb{E}\left[\|Y - \mathbb{E}[Y \mid f(X)]\|_2^2\right] + \mathbb{E}\left[\|\mathbb{E}[Y \mid f(X)] - f(X)\|_2^2\right],$$

the orthogonal decomposition associated with conditioning on $f(X)$. ◇

15.2 Measuring calibration

The first step to building a practicable theory of calibration is to define and measure the calibration of a predictor f. Defining a calibrated predictor is relatively easy, but measuring the calibration of a predictor f raises several challenges, as typical and naive measures of calibration are impossible to estimate. Thus, in this section, we develop several stable alternatives, providing a main theorem relating the different quantities to one another and demonstrate a simple estimator for one. We return in Section 15.5.2* to quantitative comparisons between the measures. Because these results require some measure-theoretic subtleties, we assume without comment that P is a Borel probability measure and that all functions are Borel measurable.

We begin with a natural candidate for calibration: the expected calibration error,

$$\mathsf{ece}(f) := \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|]. \tag{15.2.1}$$

The calibration error (15.2.1) is 0 if and only if f is perfectly calibrated. The next lemma, whose technical proof we defer to Section 15.6.1, guarantees that ece is lower semicontinuous.

Lemma 15.2.1. *The expected calibration error* ece *is lower semicontinuous with respect to* $L^1(P)$*: if* $f \in L^1(P)$ *and* $\mathbb{E}[\|f_n(X) - f(X)\|] \to 0$*, then*

$$\liminf_n \mathsf{ece}(f_n) \geq \mathsf{ece}(f).$$

The discontinuity of ece, however, arises even in very simple cases.

Example 15.2.2 (Discontinuity of the calibration error)**:** Let $Y \in \{0, 1\}$ be a Bernoulli random variable, and let $X \in \{0, 1\}$ be uniform. Take $Y = X$ with probability 1. Then the predictor that always predicts $\frac{1}{2}$ is perfectly calibrated, but if for $\epsilon \in [0, \frac{1}{2}]$ we define f_ϵ by

$$f_\epsilon(0) = \frac{1}{2} - \epsilon \quad \text{and} \quad f_\epsilon(1) = \frac{1}{2} + \epsilon$$

then we see that $\mathsf{ece}(f_\epsilon) = \frac{1}{2} - \epsilon$, while $\mathsf{ece}(f_0) = 0$. Certainly $f_\epsilon \to f_0$ in any L^p distance, while $\lim_{\epsilon \to 0} \mathsf{ece}(f_\epsilon) = \frac{1}{2}$. $\diamond$

15.2.1 The impossibility of measuring calibration

Using the tools on functional estimation and testing we develop in Chapter 13, we can show strong lower bounds for estimating the calibration error $\mathsf{ece}(f)$ unless one makes unjustifiable assumptions about the distribution of $Y \mid f(X)$. The intuition is straightforward: if $f(X)$ has a density, then even given a very large sample (X_1^n, Y_1^n), all the observations $f(X_i)$ will be distinct, and we have no *a priori* reason to assume that $\mathbb{E}[Y \mid f(X)]$ should be continuous in the predicted value $f(X)$.

To make this precise, fix a function f whose calibration error we wish to evaluate, and consider a hypothesis test of $H_0 : \mathsf{ece}(f) = 0$ against alternatives that f is miscalibrated, $H_1 : \mathsf{ece}(f) \geq \gamma$ for some $\gamma > 0$. We observe predictions $f(X_i)$ and outcomes Y_i, that is, pairs

$$Z_i = (f(X_i), Y_i)$$

drawn i.i.d. The lower bound holds if $\mathcal{X} = [0, 1]$ and $f(x) = x$, so observing X is of no help. Recall from Section 13.4 the the (worst-case) test risk

$$R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_1) := \sup_{P \in \mathcal{P}_0} P(\Psi(Z_1^n) \neq 0) + \sup_{P \in \mathcal{P}_1} P(\Psi(Z_1^n) \neq 1).$$

for testing between classes $H_0 : P \in \mathcal{P}_0$ and $H_1 : P \in \mathcal{P}_1$ of distributions. We define

$$\mathcal{P}_\gamma = \{\text{distributions } P \text{ on } (X, Y) \text{ such that } \mathsf{ece}_P(f) \geq \gamma\}$$

as the collection of distributions for which f is γ mis-calibrated.

Theorem 15.2.3. *Let $f : \mathcal{X} \to [0, 1]$ be a predictor of $Y \in \{0, 1\}$. Assume for some $0 < c < \frac{1}{2}$ that $f(\mathcal{X}) \cap [c, 1 - c]$ has cardinality at least N. Then there is a distribution P_0 such that $\mathsf{ece}_{P_0}(f) = 0$ and for any $0 < \gamma \leq c$,*

$$\inf_\Psi R_n(\Psi \mid \{P_0\}, \mathcal{P}_\gamma) \geq 1 - \frac{n\gamma^2}{2\sqrt{N}} \frac{1}{c(1 - c)}.$$

Before proving the theorem, we note the following corollary, which follows by taking $N \uparrow \infty$.

Corollary 15.2.4. *Let the conditions of Theorem 15.2.3 hold and $\mathcal{P}_0 = \{P \mid \mathsf{ece}_P(f) = 0\}$. If $f(\mathcal{X}) \cap [c, 1 - c]$ has infinite cardinality for some $c > 0$, then $\mathcal{P}_0$ is non-empty and for any $0 < \gamma \leq c$,*

$$\liminf_n \inf_\Psi R_n(\Psi \mid \mathcal{P}_0, \mathcal{P}_\gamma) = 1.$$

In brief, for all $\epsilon > 0$, no test is better than random guessing for testing between

$$H_0 : \mathsf{ece}(f) = 0 \;\text{ and }\; H_1 : \mathsf{ece}(f) \ge \frac{1}{2} - \epsilon$$

given access to the predictions $f(X_i)$ and observed outcomes Y_i.

Proof of Theorem 15.2.3 The proof relies on the convex hull testing lower bound from Proposition 13.4.1. Without loss of generality, we can assume that $\mathcal{X} \subset [0,1]$ and that $f(x) = x$ by transforming the input space. Let $M = f(X)$ be the (random) mean values that f outputs.

We construct the perfectly calibrated distribution P_0 and miscalibrated family $\mathcal{P}_\gamma$. Define P_0 so that M is uniform on distinct $\mu_1, \ldots, \mu_N \in [c, 1-c]$ and $Y \mid M = \mu \sim \mathsf{Bernoulli}(\mu)$. By construction, $\mathsf{ece}_{P_0}(f) = 0$. To construct the members of the alternative $\mathcal{P}_\gamma$, for each $j \in [N]$, define

$$\phi_j(y, \mu) := \left(\frac{y}{\mu_j} - \frac{1-y}{1-\mu_j}\right) \mathbf{1}\left\{\mu = \mu_j\right\}.$$

Then $\mathbb{E}_0[\phi_j(Y, M)] = 0$ while

$$\mathrm{Var}_0(\phi_j(Y, M)) = \frac{1}{N}\mathbb{E}_0\left[\left(\frac{Y}{\mu_j} - \frac{1-Y}{1-\mu_j}\right)^2 \mid M = \mu_j\right] = \frac{1}{N}\frac{1}{\mu_j(1-\mu_j)}.$$

Note that $|\phi_j(y, \mu)| \le \frac{1}{c}$ as $c < \frac{1}{2}$, and the vector $\phi(y, \mu) := (\phi_1(y, \mu), \ldots, \phi_N(y, \mu))$, has at most 1 non-zero entry. As $\gamma \in [0, c]$, for each $v \in \{-1, 1\}^N$ we may define the tilted distribution P_v with

$$P_v(Y = y, M = \mu) = \left(1 + \gamma\langle v, \phi(y, \mu)\rangle\right) P_0(Y = y, M = \mu),$$

which is a valid distribution whenever $\gamma \le c$, as $|\langle v, \phi(y, \mu)\rangle| \le \frac{1}{c}$. We compute the calibration error for distributions $P \in \{P_v\}$. Noting that M is still uniform on $\{\mu_1, \ldots, \mu_N\}$ under P_v, we have

$$\mathbb{E}_v[Y \mid M = \mu_j] = \mu_j + \gamma v_j \mathbb{E}[\phi_j(Y, \mu_j) Y \mid M = \mu_j] = \mu_j + \gamma v_j,$$

and so $\mathsf{ece}_{P_v}(f) = \frac{1}{N}\sum_{j=1}^N \gamma |v_j| = \gamma$. In particular, $P_v \in \mathcal{P}_\gamma$.

We compute a bound on the testing error using Lemma 13.2.3. Letting $\overline{P^n} = \frac{1}{2^N}\sum_v P_v^n$, we have

$$\begin{aligned} D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 &= \frac{1}{2^{2N}} \sum_{v, v'} \mathbb{E}_0\left[\left(1 + \gamma\langle v, \phi(Y, M)\rangle\right)\left(1 + \gamma\langle v', \phi(Y, M)\rangle\right)\right]^n \\ &= \frac{1}{2^{2N}} \sum_{v, v'} \left(1 + \gamma^2 v^\top \mathrm{Cov}_0(\phi(Y, M)) v'\right)^n \end{aligned}$$

because the sampling is i.i.d. By our variance calculation for ϕ and that each ϕ_j has disjoint support, $\mathrm{Cov}_0(\phi(Y, M)) = \frac{1}{N}\,\mathrm{diag}([\frac{1}{\mu_j(1-\mu_j)}]_{j=1}^N)$, and so

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 = \mathbb{E}\left[\left(1 + \frac{\gamma^2}{N}\sum_{j=1}^N \frac{V_j V_j'}{\mu_j(1-\mu_j)}\right)^n\right] \le \mathbb{E}\left[\exp\left(\frac{n\gamma^2}{N}\sum_{j=1}^N \frac{V_j V_j'}{\mu_j(1-\mu_j)}\right)\right],$$

where the expectation is over $V, V' \stackrel{\mathrm{iid}}{\sim} \mathsf{Uniform}(\{\pm 1\}^N)$. The i.i.d. random signs $V_j V_j'$ are 1-sub-Gaussian, so that

$$D_{\chi^2}\left(\overline{P^n} \| P_0^n\right) + 1 \le \exp\left(\frac{n^2\gamma^4}{2N^2}\sum_{j=1}^N \frac{1}{\mu_j^2(1-\mu_j)^2}\right) \le \exp\left(\frac{n^2\gamma^4}{2N}\frac{1}{c^2(1-c)^2}\right)$$

as $c \le \mu_j \le 1-c$. Apply Proposition 13.4.1 and Pinsker's inequality (Propositions 2.2.9 and 2.2.10) to see

$$\inf_{\Psi} R_n(\Psi \mid \{P_0\}, \mathcal{P}_\gamma) \ge 1 - \sqrt{\frac{1}{2}\log\left(1 + D_{\chi^2}\left(\overline{P^n} \| P_0^n\right)\right)} \ge 1 - \sqrt{\frac{n^2\gamma^4}{4N}\frac{1}{c^2(1-c)^2}}$$

as desired. □

15.2.2 Alternative calibration measures

The impossibility results in Theorem 15.2.3 and Corollary 15.2.4 motivate estimable calibration measures. We therefore introduce three such quantities here, giving a corollary to the formal results in Section 15.5 on their equivalence. We take the general setting of this chapter, where $\mathcal{Y} \subset \mathbb{R}^k$, P is a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{F}$ is a collection of (measurable) $f : \mathcal{X} \to \mathbb{R}^k$ with $\mathbb{E}[\|f(X)\|] < \infty$.

We require that a (candidate) calibration measure $\mathsf{ccm} : \mathcal{F} \to \mathbb{R}_+$ be *sound* (in analogy with proof systems, where soundness means nothing false can be proved), meaning that

$$\mathsf{ccm}(f) = 0 \;\text{ implies }\; \mathbb{E}[Y \mid f(X)] = f(X) \tag{15.2.2a}$$

and *complete* (continuing the analogy, that everything true can be proved), meaning that

$$\mathbb{E}[Y \mid f(X)] = f(X) \;\text{ implies }\; \mathsf{ccm}(f) = 0. \tag{15.2.2b}$$

We begin by considering *distance to calibration*. Let $\mathcal{C}(P)$ be the set of measurable, integrable calibrated predictors, $\mathcal{C}(P) = \{g : \mathcal{X} \to \mathbb{R}^k \mid \mathbb{E}_P[Y \mid g(X)] = g(X) \text{ a.s.}\}$. The set $\mathcal{C}(P)$ always contains the constant $g(X) = \mathbb{E}_P[Y]$ and so is non-empty. We call the minimum $L^1(P)$ distance of a function f to the set $\mathcal{C}(P)$ the *distance to calibration*

$$d_{\mathrm{cal}}(f) := \inf_g \left\{\mathbb{E}[\|g(X) - f(X)\|] \text{ s.t. } g \in \mathcal{C}(P)\right\}. \tag{15.2.3}$$

It is not always clear how to estimate the distance $d_{\mathrm{cal}}(f)$, which motivates a variational alternative.

Let $\mathcal{W} \subset \{w : \mathbb{R}^k \to \mathbb{R}^k\}$ be a symmetric collection of functions, meaning that $w \in \mathcal{W}$ implies $-w \in \mathcal{W}$. We can view any such collection as potential witnesses of miscalibration, in that

$$\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] = \mathbb{E}[\langle w(f(X)), \mathbb{E}[Y \mid f(X)] - f(X)\rangle]$$

and so if w can "witness" the portions of space where $f(X) \not\approx \mathbb{E}[Y \mid f(X)]$, it certifies miscalibration. We then arrive at what we term the *calibration error relative to the class* $\mathcal{W}$,

$$\mathrm{CE}(f, \mathcal{W}) := \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle]. \tag{15.2.4}$$

Depending on the class $\mathcal{W}$, this is sometimes called the *weak calibration error*, and with large enough classes, we can recover the classical expected calibration error (15.2.1).

Example 15.2.5 (Recovering expected calibration error)**:** For a norm $\|\cdot\|$, let $\mathcal{W}$ contain all measurable functions w with bound $\sup_\mu \|w(\mu)\|_* \le 1$. Then

$$\mathrm{CE}(f, \mathcal{W}) = \mathbb{E}\left[\sup_{\|w\|_* \le 1} \langle w, \mathbb{E}[Y \mid f(X)] - f(X)\rangle\right] = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|] = \mathsf{ece}(f),$$

the expected calibration error. ◇

Considering restricted classes is more interesting, because we can then often estimate the calibration error. The bounded Lipschitz functions provide one important collection, so we let

$$\mathcal{W}_{\|\cdot\|} := \left\{w : \mathbb{R}^k \to \mathbb{R}^k \mid \|w(\mu_0) - w(\mu_1)\|_* \le \|\mu_0 - \mu_1\| \text{ and } \|w(\cdot)\|_* \le 1 \text{ for all } \mu_0, \mu_1\right\} \quad (15.2.5)$$

denote the functions bounded by 1 in $\|\cdot\|_*$ that are 1-Lipschitz with respect to $\|\cdot\|$.

The final calibration measure quantizes the prediction space $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$. Given a finite (measurable) partition $\mathcal{A}$ of $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$, we consider the calibration error (15.2.4) for indicator functions $w(x) = \mathbf{1}\{f(x) \in A\}$, where we abuse notation slightly to define

$$\mathrm{CE}(f \mid \mathcal{A}) := \sum_{A \in \mathcal{A}} \|\mathbb{E}[(f(X) - Y)\mathbf{1}\{f(X) \in A\}]\| = \sum_{A \in \mathcal{A}} \|\mathbb{E}[f(X) - Y \mid f(X) \in A]\| \, \mathbb{P}(f(X) \in A).$$

While finer partitions (typically) approach the expected calibration error, they are difficult to estimate. We therefore trade calibration in the partition against the size of elements $A \in \mathcal{A}$, defining the *partitioned calibration error*

$$\mathsf{pce}(f) := \inf_{\mathcal{A}} \left\{\mathrm{CE}(f \mid \mathcal{A}) + \sum_{A \in \mathcal{A}} \mathrm{diam}(A)\mathbb{P}(f(X) \in A)\right\}. \quad (15.2.6)$$

Each of the three measures d_{cal}, $\mathrm{CE}(\cdot, \mathcal{W})$, and pce is equivalent to within polynomial scaling.

Corollary 15.2.6. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter and $\|\cdot\|$ be a norm. Then each of the calibration measures d_{cal}, $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$, and pce in definitions (15.2.3), (15.2.4), and (15.2.6) is sound and complete (15.2.2). Additionally, let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\|\cdot\| = \|\cdot\|_1$ be the ℓ_1-norm. Then for any $f : \mathcal{X} \to \mathcal{M} = \mathrm{Conv}(\mathcal{Y})$,*

$$\frac{1}{3}\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le d_{\mathrm{cal}}(f) \le 2 \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) + 2k\sqrt{2\,\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})}$$

and

$$d_{\mathrm{cal}}(f) \le \mathsf{pce}(f) \le d_{\mathrm{cal}}(f) + 2k\sqrt{d_{\mathrm{cal}}(f)}.$$

Corollary 15.2.6 will come as a consequence of the deeper development we pursue in Section 15.5 (see Theorem 15.5.6 and Corollary 15.5.10).

Corollary 15.2.6 motivates a basic estimation result: estimating $\mathrm{CE}(f, \mathcal{W})$ bounds the distance- and partition-based errors. In the binary case with $f : \mathcal{X} \to [0, 1]$ and $\mathcal{W} = \mathcal{W}_{\mathrm{Lip}}$ consisting of bounded Lipschitz functions $w : [0, 1] \to [-1, 1]$, the empirical variant of $\mathrm{CE}(f, \mathcal{W})$ becomes

$$\widehat{\mathrm{CE}}_n(f) := \sup_{v \in \mathbb{R}^n} \left\{\frac{1}{n}\sum_{i=1}^n v_i(Y_i - f(X_i)) \text{ s.t. } |v_i - v_j| \le |f(X_i) - f(X_j)| \text{ for } i, j \le n, \ \|v\|_\infty \le 1\right\}.$$

By combining uniform covering bounds for the class of Lipschitz functions with a standard concentration inequality, we then have the following convergence guarantee for $\widehat{\mathrm{CE}}_n$.

Proposition 15.2.7. *There exists a numerical constant C such that for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\left|\widehat{\mathrm{CE}}_n(f) - \mathrm{CE}(f, \mathcal{W}_{\mathrm{Lip}})\right| \le C\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

Proof Fix $\epsilon > 0$ and let $\mathcal{N}(\epsilon)$ be a minimal ϵ-cover of the set $\mathcal{W}_{\rm Lip}$ in uniform norm, meaning that $\left\|w - w^{(j)}\right\|_\infty \le \epsilon$ for each $w^{(j)} \in \mathcal{N}(\epsilon)$, and let $N(\epsilon)$ be its (minimal) cardinality. Then $\log N(\epsilon) \lesssim \frac{1}{\epsilon}$ (recall Proposition 12.2.3 and Eq. (12.2.4)). For shorthand, let the error vector $E \in [-1,1]^n$ have entries $E_i = Y_i - f(X_i)$, and abusing notation, for $w \in \mathcal{W}_{\rm Lip}$ let $\langle w, E\rangle_n = \frac{1}{n}\sum_{i=1}^n w(f(X_i))E_i$. Then for any $w \in \mathcal{W}_{\rm Lip}$, there exists $i \le N(\epsilon)$ such that

$$|\langle w, E\rangle_n - \langle w^{(i)}, E\rangle_n| \le \epsilon,$$

while $\widehat{\rm CE}_n(f) = \sup_{w\in\mathcal{W}_{\rm Lip}} \langle w, E\rangle_n$. In particular,

$$\left|\widehat{\rm CE}_n(f) - {\rm CE}(f, \mathcal{W}_{\rm Lip})\right| \le \sup_{w\in\mathcal{W}_{\rm Lip}} |\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]| \le \max_{w\in\mathcal{N}(\epsilon)} |\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]| + 2\epsilon.$$

Thus for any $t \ge 0$, the Azuma-Hoeffding inequality (Theorem 4.2.4) and a union bound yield

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{\rm CE}_n(f) - {\rm CE}(f, \mathcal{W}_{\rm Lip})\right| \ge t\right) &\le \mathbb{P}\left(\max_{w\in\mathcal{N}(\epsilon)} |\langle w, E\rangle_n - \mathbb{E}[\langle w, E\rangle_n]| \ge t - 2\epsilon\right) \\ &\le 2N(\epsilon)\exp\left(-\frac{n\left[t-2\epsilon\right]_+^2}{2}\right).\end{aligned}$$

Take $\epsilon = n^{-1/3}$ and $t = 2n^{-1/3} + C\sqrt{\frac{1}{n}\log\frac{1}{\delta} + n^{-2/3}}$ for an appropriate numerical constant C. □

Summarizing, while the expected calibration error is fundamentally inestimable, each of the measures (15.2.3), (15.2.4), and (15.2.6) are sound and complete. The witness measure (15.2.4) admits a direct estimator, and Corollary 15.2.6 transfers its control to the distance- and partition-based calibration measures (15.2.3) and (15.2.6). As the class size k grows, it can become statistically infeasible to estimate the calibration of predictors f, so one must consider alternative metrics; see the bibliographic section for more.

15.3 Auditing calibration with proper losses

Theorems 15.1.1 and 15.1.2 provide decompositions of the expected loss of a predictor

$$\mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_\Omega\left(\mathbb{E}[Y \mid f(X)], f(X)\right)\right]$$

into an average loss and nonnegative risk component. This decomposition allows leads to an audit: if a witness w certifies miscalibration, then post processing f can improve a proper loss; conversely, if no post-processing can improve f, then f is calibrated relative to the witnesses $\mathcal{W}$.

15.3.1 The post-processing gap and calibration audits for squared error

Consider a thought experiment: instead of using f to make predictions, we use a postprocessing $g\circ f$, where $g : \mathbb{R}^k \to \mathbb{R}^k$ has the (suggestively chosen) form $g(v) = v + w(v)$, where $w(v) = (g(v) - v)$. Then using the representation $\ell(\mu, y) = -\Omega(\mu) - \langle\nabla\Omega(\mu), y - \mu\rangle$ for the proper loss, we recall Theorem 15.1.1 and for $\mu(f(X)) := \mathbb{E}[Y \mid f(X)]$ expand

$$\begin{aligned}\mathbb{E}[\ell(g\circ f(X), Y)] &= \mathbb{E}[-\Omega(g\circ f(X)) - \langle\nabla\Omega(g\circ f(X)), Y - g\circ f(X)\rangle] \\ &= \mathbb{E}[-\Omega(\mu(f(X)))] + \mathbb{E}[\Omega(\mu(f(X))) - \langle\nabla\Omega(g\circ f(X)), Y - g\circ f(X)\rangle] \\ &= \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}[D_\Omega(\mathbb{E}[Y \mid f(X)], g\circ f(X))],\end{aligned}$$

that is, for $g(v) = v + w(v)$,

$$\mathbb{E}[\ell(g \circ f(X), Y)] = \mathbb{E}[\ell(\mathbb{E}[Y \mid f(X)], Y)] + \mathbb{E}\left[D_\Omega\left(\mathbb{E}[Y \mid f(X)], f(X) + w(f(X))\right)\right]. \tag{15.3.1}$$

This decomposes the expected loss $\mathbb{E}[\ell(g \circ f(X), Y)]$ into a term that post-processing does not change, which measures the sharpness with which $\mathbb{E}[Y \mid f(X)]$ predicts Y, and a divergence term D_Ω measuring the error in calibration of $g \circ f(X) = f(X) + w(f(X))$ for $\mathbb{E}[Y \mid f(X)]$.

The expansion (15.3.1) suggests that post-processing a prediction f can improve a proper loss if f is mis-calibrated. Leaving the distribution P on (X, Y) implicit, we thus define the *post-processing gap* for the (proper) loss ℓ and predictor f relative to a class $\mathcal{W}$ of functions by

$$\mathsf{gap}(\ell, f, \mathcal{W}) := \mathbb{E}[\ell(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\ell(f(X) + w(f(X)), Y)]. \tag{15.3.2}$$

To relate this gap to the calibration error, we first specialize to squared error, as the statements are most transparent. Let $\mathcal{W}$ be symmetric and convex, so that $w \in \mathcal{W}$ implies $-w \in \mathcal{W}$.

Proposition 15.3.1. *Let $\ell(\mu, y) = \frac{1}{2}\|y - \mu\|_2^2$ be the squared error (Brier score), and let $\mathcal{W}$ be a symmetric convex collection of functions with $R^2(f) = \sup_{w \in \mathcal{W}} \mathbb{E}[\|w(f(X))\|_2^2] > 0$. Then*

$$\frac{1}{2} \min\left\{\mathrm{CE}(f, \mathcal{W}), \frac{\mathrm{CE}(f, \mathcal{W})^2}{R^2(f)}\right\} \le \mathsf{gap}(\ell, f, \mathcal{W}) \le \mathrm{CE}(f, \mathcal{W})$$

The result immediately follows from the more general result Proposition 15.3.5 to come.

Taking $\mathcal{W}$ to consist of bounded functions gives a cleaner guarantee:

Corollary 15.3.2. *Let $w \in \mathcal{W}$ satisfy $\|w(\cdot)\|_2 \le 1$ and the conditions of Proposition 15.3.1 hold. Then*

$$\frac{1}{2} \min\left\{\mathrm{CE}(f, \mathcal{W}), \mathrm{CE}(f, \mathcal{W})^2\right\} \le \mathsf{gap}(\ell, f, \mathcal{W}) \le \mathrm{CE}(f, \mathcal{W}).$$

Thus, the calibration error upper and lower bounds the amount we could improve the expected squared loss of the function f by postprocessing f.

15.3.2 Calibration audits for losses based on conjugate linkages

Using the transformations between losses and their associated negative entropies in Section 14.3.1, we can generalize Proposition 15.3.1 to audit "nice enough" loss functions ℓ. Let Ω be 1-strongly convex with respect to a norm $\|\cdot\|$, and recall the surrogate (14.3.3) and conjugate link (14.3.2)

$$\varphi(s, y) := \Omega^*(\mu) - \langle s, y\rangle \quad \text{and} \quad \mathsf{pred}_\Omega(s) = \operatorname*{argmin}_\mu \left\{-\langle \mu, s\rangle + \Omega(\mu)\right\} = \nabla\Omega^*(s).$$

Then Proposition C.2.6 implies that Ω^* has smooth gradient, that is,

$$\|\mathsf{pred}_\Omega(s_0) - \mathsf{pred}_\Omega(s_1)\| \le \|s_0 - s_1\|_* .$$

In complete analogy with the post-processing gap (15.3.2) when we assume f makes predictions in $\mathrm{Conv}(\mathcal{Y})$, we will define a *surrogate post-processing gap*. Now, let $f : \mathcal{X} \to \mathbb{R}^k$ be a score function, and let $\mathcal{W}$ be a symmetric convex class of functions on $\mathcal{M}$. A witness w acts in score space through the linked prediction, giving the update

$$f_w(x) = f(x) + w(\mathsf{pred}_\Omega(f(x))).$$

We define the corresponding *surrogate post-processing gap* by

$$\mathsf{gap}(\varphi, f, \mathcal{W}) := \mathbb{E}[\varphi(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\varphi(f(X) + w(\mathsf{pred}_\Omega(f(X))), Y)], \tag{15.3.3}$$

The relevant calibration error is thus $\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W})$.

Example 15.3.3 (Multiclass logistic regression)**:** The negative entropy $\Omega(p) = \sum_{j=1}^k p_j \log p_j$ is strongly convex with respect to the ℓ_1 norm (this is Pinsker's inequality; see inequality (2.2.11)). Thus, the conjugate $\Omega^*(s) = \log(\sum_{j=1}^k e^{s_j})$ has Lipschitz gradient with respect to the ℓ_∞ norm, meaning that the prediction link

$$\mathsf{pred}_\Omega(s) = \left[\frac{e^{s_y}}{\sum_{j=1}^k e^{s_j}}\right]_{y=1}^k$$

satisfies $\|\mathsf{pred}_\Omega(s) - \mathsf{pred}_\Omega(s)\|_1 \le \|s - s'\|_\infty$ for all $s, s' \in \mathbb{R}^k$. $\diamond$

Example 15.3.4 (The squared error)**:** When the squared error in $\mathbb{R}^k$ corresponds to the generalized negative entropy $\Omega(\mu) = \frac{1}{2}\|\mu\|_2^2$. In this case, the norm $\|\cdot\| = \|\cdot\|_2 = \|\cdot\|_*$, and we have the self duality $\Omega^* = \Omega$, so that the prediction mapping pred_Ω is the identity. $\diamond$

With these examples as motivation, we can generalize Proposition 15.3.1.

Proposition 15.3.5. *Let the negative entropy Ω be strongly convex with respect to the norm $\|\cdot\|$ and the surrogate loss $\varphi(s, y) = \Omega^*(s) - \langle s, y\rangle$. Let $\mathcal{W}$ be a symmetric convex collection of functions with $R_*^2(f) := \sup_{w \in \mathcal{W}} \mathbb{E}[\|w(\mathsf{pred}_\Omega \circ f(X))\|_*^2] > 0$. Then*

$$\frac{1}{2}\min\left\{\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}), \frac{\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W})^2}{R_*^2(f)}\right\} \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}).$$

Proof Fix x and let $s = f(x)$ and $v = w(\mathsf{pred}_\Omega(f(x)))$, and notice that for any y we have

$$\varphi(s, y) + \langle \nabla\varphi(s, y), v\rangle \le \varphi(s + v, y) \le \varphi(s, y) + \langle \nabla\varphi(s, y), v\rangle + \frac{1}{2}\|v\|_*^2$$

by smoothness of $\nabla\Omega^*$. Recognizing that $\nabla\varphi(s, y) = \nabla\Omega^*(s) - y$, for any $w \in \mathcal{W}$, we define $w_\Omega = w \circ \mathsf{pred}_\Omega$ to obtain

$$\begin{aligned}
&-\mathbb{E}[\langle \nabla\varphi(f(X), Y), w_\Omega(f(X))\rangle] - \frac{1}{2}\mathbb{E}[\|w_\Omega(f(X))\|_*^2] \\
&\quad \le \mathbb{E}[\varphi(f(X), Y)] - \mathbb{E}[\varphi(f(X) + w_\Omega(f(X)), Y)] \le -\mathbb{E}[\langle \nabla\varphi(f(X), Y), w_\Omega(f(X))\rangle].
\end{aligned}$$

Taking suprema over w on each side and using the symmetry of $\mathcal{W}$ gives

$$\begin{aligned}
&\sup_{w \in \mathcal{W}}\left\{\mathbb{E}[\langle \nabla\Omega^*(f(X)) - Y, w_\Omega(f(X))\rangle] - \frac{1}{2}\mathbb{E}[\|w_\Omega(f(X))\|_*^2]\right\} \\
&\qquad \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \sup_{w \in \mathcal{W}} \mathbb{E}[\langle \nabla\Omega^*(f(X)) - Y, w_\Omega(f(X))\rangle] = \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}).
\end{aligned}$$

We can use the convexity of $\mathcal{W}$ and the definition $R_*^2(f) := \sup_{w \in \mathcal{W}} \mathbb{E}[\|w_\Omega(f(X))\|_*^2]$, to see that for any $\eta \in [0, 1]$, we may replace w with $\eta \cdot w \in \mathcal{W}$ and

$$\sup_{\eta \in [0,1]}\left[\eta \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}) - \frac{\eta^2}{2}R_*^2(f)\right] \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}).$$

Set $\eta = \min\{1, \frac{\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W})}{R_*^2(f)}\}$. $\square$

Specializing to the case of bounded witness functions again provides an operational interpretation: calibration error upper and lower bounds improvement possible through postprocessing.

Corollary 15.3.6. *Let the conditions of Proposition 15.3.5 hold, and additionally assume that the witness functions $\mathcal{W}$ satisfy $\|w(\mu)\|_* \le 1$ for all $\mu \in \mathbb{R}^k$. Then*

$$\frac{1}{2} \min\left\{\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}), \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W})^2\right\} \le \mathsf{gap}(\varphi, f, \mathcal{W}) \le \mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}).$$

Thus a small post-processing gap is equivalent, up to a square, to small calibration error relative to $\mathcal{W}$. The multiclass logistic loss provides a clean example.

Example 15.3.7 (Multiclass logistic losses, continued)**:** Let $\Omega(p) = \sum_{j=1}^k p_j \log p_j$ restricted to the probability simplex $\Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p\rangle = 1\}$ and $\varphi(s, y) = \log(\sum_{j=1}^k e^{s_j}) - s_y$. Then for any convex $\mathcal{W}$ whose elements satisfy $\|w(p)\|_\infty \le 1$ for $p \in \Delta_k$ and any $f : \mathcal{X} \to \mathbb{R}^k$,

$$\frac{1}{4}\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W})^2 \le \mathbb{E}[\varphi(f(X), Y)] - \inf_{w \in \mathcal{W}} \mathbb{E}[\varphi(f(X) + w(\nabla\Omega^*(f(X))), Y)]$$

because $\mathrm{CE}(\mathsf{pred}_\Omega \circ f, \mathcal{W}) \le 2$. $\diamond$

This perspective also suggests an iterative (population-level) algorithm to improve f: find a witness of miscalibration, update the scores in its direction, and repeat. Exercise 15.1 makes this construction explicit.

15.4 The power of working in quantile space

Postprocessing predictors f to improve them becomes a powerful approach on finite samples as well. The key idea is that empirical probabilities that $f(X) \in A$ or $Y - f(X) \in A$ for some region $A \subset \mathbb{R}$ must concentrate, meaning that an empirical quantile estimate is typically a very good proxy for a population quantile. We can leverage this both to guarantee predictive validity, constructing prediction sets with coverage of outcomes, and to improve calibration of a predictor, transforming any predictor f into a predictor g with small calibration error and improved loss. In each case, we assume we have a sample $(X_i, Y_i)_{i=1}^n$ of outcomes drawn i.i.d. and independent of f.

15.4.1 Conformal prediction

We first use the principle that empirical quantiles are good proxies for population quantiles to develop a guarantee slightly different from calibration. We focus on a regression setting, where $f : \mathcal{X} \to \mathbb{R}$ predicts outcomes $Y \in \mathbb{R}$, but now, instead of $\mathbb{E}[Y \mid f(X)]$, we seek a prediction set $C(X)$ containing Y with a prescribed probability. To give the idea, we consider interval-valued prediction sets, where for a threshold $\tau \ge 0$ we define

$$C_\tau(x) = \{y \in \mathbb{R} \mid |y - f(x)| \le \tau\} = [f(x) - \tau, f(x) + \tau]\,.$$

We seek τ so that on a new example (X_*, Y_*) drawn from the same distribution as $(X_i, Y_i)_{i=1}^n$,

$$\mathbb{P}\left(Y_* \in C_\tau(X_*)\right) \ge 1 - \alpha.$$

Intuitively, a new example is unlikely to have error much worse than the worst α-fraction of errors $Z_i = |Y_i - f(X_i)|$ on the sample, so we let

$$\widehat{\tau}_n = \mathsf{Quant}_{1-\alpha}\left(Z_1^n\right) = \inf\left\{t \mid P_n(Z \le t) \ge 1 - \alpha\right\} \tag{15.4.1}$$

be the empirical $1-\alpha$ quantile of the errors; we expect $C_{\widehat{\tau}_n}$ to have roughly $1-\alpha$ coverage.

The key is that quantiles on the empirical distribution P_n behave like quantiles:

Lemma 15.4.1. *Let $F(t) = P(Z \le t)$ be the CDF of a random variable Z, $0 < \beta < 1$, and let $\widehat{\tau}_n = \mathsf{Quant}_\beta(Z_1^n)$ be the empirical β-quantile of $Z_1^n \stackrel{\text{iid}}{\sim} P$. Let $0 \le \delta \le 1$ and define*

$$\gamma_n(\delta) = \sqrt{\left(\frac{4}{3n}\log\frac{1}{\delta}\right)^2 + \frac{2\beta(1-\beta)}{n}\log\frac{1}{\delta}} + \frac{4}{3n}\log\frac{1}{\delta} \le \sqrt{\frac{2\beta(1-\beta)}{n}\log\frac{1}{\delta}} + \frac{8}{3n}\log\frac{1}{\delta}.$$

Then with probability at least $1-\delta$ over the draw of Z_1^n,

$$F(\widehat{\tau}_n) \ge \beta - \gamma_n(\delta).$$

If additionally F is continuous, then $F(\widehat{\tau}_n) \le \beta + \gamma_n(\delta)$ with the same probability.

Proof Fix $0 \le \gamma \le \beta$ and let $\tau_\gamma^\star = \mathsf{Quant}_{\beta-\gamma}(Z)$ be the population $\beta - \gamma$ quantile of Z, so that $P(Z < \tau_\gamma^\star) \le \beta - \gamma$ by continuity from below of probabilities. Then $F(\widehat{\tau}_n) < \beta - \gamma$ implies that $\widehat{\tau}_n = \mathsf{Quant}_\beta(P_n)$ satisfies $\widehat{\tau}_n < \tau_\gamma^\star$, which occurs if and only if $P_n(Z < \tau_\gamma^\star) \ge \beta$. The Bernoulli $B = \mathbf{1}\left\{Z < \tau_\gamma^\star\right\}$ satisfies $P(B = 1) = \beta - t$ for some $0 \le t \le \gamma$ and has variance $\mathrm{Var}(B) = (\beta - t)(1 - \beta + t) \le \beta(1-\beta) + \gamma$, so Bernstein's inequality (4.1.10) shows

$$\begin{aligned}
\mathbb{P}\left(\widehat{\tau}_n < \tau_\gamma^\star\right) &= \mathbb{P}\left(P_n(Z < \tau_\gamma^\star) \ge \beta\right) \\
&= \mathbb{P}\left(P_n(Z < \tau_\gamma^\star) - P(Z < \tau_\gamma^\star) \ge \beta - P(Z < \tau_\gamma^\star)\right) \\
&\le \exp\left(-\frac{n\gamma^2}{2(1-\beta+\gamma)(\beta-\gamma) + \frac{2}{3}\gamma}\right) \le \exp\left(-\frac{n\gamma^2}{2\beta(1-\beta) + 8\gamma/3}\right).
\end{aligned}$$

Solving for $\delta = \exp(-\frac{n\gamma^2}{2(1-\beta)\beta + 8\gamma/3})$ gives $\gamma = \frac{4}{3n}\log\frac{1}{\delta} + \sqrt{(\frac{4}{3n}\log\frac{1}{\delta})^2 + \frac{2(1-\beta)\beta}{n}\log\frac{1}{\delta}}$. The calculation with the upper bound when F is continuous is completely analogous. □

Applying Lemma 15.4.1 to the empirical quantile (15.4.1), we see that the error $Z_* = |Y_* - f(X_*)|$ on a new example $(X_*, Y_*) \sim P$ is unlikely to exceed $\widehat{\tau}_n$:

Proposition 15.4.2. *With probability at least $1-\delta$ over P_n, the confidence set $C_{\widehat{\tau}_n}$ with empirical quantile (15.4.1) satisfies*

$$\mathbb{P}\big(Y_* \in C_{\widehat{\tau}_n}(X_*) \mid P_n\big) \ge 1 - \alpha - \sqrt{\frac{2\alpha(1-\alpha)}{n}\log\frac{1}{\delta}} - \frac{8}{3n}\log\frac{1}{\delta}.$$

Proof For $\tau \ge 0$, observe that $y \in C_\tau(x)$ if and only if $|y - f(x)| \le \tau$, and so if $Z_* = |Y_* - f(X_*)|$ we have $\mathbb{P}(Y_* \in C_\tau(X_*)) = P(Z \le \tau)$. Apply Lemma 15.4.1 with $\beta = 1 - \alpha$. □

In brief, with high probability—with no assumptions on the error distribution $Y_i - f(X_i)$ or the predictor f itself—the empirical prediction set set $C_{\widehat{\tau}_n}$ has nearly $1-\alpha$ coverage. The next section illustrates how such accurate quantile statements extend to yield improved losses and calibration.

15.4.2 Calibeating: improving squared error by calibration

Sections 15.1 and 15.3 show that at the population level, postprocessing a predictor f to guarantee its calibration improves losses. We give an example to show how working in quantile (probability) space allows such a result in finite samples. While we cannot be completely black box here—we assume Y concentrates and the range of f is bounded—we demonstrate *calibeating*: we construct g with both small calibration error and no larger squared error than f.

Thus, let $f : \mathcal{X} \to \mathbb{R}$ have bounded range and consider the squared error $\ell(\mu, y) = (\mu - y)^2$ with population loss $L(f) = \mathbb{E}[(Y - f(X))^2]$. We discretize f by binning its predictions to equalize the counts of $f(X_i)$ falling in each bin, then construct the post-processed g by averaging observed Y values over those bins. This transforms the optimal post-processing $g^\star(x) = \mathbb{E}[Y \mid f(X) = f(x)]$ in Theorem 15.1.1 into the finite sample analogue, which approximates

$$g(x) \approx \mathbb{E}[Y \mid l(x) \le f(X) \le u(x)],$$

where l and u are lower and upper bounds over which to average the predictions of f, which we choose to make $\mathbb{P}(l(X) \le f(X) \le u(X))$ approximately equal over partitions.

Assume we have a sample $(X_i, Y_i)_{i=1}^{2n}$ of size $2n$ drawn i.i.d. P (we choose $2n$ for notational convenience), which we divide into $(X_i, Y_i)_{i=1}^{n}$ and $(X_i, Y_i)_{i=n+1}^{2n}$, letting $P_n^{(1)}$ denote the empirical distribution on the first sample and $P_n^{(2)}$ that on the second. We use the first to choose the binning (quantization) of f into approximately equal probability regions and the second to choose values for the binned function g. Fix a number of bins $b \in \mathbb{N}$, for convenience assuming that b divides n, and let $r = 1/b$ be the associated (probability) resolution. Let $i_1, \ldots, i_n$ sort $f(X_i)$, so

$$f(X_{i_1}) < f(X_{i_2}) < \cdots < f(X_{i_n}),$$

and construct index partitions I_j, $j = 1, \ldots, b$, by $I_j := \{i_{n(j-1)/b+1}, \ldots, i_{nj/b}\}$. We assume that the predictions $f(X_i)$ are distinct with probability 1 (as otherwise we may randomly dither the values $f(x)$ with negligible noise). For $j = 1, \ldots, b$ define the bin boundaries

$$\widehat{u}_j = \max_{i \in I_j} f(X_i),$$

except that $\widehat{u}_b = \sup_x f(x)$ and $\widehat{u}_0 = \inf_x f(x)$. Partition $\operatorname{Im} f$ into equal empirical probability bins

$$B_1 = [\widehat{u}_0, \widehat{u}_1)\,, \quad B_2 = [\widehat{u}_1, \widehat{u}_2)\,, \quad \ldots, \quad B_b = [\widehat{u}_{b-1}, \widehat{u}_b]\,.$$

With the distinctness assumption, each bin contains exactly n/b points from the first sample.

To construct the recalibrated version g of f, define the bin mapping

$$\text{bin}(x) := \text{the bin } j \text{ such that } f(x) \in B_j,$$

which implicitly depends on $P_n^{(1)}$ and f, and the piecewise constant function

$$g(x) := \mathbb{E}_{P_n^{(2)}}[Y \mid \text{bin}(X) = \text{bin}(x)], \tag{15.4.2}$$

averaging those Y_i for which $f(X_i) \in B_j$ on the second sample (taking $g(x)$ arbitrary if $\text{bin}(x)$ is empty). Informally, g partitions X space into regions of probability $\approx r$ and for which $f(x)$ belongs to a given interval on each region. Then recalibrating f on that region should not worsen the prediction error $(Y - f(X))^2$ (much). Figure 15.1 illustrates the procedure.

To show that this procedure works, we need to demonstrate that the estimates (15.4.2) defining g are accurate; we therefore make the following assumption on the randomness in Y:

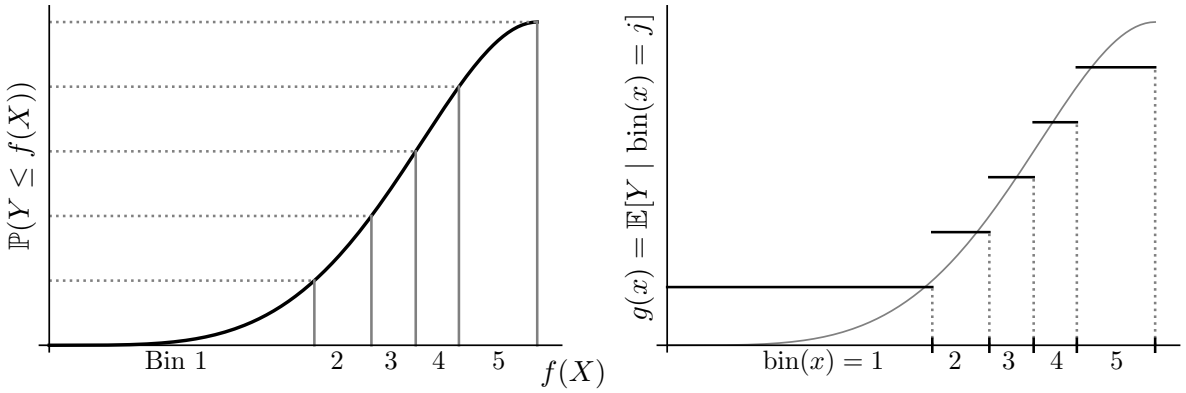


Figure 15.1. Illustration of the binning and estimation recalibration procedure (15.4.2). Left: bins $B_j = [l_j, u_j]$ equalize probabilities $\mathbb{P}(f(X) \in B_j)$. Right: recalibrated values $g(x) = \mathbb{E}[Y \mid \text{bin}(X)]$.

Assumption A.15.1. *For any (measurable) subset $A \subset \mathcal{X}$, the variables Y_i are independent and conditionally σ^2-sub-Gaussian, in that the set $I_A := \{i \in [2n] \mid X_i \in A\}$ satisfies*

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i \in I_A} (Y_i - \mathbb{E}[Y \mid X \in A])\right) \mid I_A\right] \le \exp\left(\frac{\lambda^2 \sigma^2 \, \text{card}(I_A)}{2}\right).$$

Hoeffding's inequality (Corollary 4.1.10) makes clear that the estimate (15.4.2) is accurate under the assumption when $P_n^{(2)}(\text{bin}(X) = j)$ is large enough for each j. As one example, Assumption A.15.1 holds when $Y_i \in [0, 1]$ with $\sigma^2 = \frac{1}{4}$ by Hoeffding's lemma (Example 4.1.6). Following the notation of Section 15.2.2, let $\text{CE}_2^2(f \mid \text{bin}) = \mathbb{E}[(\mathbb{E}[Y \mid \text{bin}(X)] - \mathbb{E}[f(X) \mid \text{bin}(X)])^2]$ be the squared calibration error of f conditional on the binning.

Theorem 15.4.3. *Let g be the binned and recalibrated estimator (15.4.2), let Assumption A.15.1 hold, and $D(f) = \sup_x f(x) - \inf_x f(x)$ be finite. Then there exists a numerical constant $C < \infty$ such that, for each $0 < \delta < 1$ and resolution $r = \frac{1}{b}$ satisfying $r \ge \frac{1}{n} \log \frac{n}{\delta}$, the following holds.*

(i) With probability at least $1 - \delta$, simultaneously for all $\eta > 0$,

$$L(g) \le (1 + \eta) \, L(f) + \left(1 + \frac{1}{\eta}\right) D^2(f) \cdot r + C \cdot \sigma^2 \cdot \frac{\log \frac{n}{\delta}}{nr} - \text{CE}_2^2(f \mid \text{bin})$$

and

$$\mathsf{ece}(g) \le \sqrt{\frac{C \sigma^2 \log \frac{n}{\delta}}{nr}}.$$

(ii) If additionally $\mathbb{E}[(Y - f(X))^2 \mid \text{bin}(X)] \le K^2$, then

$$L(g) \le L(f) + \frac{5}{4}(D^2(f) + K D(f)) \cdot r + C \cdot \sigma^2 \cdot \frac{\log \frac{n}{\delta}}{nr} - \text{CE}_2^2(f \mid \text{bin}).$$

Given our development, the proof of Theorem 15.4.3 is elementary: we argue (similar to Lemma 15.4.1) that the probability estimates for each bin B_j are accurate regardless of f, after which the concentration in Assumption A.15.1 guarantees the accuracy of the piecewise estimates (15.4.2); the rest is algebra. Exercises 15.5, 15.6, 15.7, and 15.8 outline the argument.

To interpret the theorem, it is instructive to specialize it to two concrete cases. First, assume that $Y \in \{0, 1\}$ and $f(x) \in [0, 1]$. Then in the second result of the theorem, we obtain $D(f) = \sup_x f(x) - \inf_x f(x) \le 1$ and $K \le 1$, so that

$$L(g) \le L(f) + O(1) \cdot \left[r + \sigma^2 \cdot \frac{\log \frac{n}{\delta}}{nr}\right] - \mathrm{CE}_2^2(f \mid \mathrm{bin})$$

Choosing resolution $r^2 = \frac{\sigma^2}{n} \log \frac{n}{\delta}$ (rounding as necessary), the calibeating predictor g guarantees

$$L(g) \le L(f) + O(1)\sqrt{\frac{\sigma^2 \log \frac{n}{\delta}}{n}} - \mathrm{CE}_2^2(f \mid \mathrm{bin}),$$

while the expected calibration error $\mathsf{ece}(g) = \widetilde{O}(n^{-1/4})$, ignoring the logarithmic factors. So this binning procedure improves the loss $L(f)$ by a factor involving the calibration error of f relative to the random binning up to a penalty tending to 0 at rate $\sqrt{\log n / n}$.

As another example, suppose that $Y = f^\star(X) + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$ and $f^\star$ has bounded range $[-\|f^\star\|_\infty, \|f^\star\|_\infty]$, as does f. Then Assumption A.15.1 holds with sub-Gaussian constant $\sigma^2 + \|f^\star\|_\infty^2$, and so the first part of Theorem 15.4.3 implies

$$L(g) \le (1 + \eta) L(f) + C \cdot \left[\frac{1 + \eta}{\eta} \|f^\star\|_\infty^2 \cdot r + \frac{(\sigma^2 + \|f^\star\|_\infty^2) \cdot \log \frac{n}{\delta}}{nr}\right] - \mathrm{CE}_2^2(f \mid \mathrm{bin}),$$

for all $\eta > 0$. Then because $L(f) = \sigma^2 + \mathbb{E}[(f(X) - f^\star(X))^2]$, we may take a universal scaling $\eta \propto \sqrt{r}$ and $r^{3/2} \propto \frac{\sigma^2}{n} \log \frac{n}{\delta}$ to obtain

$$L(g) \le L(f) + C_{f,f^\star} \cdot \left(\frac{\sigma^2 + \|f^\star\|_\infty^2}{n} \log \frac{n}{\delta}\right)^{1/3} - \mathrm{CE}_2^2(f \mid \mathrm{bin}),$$

where the constant $C_{f,f^\star}$ depends on $L(f)$ and $f^\star$. In either case, we see limited performance penalty from recalibration and an improvement relative to the miscalibration of f.

15.5 Continuous and equivalent calibration measures

We finally return to constructing a calculus and tools with which to measure calibration, addressing the discontinuity of ece in Example 15.2.2, and establishing results that imply Corollary 15.2.6. We will see that for appropriate classes $\mathcal{F}$ of predictors, several candidates $\mathsf{ccm} : \mathcal{F} \to \mathbb{R}_+$ are topologically equivalent sound and complete calibration measures. We begin with two definitions.

Definition 15.1. *A function* $\mathsf{ccm} : \mathcal{F} \to \mathbb{R}_+$ *is a* continuous calibration measure *for the distribution* P *on* $\mathcal{X} \times \mathcal{Y}$ *if*

(i) it is sound and complete (15.2.2)*, that is,* $\mathsf{ccm}(f) = 0$ *if and only if* f *is calibrated for* P*, and*

(ii) it is continuous with respect to the $L^1(P)$ *metric on* $\mathcal{F}$*, that is, for any* f*, if* f_n *is a sequence of functions with* $\mathbb{E}[\|f(X) - f_n(X)\|] \to 0$*, then*

$$\mathsf{ccm}(f) - \mathsf{ccm}(f_n) \to 0.$$

A stronger definition replaces continuity with a Lipschitz requirement.

Definition 15.2. *A function* $\mathsf{ccm} : \mathcal{F} \to \mathbb{R}_+$ *is a* Lipschitz calibration measure *for the distribution* P *on* $\mathcal{X} \times \mathcal{Y}$ *if it is sound and complete (Definition 15.1, part (i)), and instead of part (ii) satisfies*

(ii') it is Lipschitz continuous with respect to the $L^1(P)$ *metric on* $\mathcal{F}$*, that is, for some* $C < \infty$

$$|\mathsf{ccm}(f_0) - \mathsf{ccm}(f_1)| \le C \cdot \mathbb{E}_P[\|f_0(X) - f_1(X)\|]$$

for all $f_0, f_1 \in \mathcal{F}$.

If conditions (i) and (ii) (respectively (ii')) hold for all P in a collection of distributions $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, we will say that ccm is a continuous (respectively, Lipschitz) calibration measure for $\mathcal{P}$.

The desiderata (ii) and (ii') are matters of taste; the central idea is that some type of continuity is essential for efficient modeling, estimation, and analysis. We leave the norm $\|\cdot\|$ implicit in the definition and typically omit P from the calibration metric as it is clear from context. We consider two variational constructions, which are nearly dual to one another, that satisfy Definition 15.2. The first uses the distance to the nearest calibrated predictor, while the other certifies miscalibration within a witness class.

We will target a quantitative equivalence between calibration measures. Let $0 < \alpha, \beta < \infty$. Then we say that two candidate calibration measures ccm_0 and ccm_1 on $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}^k\}$ are (α, β)-*equivalent* if there exist c_0, c_1 such that

$$\mathsf{ccm}_0(f) \le c_0 \left[\mathsf{ccm}_1(f) + \mathsf{ccm}_1(f)^\alpha\right] \quad \text{and} \quad \mathsf{ccm}_1(f) \le c_1 \left[\mathsf{ccm}_0(f) + \mathsf{ccm}_0(f)^\beta\right]. \tag{15.5.1}$$

Then in a strong sense, $\mathsf{ccm}_0(f) \to 0$ if and only if $\mathsf{ccm}_1(f) \to 0$.

15.5.1 Calibration measures

We revisit the potential calibration measures in Section 15.2.2 to recapitulate definitions, providing initial results on the soundness and completeness of the distance to calibration (15.2.3) and relative calibration errors (15.2.4). Throughout this section, we assume all functions are integrable, and the partitioned calibration error (15.2.6) serves mainly as a proof device.

Distances to calibration. Recall the class $\mathcal{C}(P)$ of calibrated predictors and the distance to calibration (15.2.3). The distance (15.2.3) is the *largest* 1-Lipschitz measure of calibration: if ccm is any complete Lipschitz calibration measure with constant $C = 1$ in Definition 15.2.(ii'), then for any f and calibrated g we have $\mathsf{ccm}(f) = \mathsf{ccm}(f) - \mathsf{ccm}(g) \le \mathbb{E}[\|f(X) - g(X)\|]$. Taking an infimum over such g guarantees

$$\mathsf{ccm}(f) \le d_{\mathrm{cal}}(f).$$

To admit cleaner analytic properties, we define the *penalized calibration distance*,

$$\begin{aligned} p_{\mathrm{cal}}(f) &:= \inf_g \left\{\mathbb{E}[\|f(X) - g(X)\|] + \mathbb{E}\left[\|\mathbb{E}[Y \mid g(X)] - g(X)\|\right]\right\} \\ &= \inf_g \left\{\mathbb{E}[\|f(X) - g(X)\|] + \mathsf{ece}(g)\right\}. \end{aligned} \tag{15.5.2}$$

In the sequel (see Corollary 15.5.8), we show that

$$p_{\mathrm{cal}}(f) \le d_{\mathrm{cal}}(f) \le p_{\mathrm{cal}}(f) + C_{\mathcal{Y}} \sqrt{p_{\mathrm{cal}}(f)},$$

where $C_{\mathcal{Y}}$ is a constant depending only on the set $\mathcal{Y}$ whenever $\mathcal{Y}$ has finite diameter.

We also have the following proposition.

Proposition 15.5.1. *The functions d_{cal} and p_{cal} are Lipschitz calibration measures.*

Proof If f is calibrated, then $p_{\mathrm{cal}}(f) = d_{\mathrm{cal}}(f) = 0$. Conversely, if $p_{\mathrm{cal}}(f) = 0$, there exists a sequence g_n satisfying $\mathbb{E}[\|f(X) - g_n(X)\|] \to 0$, as each term in the definition (15.5.2) is nonnegative, and $\mathsf{ece}(g_n) = \mathbb{E}[\|\mathbb{E}[Y \mid g_n(X)] - g_n(X)\|] \to 0$. Lemma 15.2.1 gives $0 = \liminf_n \mathsf{ece}(g_n) \geq \mathsf{ece}(f)$. If $d_{\mathrm{cal}}(f) = 0$, there exists a sequence of functions g_n with $\mathsf{ece}(g_n) = 0$ and $\mathbb{E}[\|f(X) - g_n(X)\|] \to 0$. The lower semicontinuity of ece from Lemma 15.2.1 gives $0 = \liminf_n \mathsf{ece}(g_n) \geq \mathsf{ece}(f)$.

To see that p_{cal} is Lipschitz, let $f_0, f_1 : \mathcal{X} \to \mathbb{R}^k$, and let g_1 be within $\epsilon > 0$ of achieving the infima in definition (15.5.2) for f_1. Then

$$\begin{aligned}
p_{\mathrm{cal}}(f_0) - p_{\mathrm{cal}}(f_1) &\leq \inf_g \left\{\mathbb{E}[\|f_0(X) - g(X)\|] + \mathbb{E}[\|\mathbb{E}[Y \mid g(X)] - g(X)\|]\right\} \\
&\qquad - \mathbb{E}[\|f_1(X) - g_1(X)\|] - \mathbb{E}[\|\mathbb{E}[Y \mid g_1(X)] - g_1(X)\|] + \epsilon \\
&\leq \mathbb{E}[\|f_0(X) - g_1(X)\|] - \mathbb{E}[\|f_1(X) - g_1(X)\|] + \epsilon \\
&\leq \mathbb{E}[\|f_0(X) - f_1(X)\|] + \epsilon.
\end{aligned}$$

Take $\epsilon \downarrow 0$. The lower inequality is similar, as is the proof for d_{cal}. □

Weak calibration. The relative calibration error (15.2.4) also satisfies our desiderata. Taking $\mathcal{W}$ to be the class $\mathcal{W}_{\|\cdot\|}$ of bounded Lipschitz witnesses (15.2.5), we have the next two propositions.

Proposition 15.5.2. *Let $\mathcal{F}$ consist of functions with $\mathbb{E}[\|f(X)\|] < \infty$ and assume $\mathbb{E}[\|Y\|] < \infty$. Then $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ is a continuous calibration measure over $\mathcal{F}$.*

Because continuity is such a weak requirement, the proof of this result relies on measure-theoretic approximation, so we defer it to Section 15.6.2.

When $\mathcal{F}$ has finite diameter and $\mathcal{Y}$ is bounded, we obtain a stronger Lipschitz guarantee:

Proposition 15.5.3. *Assume that $\mathrm{diam}(\mathcal{Y})$ is finite and that $\mathcal{F}$ is a uniformly bounded collection of functions $f : \mathcal{X} \to \mathbb{R}^k$. Then $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ is a Lipschitz calibration measure over $\mathcal{F}$.*

Proof Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$ for shorthand. That $\mathrm{CE}(f, \mathcal{W}) = 0$ when f is calibrated is immediate, as by definition of conditional expectation we have

$$\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] = \mathbb{E}[\langle w(f(X)), \mathbb{E}[Y \mid f(X)] - f(X)\rangle] = 0.$$

To obtain the converse that $\mathrm{CE}(f, \mathcal{W}) = 0$ implies f is calibrated, we require an intermediate lemma, which leverages the density of Lipschitz functions in L^p spaces. As was the case for the lower semicontinuity lemma 15.2.1 central to the proof of the converse in Proposition 15.5.1, this lemma requires measure-theoretic approximation arguments, so we defer its proof to Section 15.6.3.

Lemma 15.5.4. *Let $Z \in \mathbb{R}^k$ be a random variable and $\mathbb{E}[\|g(Z)\|] < \infty$. If $\mathbb{E}[\langle w(Z), g(Z)\rangle] = 0$ for all bounded and 1-Lipschitz functions w, then $g(Z) = 0$ with probability 1.*

The converse follows: let $M = f(X)$, note that $\mathrm{CE}(f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(M), \mathbb{E}[Y \mid M] - M\rangle]$, and take $g(M) = \mathbb{E}[Y \mid M] - M$ in Lemma 15.5.4.

To see that CE is Lipschitz, let $w_0 \in \mathcal{W}$ be such that $\mathrm{CE}(f_0, \mathcal{W}) \ge \mathbb{E}[\langle w_0(f_0(X)), Y - f_0(X)\rangle] - \epsilon$, and set $C = \sup_{y\in\mathcal{Y}, x\in\mathcal{X}, f\in\mathcal{F}} \|y - f(x)\| < \infty$. Then

$$\begin{aligned}
&\mathrm{CE}(f_0, \mathcal{W}) - \mathrm{CE}(f_1, \mathcal{W}) \\
&\le \mathbb{E}[\langle w_0(f_0(X)), Y - f_0(X)\rangle] - \mathbb{E}[\langle w_0(f_1(X)), Y - f_1(X)\rangle] + \epsilon \\
&= \mathbb{E}[\langle w_0(f_0(X)) - w_0(f_1(X)), Y - f_0(X)\rangle] + \mathbb{E}[\langle w_0(f_1(X)), f_1(X) - f_0(X)\rangle] + \epsilon \\
&\le C\mathbb{E}[\|w_0(f_0(X)) - w_0(f_1(X))\|_*] + \mathbb{E}[\|f_1(X) - f_0(X)\|] + \epsilon \\
&\le (1 + C)\mathbb{E}[\|f_1(X) - f_0(X)\|] + \epsilon.
\end{aligned}$$

Repeating the same argument, *mutatis mutandis*, for the lower bound gives Lipschitz continuity. □

Exercises 15.3 and 15.4 explore how the weak calibration measures $\mathrm{CE}(f, \mathcal{W})$ vary as we modify the collection of potential witness functions $\mathcal{W}$ in the context of multiclass calibration problems. Naive binning strategies, such as those we use to calibeat in Section 15.4.2, cannot work in these cases, because the number of bins grows exponentially in the number of classes k, making the choice of $\mathcal{W}$ important for practicality.

15.5.2* Equivalent calibration measures

All three measures $d_{\mathrm{cal}}(f)$, $p_{\mathrm{cal}}(f)$, $\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})$ are Lipschitz calibration measures, and moreover, they are equivalent (15.5.1) for multiclass classification.

Theorem 15.5.5. *Let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\mathcal{W}_{\|\cdot\|}$ be the collection (15.2.5) of bounded Lipschitz functions for a norm $\|\cdot\|$ on $\mathbb{R}^k$. Then $d_{\mathrm{cal}}, p_{\mathrm{cal}}$, and $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ are pairwise $(\frac{1}{2}, \frac{1}{2})$-equivalent.*

The exponent is sharp: even in the binary case, the calibration measures are not (α, β)-equivalent for $\alpha, \beta > \frac{1}{2}$ (see Exercise 15.9).

We prove the theorem through two comparisons. The first, Theorem 15.5.6, shows an equivalence between the penalized calibration distance (15.5.2) and partitioned calibration error (15.2.6); this result holds for general finite-dimensional outcome sets $\mathcal{Y}$. The second, Theorem 15.5.9, applies for multiclass classification when $\mathcal{Y} = \{e_1, \ldots, e_k\}$, and gives an equivalence between the penalized distance (15.5.2) and the calibration error relative to Lipschitz functions (15.2.4). Throughout we restrict predictions to $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$.

Partition-based calibration measures and lifting to random variables

We work directly in the space of predictions $f(X) \in \mathcal{M}$ rather than addressing the underlying space $\mathcal{X}$. Let $M = f(X)$ be the random vector (use the mnemonic that M is for "mean prediction") induced by $f(X)$ and taking values in $\mathrm{Conv}(\mathcal{Y})$, which has a joint distribution (M, Y) with the label Y. When a calibration measure ccm depends only on the joint law of $(Y, f(X))$, we write $\mathsf{ccm}(M) = \mathsf{ccm}(f)$. For example, the expected calibration error of f is

$$\mathsf{ece}(f) = \mathsf{ece}(M) = \mathbb{E}[\|\mathbb{E}[Y \mid M] - M\|].$$

Once we work exclusively in with predictions $M = f(X)$, we may define alternative distances to calibration in analogy with the (penalized) distances to calibration, which will allow us to more easily relate distances to the partitioned error (15.2.6). Thus, we define

$$d_{\mathrm{cal,low}}(f) := \inf_V \{\mathbb{E}[\|M - V\|] \;\text{ s.t. }\; \mathbb{E}[Y \mid V] = V\} \tag{15.5.3a}$$

and

$$p_{\rm cal,low}(f) := \inf_V \left\{\mathbb{E}[\|M - V\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|]\right\}, \tag{15.5.3b}$$

where both infima range over couplings (M, Y, V) that preserve the joint law of (M, Y) and satisfy $V \in \mathcal{M}$; in (15.5.3a), we require V to be calibrated. We can represent this formulation as a convex optimization problem, allowing us to bring the tools of duality to bear on it, though we defer this temporarily. By restricting $V = g(M)$, we immediately see that $p_{\rm cal}(f) \ge p_{\rm cal,low}(f)$. We can also consider upper distances

$$d_{\rm cal,up}(f) := \inf_g \left\{\mathbb{E}[\|M - g(M)\|] \;\; \text{s.t.} \;\; \mathbb{E}[Y \mid g(M)] = g(M)\right\}$$

and

$$p_{\rm cal,up}(f) := \inf_{g:\mathbb{R}^k \to {\rm Conv}(\mathcal{Y})} \left\{\mathbb{E}[\|M - g(M)\|] + \mathbb{E}[\|\mathbb{E}[Y \mid g(M)] - g(M)\|]\right\},$$

which restrict the definitions (15.2.3) and (15.5.2) to compositions. We therefore have

$$d_{\rm cal,low}(f) \le d_{\rm cal}(f) \le d_{\rm cal,up}(f) \;\; \text{and} \;\; p_{\rm cal,low}(f) \le p_{\rm cal}(f) \le p_{\rm cal,up}(f). \tag{15.5.4}$$

The partitioned calibration error (15.2.6) allows us to relate these calibration errors via

$$\kappa_{\|\cdot\|,k} := \sup_{u \in \mathbb{R}^k} \left\{\|u\| \mid \|u\|_\infty \le 1\right\}, \tag{15.5.5}$$

a norm-dependent constant that for unconditional norms satisfies $\kappa_{\|\cdot\|,k} = \|\mathbf{1}\|$.

Theorem 15.5.6. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter $\operatorname{diam}(\mathcal{Y})$ in the norm $\|\cdot\|$, rescaled so that $\|\cdot\|_\infty \le \|\cdot\|$. Then for $\varepsilon > 0$,*

$$\begin{aligned} p_{\rm cal,up}(f) \le d_{\rm cal,up}(f) \le \mathsf{pce}(f) &\le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) p_{\rm cal,low}(f) + \kappa_{\|\cdot\|,k}\varepsilon \\ &\le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) d_{\rm cal,low}(f) + \kappa_{\|\cdot\|,k}\varepsilon. \end{aligned}$$

The rightmost upper bounds require care, so we defer the proof of the theorem to Section 15.6.4.

We record a few corollaries, one consequence of which is to show that the partitioned calibration error (15.2.6) is a calibration measure in the sense of Definition 15.2.(i). Theorem 15.5.6 also shows that the penalized calibration distance $p_{\rm cal}(f)$ is equivalent, up to taking a square root, to the upper and lower calibration "distances", adjusted by the constant (15.5.5).

Corollary 15.5.7. *Let the conditions of Theorem 15.5.6 hold. Then*

$$p_{\rm cal,low}(f) \le p_{\rm cal}(f) \le p_{\rm cal,low}(f) + 2\sqrt{2\kappa_{\|\cdot\|,k} k \operatorname{diam}(\mathcal{Y})}\sqrt{p_{\rm cal,low}(f)}$$

and

$$d_{\rm cal,low}(f) \le d_{\rm cal}(f) \le d_{\rm cal,low}(f) + 2\sqrt{2\kappa_{\|\cdot\|,k} k \operatorname{diam}(\mathcal{Y})}\sqrt{d_{\rm cal,low}(f)}.$$

Proof The first lower bound is immediate (recall the naive inequalitites (15.5.4)). Now set $\varepsilon = \sqrt{2k \operatorname{diam}(\mathcal{Y}) p_{\rm cal,low}(f)/\kappa_{\|\cdot\|,k}}$ in Theorem 15.5.6, and recognize that $p_{\rm cal,low}(f) \le p_{\rm cal,up}(f)$. □

We also obtain an approximate equivalence between the calibration distance $d_{\rm cal}$ and penalized calibration distance $p_{\rm cal}$ from definitions (15.2.3) and (15.5.2).

Corollary 15.5.8. *Let the conditions of Theorem 15.5.6 hold. Then*

$$p_{\rm cal}(f) \le d_{\rm cal}(f) \le p_{\rm cal}(f) + 2\sqrt{2\,\kappa_{\|\cdot\|,k} \cdot k\,\diam(\mathcal{Y})}\sqrt{p_{\rm cal}(f)}.$$

Proof The first inequality is immediate. Theorem 15.5.6 and $p_{\rm cal,low}(f) \le p_{\rm cal}(f)$ give

$$d_{\rm cal}(f) \le d_{\rm cal,up}(f) \le \mathsf{pce}(f) \le \left(1 + \frac{2k\,\diam(\mathcal{Y})}{\varepsilon}\right) p_{\rm cal}(f) + \kappa_{\|\cdot\|,k}\varepsilon$$

for $\varepsilon > 0$; set $\varepsilon = \sqrt{2k\,\diam(\mathcal{Y}) p_{\rm cal}(f)/\kappa_{\|\cdot\|,k}}$. □

Let us instantiate the theorem and its corollaries in a few special cases. If we make binary predictions with $\mathcal{Y} = \{0, 1\}$, then $\kappa_{\|\cdot\|,k} = k = \diam(\mathcal{Y}) = 1$, and Theorem 15.5.6 implies that

$$p_{\rm cal,low}(f) \le p_{\rm cal}(f) \le p_{\rm cal,low}(f) + 2\sqrt{2p_{\rm cal,low}(f)}.$$

For k-class multiclass classification, identify $\mathcal{Y} = \{e_1, \ldots, e_k\}$ with the k standard basis vectors and measure calibration errors with respect to the ℓ_1-norm. Then $\diam(\mathcal{Y}) = 2$ and $\kappa_{\|\cdot\|,k} = k$, so

$$p_{\rm cal,low}(f) \le p_{\rm cal}(f) \le p_{\rm cal,low}(f) + 4k\sqrt{p_{\rm cal,low}(f)}.$$

The equivalence between calibration error and the calibration distance

We can rewrite the calibration error $\mathrm{CE}(f \mid \mathcal{A})$ relative to partitions in the definition (15.2.6) as the supremum over a collection $\mathcal{W}_{\mathcal{A}}$ of functions of the form $w(\mu) = v\mathbf{1}\{\mu \in A\}$, where $\|v\|_* \le 1$, so $\mathrm{CE}(f, \mathcal{W}_{\mathcal{A}}) = \sum_{A \in \mathcal{A}} \mathbb{E}[\|\mathbb{E}[Y \mid M \in A] - \mathbb{E}[M \mid M \in A]\|]$. Relaxing this supremum, and removing the infimum over partitions, we might expect a similar relationship to Theorem 15.5.6 to hold. Via a duality argument that the definition (15.5.3) of the lower calibration error as an infimum over joint distributions makes possible, we can directly relate the measures.

Theorem 15.5.9. *Let $\mathcal{Y} \subset \mathbb{R}^k$ have finite diameter in the norm $\|\cdot\|$ and $\mathcal{W}_{\|\cdot\|}$ be the collection (15.2.5) of bounded Lipschitz functions. Then*

$$\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le (1 + \diam(\mathcal{Y})) \cdot p_{\rm cal,low}(f).$$

Conversely, let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and define $C_k := \kappa_{\|\cdot\|_*,k} \cdot \max\{1, \diam(\mathcal{Y})\}$. Then

$$d_{\rm cal,low}(f) \le C_k \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}).$$

Before giving the proof, we develop a few corollaries to give a fuller picture of the relationships between the calibration measures. For k-class multiclass classification with $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and the ℓ_1-norm, we have $\kappa_{\|\cdot\|_*,k} = 1$ and $\diam(\mathcal{Y}) = 2$, so the converse becomes $d_{\rm cal,low}(f) \le 2\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|_1})$. Thus, combining the theorem with Corollary 15.5.7 gives the following explicit comparisons.

Corollary 15.5.10. *Let $\mathcal{Y} = \{e_1, \ldots, e_k\}$ and $\|\cdot\| = \|\cdot\|_1$. Then for $f : \mathcal{X} \to \mathrm{Conv}(\mathcal{Y})$,*

$$\frac{1}{3}\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le p_{\rm cal}(f) \le 2 \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) + 4k\sqrt{2 \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})}$$

and

$$\frac{1}{3}\mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) \le d_{\rm cal}(f) \le 2 \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|}) + 4k\sqrt{2 \cdot \mathrm{CE}(f, \mathcal{W}_{\|\cdot\|})}.$$

Proof Let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$ for shorthand. Theorem 15.5.9 gives $\mathrm{CE}(f, \mathcal{W}) \le 3p_{\mathrm{cal,low}}(f)$, while $p_{\mathrm{cal,low}}(f) \le p_{\mathrm{cal}}(f)$ and $p_{\mathrm{cal,low}}(f) \le d_{\mathrm{cal,low}}(f)$, giving each of the leftmost inequalities. For the rightmost inequalities, Corollary 15.5.7 specialized to the ℓ_1-norm gives

$$p_{\mathrm{cal}}(f) \le p_{\mathrm{cal,low}}(f) + 4k\sqrt{p_{\mathrm{cal,low}}(f)} \quad \text{and} \quad d_{\mathrm{cal}}(f) \le d_{\mathrm{cal,low}}(f) + 4k\sqrt{d_{\mathrm{cal,low}}(f)}.$$

Use that $p_{\mathrm{cal,low}}(f) \le d_{\mathrm{cal,low}}(f)$ and the second part of Theorem 15.5.9 with $C_k = 2$. □

Proof of Theorem 15.5.9

The proof of the first upper bound is fairly straightforward. For any $w \in \mathcal{W}_{\|\cdot\|}$, we have

$$\begin{aligned}
\mathbb{E}[\langle w(M), Y - M\rangle] &= \mathbb{E}[\langle w(M), V - M\rangle] + \mathbb{E}[\langle w(M) - w(V), Y - V\rangle] + \mathbb{E}[\langle w(V), Y - V\rangle] \\
&\le \mathbb{E}[\|V - M\|] + \mathrm{diam}(\mathcal{Y})\mathbb{E}[\|V - M\|] + \mathbb{E}[\langle w(V), \mathbb{E}[Y \mid V] - V\rangle] \\
&\le (1 + \mathrm{diam}(\mathcal{Y}))\mathbb{E}[\|V - M\|] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\|].
\end{aligned}$$

For the converse, we first reduce to a discrete problem over probability mass functions rather than general distributions, as then we may apply convex duality (in particular, that linear programming duality obtains when the primal problem is feasible). Consider the value

$$d_{\mathrm{cal,low}}(M) = \inf_V \left\{\mathbb{E}[\|M - V\|] \;\text{ s.t. }\; \mathbb{E}[Y \mid V] = V\right\}.$$

Let $b \in \mathbb{N}$ and $\mathcal{M}_b$ be a $1/b$ covering $\{\mu_1, \ldots, \mu_N\}$ of $\mathcal{M} = \mathrm{Conv}(\mathcal{Y})$, and define M_b to be the projection of M to the nearest μ_i. Then $\|M - V\| = \|M_b - V\| \pm \frac{1}{b}$, and

$$\left|d_{\mathrm{cal,low}}(M) - \inf_V \left\{\mathbb{E}[\|M_b - V\|] \text{ s.t. } \mathbb{E}[Y \mid V] = V\right\}\right| \le \frac{1}{b}.$$

If we replace the infimum over joint distributions of (M_b, Y, V) that leave the marginal (M_b, Y) unchanged (with V calibrated) with, instead, an infimum over discrete distributions on V, then

$$d_{\mathrm{cal,low}}(M) \le \inf_{V \text{ finitely supported}} \left\{\mathbb{E}[\|M_b - V\|] \;\text{ s.t. }\; \mathbb{E}[Y \mid V] = V\right\} + \frac{1}{b}. \tag{15.5.6}$$

Notably, the infimum is non-empty, as we can always choose $V = Y$.

With the problem (15.5.6) in hand, we can write a finite dimensional optimization problem whose optimal value is the discretized infimum on the right side. Without loss of generality assuming that M is finitely supported, we let $p_{\mu y} = \mathbb{P}(M = \mu, Y = y)$ be the probability mass function of (M, Y). We then assume that the support $\mathcal{V}$ of V contains at least each $y \in \mathcal{Y}$ as well as the following vectors: for each $\mu \in \mathrm{supp}\, M$ and non-empty $I \subset \{i \in [n] \mid \mu_i > 0\}$, let $\mathbf{1}_I = \sum_{i \in I} e_i$ be the $\{0, 1\}$ vector supported on I. Then

$$v = \frac{1}{\langle \mu, \mathbf{1}_I\rangle}\mathbf{1}_I \odot \mu \in \mathcal{V}, \tag{15.5.7}$$

that is, the renormalization of vectors in $\mathrm{supp}\, M$ while restricting their supports. (Here $\odot$ denotes Hadamard, or elementwise, product.)

Introducing the joint distribution Q with p.m.f. $q_{\mu y v} = Q(M = s, Y = y, V = v)$, the infimum (15.5.6) has the constraint that $\sum_v q_{\mu y v} = p_{\mu y}$. Then $\mathbb{E}[\|M - V\|] = \sum_{\mu, y, v} q_{\mu y v} \|\mu - v\|$ and

the calibration constraint $\mathbb{E}[Y \mid V] = V$ becomes the equality constraint that $\sum_{\mu,y} q_{\mu yv}(y - v) = 0$ for each v. This yields the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{\mu,y,v} q_{\mu yv} \left\|\mu - v\right\| \\ \text{subject to} & \sum_v q_{\mu yv} = p_{\mu y}, \quad q \succeq 0, \quad \sum_{\mu,y} q_{\mu yv}(y - v) = 0 \text{ for all } v \end{array} \tag{15.5.8}$$

in the variable q.

We take the dual of this problem. Introduce Lagrange multipliers $\lambda_{\mu y}$ for each equality constraint that $\sum_v q_{\mu yv} = p_{\mu y}$, multipliers $\theta_{\mu yv} \geq 0$ for the nonnegativity constraints on q, and $\beta_v \in \mathbb{R}^k$ for each equality constraint that $0 = \sum_{\mu,y} q_{\mu yv}(y - v)$. Then we have Lagrangian

$$\begin{aligned} &\mathcal{L}(q, \lambda, \theta, \beta) \\ &= \sum_{\mu,y,v} q_{\mu yv} \left\|\mu - v\right\| + \sum_{\mu,y,v} q_{\mu yv} \beta_v^\top (y - v) - \sum_{\mu,y} \lambda_{\mu y} \left(\sum_v q_{\mu yv} - p_{\mu y} \right) - \langle \theta, q \rangle. \end{aligned}$$

Taking an infimum over q, we see that unless

$$\left\|\mu - v\right\| + \beta_v^\top (y - v) - \lambda_{\mu y} - \theta_{\mu yv} = 0$$

for each triple (μ, y, v), we have $\inf_q \mathcal{L}(q, \lambda, \theta, \beta) = -\infty$. The equality in the preceding display is equivalent to $\left\|\mu - v\right\| + \beta_v^\top (y - v) \geq \lambda_{\mu y}$, so that eliminating $\theta \succeq 0$ variables, we have the dual

$$\begin{array}{ll} \text{maximize} & \sum_{\mu,y} \lambda_{\mu y} p_{\mu y} \\ \text{subject to} & \lambda_{\mu y} \leq \left\|\mu - v\right\| + \beta_v^\top (y - v), \quad \text{all } \mu, y, v \end{array}$$

to problem (15.5.8). Equivalently, recognizing that at the optimum we must saturate the constraints on λ via $\lambda_{\mu y} = \min_v \{\left\|\mu - v\right\| + \beta_v^\top (y - v)\}$, we have

$$\text{maximize} \quad \sum_{\mu,y} p_{\mu y} \min_v \left\{ \left\|\mu - v\right\| + \beta_v^\top (y - v) \right\} \tag{15.5.9}$$

in the variables β_v, and strong duality obtains.

The dual problem (15.5.9) is the key to the final step in the proof. To make the functional notation clearer, let us fix any collection of vectors β_v and define $\lambda_y(\mu) = \min_v \{\left\|\mu - v\right\| + \beta_v^\top (y - v)\}$ for each $y \in \mathcal{Y}$. If we can exhibit a C-Lipschitz function $\mu \mapsto w(\mu)$ that satisfies

$$\langle w(\mu), y - \mu \rangle \geq \lambda_y(\mu) \tag{15.5.10}$$

for each $y \in \mathcal{Y}$ and $\left\|w(\mu)\right\|_* \leq C$, we will evidently have shown that

$$\sup_{w \in \mathcal{W}_{\left\|\cdot\right\|}} \mathbb{E}[\langle w(M), Y - M \rangle] \geq \frac{1}{C} d_{\mathrm{cal,low}}(M),$$

by the dual formulation (15.5.9).

The functions λ_y are each 1-Lipschitz with respect to $\left\|\cdot\right\|$, as

$$\begin{aligned} \lambda_y(\mu) - \lambda_y(\mu') &\geq \min_v \left\{ \left\|\mu - v\right\| + \beta_v^\top (y - v) - \left\|\mu' - v\right\| - \beta_v^\top (y - v) \right\} \\ &= \min_v \left\{ \left\|\mu - v\right\| - \left\|\mu' - v\right\| \right\} \geq - \left\|\mu - \mu'\right\|, \end{aligned}$$

and similarly

$$\lambda_y(\mu) - \lambda_y(\mu') \le \max_v \left\{ \|\mu - v\| + \beta_v^\top (y - v) - \|\mu' - v\| - \beta_v^\top (y - v) \right\} \le \|\mu - \mu'\|$$

by the triangle inequality. Now we specialize to the particular multiclass classification case in which the set $\mathcal{Y} = \{e_1, \ldots, e_k\}$ consists of extreme points of the probability simplex, so that $\mu \in \mathrm{Conv}(\mathcal{Y})$ means that $\langle \mathbf{1}, \mu \rangle = 1$ and $\mu \succeq 0$. Abusing notation slightly, let $\lambda_i = \lambda_{e_i}$ for $i = 1, \ldots, k$. We observe that $\lambda_i(\mu) \le \|\mu - e_i\|$, while for any $v \in \mathcal{V} \subset \Delta_k$, we have $\lambda_i(v) \le \langle \beta_v, e_i - v \rangle$, and so

$$\sum_{i=1}^k v_i \lambda_i(v) \le \sum_{i=1}^k v_i \langle \beta_v, e_i - v \rangle = \langle \beta_v, v - v \rangle = 0. \tag{15.5.11}$$

Fix the value $D = \mathrm{diam}(\mathcal{Y})$ and define the truncated values

$$\overline{\lambda}_i(\mu) := \max\{\lambda_i(\mu), -D\}.$$

These remain 1-Lipschitz, and we claim that

$$\max_\mu \sum_{i=1}^k \mu_i \overline{\lambda}_i(\mu) \le 0. \tag{15.5.12}$$

Assuming we can demonstrate inequality (15.5.12), we can define our Lipschitz function satisfying the witness condition (15.5.10). In particular, define

$$w(\mu) := \left(\overline{\lambda}_1(\mu), \ldots, \overline{\lambda}_k(\mu)\right) \in \mathbb{R}^k.$$

Then by inspection,

$$\left\|w(\mu) - w(\mu')\right\|_* \le \sup_{\|u\|_\infty \le 1} \left\|\mu - \mu'\right\| \|u\|_* = \kappa_{\|\cdot\|_*, k} \left\|\mu - \mu'\right\|.$$

Additionally, because $-D \le \overline{\lambda}_i(\mu) \le \|\mu - e_i\|$, we have $\|w(\mu)\|_* \le \kappa_{\|\cdot\|_*, k} \mathrm{diam}(\mathcal{Y})$. Finally, assuming the condition (15.5.12), we see that $w(\mu)$ satisfies inequality (15.5.10): for $y = e_j$,

$$\langle w(\mu), e_j - \mu \rangle = \overline{\lambda}_j(\mu) - \sum_{i=1}^k \overline{\lambda}_i \mu_i) \ge \overline{\lambda}_j(\mu) \ge \lambda_j(\mu).$$

It remains to demonstrate inequality (15.5.12). Fix μ and let $T = \{i \in [k] \mid \lambda_i(\mu) < -D\}$ be the truncated indices, and letting $\mathbf{1}_T = \sum_{i \in T} e_i$ be the vector with 1 in positions $i \in T$, consider the value $\langle \mathbf{1}_T, \mu \rangle$. If $\langle \mathbf{1}_T, \mu \rangle = 1$, then $\mu_i > 0$ implies $i \in T$ so $\sum_i \mu_i \overline{\lambda}_i(\mu) = -D < 0$. Otherwise, let $T^c = [k] \setminus T = \{i \in [k] \mid \lambda_i(\mu) \ge -D\}$ be the collection of non-truncated indices. Let $r = \langle \mathbf{1}_{T^c}, \mu \rangle = 1 - \langle \mathbf{1}_T, \mu \rangle > 0$ be the remainder of the probability mass, and define $\nu = \mathbf{1}_{T^c} \odot \mu / r$ to be the renormalized vector μ with support restricted to T^c. By assumption (15.5.7) on the range of $V \in \mathcal{V}$, we see that $\nu \in \mathcal{V}$. Thus

$$\sum_{i \in T^c} \mu_i \lambda_i(\mu) = r \sum_{i \in T^c} \nu_i \lambda_i(\mu) \le r \sum_{i \in T^c} \nu_i \left(\lambda_i(\nu) + \|\nu - \mu\|\right) \le r \|\nu - \mu\|$$

by the Lipschitzness of λ_i and inequality (15.5.11). Writing $\mu = r\nu + (1 - r)u$ for some vector $u \in \mathcal{M}$, we see that $\|\mu - \nu\| \le (1 - r) \|\nu - u\| \le (1 - r)D$, and so

$$\sum_i \mu_i \overline{\lambda}_i(\mu) = \sum_{i \in T^c} \mu_i \lambda_i(\mu) - (1 - r)D \le r \|\nu - \mu\| - (1 - r)D \le (r - 1)(1 - r)D \le 0.$$

Inequality (15.5.12) holds.

15.6* Deferred technical proofs

These technical proofs use the integration results that Appendix A.1 recapitulates.

15.6.1 Proof of Lemma 15.2.1

Let $\mathcal{W}_c$ be the collection of c-Lipschitz functions w with $\|w(\mu)\|_* \le 1$ for all μ, and let $\mathcal{W}$ denote the Borel measurable functions with $\|w(\cdot)\|_* \le 1$, so $\mathsf{ece}(g) = \mathrm{CE}(g, \mathcal{W})$.

Fix $c < \infty$ and $w \in \mathcal{W}_c$ and let $f_n \to f$ in $L^1(P)$. Then

$$\begin{aligned}
&|\langle w(f_n(x)), y - f_n(x)\rangle - \langle w(f(x)), y - f(x)\rangle| \\
&\le |\langle w(f_n(x)) - w(f(x)), y - f(x)\rangle| + |\langle w(f_n(x)), f_n(x) - f(x)\rangle| \\
&\le \min\{c \|f(x) - f_n(x)\|, 2\} \|y - f(x)\| + \|f_n(x) - f(x)\|.
\end{aligned}$$

The second term converges to zero in expectation, while for the first, for any $K < \infty$ we have

$$\begin{aligned}
&\mathbb{E}\left[\min\{c \|f(X) - f_n(X)\|, 2\} \|Y - f(X)\|\right] \\
&\le 2\mathbb{E}\left[\mathbf{1}\left\{\|Y - f(X)\| > K\right\}\right] + K\mathbb{E}\left[\min\{c \|f(X) - f_n(X)\|, 2\}\right].
\end{aligned}$$

Dominated convergence (Theorem A.1.3 in Appendix A.1) implies $\mathbb{E}[\min\{c \|f(X) - f_n(X)\|, 2\}] \to 0$ as $n \to \infty$, and Markov's inequality gives $\mathbb{P}(\|Y - f(X)\| > K) \to 0$ as $K \to \infty$. As a consequence,

$$\sup_{w \in \mathcal{W}_c} |\mathbb{E}\left[\langle w(f_n(X)), Y - f_n(X)\rangle\right] - \mathbb{E}\left[\langle w(f(X)), Y - f(X)\rangle\right]| \to 0 \tag{15.6.1}$$

for any $c < \infty$. The limit (15.6.1) certainly implies

$$\liminf_n \mathsf{ece}(f_n) = \liminf_n \mathrm{CE}(f_n, \mathcal{W}) \ge \liminf_n \mathrm{CE}(f_n, \mathcal{W}_c) = \sup_{w \in \mathcal{W}_c} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle].$$

Because the Lipschitz functions are dense in $L^p(P)$ for any Borel measure P (Lemma A.1.6 in in Appendix A.1), $\lim_c \sup_{w \in \mathcal{W}_c} \mathbb{E}[\langle w(f(X)), Y - f(X)\rangle] = \mathrm{CE}(f, \mathcal{W}) = \mathsf{ece}(f)$, giving the result.

15.6.2 Proof of Proposition 15.5.2

The proof that $\mathrm{CE}(\cdot, \mathcal{W}_{\|\cdot\|})$ identifies calibration (Definition 15.1, part (i)) is identical to the corresponding argument in Proposition 15.5.3. To demonstrate continuity, let $\mathcal{W} = \mathcal{W}_{\|\cdot\|}$, and consider a sequence of functions $f_n \to f$ in $L^1(P)$. The convergence (15.6.1) implies that

$$|\mathrm{CE}(f_n, \mathcal{W}) - \mathrm{CE}(f, \mathcal{W})| \le \sup_{w \in \mathcal{W}} |\mathbb{E}[\langle w(f(X)), Y - f(X)\rangle - \langle w(f_n(X)), Y - f_n(X)\rangle]| \to 0.$$

15.6.3 Proof of Lemma 15.5.4

Define $f(z) = g(z)/\|g(z)\|_2$ if $g(z) \neq 0$ and 0 otherwise, so $\|g(z)\|_2 = \langle f(z), g(z)\rangle$. Lemma A.1.6 shows that for each $n \in \mathbb{N}$ there exists a Lipschitzian function w_n with $\mathbb{E}[\|w_n(Z) - f(Z)\|] \le \frac{1}{n}$, and w.l.o.g. we may take $\|w_n(z)\|_2 \le 1$. Then because $\mathbb{E}[\langle w_n(Z), g(Z)\rangle] = 0$,

$$\mathbb{E}[\|g(Z)\|_2] = \mathbb{E}[\langle f(Z) - w_n(Z), g(Z)\rangle].$$

Recognize the domination $|\langle f(z) - w_n(z), g(z)\rangle| \le 2\|g(z)\|_2$. We also have $w_n \to f$ in $L^1(P)$, and hence in P-measure, so that $\langle f(Z) - w_n(Z), g(Z)\rangle \xrightarrow{p} 0$. Dominated convergence (Theorem A.1.3) implies that $\mathbb{E}[\|g(Z)\|_2] = \mathbb{E}[\langle f(Z) - w_n(Z), g(Z)\rangle] \to 0$, so $g(Z) = 0$ with probability 1.

15.6.4 Proof of Theorem 15.5.6

The following lemma gives the first two inequalities in the theorem.

Lemma 15.6.1. *For $M = f(X)$, we have*

$$p_{\text{cal,up}}(f) \le d_{\text{cal,up}}(f) \le \mathsf{pce}(M). \tag{15.6.2}$$

Proof Fix any partition $\mathcal{A}$, and define $\mathsf{q}_{\mathcal{A}}(\mu)$ to be the (unique) set A such that $\mu \in A$. Then set $g(\mu) = \mathbb{E}[Y \mid M \in \mathsf{q}_{\mathcal{A}}(\mu)]$. Then $g(M) = \mathbb{E}[Y \mid g(M)]$ with probability 1, and

$$\begin{aligned}
p_{\text{cal,up}}(f) \le d_{\text{cal,up}}(f) &\le \mathbb{E}[\|M - g(M)\|] \\
&= \sum_{A \in \mathcal{A}} \mathbb{E}[\|M - \mathbb{E}[Y \mid M \in A]\| \mathbf{1}\{M \in A\}] \\
&\le \sum_{A \in \mathcal{A}} \mathbb{E}\left[(\|M - \mathbb{E}[M \mid M \in A]\| + \|\mathbb{E}[M - Y \mid M \in A]\|) \mathbf{1}\{M \in A\}\right] \\
&\le \sum_{A \in \mathcal{A}} \operatorname{diam}(A) \mathbb{P}(M \in A) + \sum_{A \in \mathcal{A}} \|\mathbb{E}[(M - Y)\mathbf{1}\{M \in A\}]\|.
\end{aligned}$$

Taking an infimum gives the claim (15.6.2). □

To prove the remaining upper bounds on $\mathsf{pce}(f)$, we consider a family of partitions restricted to boxes of width $\varepsilon > 0$, the resolution in the statement of Theorem 15.5.6. Let $B = [0,1)^k$ be the half-open box, and for each $u \in [0,1]^k$ let $\mathcal{A}_u$ be the partition of $\mathbb{R}^k$ into shifted and scaled boxes

$$A_z(u) = \varepsilon u + \varepsilon(z + B) = [\varepsilon(z_1 + u_1), \varepsilon(z_1 + u_1 + 1)) \times \cdots \times [\varepsilon(z_k + u_k, \varepsilon(z_k + u_k + 1))$$

ranging over $z \in \mathbb{Z}^k$. The $\|\cdot\|$ diameter of B satisfies $\operatorname{diam}(B) \le \sup_{\|u\|_\infty \le 1} \|u\| = \kappa_{\|\cdot\|,k}$, so $\operatorname{diam}(A_z(u)) \le \kappa_{\|\cdot\|,k} \cdot \varepsilon$, and the definition (15.2.6) of the partitioned calibration error gives

$$\mathsf{pce}(f) \le \text{CE}(f \mid \mathcal{A}_u) + \kappa_{\|\cdot\|,k} \cdot \varepsilon. \tag{15.6.3}$$

We now upper bound $\text{CE}(f \mid \mathcal{A}_u)$, recalling the notation $M = f(X)$. Fix any coupling (M, Y, V) allowed in (15.5.3b). Recall the definition $\text{CE}(M \mid \mathcal{A}_u) = \sum_{A \in \mathcal{A}_u} \|\mathbb{E}\left[(Y - M)\mathbf{1}\{M \in A\}\right]\|$ of the partitioned calibration error, and consider a single element A in the sum. For any such $A \in \mathcal{A}_u$,

$$\begin{aligned}
&\mathbb{E}\left[(Y - M)\mathbf{1}\{M \in A\}\right] \\
&= \mathbb{E}\left[(V - M)\mathbf{1}\{M \in A\}\right] + \mathbb{E}\left[(Y - V)\mathbf{1}\{V \in A\}\right] + \mathbb{E}\left[(Y - V)\left(\mathbf{1}\{M \in A\} - \mathbf{1}\{V \in A\}\right)\right],
\end{aligned}$$

so by the triangle inequality,

$$\begin{aligned}
\|\mathbb{E}\left[(Y - M)\mathbf{1}\{M \in A\}\right]\| \le \mathbb{E}[\|V - M\| \mathbf{1}\{M \in A\}] + \mathbb{E}[\|\mathbb{E}[Y \mid V] - V\| \mathbf{1}\{V \in A\}] \\
+ 2 \operatorname{diam}(\mathcal{Y}) \left[\mathbb{P}(M \in A, V \notin A) + \mathbb{P}(M \notin A, V \in A)\right].
\end{aligned}$$

As in the proof of Lemma 15.6.1, let $\mathsf{q}_u(\mu)$ denote the unique set $A \in \mathcal{A}_u$ containing μ. Then

$$\text{CE}(M \mid \mathcal{A}_u) \le \mathbb{E}[\|V - M\|] + \mathbb{E}\left[\|\mathbb{E}[Y \mid V] - V\|\right] + 2 \operatorname{diam}(\mathcal{Y}) \mathbb{P}\left(\mathsf{q}_u(M) \ne \mathsf{q}_u(V)\right).$$

Now draw $U \sim \mathsf{Uniform}([0,1]^k)$ independently of (M, Y, V). For $\mu, \nu \in \mathbb{R}^k$, assume that $\mu_j < \nu_j$; then there exists a $z_j \in \mathbb{Z}$ such that $\mu_j \le \varepsilon(z_j + U_j) < \nu_j$ with probability at most $\frac{1}{\varepsilon}|\mu_j - \nu_j|$. Thus the union bound implies that

$$\mathbb{P}\left(\mathsf{q}_U(\mu) \neq \mathsf{q}_U(\nu)\right) \le \sum_{j=1}^k \frac{|\mu_j - \nu_j|}{\varepsilon} \le \frac{k}{\varepsilon} \left\|\mu - \nu\right\|_\infty \le \frac{k}{\varepsilon} \left\|\mu - \nu\right\|.$$

Integrating therefore yields

$$\int_{[0,1]^k} \mathrm{CE}(M \mid \mathcal{A}_u) du \le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) \mathbb{E}\left[\left\|V - M\right\|\right] + \mathbb{E}\left[\left\|\mathbb{E}[Y \mid V] - V\right\|\right].$$

Inequality (15.6.3) holds for each $u \in [0,1]^k$, so we see that

$$\mathsf{pce}(M) \le \left(1 + \frac{2k \operatorname{diam}(\mathcal{Y})}{\varepsilon}\right) \cdot \left(\mathbb{E}\left[\left\|V - M\right\|\right] + \mathbb{E}\left[\left\|\mathbb{E}[Y \mid V] - V\right\|\right]\right) + \kappa_{\|\cdot\|,k} \cdot \varepsilon.$$

Taking an infimum over V proves the upper bound involving $p_{\mathrm{cal,low}}(f)$. Restricting to calibrated V proves the bound involving $d_{\mathrm{cal,low}}(f)$.

15.7 Bibliography and further reading

Calibration remains an active research area involving multiple communities. In the game theory and probability literatures, work on calibration descends from Dawid's work on *prequential probabilities*, which give a game-theoretic basis for probability [132, 133, 131] based on forecasting, and move beyond traditional frequentist probability in that there no longer needs to be a notion of frequencies of "unrelated trials under constant conditions" [133]. There, in a forecasting problem of predicting values $Y_1, Y_2, \ldots$, we wish to have a calibrated forecaster in the sense that at time i, the forecaster's value F_i depends only on $Y_1, \ldots, Y_{i-1}$, and that for any sequence of predictable times $\tau_1, \tau_2, \ldots \in \mathbb{N}$, then $\frac{1}{k}\sum_{i=1}^k (Y_{\tau_i} - F_{\tau_i}) \to 0$. Foster and Vohra [192] show the striking result that even in adversarial settings, where nature chooses the values Y_i depending on the past and adversarially, this type of calibration is possible so long as the forecasts F_i may be randomized. This game-theoretic perspective continues to inspire research, even at the foundations of probability and statistics, by formulating problems as prediction games; the book [373] and survey [353] describe this perspective.

Calibration is also a central goal in statistical forecasting problems. Much of this literature focuses on predicting real-valued random variables via cumulative distribution functions. When predicting a CDF F as in Example 14.2.6, perfect calibration (15.0.1) becomes that

$$\mathbb{P}(Y \le u \mid F) = F(u) \quad \text{for all } u \in \mathbb{R}. \tag{15.7.1}$$

Gneiting and collaborators [213, 215, 212] have contributed significantly to this literature, and many questions revolve around achieving different types of calibration for sequences F_t of CDFs, as the *conditional calibration* (15.7.1) is too strict. *Probabilistic calibration* requires the realized values of probability integral transformation $u_t = F_t(Y_t)$ to be calibrated, in that $\frac{1}{T}\sum_{t=1}^T \mathbf{1}\left\{u_t \le u\right\} \to u$ with probability 1 for each u. *Marginal calibration* requires that $\frac{1}{T}\sum_{t=1}^T F_t(y) - \frac{1}{T}\sum_{t=1}^T \mathbf{1}\left\{Y_t \le y\right\} \to 0$ for all y. Questions about how diffuse differently calibrated predictors may be remain open.

In computer science and machine learning, calibration questions have taken on new urgency in the face of increasing reliance on machine learning predictions, especially with large-scale models or

in sensitive domains. Some of this work arises from the fairness literature [234, 265, 223, 52, 259]. At a high level, for binary classification, one seeks the *multicalibtration* guarantee that $\mathbb{P}(Y = 1 \mid X \in C) \approx \mathbb{E}[f(X) \mid X \in C]$ for all subsets C in a collection $\mathcal{C}$ of X-space. When $\mathcal{C}$ is not too complicated, boosting and other strategies achieve this [234, 223, 52, 259]. Kakade and Foster [260] introduce the calibration error (15.2.4) relative to a class $\mathcal{W}$ of witness functions.

Section 15.2.1 highlights that, as in the forecasting literature [213], evaluating the calibration of a predictor is challenging. These impossibility results relate Donoho's classical results on one-sided inference [145], which show that many functionals (typically measuring some type of "complexity") admit only consistent under-estimates, because it is easy to modify a well-behaved function—e.g., one with no calibration error—to introduce pathologies, but small modifications cannot drastically reduce complexity. See also Lee et al. [285]. Blasiok et al. [75] therefore develop metrics for distances from calibration, focusing on the case of binary prediction. Their results on binary prediction ($Y \in \{0, 1\}$) provide natural precursors to those we present in Section 15.5; Propositions 15.5.1, 15.5.2, and 15.5.3 formulate the corresponding continuity properties for vector-valued predictions, and they also demonstrate that the measures are indeed valid calibration measures (Definition 15.1, part (i)). Blasiok et al. [76] also introduce the post-processing gap for squared error (Proposition 15.3.1); Proposition 15.3.5 extends this perspective to losses based on conjugate linkages.

Practically applying calibration remains challenging, not least because evaluating and calibrating predictors naively requires sample complexity exponential in the number of outcomes. Thus, one looks for simpler approaches to calibration. Conformal prediction (Section 15.4.1) guarantees coverage of Y in general settings, of which we only touch the barest surface; see the survey [19] or the books by Vovk et al. [417] and Angelopoulos et al. [20]. Duchi [152] proves Proposition 15.4.2. Kumar et al. [269] develop the methods in Section 15.4.2; see also Foster and Hart [196]. For multiclass problems, weaker notions of calibration, such as calibrating only the prediction of the most probable class, appear necessary [222] (see Exercises 15.3 and 15.4). Recent work has begun to investigate approaches to calibrate predictions in large language models [317, 111].

15.8 Exercises

Exercise 15.1: Let the setting of Section 15.3.2 hold, so that $\varphi(s, y) = \Omega^*(s) - \langle s, y\rangle$. Let $\mathcal{W}$ be a collection of witness functions satisfying $R_* = \sup_{w\in\mathcal{W},\mu} \|w(\mu)\|_* < \infty$, and assume $\varphi \geq 0$. Show that the algorithm in Figure 15.2 guarantees that

$$\min_{\tau<t} \mathrm{CE}(\mathsf{pred}_\Omega \circ f_\tau, \mathcal{W}) \leq \frac{\sqrt{2R_*^2\mathbb{E}[\varphi(f_0(X), Y)]}}{\sqrt{t}},$$

and in particular terminates with $\mathrm{CE}(\mathsf{pred}_\Omega \circ f_t, \mathcal{W}) \leq \epsilon$ for some t with

$$t \leq \frac{2R_*^2\mathbb{E}[\varphi(f_0(X), Y)]}{\epsilon^2}.$$

Exercise 15.2: We say $\mathcal{W}$ is *weakly complete* if $\mathbb{E}[\langle w(S), g(S)\rangle] = 0$ implies that $g(S) = 0$ with probability 1. Let $\mathcal{W}$ be any symmetric weakly complete collection of functions. Show that

$$\mathrm{CE}(f, \mathcal{W})$$

is a sound and complete (15.2.2) calibration measure.

Exercise 15.3 (Marginal calibration)**:** A predictor $f : \mathcal{X} \to \mathbb{R}^k$ of $Y \in \mathbb{R}^k$ is *marginally calibrated* if for each $j \in \{1, \ldots, k\}$ we have $\mathbb{E}[Y_j \mid f_j(X)] = f_j(X)$ with probability 1.

Input: Population distribution P, collection of bounded witness functions $\mathcal{W}$, generalized negative entropy Ω strongly convex w.r.t. norm $\|\cdot\|$, initial predictor $f_0 : \mathcal{X} \to \mathbb{R}^k$, calibration tolerance $\epsilon > 0$
Initialize: set $R_*^2 := \sup_{w \in \mathcal{W}, \mu} \|w(\mu)\|_*^2$
Repeat: for $t = 0, 1, \ldots$

i. Find witness w_t maximizing $\mathbb{E}[\langle w(f_t(X)), Y - \mathsf{pred}_\Omega(f_t(X))\rangle]$

ii. Set $\eta_t = \frac{\mathbb{E}[\langle w_t(f_t(X)), Y - \mathsf{pred}_\Omega(f_t(X))\rangle]}{R_*^2}$

iii. Update $f_{t+1} = f_t + \eta_t \cdot w_t \circ \mathsf{pred}_\Omega \circ f_t$

iv. Terminate if
$$\mathrm{CE}(\mathsf{pred}_\Omega \circ f_t, \mathcal{W}) \le \epsilon.$$

Figure 15.2: Improving calibration relative to the class $\mathcal{W}$

(a) Show that if $f : \mathcal{X} \to \mathbb{R}^k$ is calibrated, then it is marginally calibrated.

(b) Let $\mathcal{W}$ be symmetric and weakly complete (Exercise 15.2) and $\mathcal{W}^k = \{(w_1, \ldots, w_k) \mid w_j \in \mathcal{W}\}$ be the vector-valued functions with components in $\mathcal{W}$, where $w \in \mathcal{W}^k$ satisfies $w(\mu) = (w_1(\mu_1), \ldots, w_k(\mu_k))$ for $\mu \in \mathbb{R}^k$. Show that $\mathrm{CE}(f, \mathcal{W}^k)$ is sound and complete for marginal calibration, that is, $\mathrm{CE}(f, \mathcal{W}^k) = 0$ if and only if f is marginally calibrated.

Exercise 15.4 (Top-K calibration)**:** In k-class multiclass classification of $y \in \{e_1, \ldots, e_k\} \subset \mathbb{R}^k$, let $f_{(1)}(x) \ge f_{(2)}(x) \ge \cdots \ge f_{(k)}(x)$ denote the sorted elements of $f(x) \in \mathbb{R}^k$. We say f is top-K calibrated if whenever $f_j(X)$ is among the K-highest predictions,

$$\mathbb{P}\left(Y = e_j \mid f_j(X), f_j(X) \ge f_{(K)}(X)\right) = f_j(X).$$

For a function class $\mathcal{W}$ mapping $\mathbb{R} \to \mathbb{R}$, define the function class

$$\mathcal{W}_{\text{top-}K} := \left\{ w(\mu) = \sum_{j=1}^k w_j(\mu_j) e_j \mathbf{1}\left\{\mu_j \ge \mu_{(K)}\right\} \mid w_j \in \mathcal{W}, j = 1, \ldots, k \right\}.$$

Show that if $\mathcal{W}$ is weakly complete (Exercise 15.2) and symmetric, then $\mathrm{CE}(f, \mathcal{W}_{\text{top-}K})$ is sound and complete for top-K calibration.

Exercise 15.5: Let the conditions of Theorem 15.4.3 hold, and fix the bins $\{B_j\}$ partitioning $\mathrm{Im}\, f$ and g. Let $\mathbb{E}_j[\cdot] = \mathbb{E}[\cdot \mid f(X) \in B_j]$ and $\mathrm{Var}_j(\cdot) = \mathrm{Var}(\cdot \mid f(X) \in B_j)$ indicate conditional expectation and variance. Let $L_j(f) = \mathbb{E}_j[(Y - f(X))^2]$, $D_j(f) = \widehat{u}_j - \widehat{u}_{j-1}$, $\overline{f}_j = \mathbb{E}_j[f(X)]$, $P(B_j) = P(\{x \mid f(x) \in B_j\})$, and $\overline{Y}_{B_j} = \mathbb{E}_{P_n^{(2)}}[Y \mid f(X) \in B_j]$. Show that

$$L(g) \le L(f) + \sum_j P(B_j)\left(2\sqrt{L_j(f)} D_j(f) + \mathrm{Var}_j(f(X)) + (\overline{Y}_{B_j} - \mathbb{E}_j[Y])^2 - (\overline{f}_j - \mathbb{E}_j[Y])^2\right).$$

Exercise 15.6: Let the conditions of Theorem 15.4.3 hold. Let $\epsilon = r/8$ and partition $\operatorname{Im} f$ into $m = 1/\epsilon$ disjoint intervals A_j with $P(f(X) \in A_j) = \epsilon$. Show that

$$\mathbb{P}\left(\max_{j \le m} \left|P_n^{(1)}(f(X) \in A_j) - P(f(X) \in A_j)\right| \ge \epsilon\right) \le \frac{2}{\epsilon} \exp\left(-\frac{3}{8} n\epsilon\right).$$

Conclude that with probability at least $1 - \frac{2}{\epsilon}\exp(-\frac{3}{8}n\epsilon)$ over $P_n^{(1)}$, the P-measure of the bins B_j each satisfy $r - 2\epsilon \le P(\{x \mid f(x) \in B_j\}) \le r + 2\epsilon$, and so for a numerical constant $c > 0$,

$$\mathbb{P}\left(\max_{j} \left|P_n^{(2)}(f(X) \in B_j) - P(f(X) \in B_j)\right| \ge \epsilon \mid P_n^{(1)}\right) \le \frac{2}{8\epsilon} e^{-cn\epsilon}.$$

Exercise 15.7: Let the conditions of Theorem 15.4.3 hold. Use Assumption A.15.1 and Exercise 15.6 to show that there are numerical constants $0 < c, C < \infty$ and an event $\mathcal{E}_{\text{acc}}$ with probability

$$\mathbb{P}\left(\mathcal{E}_{\text{acc}}\right) \ge 1 - \frac{C}{\epsilon}\left[e^{-cn\epsilon} + e^{-cnrt^2}\right] \tag{15.8.1}$$

such that $\frac{3}{4}r \le P(\{x \mid f(x) \in B_j\}) \le \frac{5}{4}r$ and $|\overline{Y}_{B_j} - \mathbb{E}_j[Y]| \le \sigma t$ for each j.

Exercise 15.8: Use the results of Exercises 15.5, 15.6, and 15.7 to finalize the proof of Theorem 15.4.3. *Hint.* The inequality $2ab \le \eta a^2 + \frac{1}{\eta} b^2$ is valid for $a, b \in \mathbb{R}$ and $\eta > 0$.

Exercise 15.9: Show that Theorem 15.5.5 is sharp: for any $\alpha, \beta > \frac{1}{2}$, there exist binary distributions and predictors $f : \mathcal{X} \to [0, 1]$ for which $d_{\text{cal}}, p_{\text{cal}}$, and $\text{CE}(\cdot, \mathcal{W}_{|\cdot|})$ fail to be (α, β)-equivalent.

Chapter 16

Classification, surrogate risk, and loss equivalence

While proper losses provide a principled approach to estimating conditional distributions, in many learning problems, we wish only to optimally predict a target Y rather than the full distribution of $Y \mid X$. Such contexts arise when the problem at hand does not require distributional predictions, or because Y is so complex that a distribution over Y is generally intractable to describe. The setting overlaps with that in the preceding two chapters: we assume data coming in pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, but now we consider functions $f : \mathcal{X} \to \mathbb{R}^k$ for some k, and we seek f minimizing the risk

$$L(f) := \mathbb{E}[\ell(f(X), Y)],$$

where $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ is a loss. Typically, ℓ is concomitant with a prediction function (or decoder) that transforms f into a prediction $\widehat{y}(f(x)) \in \mathcal{Y}$ of y.

Binary and multiclass classification motivate this approach. In the former, take $Y \in \{-1, 1\}$, $f : \mathcal{X} \to \mathbb{R}$ with prediction $\widehat{y}(s) = \operatorname{sign}(s)$, setting $\operatorname{sign}(0) = 1$ to break ties. For multiclass classification, take $Y \in \{1, \ldots, k\}$ and $f : \mathcal{X} \to \mathbb{R}^k$ with prediction $\widehat{y}(f(x)) = \operatorname{argmax}_j f_j(x)$ (breaking ties arbitrarily). In both cases, the zero-one error for label y and associated risk become

$$\ell_{0\text{-}1}(s, y) = \mathbf{1}\left\{\widehat{y}(s) \neq y\right\} \quad \text{and} \quad L(f) = \mathbb{P}(\widehat{y}(f(X)) \neq Y).$$

(For the multiclass case, we sometimes use the conservative score-based form $\mathbf{1}\{s_y \leq \max_{j \neq y} s_j\}$.)

In many of the cases we consider in this chapter, the loss ℓ is hard to minimize: it is non-convex, non-differentiable, NP-hard to minimize in general, and gradients provide no information to allow searching for a predictive function. As a consequence, we take a different tack, where we seek surrogate losses $\varphi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ admitting nicer properties—such as convexity and smoothness—such that minimizing φ implies minimizing the original loss ℓ. To define our desiderata, let

$$L_\varphi(f) := \mathbb{E}[\varphi(f(X), Y)]$$

be the surrogate risk of a function f, and define the *Bayes risk* (for ℓ and φ) as the minimal possible expected loss,

$$L^\star := \inf_f \mathbb{E}[\ell(f(X), Y)] \quad \text{and} \quad L_\varphi^\star := \inf_f \mathbb{E}[\varphi(f(X), Y)],$$

where the infima are over all measurable functions. We then seek *surrogate risk consistency*: a population-level guarantee that minimizing L_φ guarantees minimizing L.

Definition 16.1 (Surrogate risk consistency)**.** *The loss φ is* surrogate risk consistent *for the loss ℓ if for any sequence of functions f_n and any distribution on (X, Y),*

$$L_\varphi(f_n) \to L_\varphi^\star \quad \textit{implies} \quad L(f_n) \to L^\star.$$

This chapter develops the theory of surrogate risk consistency. The theory obtains its cleanest and most transparent form for binary classification with the zero-one error, but it extends beyond this. As we will show, the dualities we develop in Chapter 14 and connections with generalized entropies remain important; in many cases, any loss φ generating a suitably concave generalized entropy guarantees consistency. Deeper equivalence results between loss functions arise via this entropy and information-based perspective, and we close the chapter by developing these. Throughout, we elide measure-theoretic details, as they are secondary to the main arguments.

16.1 Surrogate risk consistency in binary classification

The first step, both because it offers the most complete solution and because the ideas are most transparent, is to consider consistency for binary classification. In this case, we take $Y \in \{-1, 1\}$, and consider margin-based losses, where we seek a predictor $f : \mathcal{X} \to \mathbb{R}$ predicting the correct sign of y. Thus, for $s \in \mathbb{R}$, we have the zero-one loss

$$\ell(s, y) = \ell_{0\text{-}1}(s, y) := \mathbf{1}\left\{sy \le 0\right\} \quad \text{and} \quad \varphi(s, y) = \phi(ys),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function acting on the *margin*

$$yf(x)$$

on examples (x, y). A large margin $yf(x) \gg 0$ thus corresponds to a correct prediction, while negative margin means that $f(x)$ has sign distinct from y.

Example 16.1.1 (Common convex loss functions)**:** Convex losses frequent in the literature on classification include the logistic loss

$$\phi(t) = \log(1 + e^{-t}),$$

the exponential loss

$$\phi(t) = e^{-t},$$

the squared error

$$\phi(t) = \frac{1}{2}(1 - t)^2,$$

and the hinge and squared hinge losses

$$\phi(t) = [1 - t]_+ \quad \text{and} \quad \phi(t) = [1 - t]_+^2 .$$

Each of these is nonnegative and satisfies $\phi'(0) < 0$. $\diamond$

The key to surrogate-risk consistency is to work conditionally on X: as we optimize over arbitrary measurable functions $f : \mathcal{X} \to \mathbb{R}$, we can choose $f(X)$ to pointwise optimally predict Y given X. To that end, let

$$\eta(x) := \mathbb{P}(Y = 1 \mid X = x)$$

be the conditional probability that $Y = 1$ given $X = x$, so that with the convention $\sign(0) = 1$,

$$L(f) = \mathbb{P}(\sign(f(X)) \neq Y) = \mathbb{E}\left[\eta(X)\mathbf{1}\left\{f(X) < 0\right\} + (1 - \eta(X))\mathbf{1}\left\{f(X) \geq 0\right\}\right]$$

and

$$L_\varphi(f) = \mathbb{E}\left[\eta(X)\phi(f(X)) + (1 - \eta(X))\phi(-f(X))\right].$$

Immediately, we see that $f^\star(x) = \eta(x) - \frac{1}{2}$ minimizes L. Defining the conditional risks

$$r(s, \eta) = \eta\mathbf{1}\left\{s < 0\right\} + (1 - \eta)\mathbf{1}\left\{s \geq 0\right\} \quad \text{and} \quad r_\phi(s, \eta) = \eta\phi(s) + (1 - \eta)\phi(-s),$$

we evidently have

$$L(f) = \mathbb{E}[r(f(X), \eta(X))] \quad \text{and} \quad L_\varphi(f) = \mathbb{E}[r_\phi(f(X), \eta(X))].$$

We seek relationships that relate $r(s, \eta)$ and $r_\phi(s, \eta)$ to one another to guarantee consistency.

Example 16.1.2 (Exponential loss)**:** AdaBoost and other boosting algorithms use the exponential loss $\phi(s) = e^{-s}$. For $\eta \in (0, 1)$ we have

$$\operatorname*{argmin}_s r_\phi(s, \eta) = \frac{1}{2}\log\frac{\eta}{1 - \eta} \quad \text{because} \quad \frac{\partial}{\partial s} r_\phi(s, \eta) = -\eta e^{-s} + (1 - \eta)e^s.$$

Then $f_\phi^\star(x) = \frac{1}{2}\log\frac{\eta(x)}{1-\eta(x)}$ minimizes $L_\varphi(f)$, where $f^\star$ may take infinite values if $\eta \in \{0, 1\}$. As $\sign(f_\phi^\star(x)) = \sign(2\eta(x) - 1)$, this is optimal for the zero-one loss as well. $\diamond$

To provide a more quantitative version of this pointwise identification of minimizers, we compare the conditional ϕ-risks. Intuitively, if the s minimizing $r_\phi(s, \eta)$ has the same sign as $2\eta - 1$, then we should obtain consistency. Contrapositively, if one cannot minimize r_ϕ accurately when we constrain s to have the incorrect sign, then any minimizer of L_φ should agree with the minimizers of L. Following this intuition, define the infimal conditional ϕ-risks as

$$r_\phi^\star(\eta) := \inf_s r_\phi(s, \eta) \quad \text{and} \quad r_\phi^{\text{wrong}}(\eta) := \inf_{s(\eta - 1/2) \leq 0} r_\phi(s, \eta).$$

We expect $r_\phi^\star(\eta) < r_\phi^{\text{wrong}}(\eta)$ for all $\eta \neq \frac{1}{2}$ to be sufficient to guarantee surrogate risk consistency for the zero-one loss—it is impossible to minimize $\mathbb{E}[r_\phi(f(X), \eta(X))]$ while predicting incorrect signs.

To make this rigorous, define the sub-optimality function $\Delta_\phi : [0, 1] \to \mathbb{R}$

$$\Delta_\phi(\delta) := r_\phi^{\text{wrong}}\left(\frac{1 + \delta}{2}\right) - r_\phi^\star\left(\frac{1 + \delta}{2}\right), \tag{16.1.1}$$

which measures the penalty of incorrect prediction as a function of the distance of the conditional probability $\mathbb{P}(Y = 1 \mid X = x)$ from $\frac{1}{2}$. We may now define

Definition 16.2 (Classification calibration)**.** *The margin-based loss ϕ is* classification calibrated *if $\Delta_\phi(\delta) > 0$ for all $\delta > 0$. Equivalently, for any $\eta \neq \frac{1}{2}$, we have $r_\phi^\star(\eta) < r_\phi^{\text{wrong}}(\eta)$.*

Unfortunately, the name *calibration* conflicts with that in Chapter 15, where it referred to probabilistic self-consistency, but both literatures have grown so extensive that it seems necessary to maintain both identificiations. Figure 16.1 illustrates the definition and margin-based losses, and the next three examples give explicit values for the sub-optimality function for different losses.

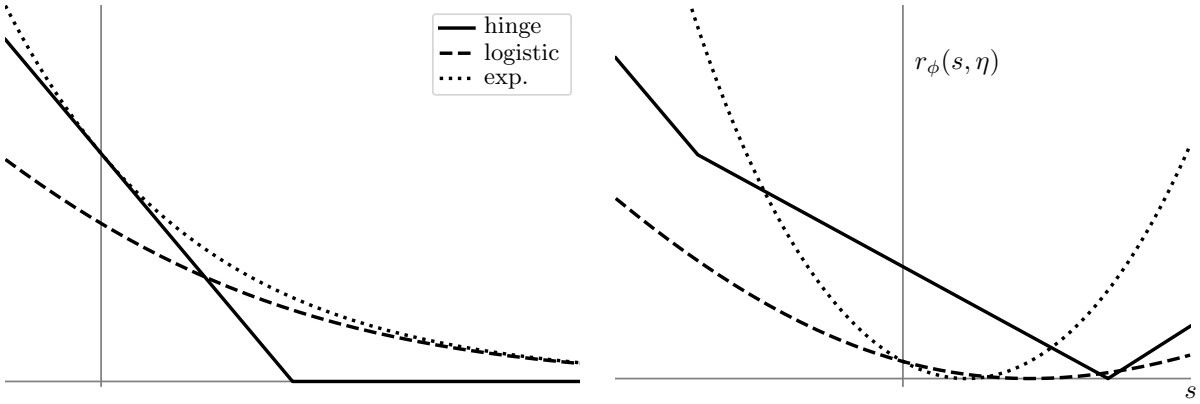


Figure 16.1. Left: the hinge $\phi(s) = [1 - s]_+$, logistic $\phi(s) = \log(1 + e^{-s})$, and exponential $\phi(s) = e^{-s}$ losses as a function of the margin s. Right: the conditional risk gaps $r_\phi(s, \eta) - r_\phi^\star(\eta)$ as a function of s for each, with $\eta = \mathbb{P}(Y = 1) = .65 > \frac{1}{2}$. Because $r'_\phi(0, \eta) = (2\eta - 1)\phi'(0) < 0$ for each, the minimizers $s_\phi^\star(\eta) = \operatorname{argmin}_s r_\phi(s, \eta) > 0$ have correct sign, and each loss is classification calibrated (Definition 16.2).

Example (Example 16.1.2 continued)**:** For the exponential loss, we have

$$r_\phi^{\text{wrong}}(\eta) = \inf_{s(2\eta-1)\le 0} \left\{\eta e^{-s} + (1 - \eta)e^s\right\} = e^0 = 1$$

while the unconstrained minimal conditional risk is

$$r_\phi^\star(\eta) = \eta\sqrt{\frac{1-\eta}{\eta}} + (1 - \eta)\sqrt{\frac{\eta}{1-\eta}} = 2\sqrt{\eta(1 - \eta)},$$

so that $\Delta_\phi(\delta) = 1 - \sqrt{1 - \delta^2} \ge \frac{1}{2}\delta^2$, where we use $\sqrt{a + x} \le \sqrt{a} + \frac{x}{2\sqrt{a}}$ for $a \ge 0, x \in \mathbb{R}$. $\diamond$

Example 16.1.3 (Hinge loss)**:** The hinge loss $\phi(s) = [1 - s]_+$ has conditional risk

$$r_\phi(s, \eta) = \eta [1 - s]_+ + (1 - \eta) [1 + s]_+ ,$$

whose unique minimizer (for $\eta \notin \{0, \frac{1}{2}, 1\}$) is $s(\eta) = \operatorname{sign}(2\eta - 1)$, while $s(\eta) = \operatorname{sign}(2\eta - 1)$ is a minimizer for all η. We thus have

$$r_\phi^\star(\eta) = 2\min\{\eta, 1 - \eta\} \quad \text{and} \quad r_\phi^{\text{wrong}}(\eta) = \eta + (1 - \eta) = 1$$

and obtain $\Delta_\phi(\delta) = 1 - \min\{1 + \delta, 1 - \delta\} = \delta$. $\diamond$

Example 16.1.4 (Squared error)**:** The squared-error surrogate $\varphi(s, y) = \frac{1}{2}(s - y)^2$ has margin-based form $\phi(t) = \frac{1}{2}(1 - t)^2$, and it satisfies

$$r_\phi^\star(\eta) = \inf_s \left\{\frac{\eta}{2}(s - 1)^2 + \frac{1 - \eta}{2}(s + 1)^2\right\} = 2\eta(1 - \eta)$$

because $s(\eta) = 2\eta - 1$ minimizes $r_\phi(s, \eta)$. On the other hand, $r_\phi^{\text{wrong}}(\eta) = \frac{1}{2}$ for all η, and so $\Delta_\phi(\delta) = \frac{1}{2} - \frac{1}{2}(1 - \delta)(1 + \delta) = \frac{1}{2}\delta^2$. $\diamond$

See Exercise 16.1 for more examples.

16.1.1 A general classification calibration result

In each of the preceding examples, the suboptimality function Δ_ϕ is convex. While this need not hold generally, we use Δ_ϕ to develop a *comparison inequality*: we construct a convex function ψ, strictly positive on $\mathbb{R}_{>0}$, such that

$$\psi(L(f) - L^\star) \le L_\phi(f) - L_\phi^\star, \tag{16.1.2}$$

where $L_\phi(f) = \mathbb{E}[\phi(Yf(X))]$. The construction proceeds via the biconjugate of Δ_ϕ, which (recall Lemma 14.1.1, or see Corollary C.2.2 in Appendix C.2) is the largest (closed) convex function below Δ_ϕ. Thus, we define the ψ-transform

$$\psi_\phi(\delta) := \Delta_\phi^{**}(\delta) \tag{16.1.3}$$

of the margin-based loss ϕ, which we call the *comparison function*. This is the largest convex function ψ satisfying $\psi \le \Delta_\phi$ (see Appendix C.2). Thus it certainly satisfies $\psi \ge 0$, $\psi(0) = 0$, and (as a one-dimensional closed convex function) is continuous; the main analytical question is whether it is strictly positive if Δ_ϕ is. The next theorem makes clear that this holds and characterizes (binary) classification calibration.

Theorem 16.1.5. *Let ϕ be a margin-based loss function and $\psi = \psi_\phi$ the associated ψ-transform. Then for any $f : \mathcal{X} \to \mathbb{R}$,*

$$\psi(L(f) - L^\star) \le L_\phi(f) - L_\phi^\star. \tag{16.1.4}$$

Moreover, if ϕ is nonnegative, the following three are equivalent:

(i) The loss ϕ is classification-calibrated (Definition 16.2).

(ii) $\psi(\delta) > 0$ for all $\delta > 0$.

(iii) For any sequence of measurable functions $f_n : \mathcal{X} \to \mathbb{R}$,

$$L_\phi(f_n) \to L_\phi^\star \quad \text{implies} \quad L(f_n) \to L^\star.$$

The proof of the theorem relies on the convex analysis we develop in the brief primer in Section 14.1.1 and Appendix B, so we defer it to Section 16.1.3.

Theorem 16.1.5 provides concrete conditions to guarantee surrogate consistency of a margin-based surrogate loss $\varphi(s, y) = \phi(ys)$. The quantitative comparison inequality (16.1.2) shows that small excess surrogate risk guarantees small excess zero-one error, though different loss functions yield different explicit upper bounds. Table 16.1 summarizes several; see also Exercise 16.1.

Loss	$\phi(s)$	Comparison function $\psi(\delta)$
Exponential	$\phi(s) = e^{-s}$	$\psi(\delta) = 1 - \sqrt{1-\delta^2} \ge \frac{1}{2}\delta^2$
Squared	$\phi(s, y) = \frac{1}{2}(s-y)^2$	$\psi(\delta) = \frac{1}{2}\delta^2$
Hinge	$\phi(s) = [1-s]_+$	$\psi(\delta) = \delta$
Logistic	$\phi(s) = \log(1+e^{-s})$	$\psi(\delta) = \log 2 - \frac{1}{2}\log\frac{4}{1-\delta^2} + \frac{\delta}{2}\log\frac{1+\delta}{1-\delta} \ge \frac{1}{2}\delta^2$

Table 16.1: Sub-optimality functions for common binary-classification losses.

16.1.2 Convex losses for binary classification

Most commonly, one takes a convex margin-based loss ϕ, because these admit computationally efficient minimization and fitting procedures. We can obtain a much simpler characterization of classification calibrated losses in this case. Intuitively, if $\phi'(0) < 0$, then any scalar minimizing the conditional ϕ-risk

$$r_\phi(s, \eta) = \eta\phi(s) + (1 - \eta)\phi(-s)$$

should necessarily satisfy $\sign(s) = \sign(2\eta - 1)$ (see the right plot of Figure 16.1). This is precisely the correct condition.

Theorem 16.1.6. *Let ϕ be convex. Then ϕ is classification calibrated if and only if $\phi'(0)$ exists and $\phi'(0) < 0$.*

Proof Let $\eta > \frac{1}{2}$ w.l.o.g. First, suppose that ϕ is differentiable at 0 and $\phi'(0) < 0$. Then $\dom \phi$ necessarily includes an interval around 0, and

$$r_\phi(s, \eta) = \eta\phi(s) + (1 - \eta)\phi(-s)$$

satisfies $r'_\phi(0, \eta) = (2\eta - 1)\phi'(0)$; if $\phi'(0) < 0$, this quantity is negative for $\eta > 1/2$, and so the minimizing $s(\eta) \in (0, \infty]$. Indeed, we have

$$r_\phi(s, \eta) \geq r_\phi(0, \eta) + (2\eta - 1)\phi'(0)s$$

by the first-order inequality for convexity, and the final term is strictly positive for $s < 0$. Moreover,

$$r_\phi(s, \eta) = r_\phi(0, \eta) + (2\eta - 1)\phi'(0)s + o(s) < r_\phi(0, \eta) = \phi(0)$$

for $s > 0$ but near 0. Thus $r_\phi^{\wrong}(\eta) = r_\phi(0, \eta) = \phi(0)$, while $r_\phi^\star(\eta) < \phi(0)$ for all $\eta \neq \frac{1}{2}$. In particular, $\Delta_\phi(\delta) > 0$ for $\delta > 0$ as desired.

To see the converse, we must prove that ϕ is differentiable at 0. Recall that a subgradient g_s of the function ϕ at $s \in \mathbb{R}$ is any g_s such that $\phi(t) \geq \phi(s) + g_s(t - s)$ for all $t \in \mathbb{R}$. Because ϕ is classification calibrated, its domain necessarily includes an interval around 0; thus that ϕ is subdifferentiable at 0 and the subgradient set $\partial\phi(0) \neq \varnothing$ (See Theorem B.2.4 or Proposition B.2.31 in Appendix B.) Let $g_1, g_2 \in \partial\phi(0)$. We show that both $g_1, g_2 < 0$ and $g_1 = g_2$, implying $\partial\phi(0) = \{g_1\}$ and so ϕ is differentiable at 0 (see Proposition B.2.35).

By convexity we have

$$\begin{aligned} r_\phi(s, \eta) &\geq \eta(\phi(0) + g_1 s) + (1 - \eta)(\phi(0) - g_2 s) \\ &= [\eta g_1 - (1 - \eta)g_2]\, s + \phi(0). \end{aligned} \tag{16.1.5}$$

We first show that $g_1 = g_2$, meaning that ϕ is differentiable. Without loss of generality, assume $g_1 > g_2$. Then for $\eta > 1/2$, we would have $\eta g_1 - (1 - \eta)g_2 > 0$, which would imply that

$$r_\phi(s, \eta) \geq \phi(0) \geq \inf_{s' \leq 0} \left\{\eta\phi(s') + (1 - \eta)\phi(-s')\right\} = r_\phi^{\wrong}(\eta),$$

for all $s \geq 0$ by (16.1.5). This contradicts classification calibration, which for $\eta > 1/2$ gives

$$\inf_s r_\phi(s, \eta) < \inf_{s \leq 0} r_\phi(s, \eta) = r_\phi^{\wrong}(\eta) \;\; \text{so} \;\; r_\phi^\star(\eta) = \inf_{s \geq 0} r_\phi(s, \eta) < r_\phi^{\wrong}(\eta).$$

We thus obtain $g_1 = g_2$, so ϕ has unique subgradient at $s = 0$ and is thus differentiable there.

Now that we know ϕ is differentiable at 0, consider

$$\eta\phi(s) + (1-\eta)\phi(-s) \geq (2\eta - 1)\phi'(0)s + \phi(0).$$

If $\phi'(0) \geq 0$, then for $s \geq 0$ and $\eta > 1/2$ the right hand side must be at least $\phi(0)$, which would again contradict classification calibration. □

Theorem 16.1.6 makes it easy to determine whether a convex margin-based loss is classification calibrated: simply take its derivative at 0. The logistic, exponential, squared, hinge, and squared-hinge losses (Example 16.1.1) each satisfy the theorem's condition.

16.1.3 Proof of Theorem 16.1.5

Throughout the proof we let $\Delta = \Delta_\phi$ and $\psi = \psi_\phi$. We begin with the first statement (16.1.4). By a calculation, the gap (for a fixed margin s) in conditional 0-1 risk is

$$\begin{aligned} r(s,\eta) - \inf_s r(s,\eta) &= \eta \mathbf{1}\left\{s < 0\right\} + (1-\eta)\mathbf{1}\left\{s \geq 0\right\} - \eta \mathbf{1}\left\{\eta \leq 1/2\right\} - (1-\eta)\mathbf{1}\left\{\eta > 1/2\right\} \\ &= \begin{cases} 0 & \text{if } \sign(s) = \sign(\eta - \frac{1}{2}) \\ \eta \vee (1-\eta) - \eta \wedge (1-\eta) = |2\eta - 1| & \text{if } \sign(s) \neq \sign(\eta - \frac{1}{2}). \end{cases} \end{aligned}$$

In particular, we obtain

$$L(f) - L^\star = \mathbb{E}\left[\mathbf{1}\left\{\sign(f(X)) \neq \sign(2\eta(X) - 1)\right\} |2\eta(X) - 1|\right]. \tag{16.1.6}$$

We use expression (16.1.6) to upper bound $L(f) - L^\star$ via the ϕ-risk, for which we use the ψ-transform (16.1.3). By Jensen's inequality,

$$\psi(L(f) - L^\star) \leq \mathbb{E}\left[\psi\big(\mathbf{1}\left\{\sign(f(X)) \neq \sign(2\eta(X) - 1)\right\} |2\eta(X) - 1|\big)\right].$$

We always have $\Delta(0) = 0$ as $r_\phi^{\rm wrong}(1/2) = \inf_{s(1-1)\leq 0} r_\phi(s, 1/2) = r_\phi^\star(1/2)$, and by construction $\Delta \geq 0$, so that its biconjugate satisfies $\psi \geq 0$ and $\psi(0) = 0$. Thus

$$\begin{aligned} \psi(L(f) - L^\star) &\leq \mathbb{E}\left[\psi\big(\mathbf{1}\left\{\sign(f(X)) \neq \sign(2\eta(X) - 1)\right\} |2\eta(X) - 1|\big)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left\{\sign(f(X)) \neq \sign(2\eta(X) - 1)\right\} \psi(|2\eta(X) - 1|)\right] \end{aligned} \tag{16.1.7}$$

Now we use the special structure of the suboptimality function. Recall that by construction, ψ is the largest convex minorant of Δ, making $\psi \leq \Delta$, and all other convex g satisfy $g \leq \psi$ (Corollary C.2.2 in Appendix C.2). Moreover, for any $s \in \mathbb{R}$

$$\begin{aligned} \mathbf{1}\left\{\sign(s) \neq \sign(2\eta - 1)\right\} \Delta(|2\eta - 1|) &= \mathbf{1}\left\{\sign(s) \neq \sign(2\eta - 1)\right\} \left[\inf_{s(2\eta-1)\leq 0} r_\phi(s,\eta) - r_\phi^\star(\eta)\right] \\ &\leq r_\phi(s,\eta) - r_\phi^\star(\eta), \end{aligned} \tag{16.1.8}$$

because $(1 + |2\eta - 1|)/2 = \max\{\eta, 1-\eta\}$. Combining inequalities (16.1.7) and (16.1.8), we see that

$$\begin{aligned} \psi(L(f) - L^\star) &\leq \mathbb{E}\left[\mathbf{1}\left\{\sign(f(X)) \neq \sign(2\eta(X) - 1)\right\} \Delta(|2\eta(X) - 1|)\right] \\ &\leq \mathbb{E}\left[r_\phi(f(X), \eta(X)) - r_\phi^\star(\eta(X))\right] \\ &= L_\phi(f) - L_\phi^\star, \end{aligned}$$

which is inequality (16.1.4).

The key to the claims (i)–(iii) in the second part of the theorem is the following lemma, which guarantees that so long as the sub-optimality function Δ_ϕ is positive, then so too is its biconjugate (16.1.3) ψ-transform.

Lemma 16.1.7. *The following hold:*

(a) The functions Δ and ψ are continuous on $[0, 1]$.

(b) We have $\Delta \geq 0$ and $\Delta(0) = 0$.

(c) If $\Delta(\delta) > 0$ for all $\delta > 0$, then $\psi(\delta) > 0$ for all $\delta > 0$.

Deferring the proof of the lemma temporarily, we turn to the equivalence of items (i)–(iii) in the theorem. Part (c) of the lemma demonstrates that classification calibration (i) implies (ii). Now, let (ii) hold: then if $L_\phi(f_n) \to L_\phi^\star$, we necessarily have $\psi(L(f_n) - L^\star) \to 0$ by inequality (16.1.4); this occurs if and only if $L(f_n) - L^\star \to 0$ by assumption, that is, the implication (iii) holds. Assume for the sake of contradiction that (iii) holds but (i) fails, that is, ϕ is not classification calibrated. Then there must exist $\eta < 1/2$ and a sequence $s_n \geq 0$ (i.e. a sequence of predictions with incorrect sign) satisfying

$$r_\phi(s_n, \eta) \to r_\phi^\star(\eta).$$

Construct the classification problem with a singleton $\mathcal{X} = \{x\}$, and set $\mathbb{P}(Y = 1) = \eta$. Then the sequence $f_n(x) = s_n$ satisfies $L_\phi(f_n) \to L_\phi^\star$ but the true 0-1 risk $L(f_n) \not\to L^\star$.

Finally, we prove the auxiliary lemma.

Proof of Lemma 16.1.7 The function $r_\phi^\star(\eta) = \inf_s\{\eta\phi(s) + (1-\eta)\phi(-s)\}$ for $\eta \in [0, 1]$, $r_\phi^\star(\eta) = -\infty$ otherwise, is concave, as it is the infimum of linear functions in η, and moreover, $-r_\phi^\star$ is closed convex. Because ϕ is nonnegative, $r_\phi^\star$ has domain $[0, 1]$. In one-dimension, closed convex functions are continuous on their domains (see Observation B.2.8 in Appendix B.2.2). Additionally,

$$r_\phi^{\rm wrong}(\eta) = \inf_{s \geq 0} \{\eta\phi(s) + (1-\eta)\phi(-s)\}$$

for $0 \leq \eta < \frac{1}{2}$, and by symmetry $r_\phi^{\rm wrong}(\eta) = r_\phi^{\rm wrong}(1-\eta)$, which is thus continuous on $\left[0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right]$. A parallel argument to that for $r_\phi^\star$ then shows that $r_\phi^{\rm wrong}$ is left- and right-continuous at $\frac{1}{2}$, and so

$$\Delta_\phi(\delta) = r_\phi^{\rm wrong}\left(\frac{1+\delta}{2}\right) - r_\phi^\star\left(\frac{1+\delta}{2}\right)$$

is continuous on $\delta \in [0, 1]$, where $\Delta_\phi(1 + \epsilon) = +\infty$. That $\psi = \Delta^{**}$ is continuous on $[0, 1]$ follows from the continuity of closed convex functions on their domains.

The nonnegativity of Δ is immediate, and to see that $\Delta(0) = 0$, note that

$$r_\phi^{\rm wrong}(1/2) := \inf_{s(1-1) \leq 0} r_\phi(s, 1/2) = \inf_s r_\phi(s, 1/2) = r_\phi^\star(1/2),$$

so $\Delta(0) = r_\phi^\star(1/2) - r_\phi^\star(1/2) = 0$.

The final claim of the lemma that $\psi = \Delta^{**}$ is strictly positive whenever Δ is is more subtle. For this, we leverage the following technical lemma (see Section 16.5.1 for the proof, which Figure 16.2 illustrates).

Lemma 16.1.8. *Let h be either (i) continuous on $[0, 1]$ or (ii) non-decreasing on $[0, 1]$, where $h(t) = +\infty$ for $t \notin [0, 1]$. If h satisfies $h(t) > 0$ for $t > 0$ and $h(0) = 0$, then $f(t) = h^{**}(t)$ satisfies $f(t) > 0$ for any $t > 0$.*

Lemma 16.1.8 immediately yields the final claim (c) of Lemma 16.1.7: we know that $\Delta(\delta) > 0$ and is continuous in δ.

16.2 General surrogate risk consistency

The conditional-risk approach for binary classification extends, with some care, to more general settings. We assume we have a loss $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ and associated surrogate $\varphi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$. To develop general consistency theory, let $\mathcal{P}_\mathcal{Y}$ denote the space of all probability distributions on $\mathcal{Y}$, and we replace the conditional probability $\eta = \eta(x)$ from the binary case with arbitrary conditional distributions $P_{Y|x} \in \mathcal{P}_\mathcal{Y}$. Thus, define the conditional (pointwise) risks on $\mathbb{R}^k \times \mathcal{P}_\mathcal{Y}$ by

$$r(s, P) = \mathbb{E}_P[\ell(s, Y)] \quad \text{and} \quad r_\varphi(s, P) = \mathbb{E}_P[\varphi(s, Y)],$$

and let $r^\star(P) = \inf_s r(s, P)$ denote the minimal conditional risk, and similarly for $r^\star_\varphi(P)$, when Y has distribution P. We may thus rewrite the risk functionals as

$$L(f) = \mathbb{E}[r(f(X), P_{Y|X})] \quad \text{and} \quad L_\varphi(f) = \mathbb{E}[r_\varphi(f(X), P_{Y|X})]. \tag{16.2.1}$$

This rewriting of the surrogate risk L_φ in terms of the conditional risks, as in Section 16.1, suggests a similar approach to surrogate risk consistency, where we control the differences in r and r_φ for any fixed $P \in \mathcal{P}_\mathcal{Y}$.

Example 16.2.1 (Multiclass classification)**:** Assume that $Y \in \{1, \ldots, k\}$. Common surrogates for the zero-one error $\ell_{0\text{-}1}(s, y) = \mathbf{1}\{s_y \le \max_{j \neq y} s_j\}$ of a score vector $s \in \mathbb{R}^k$ include the multiclass logistic loss

$$\varphi(s, y) = \log\left(1 + \sum_{j \neq y} e^{s_j - s_y}\right) = -s_y + \log\left(\sum_{j=1}^k e^{s_j}\right)$$

and, for ϕ a convex differentiable function with $\phi'(0) < 0$, the pairwise comparison loss

$$\varphi(s, y) = \sum_{j \neq y} \phi(s_y - s_j).$$

We will see these are both consistent surrogate losses. $\diamond$

16.2.1 Uniform calibration

Following the pointwise development in Section 16.1 via the expansion (16.2.1), for $\epsilon \ge 0$ define the *suboptimality gap function*

$$\Delta_\varphi(\epsilon, P) := \inf_{s \in \mathbb{R}^k} \left\{r_\varphi(s, P) - r^\star_\varphi(P) \mid r(s, P) - r^\star(P) \ge \epsilon\right\}, \tag{16.2.2}$$

which measures the penalty in surrogate risk minimization when we enforce that the true loss may be minimized to accuracy only ϵ. We can then define the *uniform suboptimality function*

$$\Delta_\varphi(\epsilon) := \inf_{s \in \mathbb{R}^k, P \in \mathcal{P}_\mathcal{Y}} \left\{r_\varphi(s, P) - r^\star_\varphi(P) \mid r(s, P) - r^\star(P) \ge \epsilon\right\}. \tag{16.2.3}$$

(Exercise 16.9 shows how this relates to r_ϕ for binary classification.) We can then make the following definition, which analogizes Definition 16.2.

Definition 16.3 (Uniform calibration)**.** *The surrogate φ is* uniformly calibrated *for the loss ℓ if $\Delta_\varphi(\epsilon) > 0$ for all $\epsilon > 0$.*

As in the binary case, letting $\Delta_\varphi^{**}(\epsilon)$ be the (convex) biconjugate of Δ_φ, we have the following proposition, which analogizes Theorem 16.1.5.

Proposition 16.2.2. *For any measurable $f : \mathcal{X} \to \mathbb{R}^k$,*

$$\Delta_\varphi^{**}(L(f) - L^\star) \le L_\varphi(f) - L_\varphi^\star.$$

Additionally, if φ is uniformly calibrated, then $\Delta_\varphi^{**}(\epsilon) > 0$ for all $\epsilon > 0$, and $L_\varphi(f_n) \to L_\varphi^\star$ implies that $L(f_n) \to L^\star$.

Proof Let $\psi = \Delta_\varphi^{**}$ for shorthand. Then by Jensen's inequality,

$$\begin{aligned}\psi(L(f) - L^\star) = \psi(\mathbb{E}[r(f(X), P_{Y|X}) - r^\star(P_{Y|X})]) &\le \mathbb{E}\left[\psi(r(f(X), P_{Y|X}) - r^\star(P_{Y|X}))\right] \\ &\le \mathbb{E}[\Delta_\varphi(r(f(X), P_{Y|X}) - r^\star(P_{Y|X}))]\end{aligned}$$

because $\psi \le \Delta_\varphi$. By definition, for any $P \in \mathcal{P}_\mathcal{Y}$ and $f(x) \in \mathbb{R}^k$, we have

$$\begin{aligned}\Delta_\varphi\left(r(f(x), P) - r^\star(P)\right) &= \inf_{s \in \mathbb{R}^k} \left\{r_\varphi(s, P) - r_\varphi^\star(P) \mid r(s, P) - r^\star(P) \ge r(f(x), P) - r^\star(P)\right\} \\ &\le r_\varphi(f(x), P) - r_\varphi^\star(P).\end{aligned}$$

Substituting into the preceding display, we obtain

$$\psi(L(f) - L^\star) \le \mathbb{E}\left[r_\varphi(f(X), P_{Y|X}) - r_\varphi^\star(P_{Y|X})\right] = L_\varphi(f) - L_\varphi^\star,$$

giving the first statement of the proposition. For the second statement, note that $\epsilon \mapsto \Delta_\varphi(\epsilon)$ is non-decreasing by construction, and uniform calibration implies it is strictly positive for $\epsilon > 0$. Lemma 16.1.8 implies its biconjugate is positive as well. □

Explicitly calculating the uniform suboptimality function Δ_φ is often non-trivial for many surrogate losses φ, but some cases admit explicit calculations.

Example 16.2.3 (Multiclass risk bounds)**:** Let $\varphi(s, y) = -s_y + \log(\sum_{j=1}^k e^{s_j})$ be the multiclass logistic loss and $\ell(s, y) = \ell_{0\text{-}1}(s, y)$ be the zero-one loss from Example 16.2.1. Then the suboptimality gap satisfies

$$\Delta_\varphi(\epsilon) \ge \frac{1}{2}\epsilon^2.$$

Proposition 16.2.2 then shows that $L(f) - L^\star \le \sqrt{2(L_\varphi(f) - L_\varphi^\star)}$ for any $f : \mathcal{X} \to \mathbb{R}^k$. Exercise 16.2 outlines a derivation of the gap bound $\Delta_\varphi(\epsilon) \ge \frac{1}{2}\epsilon^2$. ◇

16.2.2 Pointwise calibration

It can be challenging to verify uniform calibration, as it involves nested infima over potentially non-convex sets. Thus, it is frequently convenient to relax to loss calibration:

Definition 16.4. *The surrogate φ is* calibrated for the loss ℓ *if for all $P \in \mathcal{P}_\mathcal{Y}$ and $\epsilon > 0$,*

$$\Delta_\varphi(\epsilon, P) > 0.$$

Definition 16.4 relaxes the uniform calibration condition in Definition 16.3 to apply pointwise, that is, for each distribution P on $\mathcal{Y}$. While this precludes the cleanest guarantee that $\psi(L(f) - L^\star) \le L_\varphi(f) - L_\varphi^\star$, under a minor condition to address integrability issues (see also Exercise 16.10), it still is sufficient to guarantee surrogate risk consistency.

Theorem 16.2.4. *Let ℓ be bounded. Then φ is calibrated for the loss ℓ if and only if it is surrogate risk consistent for ℓ, that is, $L_\varphi(f_n) \to L_\varphi^\star$ implies that $L(f_n) \to L^\star$.*

Proof We prove the implication that φ is calibrated implies that it is surrogate risk consistent; the converse is an essentially immediate exercise by considering distributions supported on a single point x. Let $B < \infty$ be the bound on ℓ, so we may assume that $r(s, P) - r^\star(P) \le B$. Let f_n be a sequence of functions satisfying $L_\varphi(f_n) \to L_\varphi^\star$ and let $\epsilon > 0$ be arbitrary. Let $P_x = P(Y \in \cdot \mid X = x)$ be the conditional distribution of Y given $X = x$, and for shorthand, define

$$\delta_n(x) := r(f_n(x), P_x) - r^\star(P_x) \quad \text{and} \quad A_{n,\epsilon} := \{x \in \mathcal{X} \mid \delta_n(x) \ge \epsilon\}.$$

so that $\delta_n(x) \in [0, B]$. Then the risk gap $L(f_n) - L^\star$ satisfies

$$L(f_n) - L^\star = \mathbb{E}[\delta_n(X)] = \mathbb{E}[\mathbf{1}\left\{X \in A_{n,\epsilon}\right\} \delta_n(X)] + \underbrace{\mathbb{E}[\mathbf{1}\left\{X \notin A_{n,\epsilon}\right\} \delta_n(X)]}_{\le \epsilon},$$

where we have used that $\delta_n(x) \le \epsilon$ on $A_{n,\epsilon}^c$ by definition. Consider the first expectation. Let $\gamma > 0$ be otherwise arbitrary, and note that

$$\begin{aligned}
\mathbf{1}\left\{x \in A_{n,\epsilon}\right\} = \mathbf{1}\left\{\delta_n(x) \ge \epsilon\right\} &\le \frac{r_\varphi(f_n(x), P_x) - r_\varphi^\star(P_x)}{\Delta_\varphi(\epsilon, P_x)} \mathbf{1}\left\{\Delta_\varphi(\epsilon, P_x) > \gamma\right\} + \mathbf{1}\left\{\Delta_\varphi(\epsilon, P_x) \le \gamma\right\} \\
&\le \frac{r_\varphi(f_n(x), P_x) - r_\varphi^\star(P_x)}{\gamma} + \mathbf{1}\left\{\Delta_\varphi(\epsilon, P_x) \le \gamma\right\}.
\end{aligned}$$

In particular, as $\delta_n(x) \in [0, B]$, we have

$$\begin{aligned}
\mathbb{E}\left[\delta_n(X) \mathbf{1}\left\{X \in A_{n,\epsilon}\right\}\right] &\le B \mathbb{E}[\mathbf{1}\left\{X \in A_{n,\epsilon}\right\}] \\
&\le B\left[\frac{\mathbb{E}[r_\varphi(f_n(X), P_X) - r_\varphi^\star(P_X)]}{\gamma} + \mathbb{P}(\Delta_\varphi(\epsilon, P_X) \le \gamma)\right] \\
&= B\left[\frac{L_\varphi(f_n) - L_\varphi^\star}{\gamma} + \mathbb{P}(\Delta_\varphi(\epsilon, P_X) \le \gamma)\right]
\end{aligned}$$

for $\gamma > 0$. By assumption, the first term satisfies $L_\varphi(f_n) - L_\varphi^\star \to 0$, while the continuity of probability measures guarantees that $\lim_{\gamma \downarrow 0} \mathbb{P}(\Delta_\varphi(\epsilon, P_X) \le \gamma) = \mathbb{P}(\Delta_\varphi(\epsilon, P_X) = 0) = 0$. In particular, $\mathbb{E}[\mathbf{1}\left\{X \in A_{n,\epsilon}\right\} \delta_n(X)] \to 0$, and so $L(f_n) - L^\star \le \epsilon + o(1)$ as $n \to \infty$. As $\epsilon > 0$ was arbitrary, this completes the proof. □

16.2.3 Examples: multiclass surrogate risk consistency

We provide several examples of the pointwise calibration condition in Definition 16.4 for multiclass classification, where $Y \in \{1, \ldots, k\}$ represents one of k classes, predictors $f : \mathcal{X} \to \mathbb{R}^k$, and we investigate consistency for the zero-one loss. While there exists no surrogate risk consistency formulation as elegant as Theorem 16.1.5 (for binary classification), we still develop several general

results. For shorthand, we identify distributions on Y with vectors $p \in \Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p \rangle = 1\}$, so that $r(s, p) = \sum_{y=1}^k p_y \mathbf{1}\{s_y \le \max_{j \neq y} s_j\}$.

We first revisit Example 16.2.1 and consider pairwise margin-based losses of the form

$$\varphi(s, y) = \sum_{j=1}^k \phi(s_y - s_j), \tag{16.2.4}$$

for $\phi : \mathbb{R} \to \mathbb{R}$ non-increasing. The intuition mimics the binary classification calibration case: to minimize φ, we expect $s_y \gg s_j$ for $j \neq y$. Incorporating probabilities p, the following lemma formalizes this intuition:

Lemma 16.2.5. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be non-increasing, satisfy $\phi(t) < \phi(-t)$ for all $t > 0$, and the surrogate φ take the pairwise form (16.2.4). Let $p \in \Delta_k$ and $s \in \mathbb{R}^k$ satisfy $r_\varphi(s, p) = r_\varphi^\star(p)$. If $p_i > p_j$, then $s_i \ge s_j$, and if ϕ is differentiable with $\phi'(0) < 0$, then $s_i > s_j$.*

Proof Without loss of generality, we take $i = 1$ and $j = 2$, so that $p_1 > p_2$. Let $s \in \mathbb{R}^k$ and $s' = (s_2, s_1, s_3, \ldots, s_k)$ be s with its first two entries flipped. Then

$$\begin{aligned}
r_\varphi(s, p) - r_\varphi(s', p) &= p_1 \phi(s_1 - s_2) + p_1 \sum_{j \ge 3} \phi(s_1 - s_j) + p_2 \phi(s_2 - s_1) + p_2 \sum_{j \ge 3} \phi(s_2 - s_j) \\
&\quad - \left(p_2 \phi(s_1 - s_2) + p_2 \sum_{j \ge 3} \phi(s_1 - s_j) + p_1 \phi(s_2 - s_1) + p_1 \sum_{j \ge 3} \phi(s_2 - s_j) \right) \\
&= (p_1 - p_2) \Big(\phi(s_1 - s_2) - \phi(s_2 - s_1) + \sum_{j \ge 3} \left(\phi(s_1 - s_j) - \phi(s_2 - s_j) \right) \Big).
\end{aligned}$$

If $s_1 < s_2$, then $\phi(s_1 - s_j) \ge \phi(s_2 - s_j)$ and $\phi(s_1 - s_2) > \phi(s_2 - s_1)$ by assumption, so that if $p_1 > p_2$, we must have $r_\varphi(s, p) - r_\varphi(s', p) > (p_1 - p_2) \cdot (0 + 0) = 0$, contradicting that $r_\varphi(s, p) = r_\varphi^\star(p)$.

When ϕ is differentiable, if s minimizes $r_\varphi(s, p)$, then $\nabla r_\varphi(s, p) = 0$, and in particular,

$$0 = p_1 \sum_{j=1}^k \phi'(s_1 - s_j) - \sum_{j=1}^k p_j \phi'(s_j - s_1) \quad \text{and} \quad 0 = p_2 \sum_{j=1}^k \phi'(s_2 - s_j) - \sum_{j=1}^k p_j \phi'(s_j - s_2).$$

Assuming for the sake of contradiction that $s_1 = s_2 = t \in \mathbb{R}$, the preceding equalities imply

$$0 = (p_1 - p_2) \sum_{j=1}^k \phi'(t - s_j) = (p_1 - p_2) \left[\phi'(0) + \phi'(0) + \sum_{j=3}^k \phi'(t - s_j) \right].$$

Because ϕ is non-increasing, $\phi' \le 0$, while $\phi'(0) < 0$, so this is absurd if $p_1 > p_2$. □

Lemma 16.2.5 implies the following sufficient conditions for a convex margin-based loss to be surrogate risk consistent. The proposition assumes that ϕ is differentiable; without differentiability, the result may fail (see Exercise 16.11).

Proposition 16.2.6. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be non-increasing, convex, and differentiable with $\phi'(0) < 0$, and let the surrogate φ take the pairwise form (16.2.4). Then φ is surrogate risk consistent.*

Proof We use Theorem 16.2.4. Fix $p \in \Delta_k$ and without loss of generality assume that $p_1 > p_2$; by Theorem 16.2.4, it is then enough to show that

$$\inf_s \left\{r_\varphi(s,p) - r_\varphi^\star(p) \mid s_1 \le s_2\right\} > 0.$$

Let $s^{(n)}$ satisfy $r_\varphi(s^{(n)},p) \to r_\varphi^\star(p)$. First we claim that if p has strictly positive entries, the infimum $r_\varphi^\star(p)$ is attained. Indeed, if $p_y > 0$, then $\liminf_n(s_y^{(n)} - s_j^{(n)}) > -\infty$ for each $j \neq y$, as otherwise $\limsup_n \phi(s_y^{(n)} - s_j^{(n)}) = +\infty$; accordingly $s_y^{(n)} - s_j^{(n)}$ is bounded along the sequence for each pair y, j. Because the loss (16.2.4) is invariant to shifts, we can assume $s^{(n)}$ is bounded and has a limit $s^{(\infty)}$ minimizing $r_\varphi(s,p)$, so Lemma 16.2.5 gives the result.

Now consider a vector p with zero elements $J = \{j \mid p_j = 0\}$. Then w.l.o.g. we may assume $s_j^{(n)} = -\infty$ for each $j \in J$, as this minimizes each term $\phi(s_y - s_j)$ for $j \in J$. Repeat the first part of the argument restricted to indices not in J. (If only $p_1 > 0$, then $p_1 = 1$ and the result is trivial.) □

One can demonstrate other variants of Proposition 16.2.6 using similar loss functions. For example, Exercise 16.12 outlines the proof of the following consistency guarantee.

Proposition 16.2.7. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be non-increasing, convex, and differentiable with $\phi'(0) < 0$. Let the surrogate*

$$\varphi(s,y) := \phi(s_y) + \sum_{j \neq y} \phi(-s_j).$$

Then φ is surrogate risk consistent.

16.3 Generalized entropies and surrogate risk consistency

Section 14.3.1 develops exact equivalences between a proper loss ℓ and a convex surrogate φ. Relaxing this exactness to require only surrogate risk consistency, we can provide a parallel development here, showing how to translate general losses into a generalized entropy, and from any generalized entropy, we can construct a surrogate loss that (often) is guaranteed to be surrogate risk consistent. (Equality of entropies will play a role in more nuanced notions of consistency that Section 16.4 explores.)

16.3.1 Multiclass classification and strictly concave entropies

For multiclass classification problems, recall that an entropy $H : \Delta_k \to \underline{\mathbb{R}}$ is any (upper semi-continuous) concave function H with $\operatorname{dom} H \supset \operatorname{relint} \Delta_k$. Following the notation of Chapter 14, $\Omega(p) = -H(p)$ denotes the negative entropy and is closed convex function. The convex conjugate $\Omega^*(s) = \sup_p\{\langle p, s\rangle - \Omega(p)\}$ specializes the surrogate (14.3.3) to

$$\varphi(s,y) := -s_y + \Omega^*(s), \tag{16.3.1}$$

which generates the entropy H, as

$$\inf_s \mathbb{E}_p[\varphi(s,Y)] = \inf_s \left\{\Omega^*(s) - \langle p, s\rangle\right\} = -\Omega^{**}(p) = -\Omega(p) = H(p).$$

In distinction from Chapter 14, we no longer concern ourselves with proper losses and so the original loss ℓ need not act on probabilities, instead acting simply on vectors. We still, however, can recover any generalized entropy with a convex loss, so that given ℓ and associated entropy $H_\ell(p) = \inf_s \mathbb{E}_p[\ell(s,Y)]$, we can construct a convex surrogate φ with $H_\varphi = H_\ell$.

Proposition 16.3.1. *Let $H : \Delta_k \to \underline{\mathbb{R}}$ be closed concave and $\Omega = -H$. Then the losses $\varphi(s, y) = -s_y + \Omega^*(s)$ are closed, convex, and satisfy*

$$H_\varphi(p) := \inf_{s \in \mathbb{R}^k} \mathbb{E}_p[\varphi(s, Y)] = H(p).$$

The construction (16.3.1) is privileged, as from it we can derive desirable properties of the convex loss φ itself, and whenever H is symmetric and appropriately regular, the loss φ will be surrogate risk consistent for the zero-one error. Specializing Definition 16.4 to the classification case yields

Definition 16.5. *The loss $\varphi : \mathbb{R}^k \times [k] \to \overline{\mathbb{R}}$ is* classification calibrated for the zero-one loss *if for any $p \in \Delta_k$ and y such that $p_y < \max_j p_j$,*

$$\inf_s \left\{ \sum_{j=1}^k p_j \varphi(s, j) \right\} < \inf_s \left\{ \sum_{j=1}^k p_j \varphi(s, j) \mid s_y \geq \max_j s_j \right\}.$$

Theorem 16.2.4 shows that Definition 16.5 is equivalent to surrogate risk consistency.

The generalized entropy H provides a convenient route to determining the classification calibration of the surrogate φ from the construction (16.3.1). One straightforward sufficient condition is that H have curvature, so we recall uniformly convex functions (see Appendix C.1.2, equation (C.1.5)):

Definition 16.6. *A convex function $f : \mathbb{R}^k \to \overline{\mathbb{R}}$ is* $(\lambda, \kappa, \|\cdot\|)$-uniformly convex *over $C \subset \mathbb{R}^k$ if it is closed and for all $t \in [0, 1]$ and $x_1, x_2 \in C$,*

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) - \frac{\lambda}{2} t(1-t) \|x_1 - x_2\|^\kappa \left[(1-t)^{\kappa-1} + t^{\kappa-1} \right].$$

Uniform convexity weakens strong convexity (which corresponds to $\kappa = 2$), and it is sufficient to guarantee the classification calibration of the associated loss φ:

Theorem 16.3.2. *Let H be closed concave, symmetric, and have domain $\operatorname{dom} H = \Delta_k$. Let φ have the definition (16.3.1). Assume that H is strictly concave and either*

(i) for each closed convex set C and $p \in \Delta_k$, $\inf_{s \in C} \sum_{j=1}^k p_j \varphi(s, j)$ is attained,

(ii) H is uniformly concave.

Then φ is classification calibrated.

This result relies on the stability properties of the minimizers of the loss φ as well as the differentiability properties of the conjugate Ω^* of the negative entropy $\Omega = -H$. We therefore postpone the proof to Section 16.3.3, instead using it to give two concrete examples of consistent losses.

Example 16.3.3 (Multiclass logistic regression)**:** The multiclass logistic loss (Example 16.2.1), $\varphi(s, y) = -s_y + \log(\sum_{j=1}^k e^{s_j})$, has entropy $H_\varphi(p) = -\sum_{j=1}^k p_j \log p_j$, the classical entropy of Y. (Recall Examples 14.3.4 and 14.3.6.) The negative entropy $\Omega = -H$ is strongly convex over Δ_k by Pinsker's inequality, so it is classification calibrated. $\diamond$

Example 16.3.4 (Norm-based losses)**:** For $1 < \kappa < \infty$, take a loss of the form

$$\varphi(s, y) = -s_y + \frac{1}{\kappa} \|s\|_\kappa^\kappa,$$

which for the conjugate $\kappa^* = \frac{\kappa}{\kappa - 1} < \infty$ satisfies

$$H(p) = \inf_s \mathbb{E}_p[\varphi(s, Y)] = -\frac{1}{\kappa^*} \|p\|_{\kappa^*}^{\kappa^*}.$$

This entropy is strictly concave, and the infimum is always attained, so Theorem 16.3.2 gives that φ is classification calibrated. $\diamond$

16.3.2 Conjugate surrogates for a discrete loss

We have seen in Section 14.3 that the surrogate $\varphi(s, y) = -s^\top y + \Omega^*(s)$ provides a proper surrogate when Ω is a Legendre entropy, and there is a one-to-one correspondence between $s^\star = \operatorname{argmin}_s \mathbb{E}_P[\varphi(s, Y)]$ and P. Here, we consider a variant of this when we construct the surrogate directly from the generalized entropy H_ℓ associated with a discrete loss ℓ. We will develop a slight generalization of the pure classification-based loss, and assume we have a feature mapping $\tau : \mathcal{Y} \to \mathbb{R}^m$ representing $y \in \mathcal{Y}$ as an element of $\mathbb{R}^m$, and we assume the loss of a prediction $\hat{y} \in \mathcal{Y}$ for a correct value $y \in \mathcal{Y}$ is

$$\ell(\hat{y}, y) = \tau(\hat{y})^\top A \tau(y) + c(y) \tag{16.3.2}$$

where $A \in \mathbb{R}^{m \times m}$ is a matrix and $c : \mathcal{Y} \to \mathbb{R}$.

To alleviate the abstractness of the formulation (16.3.2), we rewrite a few examples in its form.

Example 16.3.5 (Multiclass classification)**:** Let the label $y \in \{1, \ldots, k\}$, and consider the zero-one loss $\ell_{0\text{-}1}(\hat{y}, y) = \mathbf{1}\{y \neq \hat{y}\}$. Then we may take $\tau(y) = e_y \in \mathbb{R}^k$, the yth standard basis vector, and let $A = \mathbf{1}\mathbf{1}^\top - I_k$ be the matrix of all-ones except on the diagonal. $\diamond$

Example 16.3.6 (Cost-weighted multiclass classification)**:** Let the label $y \in \{1, \ldots, k\}$, and let $w_{ij} \geq 0$ be the loss for predicting class i when the true class is j (and $w_{ii} = 0$). Then we take $\tau(y) = e_y$, the yth standard basis vector, and $A \in \mathbb{R}^{k \times k}$ with entries $A_{ij} = w_{ij}$, which gives $\ell(\hat{y}, y) = \tau(\hat{y})^\top A \tau(y) = w_{\hat{y}, y}$ in the formulation (16.3.2). $\diamond$

Let us consider the construction of the generalized entropy associated to the loss (16.3.2), where for simplicity we (w.l.o.g.) assume $c = 0$. Then for a distribution P on $Y \in \mathcal{Y}$, we define the entropy

$$H_\ell(P) := \min_{y \in \mathcal{Y}} \mathbb{E}_P \left[\tau(y)^\top A \tau(Y) \right].$$

Let $\Delta_\mathcal{Y}$ be the set of probability vectors (or p.m.f.s) on $\mathcal{Y}$, where $p \in \Delta_\mathcal{Y}$ assigns probability p_y to $y \in \mathcal{Y}$. Recalling Section 14.2.3, where we considered vector-valued targets y, we generalize the mean mapping by defining *marginal polytope* associated with the the features τ of y by

$$\mathcal{M} := \operatorname{Conv}(\{\tau(y)\}_{y \in \mathcal{Y}}) = \left\{ \sum_{y \in \mathcal{Y}} p_y \tau(y) \mid p \in \Delta_\mathcal{Y} \right\}$$

and the mean mapping $\mu : \Delta_\mathcal{Y} \to \mathcal{M}$ by

$$\mu(P) := \mathbb{E}_P[\tau(Y)] = \sum_{y \in \mathcal{Y}} p_y \tau(y)$$

when P has p.m.f. p. Then we immediately see that

$$H_\ell(P) = \min_{y \in \mathcal{Y}} \tau(y)^\top A \mu(P) = \inf_{\nu \in \mathcal{M}} \nu^\top A \mu(P),$$

so that H_ℓ is equivalent for all distributions P with identical mean $\mu(P) = \mathbb{E}_P[\tau(Y)]$. With some abuse of notation, we can therefore define the negative entropy mapping by

$$\Omega(\mu) := -\min_{y \in \mathcal{Y}} \tau(y)^\top A \mu + \mathbf{I}_{\mathcal{M}}(\mu),$$

where $\mathbf{I}_{\mathcal{M}}(\mu) = 0$ if $\mu \in \mathcal{M}$ and $+\infty$ otherwise, which satisfies $\Omega(\nu) = -H_\ell(P)$ for any $\nu \in \mathcal{M}$ satisfying $\nu = \mu(P)$; Ω is closed convex.

With this construction in hand, the immediate generalization of the surrogate (16.3.1) is

$$\varphi(s, y) := -s^\top \tau(y) + \Omega^*(s), \tag{16.3.3}$$

which is closed and convex. Following the link construction (14.3.2), given s, we make prediction

$$\mathsf{pred}(s) = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ \tau(y)^\top s \right\}. \tag{16.3.4}$$

For this surrogate, we observe that

$$\mathbb{E}_p[\varphi(s, Y)] = -s^\top \mu(p) + \Omega^*(s),$$

meaning we also obtain the following analogue of Proposition 16.3.1:

$$\inf_s \mathbb{E}_p[\varphi(s, Y)] = \inf_s \left\{ -s^\top \mu(p) + \Omega^*(s) \right\} = -\sup_s \left\{ s^\top \mu(p) - \Omega^*(s) \right\} = -\Omega(\mu(p)) = H_\ell(p).$$

The surrogate φ yields the same (generalized) entropy as the loss ℓ.

Example 16.3.7 (Multiclass classification)**:** For k-class classification with the zero-one loss $\ell_{0\text{-}1}(y, y') = \mathbf{1}\{y \neq y'\}$, we have $\tau(y) = e_y$ and take $A = \mathbf{1}\mathbf{1}^\top - I_k$ in the representation (16.3.2). Then for $p \in \Delta_k$,

$$H_\ell(p) = \inf_{q \in \Delta_k} q^\top (\mathbf{1}\mathbf{1}^\top - I_k) p = 1 - \sup_{q \in \Delta_k} q^\top p = 1 - \max_j p_j,$$

which of course we could have obtained by direct calculation. Exercise 16.4 asks you to show that for $\Omega(p) = \max_j p_j - 1$,

$$\Omega^*(s) = 1 + \max \left\{ s_{(1)} - 1, \frac{s_{(1)} + s_{(2)} - 1}{2}, \ldots, \frac{s_{(1)} + \cdots + s_{(k)} - 1}{k} \right\}, \tag{16.3.5}$$

where $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(k)}$ denotes the vector s sorted in decreasing order. Then the convex surrogate loss of the zero-one loss is

$$\varphi(s, y) = 1 - s_y + \max \left\{ s_{(1)} - 1, \frac{s_{(1)} + s_{(2)} - 1}{2}, \ldots, \frac{s_{(1)} + \cdots + s_{(k)} - 1}{k} \right\}.$$

In passing, we note that other losses *also* generate the 0-1 entropy $H(p) = 1 - \max_j p_j$. For example, letting $\mathcal{Z} = \{s \mid s^\top \mathbf{1} = 0\}$, the hinge-type loss

$$\varphi^{\text{hinge}}(s, y) := \sum_{j \neq y} [1 + s_j]_+ + \mathbf{I}_{\mathcal{Z}}(s)$$

satisfies $\inf_s \mathbb{E}_p[\varphi^{\text{hinge}}(s, Y)] = k(1 - \max_j p_j)$, where the restriction to vectors satisfying $\sum_j s_j = 0$ addresses scaling issues. (Exercise 16.4 asks you to prove this as well.) $\diamond$

We will require a qualification on the types of losses that we consider; while this does restrict somewhat the situations in which we demonstrate consistency, it is almost necessary (see Exercises 16.15 and 16.16):

Assumption A.16.1. *The loss ℓ is symmetric, i.e., $A = A^\top$, and satisfies the identifiability condition that if y minimizes $\mathbb{E}_P[\ell(y, Y)] = \mathbb{E}_P[\tau(y)^\top A \tau(Y)]$ then $P(Y = y) > 0$.*

For k-class multiclass classification problems with the zero-one loss as in Example 16.3.7, we see immediately that if $y \in \operatorname{argmin}_y \mathbb{E}_P[\ell_{0\text{-}1}(y, Y)]$, then $P(Y = y) \geq \frac{1}{k} > 0$.

The first question is why the surrogate (16.3.3) should be consistent for the discrete loss (16.3.2). To gain intuition, we present a few calculations before we move to the more sophisticated argument to come. Consider a $P \in \Delta_{\mathcal{Y}}$ and its induced mean $\mu = \mu(P)$. Consider a vector s minimizing $r_\varphi(s, P) = -s^\top \mu + \Omega^*(s)$ (recall (16.3.3)), so that $\Omega(\mu) - s^\top \mu + \Omega^*(s) = 0$. Using Fenchel-duality (Proposition C.2.3 in Appendix C.2.1) we necessarily have $s \in \partial\Omega(\mu)$. Now we make the unrealistic simplifying assumption that $\mu \in \operatorname{int} \mathcal{M}$. Then because $\Omega(\mu) = \max_{y \in \mathcal{Y}} -\tau(y)^\top A\mu + \mathbf{I}_{\mathcal{M}}(\mu)$

$$s \in \partial\Omega(\mu) = \operatorname{Conv}\{-A^\top \tau(y) \mid \tau(y)^\top A\mu = H_\ell(P)\} = \{-A^\top \nu \mid \nu^\top A\mu = H_\ell(P), \nu \in \mathcal{M}\}.$$

Letting $y^\star(P) = \operatorname{argmin}_y \mathbb{E}_P[\ell(y, Y)]$, the elements $\nu \in \mathcal{M}$ with $\nu^\top A\mu(P) = H_\ell(P)$ necessarily satisfy $\nu \in \operatorname{Conv}\{\tau(y)\}_{y \in y^\star(P)}$, so $s = -A^\top \sum_{y \in y^\star(P)} q_y \tau(y)$ for some vector $q \in \Delta_{y^\star(P)}$. Then the prediction function (16.3.4) becomes

$$\mathsf{pred}(s) \in \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ s^\top \tau(y) \right\} = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ -\tau(y)^\top A \sum_{y' \in y^\star(P)} q_{y'} \tau(y') \right\} \subset y^\star(P)$$

by Assumption A.16.1 on the label probabilities.

Of course, even in cases such as multiclass classification with $\tau(y) = e_y$, $\mathcal{M}$ has only a relative interior, so the construction does not quite work. Nonetheless, the basic structure of the argument works, and under the qualification Assumption A.16.1 imposes, the discrete surrogate loss is consistent. Thus we have the following theorem, whose proof we provide in Section 16.5.2.

Theorem 16.3.8. *Let Assumption A.16.1 hold for a loss ℓ of the form (16.3.2). Then $\varphi(s, y) = -s^\top \tau(y) + \Omega^*(s)$ is surrogate-risk consistent for ℓ.*

So under an identifiability assumption on the loss, the generalized entropy provides a general purpose construction that guarantees consistency. For certain special cases, such as the zero-one loss for multiclass classification with $A = \mathbf{1}\mathbf{1}^\top - I$ and $\tau(y) = e_y$, we can say more.

Example 16.3.9 (The sub-optimality gap for multiclass classification)**:** For the zero-one loss $\ell_{0\text{-}1}(s, y) = \mathbf{1}\{s_y \leq \max_{j \neq y} s_j\}$, the entropy $H_{\ell_{0\text{-}1}}(p) = 1 - \max_j p_j$. For the negative entropy $\Omega(p) = -H_{\ell_{0\text{-}1}}(p)$, whose conjugate Ω^* has explicit formula (16.3.5), the surrogate $\varphi(s, y) = -s^\top e_y + \Omega^*(s)$ satisfies

$$\mathbb{E}_p[\varphi(s, Y)] - \inf_s \mathbb{E}_p[\varphi(s, Y)] \geq \frac{1}{k} \left(\mathbb{E}_p[\ell_{0\text{-}1}(s, Y)] - \inf_s \mathbb{E}_p[\ell_{0\text{-}1}(s, Y)] \right)$$

for all $s \in \mathbb{R}^k$. This, of course, implies the uniform lower bound $\Delta_\varphi(\epsilon) \geq \frac{1}{k}\epsilon$ on the gap functional (16.2.3), and so for $L(f) = \mathbb{P}(\operatorname{argmax}_y f_y(X) \neq Y)$, we obtain

$$L(f) - L^\star \leq k \cdot \left(L_\varphi(f) - L_\varphi^\star \right)$$

for all $f : \mathcal{X} \to \mathbb{R}^k$ by Proposition 16.2.2. $\diamond$

Exercise 16.5 asks for a proof of these results.

16.3.3 Proof of Theorem 16.3.2

We will freely use the conjugate duality relationships we provide in Appendices C.1.2 and C.2.2. Let $\Omega = -H$ be the negative entropy, noting that $\Omega^*(s) = \sup_{p \in \Delta_k} \{\langle p, s \rangle + H(p)\} < \infty$, as suprema of closed concave functions over compact sets are attained, so $\dom \Omega^* = \mathbb{R}^k$. We begin with an intermediate lemma.

Lemma 16.3.10. *Under the conditions of Theorem 16.3.2, Ω^* is continuously differentiable on $\mathbb{R}^k$, satisfies $\Delta_k \supset \nabla\Omega^*(\mathbb{R}^k)$, and if $s_i \geq s_j$, then $p = \nabla\Omega^*(s)$ satisfies $p_i \geq p_j$.*

Proof That Ω^* is continuously differentiable follows immediately because Ω is strictly convex, as uniform convexity strengthens strict convexity. (See Propositions C.2.7 and C.2.8 in Appendix C.2.2.)

We show that if $s_i \geq s_j$ then $p_i \geq p_j$. We know that $p = \nabla\Omega^*(s) = \argmax_{p \in \Delta_k} \{\langle p, s \rangle - \Omega(p)\}$, and let us assume for the sake of contradiction that $p_i < p_j$. Then letting p' be the vector p with entries i and j swapped, we have

$$\Omega(p') = \Omega(p) \;\; \text{but} \;\; \Omega\Big(\frac{1}{2}(p' + p)\Big) < \frac{1}{2}\Omega(p') + \frac{1}{2}\Omega(p) = \Omega(p') = \Omega(p).$$

We also have

$$\langle s, p \rangle - \langle s, p' \rangle = (s_i - s_j)(p_i - p_j) \leq 0,$$

that is, $\langle s, p \rangle \leq \langle s, p' \rangle$. But then

$$\frac{1}{2}\langle s, p + p' \rangle - \Omega\Big(\frac{1}{2}p + \frac{1}{2}p'\Big) \geq \langle s, p \rangle - \Omega\Big(\frac{1}{2}p + \frac{1}{2}p'\Big) > \langle s, p \rangle - \Omega(p),$$

a contradiction to the assumed optimality that $p = \nabla\Omega^*(s)$. □

From this lemma, we obtain the following immediate observation:

Observation 16.3.11. *Under the conditions of Theorem 16.3.2, if $s^\star \in \mathbb{R}^k$ minimizes $\mathbb{E}_p[\varphi(s, Y)]$ then $p_i > p_j$ implies that $s_i^\star > s_j^\star$.*

Proof $s^\star$ minimizes $\mathbb{E}_p[\varphi(s, Y)]$ if and only if $p = \nabla\Omega^*(s^\star)$ by the Fenchel-Young inequality (14.1.3). Take the contrapositive of Lemma 16.3.10. □

We also require the following technical lemma, whose proof we defer temporarily:

Lemma 16.3.12. *Let f have Hölder-continuous gradient and assume $f^\star = \inf_u f(u) > -\infty$. Then $\nabla f(u_n) \to 0$ if $f(u_n) \to f^\star$.*

Now we may prove the theorem. Under assumption (i) of the theorem, that H is strictly concave and that the infimum $\inf_{s \in C} \mathbb{E}_p[\varphi(s, Y)]$ is always attained for closed convex C, the preceding observation immediately gives the result: if $p_i > p_j$, we must have

$$\inf_s \{\mathbb{E}_p[\varphi(s, Y)] \mid s_i \leq s_j\} > \inf_s \mathbb{E}_p[\varphi(s, Y)],$$

which implies classification calibration (Definition 16.5). Under assumption (ii) that Ω is uniformly convex, recall that $H(p) = \inf_s \{\sum_{j=1}^k p_j \varphi(s, j)\} > -\infty$ (because $\dom H = \Delta_k$), and let $p_i > p_j$. Let $s^{(m)}$ be any sequence such that $\sum_{j=1}^k p_j \varphi(s^{(m)}, j) \to H(p)$. Because Ω is uniformly convex, the

dual gradient $\nabla\Omega^*$ is Hölder continuous with $\operatorname{dom}\nabla\Omega^* = \mathbb{R}^k$ (Proposition C.2.7). Then because $\sum_{j=1}^k p_j\varphi(s,j) = -\langle p, s\rangle + \Omega^*(s)$, we obtain

$$\lim_{m\to\infty} \nabla\Omega^*(s^{(m)}) = p$$

by Lemma 16.3.12. If $s_i^{(m)} \geq s_j^{(m)}$, then Lemma 16.3.10 implies that $p^{(m)} = \nabla\Omega^*(s^{(m)})$ satisfies $p_i^{(m)} \geq p_j^{(m)}$, and so if $p_i > p_j$, it must be the case that eventually $s_i^{(m)} > s_j^{(m)}$. Moreover, Hölder continuity and symmetry of $\nabla\Omega^*$ show that

$$\liminf_m |s_i^{(m)} - s_j^{(m)}| = 0 \;\; \text{implies} \;\; \liminf_m |p_i^{(m)} - p_j^{(m)}| = 0,$$

and so if $p_i > p_j$, then we necessarily have $\liminf_m (s_i^{(m)} - s_j^{(m)}) > 0$. In particular, we have demonstrated that

$$\inf_s \left\{ \sum_{j=1}^k p_j\varphi(s,j) \mid s_i \leq s_j \right\} > H(p)$$

whenever $p_i > p_j$, which again implies classification calibration (Definition 16.5).

We finally return to the promised proof of Lemma 16.3.12.

Proof of Lemma 16.3.12 Let $\beta > 0$ be the Hölder constant of ∇f, so that for $\|\cdot\| = \|\cdot\|_2$ there exists $C < \infty$ such that $\|\nabla f(u) - \nabla f(v)\| \leq C\|u-v\|^\beta$ for all u, v. Then using that $h(t) = f(u+t(v-u))$ satisfies $h'(t) = \langle \nabla f(u+t(v-u)), v-u\rangle$ and $f(v) = h(1) = h(0) + \int_0^1 h'(t)dt$, we have

$$\begin{aligned} f(v) &= f(u) + \langle \nabla f(u), v-u\rangle + \int_0^1 \langle \nabla f(u+t(v-u)) - \nabla f(u), v-u\rangle dt \\ &\leq f(u) + \langle \nabla f(u), v-u\rangle + \int_0^1 Ct^\beta \|u-v\|^{\beta+1} dt = f(u) + \langle \nabla f(u), v-u\rangle + \frac{C}{\beta+1}\|u-v\|^{\beta+1}. \end{aligned}$$

For the sake of contradiction, let u_n satisfy $f(u_n) \to \inf_u f(u)$ and $g_n = \nabla f(u_n) \not\to 0$. Then there exists some $c > 0$ such that $\|g_n\| \geq c$ infinitely often; without loss of generality, assume this is the full sequence. Fixing $\delta > 0$ to be chosen, let $v_n = u_n - \delta g_n / \|g_n\|$, so that

$$f(v_n) \leq f(u_n) - \delta\|g_n\| + \frac{C}{\beta+1}\delta^{\beta+1} \leq f(u_n) + \delta\left(\frac{C}{\beta+1}\delta^\beta - c\right).$$

Take δ small enough that the last quantity is strictly negative to obtain that $\limsup_n f(v_n) < \liminf_n f(u_n) = f^\star$, a contradiction. □

16.4 Universal loss equivalence and entropy

Solving a statistical learning problem involves more than simply learning a (measurable) predictor $f : \mathcal{X} \to \mathbb{R}^k$—we frequently must estimate or learn a representation as well. Surrogate risk consistency typically requires minimizing over all measurable functions f, enabling us to focus on pointwise calculations (conditional on X), so they say little about the more elaborate process of jointly estimating a data representation and predictor. The connection between losses and entropies, however, allows us to exhibit deep equivalences.

We re-adopt the notation of the generalized entropies associated with the loss $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ we use in Section 14.1.3, so $H_\ell(Y \mid X) := \mathbb{E}[\inf_s \mathbb{E}[\ell(s, Y) \mid X]]$. The information (definition (14.1.6)) that X carries about Y is the amount that X reduces the expected loss in prediction Y,

$$I_\ell(X; Y) := H_\ell(Y) - H_\ell(Y \mid X).$$

We have seen examples in which entropies associated to different losses coincide (up to scaling), and Section 16.3 explicitly constructs surrogate risks whose entropy coincides with that of the original loss ℓ. For example, in binary classification, letting $p = P(Y = 1)$, the zero-one loss has entropy

$$H_{\ell_{0\text{-}1}}(P) = \inf_{s \in \mathbb{R}} P(sY \le 0) = 1 - \max_{y \in \{\pm 1\}} P(Y = y) = \min\{p, 1 - p\}$$

while the hinge loss $\phi(t) = [1 - t]_+$ satisfies

$$H_\phi(P) = \inf_{s \in \mathbb{R}} \left\{p \left[1 - s\right]_+ + (1 - p) \left[1 + s\right]_+\right\} = 2 \min\{p, 1 - p\}.$$

This evocative correspondence guarantees more than consistency.

To develop this, we adopt the language of quantization and data processing. We say that a quantizer of $X \in \mathcal{X}$ is any mapping $\mathsf{q} : \mathcal{X} \to \mathcal{Z}$ for some space $\mathcal{Z}$, thinking of such quantizers as a data representation, taking raw inputs x and transforming them into $z \in \mathcal{Z}$. Then for a prediction function $f : \mathcal{Z} \to \mathbb{R}^k$, we define the quantized and optimal quantized risks

$$L(f \mid \mathsf{q}) := \mathbb{E}[\ell(f(\mathsf{q}(X)), Y)] \quad \text{and} \quad L^\star(\mathsf{q}) := \inf_f L(f \mid \mathsf{q}).$$

Rewriting these in terms of entropies, we have $H_\ell(Y \mid \mathsf{q}(X)) = \mathbb{E}[\inf_s \mathbb{E}[\ell(s, Y) \mid \mathsf{q}(X)]] = L^\star(\mathsf{q})$, and

$$I_\ell(\mathsf{q}(X); Y) = H_\ell(Y) - H_\ell(Y \mid \mathsf{q}(X)) = H_\ell(Y) - L^\star(\mathsf{q}).$$

Then a quantizer q_1 outperforms q_2 when

$$L^\star(\mathsf{q}_1) < L^\star(\mathsf{q}_2) \quad \text{if and only if} \quad I_\ell(\mathsf{q}_1(X); Y) > I_\ell(\mathsf{q}_2(X); Y),$$

that is, $\mathsf{q}_1(X)$ carries more information about Y than $\mathsf{q}_2(X)$.

We can now ask for more than surrogate risk consistency, where in addition to consistency of φ for a given loss ℓ, we ask that using φ to choose a data representation (or quantizer) from a class Q of potential quantizers should be equivalent to using the original loss ℓ. We therefore define

Definition 16.7. *Losses ℓ_1 and ℓ_2 are* universally equivalent *if for all distributions on (X, Y) and all quantizers q_1 and q_2,*

$$I_{\ell_1}(\mathsf{q}_1(X); Y) \le I_{\ell_1}(\mathsf{q}_2(X); Y) \quad \textit{if and only if} \quad I_{\ell_2}(\mathsf{q}_1(X); Y) \le I_{\ell_2}(\mathsf{q}_2(X); Y).$$

Swapping the roles of q_1 and q_2 and taking contrapositives, Definition 16.7 is equivalent to the condition

$$I_{\ell_1}(\mathsf{q}_1(X); Y) < I_{\ell_1}(\mathsf{q}_2(X); Y) \quad \text{if and only if} \quad I_{\ell_2}(\mathsf{q}_1(X); Y) < I_{\ell_2}(\mathsf{q}_2(X); Y).$$

It is almost immediate that if H_{ℓ_1} and H_{ℓ_2} are generalized entropies associated with losses ℓ_1 and ℓ_2 equal to multiplicative constants, then the losses ℓ_1 and ℓ_2 are universally equivalent. The converse turns out to be true as well for classification problems with finite $\mathcal{Y}$, where $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$.

Theorem 16.4.1. *Let the multiclass losses ℓ_1 and ℓ_2 be bounded below and H_1 and H_2 be the associated generalized entropies. Then ℓ_1 and ℓ_2 are universally equivalent if and only if there exist $a > 0$, $b \in \mathbb{R}^k$, and $c \in \mathbb{R}$ such that for all distributions on $Y \in [k]$,*

$$H_1(Y) = aH_2(Y) + b^\top p + c, \tag{16.4.1}$$

where $p = [P(Y = y)]_{y=1}^k$ is the p.m.f. of Y.

One direction of the theorem, as we mention above, is easy: if the entropy equivalence (16.4.1) holds, then ℓ_1 and ℓ_2 are universally equivalent. Indeed, we see that

$$I_1(\mathsf{q}_1(X); Y) \le I_1(\mathsf{q}_2(X); Y) \text{ if and only if } H_1(Y \mid \mathsf{q}_1(X)) \ge H_1(Y \mid \mathsf{q}_2(X)),$$

and the latter occurs if and only if

$$aH_2(Y \mid \mathsf{q}_1(X)) + b^\top [P(Y = y)]_{y=1}^k + c \ge aH_2(Y \mid \mathsf{q}_2(X)) + b^\top [P(Y = y)]_{y=1}^k + c.$$

The sufficiency of condition (16.4.1) is thus immediate; for its necessity, see Section 16.5.3.

Example 16.4.2 (Universal equivalence for 0-1 losses)**:** Consider classification with the 0-1 loss $\ell_{0\text{-}1}(s, y) = \mathbf{1}\{s_y \le \max_{j \ne y} s_j\}$. Then $H_{\ell_{0\text{-}1}}(Y) = 1 - \max_y P(Y = y)$. Several convex losses are classification calibrated and universally equivalent to the zero-one loss. Recalling Example 16.3.7,

$$\varphi(s, y) := 1 - s_y + \max_{1 \le i \le k} \left\{ \frac{s_{(1)} + \cdots + s_{(i)} - 1}{i} \right\}$$

is consistent and equivalent to $\ell_{0\text{-}1}$, with $H_\varphi(Y) = H_{\ell_{0\text{-}1}}(Y)$. Similarly, for $\mathcal{Z} = \{s \mid s^\top \mathbf{1} = 0\}$, the hinge loss $\varphi^{\text{hinge}}(s, y) = \sum_{j \ne y} [1 + s_j]_+ + \mathbf{I}_{\mathcal{Z}}(s)$ satisfies $H_{\varphi^{\text{hinge}}}(Y) = k \cdot H_{\ell_{0\text{-}1}}(Y)$. $\diamond$

We complete this section with a corollary of Theorems 16.3.8 and 16.4.1, which shows that the entropy-based construction of surrogate losses is enough to achieve *both* surrogate risk consistency and universal equivalence of losses.

Corollary 16.4.3. *Let ℓ be a loss of the form (16.3.2) with associated generalized entropy $H_\ell(Y) = \inf_s \mathbb{E}[\ell(\mathsf{pred}(s), Y)]$, and let Assumption A.16.1 hold. Then the surrogate $\varphi(s, y) := -s^\top \tau(y) + \Omega^*(s)$ is both universally equivalent to and surrogate risk consistent for ℓ.*

In brief, given a (discrete) loss, we can construct its entropy, develop a convex surrogate whose associated entropy is identical to the original function, and guarantee consistency of the convex surrogate.

16.5* Deferred and technical proofs

This appendix to the chapter contains the proofs of many of the more technical results to preserve the flow of the chapter. Most of them rely on the material in Appendices B and C.

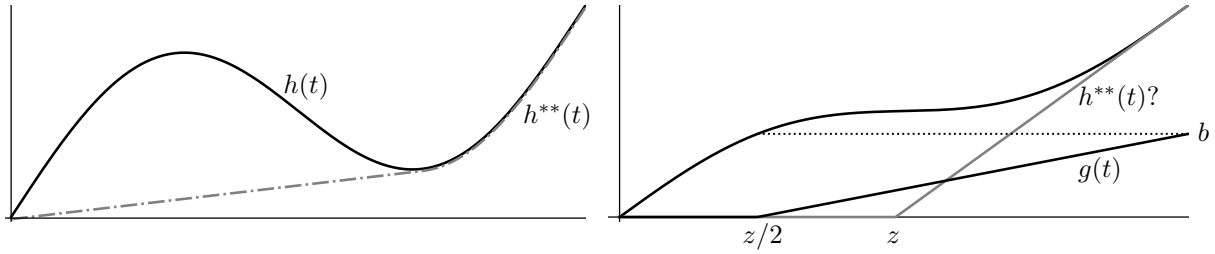


Figure 16.2. Left: illustration of Lemma 16.1.8. Right: illustration of the construction of g in the proof of Lemma 16.1.8, where the gray line represents a putative value of h^{**}.

16.5.1 Proof of Lemma 16.1.8

We begin with the case (i). Define the function $h_{\rm low}(t) := \inf_{s \geq t} h(s)$. Then because h is continuous, it attains its infimum over any compact set, so (by assumption on h) $h_{\rm low}(t) > 0$ for all $t > 0$. Moreover, $h_{\rm low}$ is non-decreasing. Define $f_{\rm low}(t) = h_{\rm low}^{**}(t)$ to be the biconjugate of $h_{\rm low}$; it is clear that $f \geq f_{\rm low}$ as $h \geq h_{\rm low}$. Thus case (ii) implies case (i), so we turn to the more general result to see that $f_{\rm low}(t) > 0$ for all $t > 0$.

For the result in case (ii), assume for the sake of contradiction there is some $z \in (0, 1)$ satisfying $h^{**}(z) = 0$. It is clear that $h^{**}(0) = 0$ and $h^{**} \geq 0$, so we must have $h^{**}(z/2) = 0$. Now, by assumption we have $h(z/2) = b > 0$ for some value b, whence $h(1) \geq b > 0$. In particular, the piecewise linear function defined by

$$g(t) = \begin{cases} 0 & \text{if } t \leq z/2 \\ \frac{b}{1 - z/2}(t - z/2) & \text{if } t > z/2 \end{cases}$$

is closed, convex, and satisfies $g \leq h$ (see the right plot in Figure 16.2). But $g(z) > 0 = h^{**}(z)$, a contradiction to the fact that h^{**} is the largest (closed) convex function below h.

16.5.2 Proof of Theorem 16.3.8

Let $\mu = \mu(P)$ and recall the gap functional (16.2.2),

$$\Delta_\varphi(\epsilon, P) := \inf_s \left\{ r_\varphi(s, P) - r_\varphi^\star(P) \mid r(s, P) - r^\star(P) \geq \epsilon \right\},$$

where $r(s, P) = \mathbb{E}_P[\ell(\mathsf{pred}(s), Y)]$ and $r^\star(P) = \inf_s \mathbb{E}_P[\ell(\mathsf{pred}(s), Y)]$. In this case, we may simplify the quantities by writing out the entropy functionals explicitly as

$$r(s, P) - r^\star(P) = \tau(\widehat{y}(s))^\top A\mu - \inf_{\nu \in \mathcal{M}} \nu^\top A\mu,$$

where $\widehat{y}(s)$ is (an arbitrary) element of the prediction set $\operatorname{argmax}_y s^\top \tau(y)$. We need only show that $\Delta_\varphi(\epsilon, P) > 0$ whenever $\epsilon > 0$, which we prove by contradiction.

Thus, assume for the sake of contradiction that $\Delta_\varphi(\epsilon, P) = 0$. As the losses φ are piecewise linear and the set of s such that $r(s, P) - r^\star(P) \geq \epsilon$ is a finite union of closed polyhedra, there must be s achieving the infimum, and so for that vector of scores s, we have

$$\Omega^*(s) - s^\top \mu + \Omega(\mu) = 0.$$

The normal cone $\mathcal{N}_{\mathcal{M}}(\mu) = \{v \mid \langle v, \nu - \mu \rangle \le 0 \text{ for all } \nu \in \mathcal{M}\}$ to $\mathcal{M}$ at μ (see Definition B.6) gives

$$\begin{aligned}\partial\Omega(\mu) &= \mathrm{Conv}\{-A^\top \tau(y) \mid \tau(y)^\top A\mu = \min_{y'} \tau(y')^\top A\mu\} + \mathcal{N}_{\mathcal{M}}(\mu) \\ &= \{-A^\top \nu \mid \nu^\top A\mu = H_\ell(P), \nu \in \mathcal{M}\} + \mathcal{N}_{\mathcal{M}}(\mu), \end{aligned} \tag{16.5.1}$$

by Proposition C.1.3, where we recall that $\mu = \mu(P)$. Following the calculation (16.5.1), we thus obtain that for some $\nu^\star \in \mathcal{M}$ satisfying $\langle \nu^\star, A\mu \rangle = \min_y \tau(y)^\top A\mu$ and a vector $w \in \mathcal{N}_{\mathcal{M}}(\mu)$, we have

$$s = -A^\top \nu^\star + w.$$

For $\nu \in \mathcal{M}$, define the set-valued mapping $y^\star(\nu) = \mathrm{argmin}_y \tau(y)^\top A\nu$. If we can show the inclusions

$$\widehat{y}(s) \subset y^\star(\nu^\star) \subset y^\star(\mu(P)), \tag{16.5.2}$$

then the proof is complete, as we would have our desired contradiction: if $\widehat{y}(s) \subset y^\star(\mu)$, then $r(s, P) = r^\star(P)$.

To see the inclusion $y^\star(\nu^\star) \subset y^\star(\mu(P))$ is relatively straightforward. Let

$$\nu \in \mathrm{Conv}\left\{\tau(y) \mid \tau(y)^\top A\mu(P) = \min_{y' \in \mathcal{Y}} \mathbb{E}_P[\ell(y', Y)] = H_\ell(P)\right\} \tag{16.5.3}$$

be otherwise arbitrary. For all P' satisfying the mean-mapping equality $\nu = \mu(P')$, the identifiability assumption A.16.1 guarantees that if $y \in y^\star(\nu)$, then $P'(Y = y) > 0$. That is, $y^\star(\nu)$ has containment

$$y^\star(\nu) \subset \bigcap_{P' : \mu(P') = \nu} \mathrm{supp}\, P'.$$

We may represent each ν in the hull (16.5.3) via P' with support $\{y \mid \tau(y)^\top A\mu(P) = H_\ell(P)\} = y^\star(\mu(P))$. Thus $y^\star(\nu) \subset y^\star(\mu(P))$ for all such ν, and in particular for $\nu^\star$.

The first inclusion $\widehat{y}(s) \subset y^\star(\nu^\star)$ in the chain (16.5.2) is more challenging. We begin with a convex analytic result that allows us to simplify maximizers of $s^\top \tau(y)$ in y.

Lemma 16.5.1. *Let $w \in \mathcal{N}_{\mathcal{M}}(\mu)$ be the element satisfying $s = -A^\top \nu^\star + w$. Then for any $y \in \mathcal{Y}$ and any $z \in \mathrm{supp}\, P$,*

$$\langle \tau(z) - \tau(y), w \rangle \ge 0.$$

Proof Fix $y \in \mathcal{Y}$ and let $z \in \mathrm{supp}\, P$, so that $p_z = P(Y = z) > 0$. Then for some vector $\alpha \in \mathrm{Conv}(\tau(y') \mid y' \notin \{y, z\})$, we can write $\mu(P) = \lambda_y \tau(y) + \lambda_z \tau(z) + (1 - \lambda_y - \lambda_z)\alpha$, where $\lambda_y \ge 0, \lambda_z \ge p_z > 0$, and $\lambda_y + \lambda_z \le 1$. The vector $\nu = (\lambda_y + \lambda_z)\tau(y) + (1 - \lambda_y - \lambda_z)\alpha$ similarly satisfies $\nu \in \mathcal{M}$. By the definition of the normal cone $\mathcal{N}_{\mathcal{M}}(\mu)$, we know that $w^\top(\mu' - \mu) \le 0$ for all $\mu' \in \mathcal{M}$, and in particular this holds for $\mu' = \nu$. As

$$\nu - \mu = \lambda_z(\tau(y) - \tau(z)),$$

we obtain

$$\lambda_z w^\top(\tau(y) - \tau(z)) \le 0,$$

and that $\lambda_z > 0$ implies the lemma. □

Consider the prediction set $\widehat{y}(s) = \operatorname{argmax}_y s^\top \tau(y)$. As $s = -A^\top \nu^\star + w$, we have

$$\widehat{y}(s) = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ -\tau(y)^\top A^\top \nu^\star + \tau(y)^\top w \right\} = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ -\tau(y)^\top A \nu^\star + \tau(y)^\top w \right\},$$

where we use the assumed symmetry of A. Let $y \in y^\star(\nu^\star)$ and $y' \notin y^\star(\nu^\star)$, so that $\tau(y')^\top A\nu^\star > \tau(y)^\top A\nu^\star$. Then by our earlier argument that $y^\star(\nu^\star) \subset y^\star(\mu(P))$ and so $P(Y = y) > 0$, Lemma 16.5.1 implies

$$\tau(y')^\top w \le \tau(y)^\top w.$$

We then see that

$$-\tau(y')^\top A\nu^\star + \tau(y')^\top w < -\tau(y)^\top A\nu^\star + \tau(y)^\top w,$$

and so $y' \notin \widehat{y}(s)$. In particular, $\widehat{y}(s) \subset y^\star(\nu^\star)$ as desired.

16.5.3 Proof of Theorem 16.4.1 for binary labels

We give the proof of the "hard" direction of Theorem 16.4.1: if ℓ_1 and ℓ_2 are universally equivalent, then their associated entropies are affine scalings (16.4.1). We prove the result in the slightly simpler case of binary classification (see Duchi et al. [162] for the full general result), as the more general result introduces few new ideas except that it requires more technical care.

We thus work with margin-based losses $\phi_i : \mathbb{R} \to \mathbb{R}_+$, $i = 1, 2$, where without loss of generality we assume $\inf_s \phi_i(s) = 0$, and we have generalized (binary) entropies

$$h_i(p) := \inf_s \left\{ p\phi_i(-s) + (1-p)\phi_i(s) \right\}.$$

Each h_i is a closed concave function (as the infimum of linear functions of p), and by symmetry they satisfy $h_i(0) = h_i(1) = 0$ and $h_i(\frac{1}{2}) = \sup_{p \in [0,1]} h_i(p)$. We show that these entropies satisfy a particular *order equivalence* property on $[0, 1]$, which will be sufficient to prove their equality.

Recall that universal equivalence (Def. 16.7) must hold for *all* distributions on (X, Y), and hence all (measurable) spaces $\mathcal{X}$ and joint distributions on $\mathcal{X} \times \{\pm 1\}$. Consider a space $\mathcal{X}$ that we can partition into sets $\{A, A^c\}$ or $\{B, B^c\}$, where we take the conditional distributions

$$Y \mid X \in A = \begin{cases} 1 & \text{w.p. } p_a \\ -1 & \text{w.p. } 1 - p_a, \end{cases} \quad Y \mid X \in A^c = \begin{cases} 1 & \text{w.p. } q_a \\ -1 & \text{w.p. } 1 - q_a, \end{cases}$$

and

$$Y \mid X \in B = \begin{cases} 1 & \text{w.p. } p_b \\ -1 & \text{w.p. } 1 - p_b, \end{cases} \quad Y \mid X \in B^c = \begin{cases} 1 & \text{w.p. } q_b \\ -1 & \text{w.p. } 1 - q_b, \end{cases}$$

where we require the consistency conditions that the marginals over Y remain constant, that is, if $P(A) = P(X \in A)$, we have

$$P(A)p_a + P(A^c)q_a = P(Y = 1) = P(B)p_b + P(B^c)q_b.$$

Defining quantizers q_1 and q_2 such that $\mathsf{q}_1(x) = \mathbf{1}\{x \in A\}$ and $\mathsf{q}_2(x) = \mathbf{1}\{x \in B\}$, we have

$$I_{\phi_1}(\mathsf{q}_1(X); Y) = h_1(P(Y = 1)) - P(A)h_1(p_a) - (1 - P(A))h_1(q_a),$$
$$I_{\phi_1}(\mathsf{q}_2(X); Y) = h_1(P(Y = 1)) - P(B)h_1(p_b) - (1 - P(B))h_1(q_b),$$

and similarly for I_{ϕ_2}. Universal equivalence then implies that

$$P(A)h_1(p_a) + (1 - P(A))h_1(q_a) \le P(B)h_1(p_b) + (1 - P(B))h_1(q_b) \text{ if and only if}$$
$$P(A)h_2(p_a) + (1 - P(A))h_2(q_a) \le P(B)h_2(p_b) + (1 - P(B))h_2(q_b)$$

whenever the consistency condition $P(A)p_a + (1 - P(A))q_a = P(B)p_b + (1 - P(B))q_b$ holds. As we may choose $\mathcal{X}$ and the probabilities, we can take $P(A) = P(B) = \frac{1}{2}$ (so that there are two equiprobable partitions), and the preceding conditions become

$$h_1(p_a) + h_1(q_a) \le h_1(p_b) + h_1(q_b) \text{ if and only if } h_2(p_a) + h_2(q_a) \le h_2(p_b) + h_2(q_b)$$

whenever $p_a + q_a = p_b + q_b$.

Generalizing this construction by taking distributions over $\mathcal{X}$ that partition it into k equiprobable sets $\{A_1, \ldots, A_k\}$ or $\{B_1, \ldots, B_k\}$, each with $P(A_i) = P(B_i) = 1/k$, we see that universal equivalence implies that for any vectors $p \in [0,1]^k$ and $q \in [0,1]^k$ satisfying $\mathbf{1}^\top p = \mathbf{1}^\top q$,

$$\sum_{i=1}^k h_1(p_i) \le \sum_{i=1}^k h_1(q_i) \quad \text{if and only if} \quad \sum_{i=1}^k h_2(p_i) \le \sum_{i=1}^k h_2(q_i). \tag{16.5.4}$$

We shall give condition (16.5.4) a name, as it implies certain equivalence properties for convex functions (we can replace h_i with $-h_i$ and obtain convex functions).

Definition 16.8. *Let $T \subset \mathbb{R}$ be a closed interval and let $f_1, f_2 : T \to \mathbb{R}$ be closed convex functions. Then f_1 and f_2 are* order equivalent *if for any $k \in \mathbb{N}$ and vectors $s \in T^k$ and $t \in T^k$ satisfying $\mathbf{1}^\top s = \mathbf{1}^\top t$, we have*

$$\sum_{i=1}^k f_1(s_i) \le \sum_{i=1}^k f_1(t_i) \quad \textit{if and only if} \quad \sum_{i=1}^k f_2(s_i) \le \sum_{i=1}^k f_2(t_i)$$

As in the brief remark following Definition 16.7, by taking complements we have as well that

$$\sum_{i=1}^k f_1(s_i) < \sum_{i=1}^k f_1(t_i) \quad \text{if and only if} \quad \sum_{i=1}^k f_2(s_i) < \sum_{i=1}^k f_2(t_i)$$

Theorem 16.4.1 will follow if we can show the following lemma.

Lemma 16.5.2. *Let f_1 and f_2 be order equivalent on T. Then there exist $a > 0$, and $b, c \in \mathbb{R}$ such that $f_1(t) = af_2(t) + bt + c$ for all $t \in T$.*

The proof of Lemma 16.5.2 proceeds in three parts. The key is that order equivalence implies a strong relationship between *affine* combinations of points in the domain of the functions f_i, not just convex combinations of points, which guarantees that we can predict values of $f_2(v)$ for any $v \in T$ by just three values of f_i evaluated in T. We state this as a lemma, whose proof we defer temporarily to Sec. 16.5.4

Lemma 16.5.3. *If f_1 and f_2 are order equivalent on T, then for any $\lambda \in \mathbb{R}^k$ satisfying $\lambda^\top \mathbf{1} = 1$ and any $u \in T^k$, if $\lambda^\top u = v \in T$ then*

$$\sum_{i=1}^k \lambda_i f_1(u_i) \le f_1(v) \quad \textit{if and only if} \quad \sum_{i=1}^k \lambda_i f_2(u_i) \le f_2(v).$$

The statement still holds with both inequalities replaced with strict inequalities.

In particular, if

$$\sum_{i=1}^k \lambda_i f_1(u_i) = f_1(v) \;\; \text{then necessarily} \;\; \sum_{i=1}^k \lambda_i f_2(u_i) = f_2(v) \tag{16.5.5}$$

whenever $\lambda \in \mathbb{R}^k$ satisfies $\lambda^\top \mathbf{1} = 1$ and $u^\top \lambda = \sum_{i=1}^k \lambda_i u_i = v$.

Second, we recognize that we may assume both f_1 and f_2 are nonlinear in the lemma; otherwise, it is immediate. Nonlinearity guarantees that

Lemma 16.5.4. *Let f be convex on $\mathbb{R}$. Let $u_0 < u_1$ and for $\lambda \in [0,1]$, define $u_\lambda = \lambda u_0 + (1-\lambda)u_1$. If there exists any $\lambda \in (0,1)$ such that $f(u_\lambda) = \lambda f(u_0) + (1-\lambda) f(u_1)$, then f is linear on $[u_0, u_1]$.*

We leave the proof (an algebraic manipulation using the definitions of convexity) as Exercise 16.14.

The last intermediate step we require in the proof of Lemma 16.5.2 is that at three particular points in the domain T, we can satisfy Lemma 16.5.2. (See Section 16.5.5 for a proof.)

Lemma 16.5.5. *Let f_1, f_2 be order equivalent on $T = [u_0, u_1]$ and $u_{1/2} = \frac{1}{2}(u_0 + u_1)$. There are $a > 0$ and $b, c \in \mathbb{R}$ such that $f_1(t) = a f_2(t) + bt + c$ for $t \in \{u_0, u_{1/2}, u_1\}$.*

We can now finalize the proof of Lemma 16.5.2:

Proof Without loss of generality by an affine rescaling, we can assume that $f_1(t) = f_2(t)$ for $t \in \{u_0, u_{1/2}, u_1\}$, and our goal will be to show that $f_1(t) = f_2(t)$ for all $t \in T$.

Let $v \in T$ with $v \notin \{u_0, u_{1/2}, u_1\}$, and $u = [u_0 \;\; u_{1/2} \;\; u_1]^\top$ for shorthand. We seek $\lambda = (\lambda_0, \lambda_{1/2}, \lambda_1) \in \mathbb{R}^3$, where (i) $\lambda^\top \mathbf{1} = 1$, such that both (ii) $\lambda^\top u = v$ and (iii) $\lambda_0 f_1(u_0) + \lambda_{1/2} f_1(u_{1/2}) + \lambda_1 f_1(u_1) = f_1(v)$. If we can find such a λ, then equality (16.5.5) guarantees that $f_1(v) = f_2(v)$, and we are done. As the points $(u_i, f_1(u_i))$, $i = 0, \frac{1}{2}, 1$, are non-collinear (Lemma 16.5.2 is trivial if they are collinear and use Lemma 16.5.4), the matrix

$$A := \begin{bmatrix} 1 & 1 & 1 \\ u_0 & u_{1/2} & u_1 \\ f_1(u_0) & f_1(u_{1/2}) & f_1(u_1) \end{bmatrix}$$

is full rank. In particular, there is a vector λ solving

$$A\lambda = [1 \;\; v \;\; f_1(v)]^\top, \;\; \text{i.e.} \;\; \lambda = A^{-1} [1 \;\; v \;\; f_1(v)]^\top,$$

which evidently satisfies our desiderata (i)–(iii). Thus $f_1(v) = f_2(v)$, completing the proof. □

16.5.4 Proof of Lemma 16.5.3

We prove the result first for λ with rational entries, as a continuity argument will give the rest. For each i, let $\alpha_i = [\lambda_i]_+$ and $\beta_i = [-\lambda_i]_+$ be the positive and negative parts of λ, so that $\lambda = \alpha - \beta$. Let $k \in \mathbb{N}$ be such that we can write $\alpha_i = \frac{a_i}{k}$ and $\beta_i = \frac{b_i}{k}$, where $a_i, b_i \in \mathbb{N}$. Then we have

$$\alpha^\top u = v + \beta^\top u \;\; \text{or} \;\; a^\top u = kv + b^\top u,$$

where $\mathbf{1}^\top a = k + \mathbf{1}^\top b$, as $\mathbf{1}^\top \lambda = \frac{1}{k} \mathbf{1}^\top (a - b) = 1$. Then we may define the two vectors

$$s = [\underbrace{u_1 \;\cdots\; u_1}_{a_1 \text{ times}} \;\cdots\; \underbrace{u_m \;\cdots\; u_m}_{a_m \text{ times}}]^\top \;\; \text{and} \;\; t = [\underbrace{v \;\cdots\; v}_{k \text{ times}} \; \underbrace{u_1 \;\cdots\; u_1}_{b_1 \text{ times}} \;\cdots\; \underbrace{u_m \;\cdots\; u_m}_{b_m \text{ times}}]^\top.$$

Then each has entries in T, and $\mathbf{1}^\top t = \mathbf{1}^\top s$. Order equivalence (Def. 16.8) guarantees that

$$\sum_{i=1}^m a_i f_1(u_i) \le k f_1(v) + \sum_{i=1}^m b_i f_1(u_i) \text{ if and only if } \sum_{i=1}^m a_i f_2(u_i) \le k f_2(v) + \sum_{i=1}^m b_i f_2(u_i)$$

and (as per the remark following the definition)

$$\sum_{i=1}^m a_i f_1(u_i) = k f_1(v) + \sum_{i=1}^m b_i f_1(u_i) \text{ if and only if } \sum_{i=1}^m a_i f_2(u_i) = k f_2(v) + \sum_{i=1}^m b_i f_2(u_i).$$

Equivalently, $\sum_{i=1}^m \lambda_i f_j(u_i) \le f_j(v)$ and $\sum_{i=1}^m \lambda_i f_j(u_i) = f_j(v)$ for $j = 1, 2$.

We have proved Lemma 16.5.3 for λ taking rational values. Closed convex functions on $\mathbb{R}$ are continuous on their domains (Observation B.2.8), so the result extends to real-valued λ.

16.5.5 Proof of Lemma 16.5.5

If either of f_1 or f_2 is linear, the other is as well, and the proof becomes trivial, so we assume w.l.o.g. they are both nonlinear. Without loss of generality, we take $u_0 = 0$, $u_1 = 1$, and $u_{1/2} = \frac{1}{2}$ by scaling. Then we must solve the three equations

$$f_1(0) = a f_2(0) + c, \quad f_1\left(\frac{1}{2}\right) = a f_2\left(\frac{1}{2}\right) + \frac{b}{2} + c, \quad f_1(1) = a f_2(1) + b + c.$$

From the first we obtain $c = f_1(0) - a f_2(0)$, and substituting into the third yields $b = f_1(1) - f_1(0) - a(f_2(1) - f_2(0))$. Substituting both equalities into the equality with $f_1(\frac{1}{2})$ yields that

$$\begin{aligned} f_1\left(\frac{1}{2}\right) &= a\left[f_2\left(\frac{1}{2}\right) - \frac{f_2(1) - f_2(0)}{2}\right] + \frac{f_1(1) - f_1(0)}{2} + f_1(0) - a f_2(0) \\ &= a\left[f_2\left(\frac{1}{2}\right) - \frac{f_2(1) + f_2(0)}{2}\right] + \frac{f_1(1) + f_1(0)}{2}. \end{aligned}$$

As f_1, f_2 are nonlinear, Lemma 16.5.4 applies to guarantee negative convexity gaps $f_1(\frac{1}{2}) - \frac{f_1(1)+f_1(0)}{2}$ and $f_2(\frac{1}{2}) - \frac{f_2(1)+f_2(0)}{2}$, and thus we take

$$a = \frac{f_1(\frac{1}{2}) - \frac{f_1(0)+f_1(1)}{2}}{f_2(\frac{1}{2}) - \frac{f_2(0)+f_2(1)}{2}} > 0.$$

16.6 Bibliography and further reading

The study of whether a loss function is consistent for estimating a particular statistical quantity traces back to Huber's introduction of robust statistics and study of M-estimation problems [243], where he was most interested in identifying a location shift in a family of densities $\{f(\cdot - \theta)\}_{\theta \in \mathbb{R}}$. When the densities are symmetric, many M-estimators are consistent for estimating θ, including those using a symmetric convex loss ϕ via $\widehat{\theta} = \operatorname{argmin}_\theta \sum_{i=1}^n \phi(X_i - \theta)$, and Huber's questions included identifying minimax-optimal losses under different perturbations [243, 244]. Within the context of statistical machine learning from which this chapter mostly draws, surrogate consistency has been a consistent subject since the early 2000s, where recognition of the centrality of convex optimization to the effectiveness of statistical procedures grew [82].

This line of work includes that by Zhang [437, 436] and Lugosi and Vayatis [296], who developed sufficient conditions for the consistency of convex losses in binary and multiclass classification, and culminated with Bartlett, Jordan, and McAuliffe's complete characterization of consistent convex losses for binary classification [49], off of which we base the treatment in Section 16.1. Their work ignited a flurry of interest on consistency: for multiclass classification [394], ranking problems [160, 358, 10], connecting with proper losses [430], and extensions to structured prediction [337]. Nowak-Vila et al. [331] inspire our proof of Theorem 16.3.8, though some of the claims in that paper appear mistaken (cf. Exercise 16.8). Steinwart [386] provides the most general abstract conditions for surrogate consistency, and Section 16.2 follows his approach. Frongillo and Waggoner [202] extend these results to identify optimal excess risk bounds, that is, the largest functions $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\psi(L(f) - L^\star) \le L_\varphi(f) - L_\varphi^\star$ possible for different surrogates φ, as in Theorem 16.1.5. They show that for most smooth surrogates φ, $\psi(\epsilon) = O(1)\epsilon^2$ is as large as possible, while consistent polyhedral surrogates typically admit linear growth via $\psi(\epsilon) \gtrsim \epsilon$. (See Exercises 16.1, 16.2, and 16.5 for examples of these phenomena in binary and multiclass classification.)

A well-known limitation of this work is that it does not address the case of restricted hypothesis classes, meaning minimizing the surrogate risk L_φ over a subset $\mathcal{H}$ of all functions $f : \mathcal{X} \to \mathbb{R}$. Long and Servedio [294] introduce $\mathcal{H}$-consistency to study some of the issues around this, identifying losses φ that are consistent so long as a perfect predictor exists in the set $\mathcal{H}$, and work on it continues [see, e.g. 300]. Alternatively, because many losses φ are surrogate risk consistent, researchers have sought more nuanced behavior. Nguyen et al. [326] demonstrate deep connections between f-divergences and the entropy measures associated with binary margin-based loss functions $\phi : \mathbb{R} \to \mathbb{R}$ (as in Section 16.1) via $H_\phi(Y) = \inf_{s \in \mathbb{R}} \mathbb{E}[\phi(sY)]$, showing a version of Theorem 16.4.1 for a modified family of f-divergences. Duchi et al. [162] extend these results to multiclass cases, providing Theorem 16.4.1, and they develop the connections between entropy and surrogate-risk consistency in Section 16.3.

16.7 Exercises

Exercise 16.1: Find the suboptimality function Δ_ϕ and ψ-transform (16.1.3) for the binary classification problem ($Y \in \{-1, 1\}$) with the following losses.

(a) Logistic loss $\phi(s) = \log(1 + e^{-s})$.

(b) The squared hinge loss $\phi(s) = [1 - s]_+^2$.

(c) Squared error with surrogate loss $\varphi(f(x), y) = \frac{1}{2}(f(x) - y)^2$.

(d) Is each of these classification calibrated?

Exercise 16.2: Consider the multiclass 0-1 loss $\ell_{0\text{-}1}(s, y) = \mathbf{1}\{s_y \le \max_{j \ne y} s_j\}$ and the multiclass logistic regression loss $\varphi(s, y) = \log(\sum_{j=1}^k e^{s_j - s_y})$. This exercise develops an explicit excess risk bound for φ over $\ell_{0\text{-}1}$, i.e., a uniform suboptimality function (16.2.3). Let $p \in \Delta_k$ be any p.m.f.

(a) Show that the entropies $H_\varphi(p) = -\sum_{j=1}^k p_j \log p_j$ and $H_{\ell_{0\text{-}1}}(p) = 1 - \max_{j \le k} p_j$.

(b) Let $\widehat{y}(s) = \operatorname{argmax}_y s_y$, breaking ties arbitrarily (so that $\widehat{y}(s)$ is a singleton), and recall the conditional risk $r_\ell(s, p) = \mathbb{E}_p[\ell(s, Y)]$. Show that

$$r_\varphi(s, p) - H_\varphi(p) = -s^\top p + \log\left(\sum_{j=1}^k e^{s_j}\right) - H_\varphi(p) \quad \text{and} \quad r_{\ell_{0\text{-}1}}(s, p) - H_{\ell_{0\text{-}1}}(p) = \|p\|_\infty - p_{\widehat{y}(s)}.$$

(c) Show that

$$r_\varphi(s,p) - \inf_s r_\varphi(s,p) \ge \frac{1}{2}\left(r_{\ell_{0\text{-}1}}(s,p) - \inf_s r_{\ell_{0\text{-}1}}(s,p)\right)^2.$$

Hint. Define $q_y = e^{s_y}/\sum_{j=1}^k e^{s_j}$ and rewrite $-s^\top p + \log(\sum_{j=1}^k e^{s_j})$ via q.

(d) Conclude that the multiclass logistic regression loss is uniformly calibrated for the 0-1 loss:

$$\sqrt{2\left(L_\varphi(f) - L_\varphi^\star\right)} \ge L_{\ell_{0\text{-}1}}(f) - L_{\ell_{0\text{-}1}}^\star \quad \text{for any } f : \mathcal{X} \to \mathbb{R}^k.$$

Exercise 16.3: Consider a regression problem with $x \in \mathcal{X}$ and responses $y \in \mathbb{R}$. We seek $f : \mathcal{X} \to \mathbb{R}$ minimizing the probability of being far from y, so for a $c > 0$, use the loss

$$\ell(f(x), y) = \mathbf{1}\left\{|y - f(x)| \ge c\right\}.$$

Show that no surrogate of the form $\varphi(s,y) = |s-y|^p$, $p \ge 1$, is consistent for ℓ, even if the distribution of Y conditioned on $X = x$ is symmetric about its mean $\mathbb{E}[Y \mid X]$. In particular, exhibit a distribution on pairs X, Y such that the set of minimizers of the surrogate

$$L_\varphi(f) := \mathbb{E}[\varphi(f(X), Y)]$$

is distinct from the set of minimizers of the true risk, $L(f) = \mathbb{P}(|Y - f(X)| \ge c)$.

Exercise 16.4 (Hinge-type losses and entropies [162])**:** Consider the setting of Example 16.3.7, where $\ell(s,y) = \mathbf{1}\left\{s_y \le \max_{j \ne y} s_j\right\}$ is the 0-1 loss.

(a) Prove the equality (16.3.5).

(b) Show that $\inf_s \mathbb{E}_p[\varphi^{\text{hinge}}(s,Y)] = k(1 - \max_j p_j)$.

Exercise 16.5* ([162])**:** Let $\Omega(p) = 1 - \max_j p_j$ be the negative entropy associated with the zero-one multiclass loss $\ell_{0\text{-}1}(s,y) = \mathbf{1}\left\{s_y \le \max_{j \ne y} s_j\right\}$, as in Example 16.3.7. Let $\varphi(s,y) = -s^\top e_y + \Omega^*(s)$. For parts (a) and (b) of the question, let $p \in \Delta_k$ satisfy $p_1 \ge p_2 \ge \cdots \ge p_k$.

(a) Show that for all $y \in [k]$,

$$\inf_{s_y \ge \max_j s_j} \mathbb{E}_p[\varphi(s,Y)] \ge 1 - p_1 + \frac{1}{k}(p_1 - p_y).$$

Hint. Whenever L is bilinear, $\inf_{p \in \Delta_k} \sup_{q \in \Delta_k} L(p,q) = \sup_{q \in \Delta_k} \inf_{p \in \Delta_k} L(p,q)$.

(b) Show that

$$\inf_{s_y \ge \max_j s_j} \mathbb{E}_p[\ell_{0\text{-}1}(s,Y)] = 1 - p_y.$$

(c) Show the claim of Example 16.3.9: for any vector $s \in \mathbb{R}^k$,

$$\mathbb{E}_p[\varphi(s,Y)] - \inf_s \mathbb{E}_p[\varphi(s,Y)] \ge \frac{1}{k}\left(\mathbb{E}_p[\ell_{0\text{-}1}(s,Y)] - \inf_s \mathbb{E}_p[\ell_{0\text{-}1}(s,Y)]\right).$$

(d) Argue that the uniform gap (16.2.3) satisfies $\Delta_\varphi(\epsilon) \ge \frac{1}{k}\epsilon$ for all $\epsilon \ge 0$.

Exercise 16.6 (General maximum margins)**:** Structured prediction problems [393, 357, 332] take the setting (16.3.2) with loss $\ell(\hat{y}, y) = \tau(\hat{y})^\top A \tau(y) + c(y)$. Define the *margin* of a vector $s \in \mathbb{R}^k$

$$\min_{\widehat{y}} \left\{ s^\top (\tau(y) - \tau(\hat{y})) - \ell(\hat{y}, y) \right\}$$

and the maximum-margin loss [120, 357]

$$\phi(s, y) := \max_{\hat{y} \in \mathcal{Y}} \left\{ \ell(\hat{y}, y) + s^\top \left(\tau(\hat{y}) - \tau(y) \right) \right\}. \tag{16.7.1}$$

(a) Assume that $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$ and is positive otherwise. Show that $\phi(s, y) = 0$ if and only if $s^\top \tau(y) \geq \ell(\hat{y}, y) + s^\top \tau(\hat{y})$ for all $\hat{y}$. Conclude that if $\phi(s, y) = 0$ then $y = \operatorname{argmax}_{\hat{y} \in \mathcal{Y}} s^\top \tau(\hat{y})$.

(b) For k-class classification with zero-one error $\ell_{0\text{-}1}(\hat{y}, y) = \mathbf{1}\{\hat{y} \neq y\}$, standard basis features $\tau(y) = e_y$, and $A = \mathbf{1}\mathbf{1}^\top - I$, show that the loss (16.7.1) is $\phi(s, y) = \max_{j \neq y} [1 + s_j - s_y]_+$.

Exercise 16.7* (Inconsistency in structured prediction [293])**:** Consider the setting of Exercise 16.6 for k-class classification with 0-1 loss, and let $\phi(s, y) = \max_{j \neq y} [1 + s_j - s_y]_+$. Let P be a distribution on $Y \in [k] = \{1, \ldots, k\}$, and $y^\star = \operatorname{argmax}_{y \in [k]} P(Y = y)$, which we take to be unique.

(a) Show that if $P(Y = y^\star) > \frac{1}{2}$, then $s = e_{y^\star}$ minimizes $\mathbb{E}_P[\phi(s, Y)]$

(b) Show that if $P(Y = y^\star) < \frac{1}{2}$, then $s = \mathbf{0}$ minimizes $\mathbb{E}_P[\phi(s, Y)]$.

Exercise 16.8 (Inconsistency of the conjugate loss)**:** Let $\mathcal{Y}$ be the collection of bipartite matchings between index sets of size k, identified with the permutations $\mathfrak{S}_k$ of $\{1, \ldots, k\}$. For $y, y' \in \mathfrak{S}_k$, where $y(i)$ denotes the assignment of item i in permutation j, define $\ell(y, y') = \sum_{i=1}^k \mathbf{1}\{y(i) \neq y'(i)\}$.

(a) Write $\ell(y, y')$ in the form (16.3.2), that is, find a feature mapping τ and matrix A such that $\ell(y, y') = \tau(y)^\top A \tau(y')$.

(b) Let M_π be the permutation matrix associated with a permutation π. Show that there is a collection of k distinct permutations $\pi_1, \ldots, \pi_k$ such that

$$\frac{1}{k} \sum_{i=1}^k M_{\pi_i} = \frac{1}{k} \mathbf{1}\mathbf{1}^\top.$$

(c) Using permutations that induce block diagonal matrices of the form

$$M = \begin{bmatrix} I_j & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{k-j-2} \end{bmatrix},$$

show that Assumption A.16.1 fails: there exist distributions on permutations for which there exists a unique minimizer of $y^\star = \operatorname{argmin}_y \mathbb{E}[\ell(y, Y)]$ but $P(Y = y^\star) = 0$.

Exercise 16.9: Show that in the case of binary classification with margin-based losses, $\varphi(s, y) = \phi(sy)$, the uniform gap (16.2.3) and binary gap (16.1.1) are equal, that is, $\Delta_\varphi(\epsilon) = \Delta_\phi(\epsilon)$, making uniform (Definition 16.3) and classification (Definition 16.2) calibration equivalent.

Exercise 16.10 (Steinwart [386])**:** We generalize Theorem 16.2.4 to the case that the loss ℓ is not uniformly bounded. Assume there exists an upper bound function $B : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}[B(X)] < \infty$ and $r(s, P_x) \le r^\star(P_x) + B(x)$ for all $x \in \mathcal{X}$ and $s \in \mathbb{R}^k$. Show that φ is calibrated for the loss ℓ (Definition 16.4) if and only if it is surrogate risk consistent (Definition 16.1).

Exercise 16.11: Let $\phi(t) = [1 - t]_+$. Show by example that the surrogate (16.2.4) with $\varphi(s, y) = \sum_{j=1}^k \phi(s_y - s_j)$ may be inconsistent.

Exercise 16.12 (Zhang [436])**:** This exercise provides one proof of Proposition 16.2.7, demonstrating the surrogate-risk consistency of losses of the form $\varphi(s, y) = \phi(s_y) + \sum_{j \ne y} \phi(-s_j)$ for convex non-increasing ϕ with $\phi'(0) < 0$.

(a) Let $t = t_q$ minimize $q\phi(t) + (1 - q)\phi(-t)$. Show that $\operatorname{sign}(t) = \operatorname{sign}(q - \frac{1}{2})$ whenever $q \ne \frac{1}{2}$.

(b) Show that if s minimizes $\mathbb{E}_p[\varphi(s, Y)]$ and $p_i < p_y$, then $s_i < s_y$.

(c) Assume that $p_1 < \max_j p_j$. Demonstrate that

$$\inf_s \left\{ \mathbb{E}_p[\varphi(s, Y)] \mid s_1 = \max_j s_j \right\} > \inf_s \mathbb{E}_p[\varphi(s, Y)].$$

(d) Prove Proposition 16.2.7.

Exercise 16.13: Consider a binary classification problem with $(X, Y) \in \mathcal{X} \times \{-1, 1\}$ and $\phi(t) = \log(1 + e^{-t})$. Show that the mutual information I satisfies

$$H_\phi(Y) - H_\phi(Y \mid X) = I(X; Y).$$

Exercise 16.14: Prove Lemma 16.5.4. *Hint:* without loss of generality, you may take $u_0 = 0$, $u_1 = 1$. Then for any $u \in [0, 1]$, write λ as a convex combination of either $\{0, u\}$ or $\{u, 1\}$ and use the definition of convexity.

Exercise 16.15 (Facial structure of the simplex)**:** For a polytope $C = \operatorname{Conv}\{z_1, \ldots, z_k\}$, define the *face exposed* by a direction v as

$$F_C(v) := \left\{ x \in C \mid \langle v, x \rangle = \sup_{y \in C} \langle v, y \rangle \right\} = \left\{ x \in C \mid \langle v, x \rangle = \max_{j \le k} \langle v, z_j \rangle \right\}.$$

We work with the particular case of the probability simplex $\Delta_k = \operatorname{Conv}\{e_1, \ldots, e_k\} = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p \rangle = 1\}$. For a set $J \subset [k]$ and $J^c = [k] \setminus J$, let $\Delta_J = \{p \in \Delta_k \mid p_{J^c} = 0\}$.

(a) Show that $F_{\Delta_k}(v) = \Delta_J$ if and only if $J = \operatorname{argmax}_{j \in [k]} v_j$.

(b) Let $v \in \mathbb{R}^k$ and $J = \operatorname{argmax}_{j \in [k]} v_j$, and let $\widehat{y}(s) = \{e_y \mid s_y \ge \max_j s_j\}$ be the (vector) prediction set associated with $s \in \mathbb{R}^k$. Show that

$$\left\{ s \in \mathbb{R}^k \mid \widehat{y}(s) \subset F_{\Delta_k}(v) \right\} = \left\{ s \in \mathbb{R}^k \mid \max_{j \in J^c} s_j < \max_{j \in J} s_j \right\}.$$

Exercise 16.16 (Faces and consistency)**:** We give necessary and sufficient conditions for surrogate risk consistency for the conjugate losses of Section 16.3.2 with standard basis features $\tau(y) = e_y \in \{0,1\}^k$. For $y, y' \in [k]$, define the multiclass loss $\ell(y, y') = e_y^\top A e_{y'}$ for a symmetric nonnegative matrix A, and for $p \in \Delta_k$, let $H_\ell(p) = \min_y \mathbb{E}_p[\ell(y, Y)] = \min_y e_y^\top A p$ and $\Omega(p) = -H_\ell(p)$. Let $\varphi(s, y) = \Omega^*(s) - e_y^\top s$ and $s^\star(p) \in \operatorname{argmin}_s \mathbb{E}_p[\varphi(s, Y)]$. Use the facial notation of Exercise 16.15.

(a) Let $y^\star(p) = \{e_y \mid e_y^\top A p = \min_j e_j^\top A p\}$, so $F_{\Delta_k}(-Ap) = \operatorname{Conv}\{y^\star(p)\}$, and $\widehat{y}(s) = \{e_y \mid s_y \geq \max_j s_j\} = \operatorname{argmax}_{e_y} e_y^\top s$ correspond to the prediction set (16.3.4). Show that the inclusion

$$\widehat{y}(s^\star(p)) \subset y^\star(p) \quad \text{for any } s^\star(p) \in \operatorname*{argmin}_s \mathbb{E}_p[\varphi(s, Y)]$$

holds for every surrogate minimizer $s^\star(p)$ and $p \in \Delta_k$ if and only if

$$\partial\Omega(p) \subset \left\{s \in \mathbb{R}^k \mid \widehat{y}(s) \subset F_{\Delta_k}(-Ap)\right\} \quad \text{for all } p \in \Delta_k.$$

(b) Call a set $J \subset [k]$ *realizable* if $J = \operatorname{argmax}_{j \in [k]} e_j^\top A p$ for some $p \in \Delta_k$. Assume that all nonempty $J_0 \subset J$ are realizable whenever J is realizable. Show that if φ is consistent, then whenever y minimizes $\mathbb{E}_P[\ell(y, Y)]$, then $P(Y = y) > 0$. That is, under a qualification, Assumption A.16.1 is necessary for consistency.

Part IV

Optimization, games, learning, and redundancy

Précis

The final part of the book turns to questions of optimization and the tools one uses to actually fit statistical models. The results here will combine many of the tools we develop in Part I on concentration inequalities and, simultaneously, are motivated by and help to motivate the focus on convexity in Part III of the book. We begin with stochastic and online convex optimization problems—stochastic optimization forming a core building block of modern statistical estimation—in Chapter 17. In Chapter 18, we will investigate scenarios where only partial information about a statistical problem is available, such as in causal inference and so-called bandit problems, necessitating tradeoffs between exploration to gather information and performance; information-based analyses will provide a core building block for the results there. In both of these chapters, we will develop a concomitant optimality theory extending the tools we develop in Part II of the book, allowing a holistic view of the problems both from statistical and information-theoretic perspectives. We conclude with a chapter that frames prediction problems as a game between nature and a player who wishes to optimally predict or encode variables, which relates sequence prediction, adaptive and online learning, and the information measures we have developed. This culminates in a duality between regret—or redundancy in the coding literature—and (generalized) information capacity of channels and experiments that we explore in Part III.

Chapter 17

Stochastic and online convex optimization

In many problems, we only wish to minimize an expected loss—recovering parameter values, a focus of Chapter 5, are unimportant—as in large-scale prediction tasks. In other cases, we have *online* problems, where data arrive in a stream, and we wish to make predictions on this stream at least as well as some reference predictor. The surrogate consistency results in Chapter 16 suggest optimality properties of non-smooth losses beyond the purview of our earlier convergence guarantees.

This chapter addresses these challenges through online and stochastic optimization. Online optimization considers a two-player sequential game: the learner plays points $\theta_1, \theta_2, \ldots$ in a space Θ, while nature plays losses $\ell_t : \Theta \to \mathbb{R}$. We wish to guarantee that the regret

$$\mathsf{Reg}_n(\theta^\star) := \sum_{t=1}^n \left[\ell_t(\theta_t) - \ell_t(\theta^\star)\right] \tag{17.0.1}$$

grows at most sub-linearly with n for any fixed $\theta^\star \in \Theta$. This defines *online convex optimization* problems, where the sequence of losses ℓ_t and domain $\Theta \subset \mathbb{R}^d$ are closed convex. The related *stochastic convex optimization* problem follows when nature cannot be so capricious: instead of losses ℓ_t chosen adversarially, for an unknown distribution P on $Z \in \mathcal{Z}$, we wish to minimize

$$L(\theta) := \mathbb{E}_P[\ell(\theta, Z)] \tag{17.0.2}$$

given a sample $Z_1^n \stackrel{\text{iid}}{\sim} P$, where for each $z \in \mathcal{Z}$, the loss $\ell(\theta, z)$ is convex in θ.

The problem (17.0.2) motivates the online problem (17.0.1), because any algorithm with a regret guarantee for (17.0.1) implies a convergence rate for optimizing (17.0.2). Indeed, let $Z_t \stackrel{\text{iid}}{\sim} P$, and in the formulation (17.0.1), set $\ell_t(\theta) = \ell(\theta, Z_t)$. Then by convexity, the average $\overline{\theta}_n := \frac{1}{n}\sum_{t=1}^n \theta_t$ satisfies

$$\mathbb{E}[L(\overline{\theta}_n)] \le \frac{1}{n}\sum_{t=1}^n \mathbb{E}[L(\theta_t)] \stackrel{(\star)}{=} \frac{1}{n}\sum_{t=1}^n \mathbb{E}[\ell_t(\theta_t)] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^n \ell_t(\theta_t)\right], \tag{17.0.3}$$

where the equality $(\star)$ follows because θ_t is a function of Z_1^{t-1}, and so conditionally independent of Z_t. Inequality (17.0.3) comprises our first *online-to-batch* conversion: any bound for the average regret in the online setting (17.0.1) implies a matching guarantee for stochastic optimization.

In the remainder of the chapter, we develop algorithms for solving online and stochastic convex optimization problems. It will turn out that *Bregman divergences*, which we have developed in the context of prediction problems, provide one of our central tools. We will also develop lower bounds showing the (minimax) optimality of these methods.

17.1 Preliminaries on convex optimization

Because this chapter develops procedures for solving convex problems based on subgradients of convex functions, we briefly recall their definition. (We refer to Appendix B for proofs.) The *subgradient set*, or *subdifferential*, of a convex function f at the point θ is

$$\partial f(\theta) := \{g \in \mathbb{R}^d \mid f(\theta') \geq f(\theta) + \langle g, \theta' - \theta \rangle \text{ for all } \theta'\},$$

and any vector $g \in \partial f(\theta)$ is a *subgradient*. For convex f, the subdifferential $\partial f(\theta)$ is non-empty for any $\theta \in \operatorname{int} \operatorname{dom} f$. (See Theorem B.2.4 in Appendix B.2.) If f is differentiable at θ, then $\partial f(\theta) = \{\nabla f(\theta)\}$ is a singleton, and we abuse notation to write $\partial f(\theta) = \nabla f(\theta)$.

We give two examples of convex losses and corresponding subgradients. The first highlights a binary classification problem of predicting labels $y \in \{-1, 1\}$ from a data point $x \in \mathbb{R}^d$.

Example 17.1.1 (Support vector machines)**:** We receive data in pairs $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$, and the loss function $\ell_t(\theta) = [1 - y_t \langle \theta, x_t \rangle]_+ = \max\{1 - y_t \langle \theta, x_t \rangle, 0\}$ is convex because it is the maximum of affine functions. The subgradient set is

$$\partial \ell_t(\theta) = \begin{cases} -y_t x_t & \text{if } y_t \langle \theta, x_t \rangle < 1 \\ -\lambda \cdot y_t x_t & \text{for } \lambda \in [0, 1] \text{ if } y_t \langle \theta, x_t \rangle = 1 \\ 0 & \text{otherwise} \end{cases}$$

as $s \mapsto [s]_+$ is differentiable except at $s = 0$. $\diamond$

Example 17.1.2 (Expert prediction and zero-one error)**:** Randomization makes it possible to cast non-convex optimization problems as convex. Consider a problem in which there are d experts, each of which makes a prediction $x_{t,j}$ (for $j = 1, \ldots, d$) at time t, represented by the vector $x_t \in \mathbb{R}^d$. Each suffers the (non-convex) zero-one loss $\ell_{0\text{-}1}(x_{t,j}, y_t) = \mathbf{1}\{x_{t,j} y_t \leq 0\}$. Assign a weight $w_j \geq 0$ to expert $x_{t,j}$, where the weights satisfy $\langle w, \mathbf{1} \rangle = 1$. Then if we choose the prediction $\widehat{y} = \operatorname{sign}(x_{t,j})$ with probability w, the probability of a mistake is

$$\ell_t(w) := \mathbb{P}(\widehat{y} \neq y_t) = \sum_{j=1}^{d} w_j \ell_{0\text{-}1}(x_{t,j}, y_t) = \langle g_t, w \rangle,$$

where we define $g_t = [\ell_{0\text{-}1}(x_{t,j}, y_t)]_{j=1}^d \in \{0, 1\}^d$. The expected zero-one loss is convex (even linear), so its minimization falls into our framework. $\diamond$

17.2 Online convex optimization methods

Our approach to online and stochastic optimization is simple: at times $t = 1, 2, \ldots$, we iteratively

i. construct a simple model of (or approximation to) the instantaneous objective ℓ_t

ii. minimize this model, regularizing so that the updated point does not move too far.

Let us make this more precise before giving derived procedures.

Definition 17.1. *A* model *of the loss ℓ at a point θ_0 is a function $\hat{\ell}(\cdot; \theta_0)$ satisfying*

i. ***Model convexity.*** *The function $\theta \mapsto \hat{\ell}(\theta; \theta_0)$ is convex and subdifferentiable*

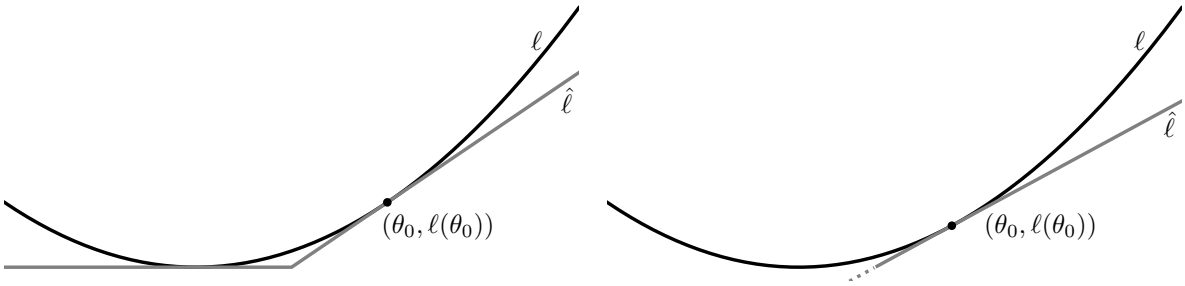


Figure 17.1. Two models (see Definition 17.1) of the loss ℓ at the point θ_0. Left: the truncated model $\hat{\ell}(\theta; \theta_0) = [\ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0)]_+$. Right: the linear model $\hat{\ell}(\theta; \theta_0) = \ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0)$.

ii. ***Lower bound.*** *The model satisfies $\hat{\ell}(\theta; \theta_0) \le \ell(\theta)$, with equality at $\theta = \theta_0$.*

The first-order model provides the most common approximation, taking $g \in \partial\ell(\theta_0)$ and setting

$$\hat{\ell}(\theta; \theta_0) := \ell(\theta_0) + \langle g, \theta - \theta_0 \rangle.$$

Convexity (because $\hat{\ell}$ is affine) and the lower bound condition are immediate. We may employ a somewhat less trivial when the losses are nonnegative; in this case, with $g \in \partial\ell(\theta_0)$ as before, the truncated model

$$\hat{\ell}(\theta; \theta_0) := [\ell(\theta_0) + \langle g, \theta - \theta_0 \rangle]_+$$

satisfies all the desired conditions as well. Figure 17.1 illustrates the definition.

The conditions in Definition 17.1 guarantee that $\hat{\ell}$ looks locally like ℓ; in particular,

$$\partial_\theta \hat{\ell}(\theta; \theta_0)\big|_{\theta=\theta_0} \subset \partial\ell(\theta_0), \tag{17.2.1}$$

so if ℓ is differentiable at θ_0 then $\hat{\ell}$ is as well, with identical derivative. To see the inclusion (17.2.1), note that if $g_0 \in \partial\hat{\ell}(\theta_0; \theta_0)$, then for any θ,

$$\ell(\theta) \ge \hat{\ell}(\theta; \theta_0) \ge \hat{\ell}(\theta_0; \theta_0) + \langle g_0, \theta - \theta_0 \rangle = \ell(\theta_0) + \langle g_0, \theta - \theta_0 \rangle$$

by repeated application of Definition 17.1, so that $g_0 \in \partial\ell(\theta_0)$.

17.2.1 Projected subgradient methods

We begin by studying *(projected) online gradient method*, a special case of the general mirror-descent-type algorithms we will consider, which requires only that we specify a sequence $\eta_t > 0$ of non-increasing stepsizes. Here, we use the simple linear model

$$\hat{\ell}_t(\theta; \theta_t) = \ell_t(\theta_t) + \langle g_t, \theta - \theta_t \rangle, \quad g_t \in \partial\ell_t(\theta_t),$$

and we analyze it via an elementary convergence proof.

The update (17.2.2) makes clear that we trade between improving performance on ℓ_t via the linear approximation $\ell_t(\theta) \approx \ell_t(\theta_t) + \langle g_t, \theta - \theta_t \rangle$ and remaining close to θ_t according to the Euclidean distance $\|\cdot\|_2$. When we use the standard linear approximation $\hat{\ell}_t(\theta) = \ell_t(\theta_t) + \langle g_t, \theta - \theta_t \rangle$, we may rewrite the update (17.2.2) as the two steps

$$\theta_{t+\frac{1}{2}} = \theta_t - \eta_t g_t, \quad \theta_{t+1} = \mathrm{Proj}_\Theta(\theta_{t+\frac{1}{2}}) = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\| \theta - \theta_{t+\frac{1}{2}} \right\|_2^2 \right\},$$

Input: Parameter space Θ, stepsize sequence η_t.
Repeat: for each iteration t,

i. predict $\theta_t \in \Theta$, receive function ℓ_t and suffer loss $\ell_t(\theta_t)$.

ii. let $g_t \in \partial \ell_t(\theta_t)$ and update

$$\theta_{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \langle g_t, \theta \rangle + \frac{1}{2\eta_t} \|\theta - \theta_t\|_2^2 \right\}. \tag{17.2.2}$$

Figure 17.2: Projected gradient descent.

where Proj_Θ denotes Euclidean projection onto Θ. The following proposition, a special case of Theorem 17.2.4 to come, highlights the main ideas present in convergence arguments.

Proposition 17.2.1 (Convergence of projected gradient descent)**.** *Let $\eta_t = \eta > 0$ for all t. Then for any $\theta \in \Theta$,*

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta)] \le \frac{1}{2\eta} \|\theta_1 - \theta\|_2^2 + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_2^2 .$$

Proof The only important step is to identify an appropriate measure of convergence: the error $\|\theta_{t+1} - \theta\|_2^2$, from which we can derive a "one-step" progress guarantee that we recurse. To that end, note that for any $\theta \in \Theta$ and any vector v,

$$\|\mathrm{Proj}_\Theta(v) - \theta\|_2^2 \le \|v - \theta\|_2^2 ,$$

because projections decrease Euclidean distance. (See Corollary B.1.19 in Appendix B.1.3.) So we have the one-step progress guarantee

$$\|\theta_{t+1} - \theta\|_2^2 \le \|\theta_t - \eta g_t - \theta\|_2^2 = \|\theta_t - \theta\|_2^2 - 2\eta \langle g_t, \theta_t - \theta \rangle + \eta^2 \|g_t\|_2^2 .$$

By the first-order conditions for convexity, $\ell_t(\theta_t) + \langle g_t, \theta - \theta_t \rangle \le \ell_t(\theta)$, or $-\langle g_t, \theta_t - \theta \rangle \le \ell_t(\theta) - \ell_t(\theta_t)$, and substituting gives

$$\|\theta_{t+1} - \theta\|_2^2 \le \|\theta_t - \theta\|_2^2 - 2\eta \left[\ell_t(\theta_t) - \ell_t(\theta)\right] + \eta^2 \|g_t\|_2^2 .$$

Rearranging and dividing by $2\eta > 0$ we obtain

$$\ell_t(\theta_t) - \ell_t(\theta) \le \frac{1}{2\eta} \left[\|\theta_t - \theta\|_2^2 - \|\theta_{t+1} - \theta\|_2^2 \right] + \frac{\eta}{2} \|g_t\|_2^2 . \tag{17.2.3}$$

Sum inequality (17.2.3) and note that the sum telescopes to obtain the proposition. □

We will discuss optimizing this bound in Section 17.2.4, where we impose various boundedness conditions on θ and $\|g_t\|_2$, which then allow more evocative guarantees.

17.2.2 Mirror descent methods

By measuring distances in ways that better reflect problem structure, we can achieve sharper convergence guarantees. With this in mind, we present an algorithm generalizing the projected

gradient method 17.2 by replacing the squared Euclidean distance in the update (17.2.2) with a *Bregman divergence*. Recall that for a convex differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$, the *Bregman divergence* associated with ψ is

$$D_\psi(w, v) = \psi(w) - \psi(v) - \langle \nabla\psi(v), w - v\rangle. \tag{17.2.4}$$

The Bregman divergence is non-negative, as $D_\psi(w, v)$ is the gap between the function value $\psi(w)$ and its linear approximation at the point v (recall Figure 3.1). Taking $\psi(w) = \frac{1}{2}\|w\|_2^2$ yields squared Euclidean distance, as $D(w, v) = \frac{1}{2}\|w - v\|_2^2$. The KL-divergence arises via negative entropy: for $\psi(w) = \sum_{j=1}^d w_j \log w_j$ on $\mathbb{R}^d_+$, if $w, v \in \Delta_d = \{u \in \mathbb{R}^d_+ \mid \langle \mathbf{1}, u\rangle = 1\}$, then $D_\psi(w, v) = D_{\rm kl}(w\|v)$.

We now present the mirror descent algorithm. We call ψ the *distance generating function* as it yields the Bregman divergence in the method, functioning analogously to the squared ℓ_2-distance in the earlier Algorithm 17.2.

Input: distance-generating function ψ, parameter space Θ, and non-increasing stepsize sequence $\eta_1, \eta_2, \ldots$.
Repeat: for each iteration t,

i. predict $\theta_t \in \Theta$, receive function ℓ_t and suffer loss $\ell_t(\theta_t)$.

ii. construct model $\hat{\ell}_t$ of ℓ_t at θ_t, and perform non-Euclidean update

$$\theta_{t+1} = \operatorname*{argmin}_{\theta\in\Theta} \left\{ \hat{\ell}_t(\theta; \theta_t) + \frac{1}{\eta_t} D_\psi(\theta, \theta_t) \right\}. \tag{17.2.5}$$

Figure 17.3: The online model-based *mirror descent* algorithm.

The mirror descent update (17.2.5) often admits easy-to-compute solutions. Assume for simplicity that we use the linear model $\hat{\ell}_t(\theta) = \ell_t(\theta_t) + \langle g_t, \theta - \theta_t\rangle$, where $g_t \in \partial\ell_t(\theta_t)$. By taking $\Theta = \mathbb{R}^d$ and $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, mirror descent recovers the gradient update $\theta_{t+1} = \theta_t - \eta_t g_t$. We can also recover the *exponentiated gradients*, also called *entropic mirror descent* or *multiplicative weights*.

Example 17.2.2 (Exponentiated gradients)**:** Let $\Theta = \Delta_d = \{v \in \mathbb{R}^d_+ \mid \langle \mathbf{1}, v\rangle = 1\}$, the probability simplex, and use negative entropy $\psi(w) = \sum_j w_j \log w_j$ with $D_\psi(w, v) = \sum_j w_j \log \frac{w_j}{v_j}$. To compute the update (17.2.5), fixing $v = \theta_t$ for notational simplicity, we must solve

$$\operatorname*{minimize}_\theta \ \langle g, \theta\rangle + \frac{1}{\eta}\sum_j \theta_j \log\frac{\theta_j}{v_j} \ \text{ subject to } \theta \in \Delta_d.$$

Writing the Lagrangian for this problem after introducing multipliers $\tau \in \mathbb{R}$ for the constraint that $\langle \mathbf{1}, \theta\rangle = 1$ and $\lambda \in \mathbb{R}^d_+$ for $\theta \succeq 0$, we have

$$\mathcal{L}(\theta, \lambda, \tau) = \langle g, \theta\rangle + \frac{1}{\eta}\sum_{j=1}^d \theta_j \log\frac{\theta_j}{v_j} - \langle \lambda, \theta\rangle + \tau(\langle \mathbf{1}, \theta\rangle - 1),$$

which we minimize by taking

$$\theta_j = v_j \exp(-\eta g_j + \lambda_j\eta - \tau\eta - 1) > 0.$$

The constraint $\theta \succeq 0$ is inactive so $\lambda_j = 0$, and choosing τ to normalize θ, we obtain the *exponentiated gradient* update

$$\theta_{t+1,i} = \frac{\theta_{t,i} e^{-\eta_t g_{t,i}}}{\sum_j \theta_{t,j} e^{-\eta_t g_{t,j}}} \quad \text{for } i = 1, \ldots, d,$$

the explicit form of the step (17.2.5) when using the linear approximation for $\hat{\ell}$. $\diamond$

17.2.3 Convergence analysis of mirror descent

We turn to an analysis of the mirror descent algorithm. Before presenting the analysis, we require two more definitions to allow us to relate Bregman divergences to norms.

Definition 17.2. *Let $\|\cdot\|$ be a norm. The* dual norm $\|\cdot\|_*$ *associated with $\|\cdot\|$ is*

$$\|y\|_* := \sup_{x : \|x\| \le 1} x^\top y.$$

A straightforward calculation shows that the ℓ_1-norm is dual to the ℓ_∞-norm, and the Euclidean norm $\|\cdot\|_2$ is self-dual. Lastly, we require functions of suitable curvature.

Definition 17.3. *A convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ is* strongly convex with respect to the norm $\|\cdot\|$ over the set Θ *if for all $w, v \in \Theta$ and $g \in \partial\psi(w)$ we have*

$$\psi(v) \ge \psi(w) + \langle g, v - w \rangle + \frac{1}{2} \|w - v\|^2 .$$

The function ψ is strongly convex if it grows at least quadratically everywhere on its domain. Strongly convex functions play an important role in the stability properties of minimizers and enjoy a number of equivalent characterizations; see Appendix C.1.

The definition (17.2.4) of the divergence shows that ψ is strongly convex for $\|\cdot\|$ if and only if

$$D_\psi(w, v) \ge \frac{1}{2} \|w - v\|^2 .$$

Three examples form the standard library of strongly convex functions: Euclidean distance, entropy, and p-norms for $1 < p \le 2$. The first two are straightforward; $\psi(w) = \frac{1}{2}\|w\|_2^2$ satisfies $D_\psi(w, v) = \frac{1}{2}\|w - v\|_2^2$, while for $\psi(w) = \sum_j w_j \log w_j$ we have $D_\psi(w, v) \ge \frac{1}{2}\|w - v\|_1^2$ on the simplex $\Delta = \{w \succeq 0 \mid \langle \mathbf{1}, w \rangle = 1\}$ by Pinsker's inequality (Proposition 2.2.9). It can be convenient to have divergences strongly convex with respect to the ℓ_1-norm over all of $\mathbb{R}^d$ rather than just the probability simplex, and squared ℓ_p-norms provide this guarantee:

Example 17.2.3: Let $\psi(w) = \frac{1}{2(p-1)}\|w\|_p^2$, where $1 < p \le 2$. Then ψ is strongly convex with respect to the ℓ_p-norm $\|\cdot\|_p$. (Exercise 17.7 asks you to prove this.)

In dimension $d \ge 3$, consider the choice $p = 1 + \frac{1}{\log d}$. Then because $\|w\|_1 \le d^{\frac{p-1}{p}} \|w\|_p$ for $p \in [1, \infty]$, this choice of p gives $\|w\|_p \ge d^{\frac{1-p}{p}} \|w\|_1 = d^{-\frac{1}{1+\log d}} \|w\|_1 \ge e^{-1} \|w\|_1$, so

$$D_\psi(w, v) \ge \frac{1}{2}\|w - v\|_p^2 \ge \frac{1}{2e^2}\|w - v\|_1^2 .$$

So $\psi(w) = \frac{e^2}{2(p-1)}\|w\|_p^2$ is strongly convex with respect to the ℓ_1-norm. $\diamond$

With these examples in place, we present the main convergence guarantee for mirror descent.

Theorem 17.2.4 (Regret of mirror descent)**.** *Let ℓ_t be an arbitrary sequence of convex functions, and let θ_t be generated via the mirror descent algorithm 17.3. Assume the distance generating ψ is strongly convex with respect to the norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. Then for a sequence of subgradients $g_t \in \partial\ell_t(\theta_t)$,*

(a) If $\eta_t = \eta$ for all t, then for any $\theta^\star \in \Theta$,

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{1}{\eta} D_\psi(\theta^\star, \theta_1) + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_*^2 .$$

(b) If Θ is compact and $D_\psi(\theta^\star, \theta) \le R^2$ for any $\theta \in \Theta$, then

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{1}{\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} \|g_t\|_*^2 .$$

We defer the proof temporarily to Section 17.2.5.

17.2.4 Instantiations of the regret guarantee

Before proving Theorem 17.2.4, we exhibit its guarantees. Note that Lipschitz continuity properties of the losses are equivalent to bounds on the norms of the subgradients g_t: indeed, the following two conditions are equivalent: (a) ℓ_t is G-Lipschitz with respect to the norm $\|\cdot\|$, meaning

$$|\ell_t(\theta) - \ell_t(\theta')| \le G \left\|\theta - \theta'\right\|$$

for all θ, θ', and (b)

$$\|g\|_* \le G \quad \text{for all } g \in \partial\ell_t(\theta) \text{ and } \theta.$$

(See Exercise 17.8.) When designing a particular instantiation of the online mirror descent algorithm 17.3, we therefore carefully consider the strong convexity properties of ψ that we employ, as these control the gradient magnitudes in Theorem 17.2.4 via the dual norms $\|g_t\|_*$.

Example 17.2.5 (Lipschitz constants and linear prediction)**:** Consider margin-based binary classification, where for data $(x_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$ we have loss $\ell_t(\theta) = \phi(y_t \langle \theta, x_t \rangle)$ for some convex ϕ with $\phi'(0) < 0$ and $|\phi'(s)| \le 1$ for any $\phi'(s) \in \partial\phi(s)$, $s \in \mathbb{R}$. Choices include

i. the logistic loss, with $\phi(t) = \log(1 + e^{-t})$, and

ii. the hinge loss, with $\phi(t) = [1 - t]_+$.

Then $\partial\ell_t(\theta) = y_t x_t \partial\phi(y_t \langle \theta, x_t \rangle)$ and the losses ℓ_t are G-Lipschitz with respect to the norm $\|\cdot\|$ if and only if $\|x\|_* \le G$ for all $x \in \mathcal{X}$. $\diamond$

In the Euclidean case, when $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$, we assume that the loss functions ℓ_t are all G-Lipschitz with respect to the ℓ_2-norm. In Example 17.2.5, this corresponds to the data x belonging to an ℓ_2-ball of radius G. In this case, the regret bounds in Theorem 17.2.4 become

$$\frac{1}{2\eta}\|\theta^\star - \theta_1\|_2^2 + \frac{\eta}{2} n G^2 \quad \text{and} \quad \frac{1}{2\eta_n} R^2 + \sum_{t=1}^n \frac{\eta_t}{2} G^2, \tag{17.2.6}$$

respectively, where the second case assumes that $\|\theta^\star - \theta_t\|_2 \le R$ for all t. In the former, we take $\eta = \frac{R}{G\sqrt{n}}$; in the latter we take $\eta_t = \frac{R}{G\sqrt{t}}$, which requires no knowledge of n.

Corollary 17.2.6. *Assume that $\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le R\}$ and that the losses ℓ_t are G-Lipschitz with respect to the Euclidean norm. Take $\eta_t = \frac{R}{G\sqrt{t}}$. Then for all $\theta^\star \in \Theta$,*

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le 2RG\sqrt{n}.$$

Proof For any $\theta, \theta^\star \in \Theta$, we have $\|\theta - \theta^\star\|_2 \le 2R$, so that $D_\psi(\theta^\star, \theta) \le 2R^2$. The integral bound

$$\sum_{t=1}^n t^{-\frac{1}{2}} \le \int_0^n t^{-\frac{1}{2}} dt = 2\sqrt{n}$$

gives the result. □

Other geometries suggest alternative choices of the distance-generating function ψ. For example, in high-dimensional settings, when the underlying domain Θ is the probability simplex, we can achieve bounds that depend only on the ℓ_∞ norm of the gradients. (Recall Example 17.1.2 for motivation.) In this case, we have the following corollary to Theorem 17.2.4.

Corollary 17.2.7. *Let $\Theta = \Delta_d = \{\theta \in \mathbb{R}^d_+ : \langle \mathbf{1}, \theta\rangle = 1\}$ and take the distance generating $\psi(\theta) = \sum_j \theta_j \log \theta_j$ in the mirror descent procedure 17.3. Then with fixed stepsize η and initial point $\theta_1 = \mathbf{1}/d$, for any sequence of convex losses ℓ_t,*

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_\infty^2 .$$

Proof Pinsker's inequality shows that ψ is strongly convex with respect to $\|\cdot\|_1$. Consequently, taking the dual norm to be the ℓ_∞-norm, part (a) of Theorem 17.2.4 gives

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{1}{\eta} \sum_{j=1}^d \theta^\star_j \log \frac{\theta^\star_j}{\theta_{1,j}} + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_\infty^2 .$$

Note that with $\theta_1 = \mathbf{1}/d$, we have $D_\psi(\theta^\star, \theta_1) \le \log d$ for any $\theta^\star \in \Theta$. □

Corollary 17.2.7 yields sharper results than Corollary 17.2.6 when Θ is the probability simplex in $\mathbb{R}^d$. Let us assume that the subgradients $g_t \in [-1, 1]^d$, the hypercube. In this case, the tightest possible bound on their ℓ_2-norm is $\|g_t\|_2 \le \sqrt{d}$, while $\|g_t\|_\infty \le 1$. Similarly, if $\Theta = \Delta_d$, then we may only bound $\|\theta^\star - \theta_1\|_2 \le \sqrt{2}$. Thus, Euclidean methods (17.2.6) only guarantee regret

$$\frac{1}{2\eta} \|\theta^\star - \theta_1\|_2^2 + \frac{\eta}{2} nd \le \sqrt{2nd} \quad \text{with the choice } \eta = \frac{1}{\sqrt{2nd}},$$

while entropic mirror descent (Alg. 17.3 with $\psi(\theta) = \sum_j \theta_j \log \theta_j$) guarantees

$$\frac{\log d}{\eta} + \frac{\eta}{2} n \le \sqrt{2n \log d} \quad \text{with the choice } \eta = \frac{\sqrt{2 \log d}}{2\sqrt{n}}. \tag{17.2.7}$$

The latter is exponentially better in the dimension.

As a final example, we revisit the p-norm algorithms in high dimensions (at least, when $d \ge 3$).

Example 17.2.8 (p-norm algorithms)**:** Let $1 < p \le 2$, so $\psi(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2$ is strongly convex with respect to $\|\cdot\|_p$ (Example 17.2.3). Exercise 17.9 explores the computation of the update (17.2.5), so we focus exclusively on the regret guarantee. Assume w.l.o.g. that $\mathbf{0} \in \Theta$, and let $\theta_1 = \mathbf{0}$ (otherwise recenter $\psi(\theta) = \frac{1}{2(p-1)} \|\theta - \theta_1\|_p^2$). Consider the fixed stepsize result in Theorem 17.2.4. Then for $q = \frac{p}{p-1}$ we obtain

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \le \frac{\|\theta^\star\|_p^2}{2(p-1)\eta} + \frac{\eta}{2} \sum_{t=1}^n \|g_t\|_q^2 .$$

Choose $p = 1 + \frac{1}{\log d} < 2$, whose conjugate $q = 1 + \log d$ satisfies $\|g\|_q \le d^{1/q} \|g\|_\infty \le e \|g\|_\infty$. As $\|\theta\|_p \le \|\theta\|_1$,

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{\log d}{2\eta} \|\theta^\star\|_1^2 + \frac{e^2 \eta}{2} \sum_{t=1}^n \|g_t\|_\infty^2 .$$

Assume each ℓ_t is G_∞-Lipschitz with respect to the ℓ_1-norm (which is easier to satisfy than Lipschitzness with respect to any other ℓ_p-norm, as it is equivalent to boundedness of the gradients in ℓ_∞), and that the domain Θ is contained in an ℓ_1-ball of ℓ_1-radius R_1. Then the choice $\eta = R_1 \sqrt{\log d}/(G_\infty \sqrt{n})$ implies the bound

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le O(1) G_\infty R_1 \sqrt{n \log d}.$$

As in Corollary 17.2.7, dimension dependence improves over Euclidean methods. $\diamond$

17.2.5 Proof of Theorem 17.2.4

The proof of the theorem proceeds in three lemmas, which distill optimality conditions for convex optimization problems, and then an inductive application.

The first (see Proposition C.1.3 in Appendix C.1) explicitly characterizes optimality for a convex optimization problem.

Lemma 17.2.9. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function and Θ be a closed convex set. Then $\theta^\star$ minimizes h over Θ if and only if there exists $g \in \partial h(\theta^\star)$ such that*

$$\langle g, \theta - \theta^\star \rangle \ge 0 \quad \text{for all } \theta \in \Theta.$$

The next result gives the key progress guarantee for any model-based minimization strategy.

Lemma 17.2.10. *Let $\hat{\ell}$ be any valid model (Definition 17.1) of $\ell : \Theta \to \mathbb{R}$ at the point θ_0, and for $\eta > 0$ define*

$$\theta_\eta := \operatorname*{argmin}_{\theta \in \Theta} \left\{ \hat{\ell}(\theta) + \frac{1}{\eta} D_\psi(\theta, \theta_0) \right\}.$$

Then for some $g_0 \in \partial \ell(\theta_0)$ and any $\theta \in \Theta$,

$$\ell(\theta_0) - \ell(\theta) \le \frac{1}{\eta} \left[D_\psi(\theta, \theta_0) - D_\psi(\theta, \theta_\eta) - D_\psi(\theta_\eta, \theta_0) \right] + \langle g_0, \theta_0 - \theta_\eta \rangle.$$

Proof By the optimality conditions in Lemma 17.2.9, there exists $g_\eta \in \partial\hat{\ell}(\theta_\eta)$ such that

$$\left\langle g_\eta + \eta^{-1}(\nabla\psi(\theta_\eta) - \nabla\psi(\theta_0)), \theta - \theta_\eta \right\rangle \geq 0 \quad \text{for all} \quad \theta \in \Theta.$$

Rearranging, we have

$$\frac{1}{\eta}\langle \nabla\psi(\theta_\eta) - \nabla\psi(\theta_0), \theta - \theta_\eta\rangle \geq \langle g_\eta, \theta_\eta - \theta\rangle.$$

The magical step is the "three term identity," valid for any u, v, w, that

$$D_\psi(u, v) - D_\psi(u, w) - D_\psi(w, v) = \langle \nabla\psi(w) - \nabla\psi(v), u - w\rangle,$$

and substituting $u = \theta$, $v = \theta_0$, and $w = \theta_\eta$, we find that

$$\langle g_\eta, \theta_\eta - \theta\rangle \leq \frac{1}{\eta}\left[D_\psi(\theta, \theta_0) - D_\psi(\theta, \theta_\eta) - D_\psi(\theta_\eta, \theta_0)\right]. \tag{17.2.8}$$

Finally, we use the modeling conditions in Definition 17.1 to lower bound the left hand side of inequality (17.2.8). If $g_0 \in \partial\hat{\ell}(\theta_0)$, then by convexity we obtain

$$\begin{aligned}\langle g_\eta, \theta_\eta - \theta\rangle \overset{(i)}{\geq} \hat{\ell}(\theta_\eta) - \hat{\ell}(\theta) &\overset{(ii)}{\geq} \hat{\ell}(\theta_\eta) - \ell(\theta) \overset{(iii)}{\geq} \hat{\ell}(\theta_0) + \langle g_0, \theta_\eta - \theta_0\rangle - \ell(\theta) \\ &= \ell(\theta_0) + \langle g_0, \theta_\eta - \theta_0\rangle - \ell(\theta)\end{aligned}$$

where inequalities (i) and (iii) are first-order convexity, and inequality (ii) follows from the lower bound condition in Definition 17.1. Substituting and rearranging in inequality (17.2.8) yields the inequality in the lemma once we use the inclusion (17.2.1) that if $g_0 \in \partial\hat{\ell}(\theta_0)$ then $g_0 \in \partial\ell(\theta_0)$. □

Summing the inequality in Lemma 17.2.10 implies the following regret bound.

Lemma 17.2.11. *Let $\ell_t : \Theta \to \mathbb{R}$ be any sequence of convex loss functions and η_t be a non-increasing sequence, where $\eta_0 = \infty$. Then for the mirror descent strategy (17.2.5), there are sub-gradients $g_t \in \partial\ell_t(\theta_t)$ such that for any $\theta^\star \in \Theta$ we have*

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \leq \sum_{t=1}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) D_\psi(\theta^\star, \theta_t) + \sum_{t=1}^n \left[-\frac{1}{\eta_t} D_\psi(\theta_{t+1}, \theta_t) + \langle g_t, \theta_t - \theta_{t+1}\rangle\right].$$

Proof Summing the inequality in Lemma 17.2.10 gives

$$\begin{aligned}\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) &\leq \sum_{t=1}^n \frac{1}{\eta_t}\left[D_\psi(\theta^\star, \theta_t) - D_\psi(\theta^\star, \theta_{t+1}) - D_\psi(\theta_{t+1}, \theta_t)\right] + \sum_{t=1}^n \langle g_t, \theta_t - \theta_{t+1}\rangle \\ &= \sum_{t=2}^n \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) D_\psi(\theta^\star, \theta_t) + \frac{1}{\eta_1} D_\psi(\theta^\star, \theta_1) - \frac{1}{\eta_n} D_\psi(\theta^\star, \theta_{n+1}) \\ &\quad + \sum_{t=1}^n \left[-\frac{1}{\eta_t} D_\psi(\theta_{t+1}, \theta_t) + \langle g_t, \theta_t - \theta_{t+1}\rangle\right]\end{aligned}$$

as desired. □

It remains to use the negative terms $-D_\psi(\theta_t, \theta_{t+1})$ to cancel the gradient terms $\langle g_t, \theta_t - \theta_{t+1}\rangle$. To that end, recall Definition 17.2 of the dual norm $\|\cdot\|_*$ and the strong convexity assumption on ψ. Using the Fenchel-Young inequality that $ab \le \frac{1}{2\eta}a^2 + \frac{\eta}{2}b^2$, valid for any $\eta > 0$, we have

$$\langle g_t, \theta_t - \theta_{t+1}\rangle \le \|g_t\|_* \|\theta_t - \theta_{t+1}\| \le \frac{\eta_t}{2}\|g_t\|_*^2 + \frac{1}{2\eta_t}\|\theta_t - \theta_{t+1}\|^2 .$$

Now, we use the strong convexity condition, which gives

$$-\frac{1}{\eta_t}D_\psi(\theta_{t+1}, \theta_t) \le -\frac{1}{2\eta_t}\|\theta_t - \theta_{t+1}\|^2 .$$

Combining the preceding two displays in Lemma 17.2.11 gives Theorem 17.2.4.

17.3 Optimality guarantees and fundamental limits

Developing lower bounds for stochastic optimization requires some additional technology beyond the minimax bounds we have already developed, as we no longer presume there is a parameter θ to actually estimate. Instead, we only wish to achieve small expected loss $L(\theta) = \mathbb{E}[\ell(\theta; Z)]$ (in the stochastic optimization case (17.0.2)), or small regret (17.0.1). Thus, we redefine the minimax risk to consider the gap in the population losses, so that for loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, we use the shorthand $L_P(\theta) = \mathbb{E}_P[\ell(\theta, Z)]$ and define the *minimax optimization risk*

$$\mathfrak{M}_n(\mathcal{P}, \Theta, \ell) := \inf_{\widehat{\theta}: \mathcal{Z}^n \to \Theta} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_{P^n}\left[L_P(\widehat{\theta}(Z_1^n))\right] - \inf_{\theta \in \Theta} L_P(\theta) \right\}. \tag{17.3.1}$$

Considering the convergence guarantees in the preceding section, we wish to provide minimax lower bounds that capture the geometric aspects of the problems—their Lipschitz continuity properties and relationship with the underlying domain $\Theta \subset \mathbb{R}^d$. It can be tedious to address the interaction between probability distributions and losses ℓ, so we modify the minimax definition (17.3.1). Let $\mathcal{L}$ consist of functions $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{P}(\mathcal{L})$ consist of all probability distributions on $\ell \in \mathcal{L}$, so that we identify the sample space $\mathcal{Z}$ with losses themselves, and make the natural definition $L_P(\theta) := \mathbb{E}_P[\ell(\theta)] = \int \ell(\theta) dP(\ell)$. This makes it easy, for example, to consider the class of convex functions Lipschitz continuous with respect to the p-norm

$$\mathcal{L}_p := \left\{ \text{convex } \ell : \mathbb{R}^d \to \mathbb{R}_+ \text{ s.t. } |\ell(\theta) - \ell(\theta')| \le \left\|\theta - \theta'\right\|_p \text{ for all } \theta, \theta' \in \mathbb{R}^d \right\}. \tag{17.3.2}$$

Then we let

$$\mathfrak{M}_n(\Theta, \mathcal{L}) := \inf_{\widehat{\theta}_n} \sup_{P \in \mathcal{P}(\mathcal{L})} \left\{ \mathbb{E}_{P^n}\left[L_P(\widehat{\theta}_n)\right] - \inf_{\theta \in \Theta} L_P(\theta) \right\}, \tag{17.3.3}$$

where the infimum is over estimators $\widehat{\theta}_n$ observing n observations i.i.d. from P.

17.3.1 From optimization to testing

Our lower bounding technique employs the same idea as our reductions from estimation to testing in Chapter 9, except now we wish to reduce optimization to testing by constructing problems where optimizing well implies that we can solve hypothesis tests. Thus, we define new analogues of packings to show that optimizing well implies accurately testing between different probability

distributions. We define this separation in terms of optimality gaps. For a parameter space Θ and collection $\mathcal{P}$ of distributions,

$$r : \Theta \times \mathcal{P} \to \mathbb{R}_+$$

is a valid *risk gap* if

$$\inf_{\theta \in \Theta} r(\theta, P) = 0 \quad \text{for each } P \in \mathcal{P}. \tag{17.3.4}$$

Example 17.3.1 (A risk gap in stochastic optimization)**:** Given loss $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and associated population losses $L_P(\theta) := \mathbb{E}_P[\ell(\theta, Z)]$, the functional

$$r(\theta, P) := L_P(\theta) - \inf_{\theta \in \Theta} L_P(\theta) \tag{17.3.5}$$

immediately satisfies the condition (17.3.4). $\diamond$

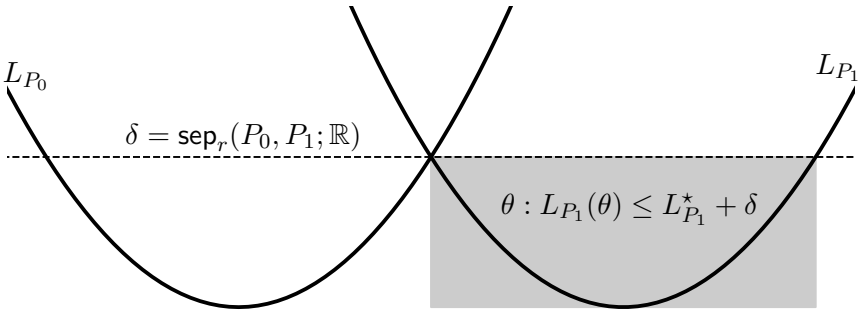


Figure 17.4. The risk separation for the losses L_{P_0} and L_{P_1}. When θ belongs to the shaded region of points with $L_{P_1}(\theta) \le L^\star_{P_1} + \delta$, then necessarily $L_{P_0}(\theta) \ge L^\star_{P_0} + \delta$.

We mimic the reduction from estimation to testing in Chapter 9.2.1. For distributions P_0, P_1, define their *separation* (for the risk gap r) by

$$\mathsf{sep}_r(P_0, P_1; \Theta) := \sup \left\{ \delta \ge 0 \mid \begin{array}{l} r(\theta, P_0) \le \delta \text{ implies } r(\theta, P_1) \ge \delta \\ r(\theta, P_1) \le \delta \text{ implies } r(\theta, P_0) \ge \delta \end{array} \text{ for any } \theta \in \Theta \right\}. \tag{17.3.6}$$

This quantity, which Figure 17.4 illustrates, captures the degree to which having small loss on P_0 implies large loss on P_1 and vice versa. The next examples show simple ways to construct separated objectives using absolute and squared losses. Note that they exhibit different scaling in the natural underlying "separation," which suggests—correctly—that optimizing one loss may be easier than the other.

Example 17.3.2 (A separation for one-dimensional objectives)**:** Consider data $x \in \{-1, 1\}$, and losses $\ell(\theta, x) = |\theta - x|$. Define the distributions P_{-1} and P_1 on X via $P_v(X = v) = \frac{1+\delta}{2}$ and $P_v(X = -v) = \frac{1-\delta}{2}$ for $v \in \{-1, 1\}$, where $\delta \in [0, 1]$. Then by inspection,

$$L_v(\theta) := \mathbb{E}_{P_v}[\ell(\theta, X)] = \frac{1+\delta}{2}|\theta - v| + \frac{1-\delta}{2}|\theta + v|$$

satisfies $\operatorname{argmin}_\theta L_v(\theta) = v$, and $L^\star_v := \inf_\theta L_v(\theta) = (1 - \delta)$. For the risk gap (17.3.5), we use that if $\theta v \le 0$ then $L_v(\theta) - L^\star_v \ge 1 - (1 - \delta) = \delta$, so that

$$\mathsf{sep}_r(P_{-1}, P_1; \mathbb{R}) = \mathsf{sep}_r(P_{-1}, P_1; [-1, 1]) = \delta.$$

We have identical separation $\mathsf{sep}_r(P_{-1}, P_1, \Theta) = \delta$ for any domain $\Theta \supset [-1, 1]$. $\diamond$

To complete the reduction of optimization to testing, we require separation between collections of distributions, so, for a collection $\{P_v\}_{v\in\mathcal{V}}$ indexed by $\mathcal{V}$, we say $\{P_v\}$ is *δ-separated* if

$$\mathsf{sep}_r(P_v, P_{v'}; \Theta) \geq \delta \quad \text{for } v \neq v' \in \mathcal{V}.$$

Then by an argument similar to Proposition 9.2.1, we have the following reduction from *optimization* to testing.

Proposition 17.3.3. *Let $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ be δ-separated for the risk gap $r : \Theta \times \mathcal{P} \to \mathbb{R}_+$. Let V be uniform on $\mathcal{V}$, and, conditional on $V = v$, draw $Z \sim P_v$. Then for any estimator $\widehat{\theta} : \mathcal{Z} \to \Theta$,*

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}[r(\widehat{\theta}, P_v)] \geq \delta \inf_{\widehat{v}} \mathbb{P}(\widehat{v} \neq V), \tag{17.3.7}$$

where $\mathbb{P}$ is the joint distribution over V and Z. If additionally $r(\theta, P) \geq 0$ for all θ, inequality (17.3.7) holds for any estimator rather than estimators taking values only in Θ.

Proof Let $\{P_v\}$ be δ-separated. Then

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}[r(\widehat{\theta}, P_v)] \geq \delta\, \mathbb{P}(r(\widehat{\theta}, P_V) \geq \delta).$$

Define $\Psi : \Theta \to \mathcal{V}$ by

$$\Psi(\theta) = \begin{cases} v & \text{if } r(\theta, P_v) < \delta \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

If $r(\widehat{\theta}, P_v) < \delta$, we must have $\Psi(\theta) = v$, and so $\mathbb{P}(r(\widehat{\theta}, P_V) < \delta) \leq \mathbb{P}(\Psi(\widehat{\theta}) = V)$, that is,

$$\mathbb{P}(r(\widehat{\theta}, P_V) \geq \delta) = 1 - \mathbb{P}(r(\widehat{\theta}, P_V) < \delta) \geq 1 - \mathbb{P}(\Psi(\widehat{\theta}) = V) = \mathbb{P}(\Psi(\widehat{\theta}) \neq V).$$

Take an infimum over tests to get the result. □

As a corollary, we obtain a lower bound on the minimax risk for stochastic optimization by leveraging that $r(\theta, P) = L_P(\theta) - \inf_{\theta^\star\in\Theta} L_P(\theta^\star)$ is a valid risk gap (Example 17.3.1).

Corollary 17.3.4. *Assume that $\mathcal{P}$ has a δ-separated subset $\{P_v\}_{v\in\mathcal{V}}$ for the loss gap (17.3.5). Then the minimax risk (17.3.1) satisfies*

$$\mathfrak{M}_n(\mathcal{P}, \Theta, \ell) \geq \delta \inf_{\widehat{v}} \mathbb{P}(\widehat{v}(Z_1^n) \neq V),$$

where $\mathbb{P}$ denotes any joint distribution over V for which, conditional on $V = v$, $Z_1^n \stackrel{\text{iid}}{\sim} P_v$.

In some cases, the separation (17.3.6) can be hard to compute explicitly, so we introduce the *optimization distance*: for functions L_0, L_1, let $\theta^v \in \operatorname{argmin}_{\theta\in\Theta} L_v(\theta)$, define

$$d_{\text{opt}}(L_0, L_1; \Theta) := \inf_{\theta\in\Theta} \left\{L_0(\theta) + L_1(\theta) - L_0(\theta^0) - L_1(\theta^1)\right\}. \tag{17.3.8}$$

This quantity always lower bounds the risk separation for stochastic optimization.

Lemma 17.3.5. *Let P_0, P_1 be distributions and $L_v(\theta) = \mathbb{E}_{P_v}[\ell(\theta, Z)]$ for $v \in \{0, 1\}$. Then the risk gap $r(\theta, P) := L_P(\theta) - \inf_{\theta \in \Theta} L_P(\theta)$ satisfies*

$$\mathsf{sep}_r(P_0, P_1; \Theta) \geq \frac{1}{2} d_{\rm opt}(L_0, L_1; \Theta).$$

Proof Let $L_v^\star = \inf_{\theta \in \Theta} L_v(\theta)$ for shorthand. Let $\delta \geq 0$ be such that $d_{\rm opt}(L_0, L_1; \Theta) \geq \delta$, and assume that $L_0(\theta) - L_0^\star \leq \delta/2$. Then

$$L_1(\theta) - L_1^\star = L_0(\theta) + L_1(\theta) - L_0^\star - L_1^\star - (L_0(\theta) - L_0^\star) \geq d_{\rm opt}(L_0, L_1; \Theta) - \delta/2 \geq \delta/2.$$

The symmetric case with $L_1(\theta) - L_1^\star$ is similar, so $\mathsf{sep}_r(P_0, P_1, \Theta) \geq \frac{\delta}{2}$. □

17.3.2 Constructing the optimization lower bound

Given Corollary 17.3.4, the program is clear: for a loss ℓ, we take the following three steps.

1. Construct a collection of distributions $\{P_v\}$ whose separation scales with a parameter $\delta \geq 0$.
2. Show that observations $Z_i \stackrel{\rm iid}{\sim} P_v$ contain little information about a random element $v \in \mathcal{V}$.
3. Apply one of the Assouad, Fano, or Le Cam methods to lower bound the testing error, and choose $\delta > 0$ as large as possible while keeping the probability of error $\mathbb{P}(\widehat{v} \neq V)$ a constant.

We follow this program to give a base construction that we can adapt to derive optimal (to within numerical constants) lower bounds for the minimax risk for several optimization problems.

The basic construction We consider distributions $\mathcal{P}$ on the sample space $\mathcal{X} = \{\pm e_j\}_{j=1}^d$ of signed standard basis vectors. For such 1-sparse $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^d$, define the loss

$$\ell(\theta; x) := \begin{cases} |\theta_j - 1| & \text{if } x = e_j \\ |\theta_j + 1| & \text{if } x = -e_j. \end{cases} \tag{17.3.9}$$

With these losses, we have the following building block result.

Theorem 17.3.6. *Let $\Theta \supset [-1, 1]^d$ and $\mathcal{P}$ contain distributions on $\{\pm e_j\}_{j=1}^d$. Then for the loss ℓ defined in (17.3.9), there exists a numerical constant $c > 0$ such that*

$$\mathfrak{M}_n(\mathcal{P}, \Theta, \ell) \geq c \min\left\{\frac{\sqrt{d}}{\sqrt{n}}, 1\right\}.$$

We show how to instantiate Steps 1–3. Beginning with Step 1, let $v \in \{-1, 1\}^d$. For $0 \leq \delta \leq \frac{1}{4}$ to be chosen, define the distribution P_v on $X \in \mathcal{X}$ by

$$X = \begin{cases} v_j e_j & \text{w.p. } \frac{1+\delta}{2d} \\ -v_j e_j & \text{w.p. } \frac{1-\delta}{2d}. \end{cases} \tag{17.3.10}$$

Then the expected loss has the explicit form

$$L_v(\theta) := \mathbb{E}_{P_v}[\ell(\theta, X)] = \frac{1}{d} \sum_{j=1}^d \left[\frac{1+\delta}{2}|\theta_j - v_j| + \frac{1-\delta}{2}|\theta_j + v_j|\right],$$

The distance between v and v' lower bounds the optimization distance between L_v and $L_{v'}$:

Lemma 17.3.7. *Let $\Theta \supset [-1,1]^d$. For any $v, v' \in \{\pm 1\}^d$,*

$$d_{\rm opt}(L_v, L_{v'}; \Theta) = \frac{\delta}{d} \left\|v - v'\right\|_1 .$$

Proof By inspection, $\theta_v := \argmin_\theta L_v(\theta) = v$ and $L_v^\star := \inf_\theta L_v(\theta) = 1 - \delta$. In service of computing the infimum (17.3.8), note that

$$\begin{aligned} & L_v(\theta) + L_{v'}(\theta) - L_v^\star - L_{v'}^\star \\ & = \frac{1}{d} \sum_{j : v_j \neq v'_j} [|\theta_j - 1| + |\theta_j + 1|] + \frac{1}{d} \sum_{j : v_j = v'_j} [(1+\delta)|\theta_j - v_j| + (1-\delta)|\theta_j + v_j|] - 2(1-\delta). \end{aligned}$$

We thus obtain

$$\inf_\theta \{L_v(\theta) + L_{v'}(\theta) - L_v^\star - L_{v'}^\star\} = \frac{1}{d} \left\|v - v'\right\|_1 + \frac{1-\delta}{d}(2d - \left\|v - v'\right\|_1) - 2(1-\delta) = \frac{\delta}{d} \left\|v - v'\right\|_1$$

as claimed. □

Step 2, to bound information, is now relatively straightforward:

Lemma 17.3.8. *Let $X \sim P_v$ according to the coordinate sampling scheme (17.3.10). For $\delta \le \frac{1}{2}$,*

$$D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) \le \frac{9n}{8d} \left\|v - v'\right\|_1 \delta^2.$$

For any $\mathcal{V} \subset \{-1,1\}^d$, if $V \sim \mathsf{Uniform}(\mathcal{V})$ and conditional on $V = v$ we draw $X_1^n \stackrel{\rm iid}{\sim} P_v$,

$$I(X_1^n; V) \le \frac{9n}{4} \delta^2.$$

Proof We demonstrate the first inequality; the second follows from the trivial observation that $I(X_1^n; V) \le \max_{v,v'} D_{\rm kl}\left(P_v^n \| P_{v'}^n\right)$. For any $v, v' \in \{-1,1\}^d$, the sampling scheme (17.3.10) gives

$$\begin{aligned} D_{\rm kl}\left(P_v \| P_{v'}\right) & = \frac{1}{2d} \sum_{j : v_j \neq v'_j} \left[(1+\delta) \log \frac{1+\delta}{1-\delta} + (1-\delta) \log \frac{1-\delta}{1+\delta}\right] \\ & = \frac{\|v - v'\|_1}{4d} \left[(1+\delta) \log \frac{1+\delta}{1-\delta} + (1-\delta) \log \frac{1-\delta}{1+\delta}\right] \le \frac{\|v - v'\|_1}{4d} \left(4\delta^2 + \delta^3\right), \end{aligned}$$

the final inequality valid for $0 \le \delta \le \frac{1}{2}$. Because $D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) = n D_{\rm kl}\left(P_v \| P_{v'}\right)$, we obtain $D_{\rm kl}\left(P_v^n \| P_{v'}^n\right) \le \frac{9n}{8d} \|v - v'\|_1 \delta^2$ for $\delta \le \frac{1}{2}$. □

Now we implement Step 3 to prove the lower bound on testing and final minimax lower bound. As in Section 9.4.1, the Gilbert-Varshamov bound (Lemma 9.2.4) implies that for each $d \in \mathbb{N}$ there exists a $d/2$-packing $\mathcal{V} \subset \{-1,1\}^d$ in ℓ_1 distance with cardinality at least $\exp(d/8)$. Using the loss (17.3.9) and distribution (17.3.10), we therefore have

$$d_{\rm opt}(L_v, L_{v'}; \Theta) \ge \frac{\delta}{2} \quad \text{for } v \neq v' \in \mathcal{V}.$$

Writing $\mathfrak{M}_n = \mathfrak{M}_n(\mathcal{P}, \Theta, \ell)$ for simplicity, we consider two cases: that $d \geq 14$ and $d < 14$. In the former, take the preceding packing $\mathcal{V} \subset \{-1, 1\}^d$ and distribution (17.3.10), which gives separation $\frac{\delta}{4}$ via Lemma 17.3.5. Then Corollary 17.3.4 and Fano's inequality imply that

$$\mathfrak{M}_n \geq \frac{\delta}{4}\left(1 - \frac{I(X_1^n; V) + \log 2}{d/8}\right).$$

By Lemma 17.3.8, we obtain

$$\mathfrak{M}_n \geq \frac{\delta}{4}\left(1 - \frac{8\log 2}{d} - \frac{36n\delta^2}{d}\right) \geq \frac{\delta}{4}\left(\frac{3}{5} - \frac{36n\delta^2}{d}\right),$$

where we used that $d \geq 14$. Choose $\delta^2 = \min\{\frac{1}{5} \cdot \frac{d}{36n}, \frac{1}{4}\}$ to obtain the theorem when $d \geq 14$.

When $d < 14$, we simply take the packing $\mathcal{V} = \{-\mathbf{1}_d, \mathbf{1}_d\}$, which by Lemma 17.3.7 implies that $d_{\rm opt}(L_v, L_{v'}; \Theta) = 2\delta$. Applying Le Cam's method and Corollary 17.3.4, we have

$$\mathfrak{M}_n \geq \frac{\delta}{2}\left(1 - \left\|P_{\mathbf{1}}^n - P_{-\mathbf{1}}^n\right\|_{\rm TV}\right) \geq \frac{\delta}{2}\left(1 - \sqrt{\frac{9n}{4}\delta^2}\right) \tag{17.3.11}$$

by Pinsker's inequality and Lemma 17.3.8. Choose $\delta^2 = \min\{\frac{1}{4}, \frac{9}{16n}\}$ and treat $d < 14$ as a numerical constant.

17.3.3 Instantiations and optimality

By rescaling the sampling distribution (17.3.10) and domain Θ, we can obtain minimax lower bounds for stochastic optimization with different Lipschitz constants and underlying parameter spaces.

Corollary 17.3.9. *Let the collection of losses $\mathcal{L}$ consist of all losses G_1-Lipschitz with respect to the ℓ_∞-norm, meaning $g \in \partial\ell(\theta)$ implies $\|g\|_1 \leq G_1$, and assume $\Theta \supset [-R_\infty, R_\infty]$. Then*

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \geq cR_\infty G_1 \min\left\{\frac{\sqrt{d}}{\sqrt{n}}, 1\right\}.$$

Proof Scale the sampling distribution (17.3.10) to

$$X = \begin{cases} R_\infty v_j e_j & \text{w.p. } \frac{1+\delta}{2d} \\ -R_\infty v_j e_j & \text{w.p. } \frac{1-\delta}{2d} \end{cases}$$

under P_v and scale the losses (17.3.9) via

$$\ell(\theta; x) = \begin{cases} G_1|\theta_j - R_\infty| & \text{if } X_j > 0 \\ G_1|\theta_j + R_\infty| & \text{if } X_j < 0. \end{cases}$$

The proof of Theorem 17.3.6 then proceeds *mutatis mutandis*. □

The main foci in our development of stochastic gradient- and mirror-descent-type methods was to develop algorithms whose convergence guarantees required only boundedness conditions on Θ and the subgradients $g \in \partial\ell$. We can extend Theorem 17.3.6 to address this as well. For $C > 0$, let $C \cdot \mathcal{L} = \{C \cdot \ell \mid \ell \in \mathcal{L}\}$ denote the multiplicative scaling of a collection of losses (so, e.g., if $\mathcal{L}$ consists of 1-Lipschitz functions, then $C \cdot \mathcal{L}$ consists of C-Lipschitz functions). The following corollary provides minimax lower bounds for such collections and domains Θ other than the ℓ_∞-ball.

Corollary 17.3.10. *Let $1 \le q \le 2 \le p \le \infty$. For $q^* = \frac{q}{q-1} \ge 2$, let $\mathcal{L}_{q^*}$ be the collection (17.3.2), and assume Θ contains the ℓ_p-ball of radius R_p. Then there is a numerical constant $c > 0$ such that for all $n \ge d$ and $G > 0$,*

$$\mathfrak{M}_n(\Theta, G \cdot \mathcal{L}_{q^*}) \ge cR_pG \cdot \frac{d^{1/2-1/p}}{\sqrt{n}}.$$

Proof The ℓ_p-ball of radius R_p contains the ℓ_∞ ball of radius $R_\infty := R_p/d^{1/p}$. The modification of the losses (17.3.9) in Corollary 17.3.9 guarantees they are G-Lipschitz: for 1-sparse $x \in \mathbb{R}^d$, the loss $\ell(\theta; x) = G|\theta_j - \sign(x_j)R_p|$ for the coordinate j such that $x_j \neq 0$ satisfies $\|g\|_q \le \|g\|_1 \le G$ for $g \in \partial\ell(\theta; x)$. Applying Corollary 17.3.9 gives $\mathfrak{M}_n(\Theta, \mathcal{L}) \ge c\frac{R_p}{d^{1/p}}G\frac{\sqrt{d}}{\sqrt{n}}$ for $n \ge d$. □

Equivalently, if $\mathcal{L}$ consists of losses ℓ whose subgradients $g \in \partial\ell(\theta)$ satisfy $\|g\|_q \le G_q$ for some $1 \le q \le 2$, and the domain Θ contains the ℓ_p-ball of radius R_p, then

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \ge cR_pG_q\frac{d^{1/2-1/p}}{\sqrt{n}} \quad \text{for } n \ge d.$$

Comparing to the regret bounds for online gradient methods (Corollary 17.2.6) shows that these results are sharp. Indeed, let $\Theta \subset \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le R_2\}$ and assume $\|g\|_2 \le G_2$ for all subgradients g, and let θ_t be the iterates of the stochastic gradient method (17.2.2) with stepsize $\eta = \frac{R_2}{G_2}\frac{1}{\sqrt{t}}$, as in Corollary 17.2.6. Inequality (17.0.3) shows that $\widehat{\theta}_n = \frac{1}{n}\sum_{t=1}^n \theta_t$ satisfies

$$\mathbb{E}[L_P(\widehat{\theta}_n)] \le L_P(\theta^\star) + \frac{2G_2R_2}{\sqrt{n}} \quad \text{for any } \theta^\star \in \Theta.$$

Because $\|g\|_2 \le \|g\|_q$ for $1 \le q \le 2$, and $\|\theta\|_2 \le d^{1/2-1/p}\|\theta\|_p$ for $p \in [2, \infty]$, the ℓ_2-ball of radius $d^{1/2-1/p}R_p$ contains the ℓ_p-ball of radius R_p. The stochastic gradient method therefore guarantees

$$\mathbb{E}[L_P(\widehat{\theta}_n)] - L_P(\theta^\star) \le 2G_qR_p \cdot \frac{d^{1/2-1/p}}{\sqrt{n}}$$

whenever $\Theta \subset \{\theta \mid \|\theta\|_p \le R_p\}$ and the losses satisfy $\|g\|_q \le G_q$ for $g \in \partial\ell(\theta)$. Comparing to Corollary 17.3.10, these rates of convergence are evidently minimax optimal.

These results provide sharp lower bounds for certain cases and gradient/parameter space geometries, and in particular, the dual geometries when Θ is an ℓ_p-ball for some $p \ge 2$ and $\mathcal{L}$ consists of functions Lipschitz with respect to the ℓ_p-norm. Here, we state (leaving the proof to the exercises) a result that naturally handles the geometry in which Θ is contained in an ℓ_p-ball for $p \le 2$. We say a convex set $\Theta \subset \mathbb{R}^d$ is *orthosymmetric* if for $\theta \in \Theta$ and S a diagonal sign matrix, $S\theta \in \Theta$. Any ℓ_p-ball is orthosymmetric. For a vector of nonnegative parameters $\{G_j\}_{j=1}^d$, we let

$$\mathcal{L}(\{G_j\}) := \left\{\text{convex } \ell : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } g \in \partial\ell(\theta) \text{ satisfies } |g_j| \le G_j \text{ for } j = 1, \ldots, d\right\}, \tag{17.3.12}$$

that is, losses whose subgradients have the specified coordinate bounds. Any loss of the form $\ell(\theta; x) = \sum_{j=1}^d G_j|\theta_j - x_j|$ evidently belongs to the class (17.3.12), as do linear functions $\ell(\theta; x) = \langle x, \theta\rangle$ for vectors x satisfying $|x_j| \le G_j$. The following theorem provides a lower bound for this class.

Theorem 17.3.11. *Let $\Theta \subset \mathbb{R}^d$ be an orthosymmetric convex set and $\mathcal{L}$ be the class (17.3.12) of losses. Then*

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \geq \frac{1}{8\sqrt{n}} \sup_{\theta \in \Theta} \sum_{j=1}^d G_j |\theta_j|.$$

Exercise 17.13 steps through one potential proof using Assouad's method.

Theorem 17.3.11, implies minimax lower bounds for other norms on Θ and the gradients. We say $\|\cdot\|$ is *absolute* if $\|Sv\| = \|v\|$ for any vector v and diagonal matrix of signs S, which by inspection implies that the associated dual norm $\|\cdot\|_*$ is absolute. Then carefully choosing the coordinate-Lipschitz parameters G_j in Theorem 17.3.11 gives the following corollary.

Corollary 17.3.12. *Let $\|\cdot\|$ be absolute and $\Theta \subset \mathbb{R}^d$ an orthosymmetric convex set. Let $\mathcal{L}$ consist of losses G-Lipschitz with respect to the norm $\|\cdot\|$. Then*

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \geq \frac{G}{8\sqrt{n}} \sup_{\theta \in \Theta} \|\theta\| .$$

The norms $\|\cdot\|$ on Θ and dual norms $\|\cdot\|_*$ of gradients thus necessarily appear in any lower bound.

When $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_p \leq R_p\}$ is an ℓ_p-ball of radius R_p and $\mathcal{L}$ consists of losses with subgradients satisfying $\|g\|_q \leq G_q$ for $q = \frac{p}{p-1}$, Corollary 17.3.12 implies the lower bound

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \geq \frac{1}{8\sqrt{n}} G_q R_p.$$

For $1 < p \leq 2$, assuming that $\theta_1 = \mathbf{0}$ and we use a fixed stepsize $\eta > 0$, the p-norm algorithms as in Example 17.2.8 guarantee regret

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \leq \frac{\|\theta^\star\|_p^2}{2(p-1)\eta} + \frac{\eta}{2} n G_q^2 \leq \frac{R_p^2}{2(p-1)\eta} + \frac{\eta}{2} n G_q^2$$

for any sequence of losses $\ell_t \in \mathcal{L}$. Choosing $\eta = \frac{R_p}{G_q} \frac{1}{\sqrt{(p-1)n}}$ then yields the minimax upper bound

$$\mathbb{E}\left[L_P(\widehat{\theta}_n)\right] - \inf_{\theta^\star \in \Theta} L_P(\theta^\star) \leq \frac{G_q R_p}{\sqrt{(p-1)n}},$$

where $\widehat{\theta}_n = \frac{1}{n}\sum_{t=1}^n \theta_t$ is the average of the iterates of the mirror descent method 17.3 and we have applied Jensen's inequality as in (17.0.3). To within a factor of $\sqrt{p-1}$, these methods achieve minimax optimal convergence. (The factor $\sqrt{p-1}$ is necessary, though this is beyond our scope [110].)

17.3.4 A lower bound for high-dimensional stochastic optimization

We conclude our discussion of lower bounds for stochastic optimization by considering the particular geometry that arises in sparse problems, where Θ is an ℓ_1 ball and gradients g lie in ℓ_∞ balls. We have seen (inequalities (17.2.7) and Example 17.2.8) that $\sqrt{\log d/n}$ upper bounds the average regret and stochastic convergence rate. This, it turns out, is sharp.

Proposition 17.3.13. *Let $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq R_1\}$ be a scaled ℓ_1-ball and $\mathcal{L}$ consist of the losses satisfying $\|g\|_\infty \leq G_\infty$ for any $g \in \partial \ell(\theta)$. Then*

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \geq \frac{1}{12} G_\infty R_1 \min\left\{\sqrt{\frac{\log(2d)}{n}}, 1\right\}.$$

Proof For shorthand, let $G = G_\infty$ and $R = R_1$, and let $0 \le \delta \le \frac{1}{2}$ to be chosen. Consider linear losses

$$\ell(\theta; x) := G\langle \theta, x \rangle \quad \text{for } x \in \mathcal{X} = \{-1, 1\}^d,$$

which satisfy $\|\nabla \ell(\theta; x)\|_\infty = G \|x\|_\infty = G$. Define the packing set $\mathcal{V} := \{\pm e_j\}_{j=1}^d$ of signed standard basis vectors and let P_v be the distribution on $X \in \{\pm 1\}^d$ with independent coordinates

$$X_j = \begin{cases} 1 & \text{w.p. } \frac{1+\delta v_j}{2} \\ -1 & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases} \tag{17.3.13}$$

We demonstrate loss separation, then apply Fano's method. Define $L_v(\theta) = \mathbb{E}_{P_v}[\ell(\theta; X)] = G\delta\langle v, \theta\rangle$, so that

$$\theta^v := \operatorname*{argmin}_{\theta \in \Theta} L_v(\theta) = -Rv \quad \text{and} \quad L_v^\star := \inf_{\theta \in \Theta} L_v(\theta) = -GR\delta.$$

For any two losses with $v \neq v' \in \{\pm e_j\}_{j=1}^d$, because $\|v + v'\|_\infty \in \{0, 1\}$, we see that

$$\inf_{\theta \in \Theta} \{L_v(\theta) + L_{v'}(\theta)\} = \inf_{\|\theta\|_1 \le R} \left\{G\delta\langle v + v', \theta\rangle\right\} = -GR\delta \left\|v + v'\right\|_\infty \ge -GR\delta.$$

Thus, the optimization distance (17.3.8) has lower bound

$$d_{\rm opt}(L_v, L_{v'}; \Theta) \ge -GR\delta + 2GR\delta = GR\delta.$$

Substituting this into the optimization-to-testing lower bound (17.3.7), we obtain that if we draw $V \sim \mathsf{Uniform}(\mathcal{V})$ and, conditional on $V = v$, draw $X_1^n \stackrel{\rm iid}{\sim} P_v$, then

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \ge \frac{GR\delta}{2} \inf_{\widehat{v}} \mathbb{P}(\widehat{v}(X_1^n) \neq V).$$

Fano's method implies

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \ge \frac{GR\delta}{2} \left(1 - \frac{I(X_1^n; V) + \log 2}{\log(2d)}\right).$$

By a careful accounting of indices in the upper bound $I(X_1^n; V) \le n \cdot \frac{1}{(2d)^2} \sum_{v,v'} D_{\rm kl}\left(P_v \| P_{v'}\right)$, we obtain $I(X_1^n; V) \le \frac{9n}{8}\delta^2$ for $\delta \le \frac{1}{2}$ (see Exercise 17.15). Thus

$$\mathfrak{M}_n(\Theta, \mathcal{L}) \ge \frac{GR\delta}{2} \left(1 - \frac{\log 2}{\log(2d)} - \frac{9n\delta^2}{8\log(2d)}\right).$$

Assuming $d \ge 4$, we have $\log 2 / \log(2d) \le \frac{1}{3}$, so taking $\delta^2 = \frac{8\log(2d)}{27n} \wedge \frac{1}{4}$ to make the testing-error component at least $\frac{1}{3}$ implies the lower bound $\mathfrak{M}_n(\Theta, \mathcal{L}) \ge GR\delta/6$. Substituting the value of δ gives the proposition when $d \ge 4$. In the case that $d \le 3$, a straightforward argument via Le Cam's two-point method gives the claimed lower bound. □

17.4 A miscellany of online optimization guarantees

We provide two extended examples of techniques important in online and stochastic optimizaton. The first details more refined online-to-batch conversions, important when arguing about generalization properties of an online optimization algorithm, while the second shows a refined convergence rate for (essentially) the mixture of experts problems in Example 17.1.2. These particular bounds will prove useful in our discussion of bandit problems in the coming chapters.

17.4.1 An online-to-batch conversion

The expectation-based online-to-batch guarantee (17.0.3) has a high-probability analog, which the Azuma-Hoeffding inequality (Theorem 4.2.4) yields. In the next theorem, a prototypical result, we define the observed regret

$$\mathsf{Reg}_n(\theta) = \sum_{i=1}^n \left[\ell(\theta_i, Z_i) - \ell(\theta, Z_i)\right]$$

for a sequence $Z_1^n \stackrel{\text{iid}}{\sim} P$. The theorem shows that this quantity provides a high-probability bound on the population gap of the averaged iterate relative to a fixed comparator, because for any online procedure, the ith point θ_i is a function only of the past Z_1^{i-1}, so that Z_i provides independent randomness at each iteration.

Theorem 17.4.1. *Let $\mathcal{L}$ consist of convex losses, G-Lipschitz continuous with respect to the norm $\|\cdot\|$, and assume $\sup_{\theta,\theta' \in \Theta} \|\theta - \theta'\| \le R$. Let θ_i be any sequence predicted by an online procedure. Then for any distribution P, $\theta \in \Theta$, and $t \ge 0$,*

$$\mathbb{P}\left(L_P(\overline{\theta}_n) - L_P(\theta) \ge n^{-1}\mathsf{Reg}_n(\theta) + 2\sqrt{2}GR\, t\right) \le \exp(-nt^2).$$

Proof Let $\mathcal{H}_i := \{Z_1^i\}$ be the history at time i, and let $\ell_i(\theta) = \ell(\theta; Z_i)$. We construct a sub-Gaussian martingale difference sequence (Definition 4.5), after which we may apply the Azuma-Hoeffding concentration bound (Theorem 4.2.4). Thus, define

$$D_i := \left(L_P(\theta_i) - L_P(\theta)\right) - \left(\ell_i(\theta_i) - \ell_i(\theta)\right),$$

which evidently are functions of $\mathcal{H}_i$ and martingale differences, as

$$\mathbb{E}[D_i \mid \mathcal{H}_{i-1}] = \left(L_P(\theta_i) - L_P(\theta)\right) - \mathbb{E}[\ell(\theta_i, Z_i) - \ell(\theta, Z_i) \mid Z_1^{i-1}] = 0,$$

because Z_i is independent of Z_1^{i-1} and θ_i is a function of Z_1^{i-1}. Under the Lipschitz and boundedness conditions on ℓ and Θ, we have $|D_i| \le 2G\|\theta_i - \theta\| \le 2GR$. The D_i are thus a $4G^2R^2$-sub-Gaussian martingale difference sequence, and Theorem 4.2.4 implies that $\overline{D}_n = n^{-1}\sum_{i=1}^n D_i$ satisfies $\mathbb{P}(\overline{D}_n \ge t) \le \exp(-\frac{nt^2}{8G^2R^2})$ for all $t \ge 0$. Jensen's inequality gives

$$L_P(\overline{\theta}_n) - L_P(\theta) \le \frac{1}{n}\sum_{i=1}^n \left(L_P(\theta_i) - L_P(\theta)\right) = \frac{1}{n}\sum_{i=1}^n D_i + \frac{1}{n}\mathsf{Reg}_n(\theta),$$

which implies the result. □

It is instructive to compare the generalization bound in Theorem 17.4.1 to the uniform convergence and PAC-Bayesian guarantees in Chapters 5 and 6. Typically, its guarantees are as strong as those from uniform concentration; for example, Example 5.2.12 yields a uniform convergence guarantee that, translated to the notation here, provides the looser upper bound $L_P(\hat{\theta}_n) \le L_P(\theta) + O(1)\frac{GR\sqrt{d}}{\sqrt{n}}t$ with probability at least $1 - ne^{-t^2}$. It is, however, challenging to extend online-to-batch results to provide parameter *recovery* guarantees as strong as those that we obtain from M-estimation problems in Section 5.3, as applications of the Azuma-Hoeffding bound fundamentally yield $\sqrt{n}$ terms. (Some extensions via localization are possible; see the bibliographic section.)

17.4.2 A more refined convergence guarantee for entropic mirror descent

It is sometimes possible to give more refined regret bounds than those we have so far provided. As motivation, let us revisit Example 17.1.2, but suppose that one of the experts has no loss—that is, it makes perfect predictions. We might expect—accurately!—that we should attain better convergence guarantees using exponentiated weights, as exponentiation should quickly eliminate non-optimal experts.

To that end, we present a refined regret bound for the mirror descent algorithm 17.3 with the entropic regularization $\psi(w) = \sum_j w_j \log w_j$. Let $\theta_t \in \Delta_d$ denote the weights at time t.

Proposition 17.4.2. *Let $\psi(w) = \sum_j w_j \log w_j$ and $\theta_t \in \Delta_d$ be generated by mirror descent (Fig. 17.3). Assume that the losses ℓ_t have subgradients with non-negative entries, that is, $g_t \in \partial \ell_t(w)$ implies $g_t \succeq 0$. For any such sequence of loss functions ℓ_t and any $\theta^\star \in \Theta = \Delta_d$,*

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^d \theta_{t,j} g_{t,j}^2.$$

We defer the proof temporarily to Section 17.4.3.

Proposition 17.4.2 replaces the $\|g_t\|_\infty^2$ terms in Corollary 17.2.7 with $\sum_{j=1}^d \theta_{t,j} g_{t,j}^2 \le \|g_t\|_\infty^2$, so it is always at least as strong. Sometimes this can be made tighter, however, as the next corollary exhibits, which applies (for example) to the experts setting of Example 17.1.2.

Corollary 17.4.3. *In addition to the conditions of Proposition 17.4.2, assume the online linear optimization setting, where $\ell_t(\theta_t) = \langle g_t, \theta_t \rangle$ for vectors g_t, and that $g_t \in \mathbb{R}_+^d$ with $\|g_t\|_\infty \le 1$. Let $\sum_{t=1}^n \ell_t(\theta^\star) \le L_n^\star$. Then taking the stepsize $\eta = \min\{1, \sqrt{\log d}/\sqrt{L_n^\star}\}$, we have*

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le 3 \max\left\{\log d, \sqrt{L_n^\star \log d}\right\}.$$

When $\ell_t(\theta^\star) = 0$ for all t, which corresponds to a perfect expert in Example 17.1.2, the upper bound becomes constant in n, yielding regret bound $3 \log d$. Unfortunately, Corollary 17.4.3 unrealistically assumes we know a bound on the loss of the best predictor $\theta^\star$. A standard approach in online learning known as the *doubling* trick replaces a fixed stepsize η with a guessed stepsize, then doubles it if it becomes apparent if it is too small. Exercises 17.1 and 17.2 explore this.

Proof Observe that $\sum_j \theta_j g_{t,j}^2 \le \langle \theta, g_t \rangle$ for any nonnegative θ, as $g_{t,j} \in [0, 1]$. Proposition 17.4.2 gives

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \langle \theta_t, g_t \rangle = \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \ell_t(\theta_t).$$

Rearranging via an algebraic manipulation, this is equivalent to

$$\left(1 - \frac{\eta}{2}\right) \sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \le \frac{\log d}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \ell_t(\theta^\star). \tag{17.4.1}$$

Take $\eta = \min\{1, \sqrt{\log d / L_n^\star}\}$. If $\sqrt{\log d / L_n^\star} \le 1$, the right hand side of the above inequality becomes $\sqrt{L_n^\star \log d} + \frac{1}{2}\sqrt{L_n^\star \log d}$. If $L_n^\star < \log d$, then the right hand side of the inequality becomes $\log d + \frac{1}{2} L_n^\star \le \frac{3}{2} \log d$. In either case, we obtain the desired result by noting that $1 - \frac{\eta}{2} \ge \frac{1}{2}$. □

17.4.3 Proof of Proposition 17.4.2

Our proof relies on a technical lemma, which applies to the exponentiated gradient update, after which the derivation is a straightforward consequence of Lemma 17.2.11.

Lemma 17.4.4. *Let $\psi(x) = \sum_j x_j \log x_j$, $\eta > 0$, and let $x, y \in \Delta_d$ be defined by*

$$y_i = \frac{x_i \exp(-\eta g_i)}{\sum_j x_j \exp(-\eta g_j)},$$

where $g \in \mathbb{R}^d_+$ is non-negative. Then

$$-\frac{1}{\eta} D_\psi(y, x) + \langle g, x - y \rangle \le \frac{\eta}{2} \sum_{i=1}^d g_i^2 x_i.$$

Deferring the proof of the lemma, we note that it precisely applies to the setting of Lemma 17.2.11. Indeed, with a fixed stepsize η, we have

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \le \frac{1}{\eta} D_\psi(\theta^\star, \theta_1) + \sum_{t=1}^n \left[-\frac{1}{\eta} D_\psi(\theta_{t+1}, \theta_t) + \langle g_t, \theta_t - \theta_{t+1} \rangle \right].$$

In the proof of Theorem 17.2.4 following Lemma 17.2.11, we used the strong convexity of ψ to eliminate the gradient terms $\langle g_t, \theta_t - \theta_{t+1} \rangle$ through the Bregman divergence D_ψ. This time, we use Lemma 17.2.11: setting $y = \theta_{t+1}$ and $x = \theta_t$ yields the bound

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \le \frac{1}{\eta} D_\psi(\theta^\star, \theta_1) + \sum_{t=1}^n \frac{\eta}{2} \sum_{i=1}^d g_{t,i}^2 \theta_{t,i}$$

as desired.

Proof of Lemma 17.4.4 Substituting the values for x and y into $D_\psi(y, x) = D_{\rm kl}(y \| x)$, we have

$$\sum_i y_i \log \frac{y_i}{x_i} = \sum_i y_i \log \left(\frac{x_i \exp(-\eta g_i)}{x_i (\sum_j \exp(-\eta g_j) x_j)} \right) = -\eta \langle g, y \rangle - \sum_i y_i \log \Big(\sum_j x_j e^{-\eta g_j} \Big).$$

Now we use a Taylor expansion of the function $g \mapsto \log(\sum_j x_j e^{-\eta g_j})$ around the point 0. If we define the vector $p(g)$ by $p_i(g) = x_i e^{-\eta g_i} / (\sum_j x_j e^{-\eta g_j})$, then

$$\log \Big(\sum_j x_j e^{-\eta g_j} \Big) = \log(\langle \mathbf{1}, x \rangle) - \eta \langle p(0), g \rangle + \frac{\eta^2}{2} g^\top (\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g}) p(\widetilde{g})^\top) g,$$

where $\widetilde{g} = \lambda g$ for some $\lambda \in [0, 1]$. Noting that $p(0) = x$ and $\langle \mathbf{1}, x \rangle = \langle \mathbf{1}, y \rangle = 1$, we obtain

$$D_\psi(y, x) = -\eta \langle g, y \rangle + \log(1) + \eta \langle g, x \rangle - \frac{\eta^2}{2} g^\top (\mathrm{diag}(p(\widetilde{g})) - p(\widetilde{g}) p(\widetilde{g})^\top) g,$$

whence

$$-\frac{1}{\eta} D_\psi(y, x) + \langle g, x - y \rangle \le \frac{\eta}{2} \sum_{i=1}^d g_i^2 p_i(\widetilde{g}). \tag{17.4.2}$$

Lastly, we claim that the function

$$s(\lambda) = \sum_{i=1}^{d} g_i^2 \frac{x_i e^{-\lambda g_i}}{\sum_j x_j e^{-\lambda g_j}}$$

is non-increasing on $\lambda \in [0, 1]$. Indeed,

$$s'(\lambda) = \frac{(\sum_i g_i x_i e^{-\lambda g_i})(\sum_i g_i^2 x_i e^{-\lambda g_i})}{(\sum_i x_i e^{-\lambda g_i})^2} - \frac{\sum_i g_i^3 x_i e^{-\lambda g_i}}{\sum_i x_i e^{-\lambda g_i}} = \frac{\sum_{ij} g_i g_j^2 x_i x_j e^{-\lambda g_i - \lambda g_j} - \sum_{ij} g_i^3 x_i x_j e^{-\lambda g_i - \lambda g_j}}{(\sum_i x_i e^{-\lambda g_i})^2}.$$

Using the Fenchel-Young inequality, we have $ab \le \frac{1}{3}|a|^3 + \frac{2}{3}|b|^{3/2}$ for any a, b, so $g_i g_j^2 \le \frac{1}{3} g_i^3 + \frac{2}{3} g_j^3$. This implies that the numerator in the expression for $s'(\lambda)$ is non-positive, so $s(\lambda) \le s(0) = \sum_{i=1}^d g_i^2 x_i$, which gives the result when combined with inequality (17.4.2). □

17.5 Bibliography and further reading

This chapter provides a view into the small slice of stochastic and online optimization most connected to the interests of this book, with a view toward information-theoretic techniques and further uses in the coming chapters. Stochastic approximation and gradient methods originate with Robbins and Monro's introduction of the stochastic gradient method [362], and their importance exploded in the 2000s with the advent of large-scale optimization problems central to machine learning.

A major vein of work develops convergence guarantees for stochastic approximation methods. In this line, Polyak and Juditsky [344] demonstrate the importance of averaging to optimal convergence guarantees, re-igniting interest in effective stochastic gradient methods. In the early 2000s, Nedić and Bertsekas [319] introduced and analyzed incremental gradient methods, close cousins to stochastic gradient methods, and Zinkevich [440] gave the important connection between online learning and (online) convex optimization. Nemirovski et al. [323] contribute to this line of work by showing the robustness of various stochastic and mirror-descent methods based on averaging, with complementary high-probability convergence guarantees, and Shapiro et al.'s book [378] provides an excellent overview of the stochastic optimization literature. Mirror descent methods originate with Nemirovski and Yudin [322], who introduce them to address high-dimensional convex optimization problems, and Beck and Teboulle [55] provide the elegant interpretation of them as non-Euclidean projected gradient methods, as in Figure 17.3. The approach we develop for convergence based on models of the loss, as in Definition 17.1, follows from work of Davis and Drusvyatskiy [129] and Asi and Duchi [28, 29], who show how the approach provides convergence guarantees for problems beyond the standard purview of stochastic optimization.

The online learning and optimization literature begins with Rosenblatt's perceptron [364]. This literature developed mostly independently from the optimization literature, with analogues of mirror descent appearing with Littlestone and Warmuth's weighted majority (multiplicative weights) algorithms [291] and Gentile and Littlestone [208], who provide p-norm-based approaches. The multiplicative-weights methods (i.e. entropic mirror descent, as in Example 17.2.2) form an important building block for algorithms in theoretical computer science [25]. Cesa-Bianchi, Conconi, and Gentile [106] demonstrate the online-to-batch conversions in Section 17.4.1, and extensions of these ideas to problems (such as strongly convex loss functions) that allow sharper convergence guarantees (even $O(1/n)$), often based on Freedman's martingale extensions [199] of Bernstein-type

concentration inequalities in Section 4.1.2, appear [261, 233, 273, 8]. Cesa-Bianchi and Lugosi [105] provide an overview of online learning as of the mid 2000s, and Shalev-Shwartz [374] and Hazan [232] develop the connections between convex optimization and online learning. There is an important line of work on adaptive algorithms, which attempt to dynamically set method parameters such as stepsize or other quantities (see Exercise 17.12), which we have not addressed; the AdaGrad method [157, 312] inspires much of this, and more recent work attempts to adapt to ever more problem parameters (as well as identify limitations of adaptation) [125, 126, 135].

We do not cover substantial portions of the optimization literature, even as it connects to information theory, especially in deterministic optimization. Nemirovski and Yudin [322] forms the foundational text, treating optimality within both deterministic and stochastic optimization problems; see also Nesterov [324]. Braun, Guzmán, and Pokutta [83] develop information-theoretic tools to prove fundamental limits for randomized optimization procedures, which inspires a line of work on optimality results in optimization by Carmon and collaborators [101, 102, 24].

17.6 Exercises

Exercise 17.1: The *doubling trick* allows good performance of online learning procedures even without knowledge of the number of steps n they will run. Define the regret

$$\mathsf{Reg}_n := \sup_{\theta^\star \in \Theta} \sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)].$$

Let an algorithm $\mathsf{A}(\eta)$ parameterized by $\eta \geq 0$ (usually a stepsize) achieve the regret bound

$$\mathsf{Reg}_n \leq \frac{r^2}{2\eta} + \frac{\eta}{2} L^2 n,$$

where r and L are known. Consider a procedure proceeding in epochs $k = 1, 2, \ldots$ lasting for $n_k = 2^k$ steps. At the start of epoch k, restart the algorithm $\mathsf{A}(\eta)$ with parameter $\eta_k = \frac{r}{L\sqrt{2^k}}$, and run the algorithm with this choice for 2^k steps. Show that

$$\mathsf{Reg}_n \leq C \cdot Lr\sqrt{n},$$

where C is a numerical constant (our solution has $C \leq 2/(\sqrt{2} - 1)$).

Exercise 17.2 (The doubling trick and problem difficulty)**:** Consider the mixture of experts problem in Example 17.1.2, where d experts suffer losses in $[0, 1]$ at each timestep and let $g_t \in [0, 1]^d$ denote the loss vector. We play a mixture of experts $\theta_t \in \Theta = \Delta_d$, suffering (expected) loss $\ell_t(\theta_t) = \langle g_t, \theta_t \rangle$. Consider the following procedure, which proceeds in epochs $k = 1, 2, \ldots$, within each of which we perform exponential weights with stepsize $\eta_k = \min\{1, \sqrt{\log d}/2^k\}$. Let E_k denote the times t belonging to epoch k. Define $L^{(k)} = \min_j \sum_{t \in E_k} g_{t,j}$ to be the loss incurred by the best expert in epoch k as the procedure runs, and continue epoch k until $L^{(k)} \geq 4^k$. Then begin a new epoch. Use the regret bounds in Section 17.4.2 to show that with this procedure,

$$\sum_{t=1}^n [\ell_t(\theta_t) - \ell_t(\theta^\star)] \leq C_1 \log\log d \cdot \log d + C_2 \sqrt{\log d \cdot \sum_{t=1}^n \ell_t(\theta^\star)}$$

for numerical constants C_1 and C_2 (we obtain $C_1 \leq 3$ and $C_2 \leq 8\sqrt{2}$).

Exercise 17.3 (Min-max games and regret)**:** The *saddle point problem* considers solving

$$\minimize_{x \in X} \sup_{y \in Y} L(x, y), \tag{17.6.1}$$

where L is convex in its first argument and concave in its second, and X, Y are convex sets. A point $(x^\star, y^\star) \in X \times Y$ is a *saddle point* if

$$\sup_{y \in Y} L(x^\star, y) \le L(x^\star, y^\star) \le \inf_{x \in X} L(x, y^\star).$$

(a) Show that if a saddle point exists, then

$$\inf_{x \in X} \sup_{y \in Y} L(x, y) = \sup_{y \in Y} \inf_{x \in X} L(x, y) = L(x^\star, y^\star).$$

Online convex optimization can identify saddle points. Assume X and Y are compact convex and that for each $x_0 \in X$, $L(x_0, y)$ is 1-Lipschitz and concave in y, and for each $y_0 \in Y$, $L(x, y_0)$ is 1-Lipschitz and convex in x. Consider the best-response game: at iteration t, the X player chooses $x_t \in X$, and then the Y player chooses

$$y_t \in \argmax_{y \in Y} L(x_t, y).$$

(b) Define $f_t(x) = L(x, y_t)$. Give a strategy for the x player so that for any T and any $x^\star \in X$,

$$\sum_{t=1}^T f_t(x_t) - f_t(x^\star) \le O(1)\sqrt{T}.$$

(c) Show that for any $x^\star \in X$, $\overline{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$ and $\overline{y}_T = \frac{1}{T} \sum_{t=1}^T y_t$ satisfy

$$\sup_{y \in Y} L(\overline{x}_T, y) \le L(x^\star, \overline{y}_T) + O(1)/\sqrt{T}.$$

(d) Show there exists a saddle point for the min-max problem (17.6.1) under the above conditions.

Exercise 17.4 (von Neumann's minimax theorem)**:** Let $A \in \mathbb{R}^{m \times n}$ and X and Y be compact convex sets. Show that

$$\inf_{x \in X} \sup_{y \in Y} \langle x, Ay \rangle = \sup_{y \in Y} \inf_{x \in X} \langle x, Ay \rangle,$$

and that there exists a saddle point $(x^\star, y^\star) \in X \times Y$ for which

$$\sup_{y \in Y} \langle x, Ay \rangle \le \langle x^\star, Ay^\star \rangle \le \inf_{x \in X} \langle x, Ay^\star \rangle.$$

In the language of games, there are strategies $(x^\star, y^\star)$ for which play order does not matter.

Exercise 17.5 (Expert prediction of an individual sequence)**:** Let $x_t \in \{-1, 1\}^d$ be a sequence of vectors and $y_t \in \{-1, 1\}$, where $x_{t,j} \in \{-1, 1\}$ indicates an "expert" prediction of y_t. Define the 0-1 regret Reg_n of a sequence of predictions $\hat{y}_t$ by

$$\mathsf{Reg}_n := \sum_{t=1}^n \mathbf{1}\left\{\hat{y}_t \neq y_t\right\} - \min_j \sum_{t=1}^n \mathbf{1}\left\{x_{t,j} \neq y_t\right\}.$$

(a) Let $\theta \in \Delta_k$ and conditional on x, draw $\hat{y} = 1$ with probability $\frac{1+\theta^\top x}{2}$ and $\hat{y} = -1$ with probability $\frac{1-\theta^\top x}{2}$. Show that $2\mathbb{E}[\mathbf{1}\{\hat{y} \neq y\}] = |x^\top \theta - y|$.

(b) Let θ_t a function of $(x_\tau, y_\tau)_{\tau<t}$ and at time t, conditional on $x_t \in \{-1, 1\}$, draw $\hat{y}_t = 1$ with probability $\frac{1}{2}(1 + \theta_t^\top x_t)$, $\hat{y}_t = -1$ otherwise. Show that with probability at least $1 - \delta$,

$$\sum_{t=1}^n \mathbf{1}\{\hat{y}_t \neq y_t\} \leq \frac{1}{2}\sum_{t=1}^n |\theta_t^\top x_t - y_t| + \sqrt{\frac{n}{2}\log\frac{1}{\delta}}.$$

(c) Let θ_t be generated according to the mirror descent method with $\psi(\theta) = \sum_j \theta_j \log \theta_j$ over the simplex Δ_k with losses $\ell(\theta; x, y) = |\theta^\top x - y|$, and let (x_t, y_t) be an arbitrary sequence of expert predictions and labels y_t, where y_t depends only on $(x_\tau, y_\tau, \hat{y}_\tau)_{\tau<t}$ and x_t. Show that with probability at least $1 - \delta$ over the random sampling of $\hat{y}_t$ in part (b),

$$\mathsf{Reg}_n = \sum_{t=1}^n \mathbf{1}\{\hat{y}_t - y_t\} - \min_j \sum_{t=1}^n \mathbf{1}\{x_{t,j} \neq y_t\} \leq O(1) \cdot \sqrt{n \log \frac{d}{\delta}}.$$

Exercise 17.6 (Second-order strong convexity conditions)**:** Let f be twice continuously differentiable on its domain, which you may assume to be open.

(a) Assume that f is strongly convex with respect to the norm $\|\cdot\|$. Show that for all $x \in \operatorname{dom} f$, $u^\top \nabla^2 f(x) u \geq \|u\|^2$ for all u.

(b) Show that if $u^\top \nabla^2 f(x) u \geq \|u\|^2$ for all u and $x \in \operatorname{dom} f$, then f is strongly convex with respect to the norm $\|\cdot\|$.

Hint. See Proposition C.1.4 in Appendix C.1.

Exercise 17.7 (The strong convexity of p-norms)**:** We use the results of Exercise 17.6 to show that for $1 < p < 2$, $f_p(w) := \frac{1}{2(p-1)}\|w\|_p^2$ is strongly convex with respect to the norm $\|\cdot\|_p$.

(a) Define $\psi(t) = \frac{1}{2(p-1)} t^{2/p}$ and $\phi(t) = |t|^p$. Let $H(w) = \nabla^2 f_p(w)$. Show that

$$H_{ii}(w) = \psi''\left(\|w\|_p^p\right)\left(\phi'(w_i)\right)^2 + \psi'\left(\|w\|_p^p\right)\phi''(w_i)$$

and for $i \neq j$,

$$H_{ij}(w) = \psi''\left(\|w\|_p^p\right)\phi'(w_i)\phi'(w_j).$$

(b) Show that for all u,

$$u^\top \nabla^2 f_p(w) u \geq \|w\|_p^{2-p} \sum_{j=1}^d \frac{u_j^2}{|w_j|^{2-p}}.$$

(c) Use Hölder's inequality to demonstrate that $u^\top \nabla^2 f_p(w) u \geq \|u\|_p^2$ for all u.

Exercise 17.8: Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex.

(a) Show that if f is G-Lipschitz with respect to the norm $\|\cdot\|$, meaning $|f(x) - f(y)| \leq G\|x - y\|$, then for all x and $g \in \partial f(x)$, $\|g\|_* \leq G$.

(b) Show that if for all x and $g \in \partial f(x)$, $\|g\|_* \le G$, then f is G-Lipschitz with respect to $\|\cdot\|$.

Hint. Use Corollary B.2.30 and Proposition B.2.31 in Appendix B.2.6.

Exercise 17.9: We show how to compute the p-norm-based mirror descent update (17.2.5).

(a) Let $h(x) = \frac{1}{2\eta}\|x\|^2$. Show that the conjugate $h^*(y) = \frac{\eta}{2}\|y\|_*^2$.

(b) Let $\psi(\theta) = \frac{1}{2}\|\theta\|_p^2$, where $1 < p \le 2$. Let $\theta_0 \in \mathbb{R}^d$, and for a fixed $g \in \mathbb{R}^d$ define

$$\theta_\eta = \operatorname*{argmin}_\theta \left\{ \langle g, \theta \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_0) \right\}.$$

Show that if we let w have entries

$$w_j = \frac{1}{\|\theta_0\|^{p-2}} |\theta_{0,j}|^{p-1} \operatorname{sign}(\theta_{0,j}) - \eta g_j$$

then for $q = \frac{p}{p-1}$ conjugate to p,

$$\theta_\eta = \frac{1}{\|w\|_q^{q-2}} \left[|w_j|^{q-1} \operatorname{sign}(w_j) \right]_{j=1}^d.$$

Hint. Use Exercise 17.10.

Exercise 17.10 (The naming of mirror descent)**:** Let $\psi : \mathbb{R}^d \to \overline{\mathbb{R}}$ be *Legendre*, meaning that it is strictly convex, continuously differentiable, and satisfies the conditions (14.3.4). Show that for $g \in \mathbb{R}^d$, $\eta > 0$, and $\theta_0 \in \operatorname{dom} \psi$,

$$\theta_\eta = \operatorname*{argmin}_\theta \left\{ \langle g, \theta \rangle + \frac{1}{\eta} D_\psi(\theta, \theta_0) \right\}$$

satisfies

$$\theta_\eta = \nabla \psi^* \left(\nabla \psi(\theta_0) - \eta g \right).$$

We therefore may think of the update (17.2.5) as performing a "mirroring" operation: we transform θ_0 into a dual version $\nabla\psi(\theta_0)$, where gradients belong to the dual space, update the parameter, then "reflect" back into the primal space via $\nabla\psi^* : \mathbb{R}^d \to \operatorname{dom}\psi$. *Hint.* Use Corollary 14.3.1.

Exercise 17.11: Let $\psi(u) = \frac{1}{2}\|u\|_p^2$, where $1 < p < \infty$.

(a) Show that for any u, v,

$$D_\psi(u, v) \le \frac{1}{2}\|u\|_p^2 + \frac{1}{2}\|u - v\|_p^2.$$

(b) Let $\Theta \subset \mathbb{R}^d$ satisfy that $\|\theta - \theta_0\|_p \le R_p$ for all $\theta \in \Theta$. Show that if $\psi(u) = \frac{1}{2}\|u - \theta_0\|_p^2$, then

$$\sup_{\theta, \theta' \in \Theta} D_\psi(\theta, \theta') \le \frac{5}{2} R_p^2.$$

Exercise 17.12 (Adaptive stepsizes)**:** Let ψ be strongly convex with respect $\|\cdot\|$ over Θ and let θ_t be generated by the iteration (17.2.5). Fix $\eta > 0$, and define the tth stepsize

$$\eta_t = \frac{\eta}{\sqrt{\sum_{\tau=1}^t \|g_\tau\|_*^2}}, \tag{17.6.2}$$

which is computable at iteration t. Assume that $D_\psi(\theta^\star, \theta) \le R^2$ for all $\theta \in \Theta$.

(a) Show that

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \le \frac{1}{\eta} R^2 \sqrt{\sum_{t=1}^n \|g_t\|_*^2} + \eta \sqrt{\sum_{t=1}^n \|g_t\|_*^2}.$$

(b) Show that with the fixed stepsize multiplier $\eta = R$,

$$\sum_{t=1}^n \ell_t(\theta_t) - \ell_t(\theta^\star) \le \sqrt{2} \inf_\alpha \left\{ \frac{R^2}{\alpha} + \frac{\alpha}{2} \sum_{t=1}^n \|g_t\|_*^2 \right\},$$

so that the adaptive steps (17.6.2) enjoy a type of post-hoc optimality guarantee.

Exercise 17.13 (Coordinate-based lower bounds [158])**:** This question explores using Assouad's method for lower bounds in stochastic convex optimization, providing a proof of Theorem 17.3.11. Let $G_j \ge 0$ and $\mathcal{L}$ be the coordinate-bounded loss class (17.3.12).

(a) Fix a vector $a \in \mathbb{R}^d$. Define

$$\ell(\theta; x) := \sum_{j=1}^d G_j |\theta_j - a_j x_j|.$$

Show that for each $x \in \mathbb{R}^d$, $\ell(\cdot, x)$ belongs to the class $\mathcal{L}$.

(b) Use a variant of Assouad's method to prove Theorem 17.3.11, at least to within a numerical constant. *Hint.* Define P_v on $X \in \{-1, 1\}^d$, indexed by $v \in \{-1, 1\}^d$, so that $X \sim P_v$ has independent coordinates $X_j \sim \mathsf{Bernoulli}(\frac{1+v_j\delta}{2})$. Take $a \in \Theta$ and use the loss above.

Exercise 17.14 (A geometric lower bound)**:** Prove Corollary 17.3.12.

Exercise 17.15: Let $V \in \mathcal{V} = \{\pm e_j\}_{j=1}^d$ be uniform on $\mathcal{V}$, and conditional on $V = v$, let $X_i \in \{-1, 1\}^d$ be i.i.d. with coordinate distribution (17.3.13), where $0 \le \delta \le \frac{1}{2}$. Show that $I(X_1^n; V) \le \frac{9n}{8}\delta^2$. *Hint.* Use that $I(X_1^n; V) \le \frac{1}{(2d)^2} \sum_{v \ne v'} D_{\rm kl}\left(P_v^n \| P_{v'}^n\right)$ and count indices.

Chapter 18

Exploration, exploitation, and bandit problems

Consider evaluating a possible treatment for a disease, where we wish to allocate resources to attain the best outcomes while simultaneously testing treatment efficacy. Because for each patient we may only observe the outcome in one of treatment or control, this sequential treatment assignment illustrates a balancing act between exploration—observing the effects of treatment or non-treatment—and exploitation—giving treatment or not as we decide which has better responses. We focus mainly on variants of the K-armed bandit problem, so named because we imagine a gambler choosing between K different slot machines (bandits). Each machine has a different and unknown reward distribution. The player wishes to put as much money as possible into the machine with the greatest expected reward. This chapter will provide the tools to analyze such games, and we also will develop techniques for demonstrating their information-theoretic limits.

18.1 The multi-armed bandit problem

In the most basic setting we consider the following sequential decision making scenario. We assume that there are K distributions $P_1, \ldots, P_K$ on $\mathbb{R}$, which we identify with K random variables $Y(1), \ldots, Y(K)$. Random variable $Y(i)$ has mean μ_i and is σ^2-sub-Gaussian, meaning that

$$\mathbb{E}\left[\exp\left(\lambda(Y(i) - \mu_i)\right)\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right). \tag{18.1.1}$$

The goal is to find the index i with the maximal mean μ_i without sampling sub-optimal distributions P_i (or random variables $Y(i)$) too often.

We consider this in an online game-playing setting, proceeding iteratively for $t = 1, 2, \ldots$, where at each iteration t of the process, the "player" (or learner) takes an action $A_t \in \{1, \ldots, K\}$, then, conditional on $i = A_t$, observes a reward $Y_t(i)$ drawn independently from the distribution P_i. Given the past, the action A_t is independent of the vector $Y_t := (Y_t(1), \ldots, Y_t(K))$, and at time t, we can only observe feedback for the action A_t, i.e., $Y_t(A_t)$. The goal is to minimize the n-step regret

$$\mathsf{Reg}_n := \sum_{t=1}^{n} \mu_{i^\star} - \mu_{A_t}, \tag{18.1.2}$$

where $i^\star \in \operatorname{argmax}_i \mu_i$ so $\mu_{i^\star} = \max_i \mu_i$. While the regret (18.1.2) involves only the means of the $Y(i)$, its dependence on A_t makes it random, so we generally seek to bound its expectation or

deviation probabilities. For now, we focus on the expected regret,

$$\overline{\mathsf{Reg}}_n := \mathbb{E}\left[\sum_{t=1}^n \mu_{i^\star} - \mu_{A_t}\right], \tag{18.1.3}$$

where the expectation is taken over any randomness in the player's actions A_t and in the repeated observations of the random variables $Y(1), \ldots, Y(K)$. We revisit Example 11.2.5 from Section 11.2.3 in the language we adopt here.

Example 18.1.1 (Potential outcomes)**:** Consider estimating the effect of a binary treatment, so the set $\{0, 1\}$ of actions corresponds to treatment or control. In the *Neyman-Rubin causal model*, an individual i has *potential outcomes* $Y_i := (Y_i(0), Y_i(1))$, where $Y_i(0)$ indicates the individual's response under the control, while $Y_i(1)$ indicates the response to treatment. These outcomes are *potential* because we may never observe both.

When the action A_i (treatment assignment) for patient i is independent of the patient, the expected response and treatment effect become

$$\mathbb{E}[Y_i(a) \mid A_i = a] = \mathbb{E}[Y_i(a)] =: \mu_a \quad \text{and} \quad \tau := \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{18.1.4}$$

(We assume individuals are exchangeable.) If we choose A_i in a way that depends on the individual i, we may have confounding: $\mathbb{E}[Y_i(a) \mid A_i = a] \neq \mathbb{E}[Y_i(a)]$.

A fully randomized trial to estimate the treatment effect chooses $n/2$ individuals to receive treatment and $n/2$ to be in the control, uniformly at random. Then

$$\widehat{\tau} := \frac{1}{n/2} \sum_{i:A_i=1} Y_i(A_i) - \frac{1}{n/2} \sum_{i:A_i=0} Y_i(A_i)$$

is an unbiased estimator for $\tau = \mathbb{E}[Y(1) - Y(0)]$, with $\mathbb{P}(|\widehat{\tau} - \tau| \geq t) \leq \exp(-c\frac{nt^2}{\sigma^2})$ in the sub-gaussian setting (18.1.1). This assignment strategy, while effective for estimating the treatment effect, incurs high regret (18.1.3) if the treatment is effective or has strong negative outcomes, because it allocates too many individuals to one arm of the study. $\diamond$

Example 18.1.1 highlights many of the central aspects of bandit-type problems. First, they have deep connections with causal reasoning: by selecting an action (or arm) A, we are intervening in a system, with a goal to identify the best intervention. Second, the online nature can be critical: in medical settings, for example, it would be unethical to continue a drug trial if it were ineffective. Finally, the example implicitly shows one of the major benefits of the online scenario: because we *select* the action A_t without any information about the outcomes $Y_t = (Y_t(1), \ldots, Y_t(K))$, we can avoid issues of confounding that would arise trying to estimate or learn from offline data.

18.2 Confidence-based algorithms

We first consider a strategy using confidence intervals with slight optimism. If we believe the true mean μ_i for an arm i lies within $[\widehat{\mu}_i - c_i, \widehat{\mu}_i + c_i]$, where the interval length $2c_i$ decreases with time t, then we optimistically act as if the value of arm i is $\widehat{\mu}_i + c_i$. Then at iteration t we choose the arm whose optimistic mean is the highest, hoping to maximize our received reward.

This strategy lies at the heart of the Upper Confidence Bound (UCB) family of algorithms [35], a simple variant of which we describe here. We revisit concentration of sub-Gaussian random variables, though we require careful calculation because of the sequential nature of our process. Let

$$N_t(i) = \mathrm{card}\{\tau \le t \mid A_\tau = i\}$$

denote the number of times that arm i has been pulled by time t of the bandit process, and define

$$\widehat{\mu}_t(i) := \frac{1}{N_t(i)} \sum_{\tau \le t, A_\tau = i} Y_\tau(i),$$

to be the running average of the rewards of arm i at time t (computed only on those instances in which arm i was selected). The conditionally independent randomness in sample $Y_t(i) \sim P_i$ when $A_t = i$ in the bandit game implies a distributional equality underpinning our analyses:

Lemma 18.2.1. *Let $Y'_\tau(i) \stackrel{\mathrm{iid}}{\sim} P_i$, $\tau = 1, 2, \ldots$, be independent copies of the random variables $Y_\tau(i)$. Then $Y'_\tau(i)$ is independent of $N_t(i)$ for all t and τ. If $\widehat{\mu}'_t(i) = \frac{1}{N_t(i)} \sum_{\tau : A_\tau = i} Y'_\tau(i)$ is the empirical mean of the copies $Y'_\tau(i)$ for those steps when arm i is selected, then*

$$\left(\widehat{\mu}_t(i), N_t(i)\right) \stackrel{\mathrm{dist}}{=} \left(\widehat{\mu}'_t(i), N_t(i)\right), \tag{18.2.1}$$

We prove the lemma in the technical appendix (Section 18.6.1) to this chapter.

The distributional equality (18.2.1) implies that the means $\widehat{\mu}_t(i)$ are sub-Gaussian, even conditional on the number of pulls $N_t(i)$. Indeed, we observe that for $\lambda \in \mathbb{R}$ and any $m \in \mathbb{N}$,

$$\mathbb{E}\left[\exp\left(\lambda m \widehat{\mu}_t(i)\right) \mid N_t(i) = m\right] \stackrel{(\star)}{=} \mathbb{E}\left[\exp\left(\lambda m \widehat{\mu}'_t(i)\right) \mid N_t(i) = m\right] \le \exp\left(\frac{m\lambda^2\sigma^2}{2}\right), \tag{18.2.2}$$

where equality $(\star)$ follows from Lemma 18.2.1. Taking conditional expectations over the value $N_t(i)$, we see via Corollary 4.1.10 that for all t,

$$\mathbb{P}\left(\widehat{\mu}_t(i) \ge \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{N_t(i)}}\right) \vee \mathbb{P}\left(\widehat{\mu}_t(i) \le \mu_i - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{N_t(i)}}\right) \le \delta. \tag{18.2.3}$$

That is, so long as we pull the arms sufficiently many times, we are unlikely to pull the wrong arm.

This leads to the UCB procedure:

Input: Sub-gaussian parameter σ^2 and sequence of deviation probabilities $\delta_1, \delta_2, \ldots$.
Initialization: Play each arm $i = 1, \ldots, K$ once
Repeat: for each iteration t, play the arm maximizing

$$\widehat{\mu}_t(i) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}}.$$

Figure 18.1: The Upper Confidence Bound (UCB) Algorithm

If we define

$$\Delta_i := \mu_{i^\star} - \mu_i$$

to be the gap in means between the optimal arm and any sub-optimal arm, we then obtain the following guarantee on the expected number of pulls of any sub-optimal arm i after n steps.

Proposition 18.2.2. *Assume that each of the K arms is σ^2-sub-Gaussian and let the sequence $\delta_1 \geq \delta_2 \geq \cdots$ be non-increasing and positive. Then for any T and any arm $i \neq i^\star$,*

$$\mathbb{E}[N_T(i)] \leq \left\lceil \frac{4\sigma^2 \log \frac{1}{\delta_T}}{\Delta_i^2} \right\rceil + 2\sum_{t=2}^{T} \delta_t.$$

Proof Without loss of generality, we assume arm 1 satisfies $\mu_1 = \max_i \mu_i$, and let arm i be any sub-optimal arm. The key insight is to carefully consider what occurs if we play arm i in the UCB procedure of Figure 18.1. In particular, if we play arm i at time t, then we certainly have

$$\widehat{\mu}_t(i) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}} \geq \widehat{\mu}_t(1) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(1)}}.$$

For this to occur, at least one of the following three events must occur (we suppress the dependence on i for each of them):

$$\mathcal{E}_{1,t} := \left\{ \widehat{\mu}_t(i) \geq \mu_i + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}} \right\}, \quad \mathcal{E}_{2,t} := \left\{ \widehat{\mu}_t(1) \leq \mu_1 - \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(1)}} \right\},$$

$$\mathcal{E}_{3,t} := \left\{ \Delta_i \leq 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}} \right\}.$$

Indeed, suppose that none of the events $\mathcal{E}_{1,t}, \mathcal{E}_{2,t}, \mathcal{E}_{3,t}$ occur at time t. Then we have

$$\widehat{\mu}_t(i) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}} < \mu_i + 2\sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(i)}} < \mu_i + \Delta_i = \mu_1 < \widehat{\mu}_t(1) + \sqrt{\frac{\sigma^2 \log \frac{1}{\delta_t}}{N_t(1)}},$$

the inequalities following by $\mathcal{E}_{1,t}$, $\mathcal{E}_{3,t}$, and $\mathcal{E}_{2,t}$, respectively.

Now, for any $l \in \{1, \ldots, T\}$, we see that

$$\begin{aligned}\mathbb{E}[N_T(i)] &= \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}\left\{A_t = i\right\}] = \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}\left\{A_t = i, N_t(i) > l\right\} + \mathbf{1}\left\{A_t = i, N_t(i) \leq l\right\}] \\ &\leq l + \sum_{t=l+1}^{T} \mathbb{P}(A_t = i, N_t(i) > l).\end{aligned}$$

Using that δ_t is non-increasing, set

$$l^\star = \left\lceil 4\frac{\sigma^2 \log \frac{1}{\delta_T}}{\Delta_i^2} \right\rceil.$$

If $N_t(i) > l^\star$, then $\mathcal{E}_{3,t}$ cannot occur: we have $2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/N_t(i)} < 2\sqrt{\sigma^2 \log \frac{1}{\delta_t}/l^\star} \leq \Delta_i$, so if $A_t = i$ and $N_t > l^\star$, then $\mathcal{E}_{1,t}$ or $\mathcal{E}_{2,t}$ must occur. Thus

$$\begin{aligned}\mathbb{E}[N_T(i)] &\leq l^\star + \sum_{t=l^\star+1}^{T} \mathbb{P}(A_t = i, N_t(i) > l^\star) \\ &\leq l^\star + \sum_{t=l^\star+1}^{T} \mathbb{P}(\mathcal{E}_{1,t} \text{ or } \mathcal{E}_{2,t}) \leq l^\star + \sum_{t=l^\star+1}^{T} 2\delta_t.\end{aligned}$$

This implies the desired result. □

The number of times arm i is selected in the sequential game relates to the regret; indeed,

$$\mathsf{Reg}_n = \sum_{t=1}^n (\mu_{i^\star} - \mu_{A_t}) = \sum_{i=1}^K (\mu_{i^\star} - \mu_i) N_n(i) = \sum_{i=1}^K \Delta_i N_n(i).$$

This identity immediately implies two theorems on the (expected) regret of the UCB algorithm.

Theorem 18.2.3. *Let $\delta_t = \delta/t^2$ for all t. Then for any $n \in \mathbb{N}$ the UCB algorithm attains*

$$\overline{\mathsf{Reg}}_n \le \sum_{i \neq i^\star} \frac{4\sigma^2[2\log n - \log \delta]}{\Delta_i} + \frac{\pi^2 - 2}{3}\left(\sum_{i=1}^K \Delta_i\right)\delta + \sum_{i=1}^K \Delta_i.$$

Proof By Proposition 18.2.2,

$$\mathbb{E}[\Delta_i N_n(i)] \le \Delta_i \left\lceil 4\sigma^2 \log \frac{1}{\delta_n}/\Delta_i^2 \right\rceil + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2} \le \frac{4\sigma^2 \log \frac{1}{\delta_n}}{\Delta_i} + \Delta_i + 2\Delta_i \sum_{t=2}^n \frac{\delta}{t^2}.$$

Summing over $i \neq i^\star$ and noting that $\sum_{t \ge 2} t^{-2} = \pi^2/6 - 1$ gives the result. □

Let us unpack the bound of Theorem 18.2.3 with $\delta = 1$. Let $\Delta = \min_{i \neq i^\star} \Delta_i$. In this case, we have expected regret

$$\overline{\mathsf{Reg}}_n \le 8\frac{K\sigma^2 \log n}{\Delta} + \frac{\pi^2 + 1}{3}\sum_{i=1}^K \Delta_i \le O(1)\left[\frac{K\sigma^2}{\Delta}\log n + \sum_{i=1}^K \Delta_i\right]. \tag{18.2.4}$$

So the asymptotic regret with this choice of δ scales at worst as $(K\sigma^2/\Delta)\log n$, roughly linear in the number of arms, logarithmic in n, and inversely proportional to the gap in means.

If any of the gaps $\Delta_i \to 0$ in the bound of Theorem 18.2.3, the bound becomes vacuous. Intuitively, however, pulling a *slightly* sub-optimal arm should be insignificant for the regret. With that in mind, we present a variant of the above bounds, which has a worse scaling with n but is independent of the gaps Δ_i.

Theorem 18.2.4. *Let $\delta_t = 1/t^2$. Then for any $n \in \mathbb{N}$, UCB attains*

$$\overline{\mathsf{Reg}}_n \le \sqrt{8K\sigma^2 n \log n} + 4\sum_{i=1}^K \Delta_i.$$

Proof Fix $\gamma > 0$. Then we may write the regret with the standard identity

$$\mathsf{Reg}_n = \sum_{i \neq i^\star} \Delta_i N_n(i) = \sum_{i:\Delta_i \ge \gamma} \Delta_i N_n(i) + \sum_{i:\Delta_i < \gamma} \Delta_i N_n(i) \le \sum_{i:\Delta_i \ge \gamma} \Delta_i N_n(i) + n\gamma,$$

where the final inequality uses that certainly $\sum_{i=1}^K N_n(i) \le n$. Taking expectations with the UCB procedure and $\delta = 1$, we have by Theorem 18.2.3 that

$$\overline{\mathsf{Reg}}_n \le \sum_{i:\Delta_i \ge \gamma} \Delta_i \frac{8\sigma^2 \log n}{\Delta_i^2} + \frac{\pi^2 + 1}{3}\sum_{i=1}^K \Delta_i + n\gamma \le K\frac{8\sigma^2 \log n}{\gamma} + n\gamma + \frac{\pi^2 + 1}{3}\sum_{i=1}^K \Delta_i,$$

Optimizing over γ by taking $\gamma = \frac{\sqrt{8K\sigma^2 \log n}}{\sqrt{n}}$ gives the result. $\square$

Combining the two theorems, we see that the UCB algorithm with parameters $\delta_t = 1/t^2$ automatically achieves the expected regret guarantee

$$\overline{\mathsf{Reg}}_n \le C \cdot \min \left\{ \sum_{i:\Delta_i > 0} \frac{\sigma^2 \log n}{\Delta_i}, \sqrt{K \sigma^2 n \log n} \right\}. \tag{18.2.5}$$

This regret is sub-optimal; there are algorithms, including the MOSS (Minimax Optimal in the Stochastic Case) bandit procedure [34], which achieve regret

$$\overline{\mathsf{Reg}}_n \le C \cdot \min \left\{ \sqrt{Kn}, \frac{K}{\Delta} \log \frac{n\Delta^2}{K} \right\}.$$

This essentially recovers the bound specified (18.2.4) (which required knowledge of the Δ_i) and improves the analysis of Theorem 18.2.4 by $\log n$. High-probability guarantees for UCB algorithms follow more or less immediately from the proof techniques of Proposition 18.2.2; see Exercise 18.8.

18.3 General losses and information-based bounds

To move beyond the simplest K-armed bandit settings will require more than per-arm confidence bounds. We now assume there is an abstract set of actions (arms) $\mathcal{A}$, which may or may not be finite, and we have a collection of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by a set Θ (we could just let Θ be distributions, but parametric scenarios remain interesting). Often, $\mathcal{A}$ is finite, as in the case of K-armed bandit problems, and Θ is a subset of $\mathbb{R}^K$ (e.g., the arm means). We generalize things slightly to address more than just the responses $Y_t(a) \in \mathcal{Y}$, so that we have a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ that measures the quality of an action $a \in \mathcal{A}$ for the observation $y \in \mathcal{Y}$.

Example 18.3.1 (Bernoulli bandits)**:** The Bernoulli bandit problem has actions $\mathcal{A} = \{1, \ldots, K\}$, and the parameter space $\Theta = [0,1]^K$, and P_θ is a distribution on $Y \in \{0,1\}^K$, where Y has independent coordinates $Y(j) \stackrel{\text{ind}}{\sim} \mathsf{Bernoulli}(\theta_j)$. The goal is to find the arm $a^\star \in \operatorname{argmax}_j \theta_j$ with highest mean reward. The loss function $\ell(a, y) = -y(a)$ satisfies $\ell(a, y) \in [-1, 0]$ and $\mathbb{E}_\theta[\ell(a, Y(a))] = -\theta_a$, so $a^\star$ minimizes $\mathbb{E}_\theta[\ell(a, Y(a))]$ over actions $a \in \mathcal{A}$. $\diamond$

In this general setting, for a given θ, we consider the expected regret

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \theta) = \mathbb{E}_\theta \left[\sum_{t=1}^n \ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star)) \right], \tag{18.3.1}$$

where

$$A^\star = A^\star(\theta) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_\theta[\ell(a, Y(a))]$$

minimizes the expected loss of taking action $a \in \mathcal{A}$ when θ is the parameter (and so is a function of θ), while the player takes action $A_t \in \mathcal{A}$ at time t. Figure 18.2 captures the framework.

In Figure 18.2, we left the choice of θ unspecified; in worst case settings, it may be adversarial. Moving to a Bayesian setting in which $\theta \sim \pi$ for some prior distribution π on the space Θ allows us to leverage information-theoretic techniques to obtain regret bounds and develop intuitive

Initialize: choose parameter $\theta \in \Theta$, yielding sampling distribution P_θ
Repeat: for each iteration $t = 1, 2, \ldots$
Choose distribution ρ_t on actions $\mathcal{A}$ based on the history

$$\mathcal{H}_{t-1} := \{(A_1, Y_1(A_1)), (A_2, Y_2(A_2)), \ldots, (A_{t-1}, Y_{t-1}(A_{t-1}))\}$$

Draw random action

$$A_t \sim \rho_t.$$

Observe feedback $Y_t(A_t) \sim P_\theta$ and loss $\ell(A_t, Y_t(A_t))$; suffer (unobserved) regret

$$\ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star))$$

Figure 18.2: The generic exploration/exploitation algorithm

algorithms. Bayesian strategies allow us to quantify a value for information by trading between reducing uncertainty in the prior π and regret, and they often yield simple procedures. In this setting where we have a prior distribution π on $\theta \in \Theta$, we then define the Bayesian regret

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) = \mathbb{E}_\pi\left[\sum_{t=1}^n \ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star))\right], \tag{18.3.2}$$

where the optimal action is random (as it is a function of $\theta \sim \pi$):

$$A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}[\ell(a, Y(a)) \mid \theta].$$

We take the expectation (18.3.2) over both the randomness in θ according to the prior π and any randomness in the player's strategy for choosing the actions A_t. The setting (18.3.2) also encapsulates settings in which we obtain observations $Y_t(a)$ by taking action $a \in \mathcal{A}$ and suffer a random loss $L_{t,a}$ that depends on the action a: we incorporate these random losses into the observation $Y_t(a)$, making $\ell(a, Y_t(a))$ identify the component of the observation corresponding to the loss.

18.3.1 An information-based regret bound

We can bound the Bayesian regret (18.3.2) via the ratio of the expected instantaneous regret at time t and the information that taking a particular action A provides about the optimal action $A^\star$. The rough idea is simple: if we can bound the ratio of squared expected instantaneous regret to the information action A_t provides about the optimal action, then each step of the Algorithm 18.2 gains information or suffers small regret, but never both. Thus, once we have collected information commensurate with the entropy of the prior π, the remaining regret must be small.

To make things rigorous, we condition mutual information and entropy on particular realizations of variables. Define the history σ-field

$$\mathcal{H}_t := \sigma\left(\{A_1, Y_1(A_1), A_2, Y_2(A_2), \ldots, A_t, Y_t(A_t)\}\right)$$

of actions and observations through iteration t of the bandit process. (For readers unfamiliar with measure theory, treat $\mathcal{H}_t$ as the realized value of $(A_i, Y_i(A_i))_{i=1}^t$ when conditioning.) Then for $a \in \mathcal{A}$, define

$$I_t(a) := I(A^\star; Y_t(a) \mid \mathcal{H}_{t-1}) \tag{18.3.3}$$

to be the mutual information between $A^\star$ and the observation $Y_t(a)$ conditional on the realized history $\mathcal{H}_{t-1}$; this corresponds to conditional mutual information $I(X; Y \mid Z = z)$ in Section 2.1.1. By this we mean that we draw θ from its posterior distribution conditional on $\mathcal{H}_{t-1}$, then draw $Y_t(a)$ conditional on θ and set $A^\star = \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}[\ell(a, Y(a)) \mid \theta]$. The mutual information (18.3.3) is random, as it is a function of the realization $\mathcal{H}_{t-1}$. Without loss of generality (by approximation), we can discretize and assume $\mathcal{A}$ is countable, allowing us to write

$$I_t(a) = H(A^\star \mid \mathcal{H}_{t-1}) - H(A^\star \mid Y_t(a), \mathcal{H}_{t-1}),$$

the reduction in entropy on the optimal action $A^\star$ conditional on observing $Y_t(a)$. Then for a distribution ρ on actions $\mathcal{A}$, define the information gained about $A^\star$ by taking action $A \sim \rho$, conditional on the realization $\mathcal{H}_{t-1}$ of the history by

$$I_t(\rho) := \mathbb{E}_\rho[I_t(A)] = \sum_{a \in \mathcal{A}} \rho(a) I_t(a) = I(A^\star; Y_t(A), A \mid \mathcal{H}_{t-1}). \tag{18.3.4}$$

In the average (18.3.4), we replace the fixed action a in the definition (18.3.3) with a random $A \sim \rho$, and conditional on $A = a$ draw $Y_t(a)$ as before.

With the random averaged information (18.3.4), we define the ratio between the expected regret and information gained by sampling an action $A \sim \rho$ via

$$R_t(\rho) := \frac{\mathbb{E}_\rho[\ell(A, Y_t(A)) - \ell(A^\star, Y_t(A^\star)) \mid \mathcal{H}_{t-1}]^2}{I_t(\rho)}. \tag{18.3.5}$$

As is I_t, R_t is also a random variable. If we can choose the distribution ρ to make the ratio (18.3.5) small, then for $A \sim \rho$, either A should have losses close to those of $A^\star$, or A and the response $Y_t(A)$ should provide substantial information about $A^\star$.

That squared expected loss gaps should be related to information might seem *a priori* unmotivated. But throughout the book, the connection between squared distances and KL-divergences has underpinned many of the concentration inequalities we develop via the Donsker-Varadhan variational representation of KL-divergence. Particular examples arise in Chapter 6 on PAC-Bayes and adaptive data analysis, as well as in transportation inequalities for concentration in Chapter 7.

By an application of the Cauchy-Schwarz inequality and using the definition of the mutual information, we can upper bound the expected regret by the loss/information ratio and the information gained throughout the process about the optimal arm:

Theorem 18.3.2. *For any procedure and any prior π on $\theta \in \Theta$,*

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le \sqrt{\mathbb{E}_\pi\left[\sum_{t=1}^n R_t(\rho_t)\right]} \sqrt{I\left(A^\star; \{A_t, Y_t(A_t)\}_{t=1}^n\right)}.$$

Proof We apply the Cauchy-Schwarz inequality:

$$\begin{aligned}
\mathbb{E}_\pi\left[\sum_{t=1}^n \ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star))\right] &= \mathbb{E}_\pi\left[\sum_{t=1}^n \frac{\ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star))}{\sqrt{I_t(\rho_t)}}\sqrt{I_t(\rho_t)}\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=1}^n \frac{\mathbb{E}_{\rho_t}[\ell(A_t, Y_t(A_t)) - \ell(A^\star, Y_t(A^\star)) \mid \mathcal{H}_{t-1}]}{\sqrt{I_t(\rho_t)}}\sqrt{I_t(\rho_t)}\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=1}^n \sqrt{R_t(\rho_t)}\sqrt{I_t(\rho_t)}\right] \\
&\le \sqrt{\mathbb{E}_\pi\left[\sum_{t=1}^n R_t(\rho_t)\right]}\sqrt{\mathbb{E}_\pi\left[\sum_{t=1}^n I_t(\rho_t)\right]},
\end{aligned}$$

where the second equality uses that the tth action sampling distribution ρ_t is a function of the history $\mathcal{H}_{t-1}$. Then observe that

$$\mathbb{E}_\pi[I_t(\rho_t)] = \mathbb{E}_\pi[I(A^\star; A_t, Y_t(A_t) \mid \mathcal{H}_{t-1})] = I\left(A^\star; A_t, Y_t(A_t) \mid \{A_\tau, Y_\tau(A_\tau)\}_{\tau<t}\right)$$

by definition of the conditional mutual information. Using the chain rule for mutual information

$$\sum_{t=1}^n I\left(A^\star; A_t, Y_t(A_t) \mid \{A_\tau, Y_\tau(A_\tau)\}_{\tau<t}\right) = I\left(A^\star; \{A_t, Y_t(A_t)\}_{t=1}^n\right),$$

as in Section 2.1.2, then yields the theorem. □

As an immediate corollary, note that whenever $\mathcal{A}$ is finite or countable, the (Shannon) entropy of $A^\star$ trivially upper bounds the mutual information:

$$I\left(A^\star; \{A_t, Y_t(A_t)\}_{t=1}^n\right) = H(A^\star) - H(A^\star \mid \{A_t, Y_t(A_t)\}_{t=1}^n) \le H(A^\star).$$

When $\mathcal{A}$ is finite, $H(A^\star) \le \log \operatorname{card}(\mathcal{A})$, and we record these bounds as a corollary.

Corollary 18.3.3. *Assume that action distributions ρ_t are chosen so that for some $R_\pi < \infty$, the average loss/information ratio satisfies*

$$\frac{1}{n}\sum_{t=1}^n \mathbb{E}_\pi[R_t(\rho_t)] \le R_\pi$$

and that $\mathcal{A}$ is countable. Then

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{R_\pi H(A^\star)} \cdot \sqrt{n}.$$

In particular, if $\mathcal{A}$ is finite, then

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{R_\pi \log \operatorname{card}(\mathcal{A})} \cdot \sqrt{n}.$$

Corollary 18.3.3 suggests a design principle for (Bayesian) bandit algorithms: at step t, choose the distribution ρ_t on actions to minimize the loss/information ratio $R_t(\rho) = \mathbb{E}_\rho[\ell(A, Y_t(A)) - \ell(A^\star, Y_t(A^\star)) \mid \mathcal{H}_{t-1}]^2 / I_t(\rho)$. Though this particular choice may be impossible, specific choices ρ frequently admit guarantees that the ratio is bounded, which is sufficient for a regret bound that only grows as $\sqrt{n}$. We exploit this design principle for the remainder of this section.

18.3.2 Posterior (Thompson) sampling

One natural sampling strategy in the Bayesian setting instantiates Algorithm 18.2 by maintaining a posterior distribution over θ and drawing actions that reflect this posterior. Algorithm 18.3, which describes Thompson sampling, captures this.

Input: Prior distribution π on space Θ, family of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$
Repeat: for each iteration t,
Choose distribution π_t to be posterior on θ given history $\mathcal{H}_{t-1} = \{A_\tau, Y_\tau(A_\tau)\}_{\tau < t}$
Draw parameter $\theta_t \sim \pi_t$ and for $Y \sim P_{\theta_t}$, choose action

$$A_t \in \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_{\theta_t}[\ell(a, Y(a))]$$

Observe feedback $Y_t(A_t)$ and loss $\ell(A_t, Y_t(A_t))$

Figure 18.3: Generic Thompson sampling algorithm

At iteration t, Thompson sampling uses the posterior

$$\pi_t(\theta) = \pi(\theta \mid \mathcal{H}_{t-1}),$$

the distribution on θ conditional on $\mathcal{H}_{t-1}$. We let

$$\rho_t^{\text{ts}} = \text{distribution of } \left\{A_t := \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}[\ell(a, Y(a)) \mid \theta_t]\right\} \tag{18.3.6}$$

be the induced distribution on the actions A_t when sampling from the posterior π_t.

We provide a few more concrete specifications of Algorithm 18.3 for Thompson (posterior) sampling, beginning with the case of Bernoulli rewards.

Example 18.3.4 (Thompson sampling for a K-armed Bernoulli bandit)**:** Let $\theta \in [0,1]^K$ parameterize K independent $\mathsf{Bernoulli}(\theta_a)$ distributions, where on action $a \in \mathcal{A} = \{1, \ldots, K\}$, we observe $Y(a) \sim \mathsf{Bernoulli}(\theta_a)$. Place a beta prior on the coordinates θ, $\theta_a \sim \mathsf{Beta}(1,1)$, which corresponds to the uniform distribution on $[0,1]^d$. Let

$$N_t^1(a) = \text{card}\{\tau \le t : A_\tau = a, Y_\tau(a) = 1\}$$

be the number of times arm a has been pulled by time t while receiving response 1, and let $N_t^0(a) = \text{card}\{\tau \le t : A_\tau = a, Y_\tau(a) = 0\}$. Then, peeking ahead to Example 19.5.2 on Beta-Bernoulli distributions, Thompson sampling with the loss $\ell(a, y) = -y(a)$ proceeds in round t as follows:

(i) For each arm $a \in \mathcal{A} = \{1, \ldots, K\}$, draw $\theta_t(a) \sim \mathsf{Beta}(1 + N_{t-1}^1(a), 1 + N_{t-1}^0(a))$.

(ii) Play the action $A_t = \operatorname{argmax}_a \theta_t(a)$.

(iii) Observe $Y_t(A_t) \in \{0,1\}$, and increment the appropriate count.

In this case, we may implement Thompson sampling with just a few counters. $\diamond$

Example 18.3.4 extends to rewards supported on $[0,1]$: randomize the observation $Y(a) \in [0,1]$ to $\widetilde{Y}(a) \in \{0,1\}$ by taking $\mathbb{P}(\widetilde{Y}(a) = 1 \mid Y(a)) = Y(a)$. We may then apply the Beta-Bernoulli update.

The key to our analysis of Thompson sampling is that whenever the losses are sub-Gaussian, (a multiple of) the mutual information between $A^\star$ and $(Y_t(A_t), A_t)$ upper bounds the squared regret. There is a subtlety that we note only in passing: when we say that for each $a \in \mathcal{A}$,

$$\ell(a, Y(a)) \text{ is } \sigma^2\text{-sub-Gaussian}, \tag{18.3.7}$$

we mean the following: for any distribution π on $\theta \in \Theta$, the observed loss $\ell(a, Y(a))$ in the following process is σ^2-sub-Gaussian: (i) choose $a \in \mathcal{A}$, (ii) draw $\theta \sim \pi$, and conditional on θ draw $Y \sim P_\theta$, then (iii) observe the pair $(Y(a), \ell(a, Y(a)))$. Alternatively, we may say there is some σ^2, which is a function of the history $\mathcal{H}_{t-1}$, such that for each $a \in \mathcal{A}$, $\ell(a, Y_t(a))$ is σ^2-sub-Gaussian conditional on $\mathcal{H}_{t-1}$. A sufficient condition for all of this is the boundedness

$$\sup_y \ell(a, y) - \inf_y \ell(a, y) \le B \text{ for all } a \in \mathcal{A},$$

in which case the losses are $\frac{1}{4}B^2$-sub-Gaussian (by Hoeffding's lemma, inequality (4.1.2)).

The following lemma uses sub-Gaussianity to show the information bound on the instantaneous (single step) regret. The result follows as a consequence of the Donsker-Varadhan variational representation of the KL-divergence (Theorem 6.1.1).

Lemma 18.3.5. *Assume that for each $a \in \mathcal{A}$, the loss $\ell(a, Y(a))$ is σ^2-sub-Gaussian (18.3.7) and that $\mathcal{A}$ is finite. For $\theta \sim \pi$, let the optimal action $A^\star = \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_\theta[\ell(a, Y(a))]$ have distribution ρ, and let $A \sim \rho$ independent of $A^\star$. Then*

$$\sum_{a \in \mathcal{A}} \rho(a) \left(\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a]\right) \le \sqrt{2\sigma^2 \operatorname{card}(\mathcal{A})} \sqrt{I(A, Y(A); A^\star)}.$$

Proof Let $L(a) := \ell(a, Y(a))$ be shorthand for the random realization of the loss. Then

$$\sum_{a \in \mathcal{A}} \rho(a) \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a]\right) \le \left(\sum_a \rho(a)^2 \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a]\right)^2\right)^{1/2} \sqrt{\operatorname{card}(\mathcal{A})}.$$

Then we observe that for each $a \in \mathcal{A}$,

$$\rho(a)^2 \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a]\right)^2 \le \rho(a) \sum_{a^\star \in \mathcal{A}} \rho(a^\star) \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a^\star]\right)^2, \tag{18.3.8}$$

because we have only added nonnegative terms. Now we use that the losses are sub-Gaussian (18.3.7). If P_a and $P_{a|A^\star = a^\star}$ denote the marginal distribution of $L(a) = \ell(a, Y(a))$ and $L(a) = \ell(a, Y(a))$ conditional on $A^\star = a$, respectively, then Theorem 7.1.2 implies that

$$\left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a^\star]\right)^2 \le 2\sigma^2 D_{\rm kl}\left(P_{a|A^\star = a^\star} \| P_a\right).$$

Summing over $a^\star$, we recognize that the marginal $P_a = \sum_{a^\star} \rho(a^\star) P_{a|A^\star = a^\star}$, and so the familiar representation (9.4.4) of the mutual information as a mixture of KL-divergences implies

$$\sum_{a^\star \in \mathcal{A}} \rho(a^\star) D_{\rm kl}\left(P_{a|A^\star = a^\star} \| P_a\right) = I(\ell(A, Y(A)); A^\star \mid A = a) \le I(A, Y(A); A^\star \mid A = a)$$

by the data processing inequality. Finally, we use the independence that $A, A^\star$ are identically distributed to observe that

$$\sum_{a \in \mathcal{A}} \rho(a) I(A, Y(A); A^\star \mid A = a) = I(A, Y(A); A^\star).$$

Substituting this into the bound (18.3.8), we have

$$\sum_{a \in \mathcal{A}} \rho(a)^2 \left(\mathbb{E}[L(a)] - \mathbb{E}[L(a) \mid A^\star = a]\right)^2 \le 2\sigma^2 I(A, Y(A); A^\star),$$

implying the lemma. □

Lemma 18.3.5 immediately extends to the instantaneous expected loss gaps conditional on the history, because in the sampling procedure in Alg. 18.3, we have the distributional equality

$$A^\star \mid \mathcal{H}_{t-1} \stackrel{\text{dist}}{=} A_t \mid \mathcal{H}_{t-1}$$

by construction of the posterior π_t on θ given the history $\mathcal{H}_{t-1}$. Additionally, A_t and $A^\star$ are independent given $\mathcal{H}_{t-1}$, so Lemma 18.3.5 implies that for Thompson sampling,

$$\left(\mathbb{E}[\ell(A_t, Y_t(A_t)) \mid \mathcal{H}_{t-1}] - \mathbb{E}[\ell(A^\star, Y_t(A^\star)) \mid \mathcal{H}_{t-1}]\right)^2 \le 2\sigma^2 \operatorname{card}(\mathcal{A}) \cdot I(A_t, Y_t(A_t); A^\star \mid \mathcal{H}_{t-1}).$$

Recalling the loss/information ratio (18.3.5), Thompson enjoys the uniform bound

$$R_t(\rho_t^{\text{ts}}) \le 2\sigma^2 \operatorname{card}(\mathcal{A}) \tag{18.3.9}$$

for all times t. Thus, as a corollary to the main Theorem 18.3.2 bounding expected regret by the mutual information, we have the following Bayesian regret bound.

Theorem 18.3.6. *Let $\mathcal{A}$ be finite, and assume that the loss $\ell(a, Y_t(a))$ is σ^2-sub-Gaussian (18.3.7) for each action $a \in \mathcal{A}$. Then Thompson sampling has regret*

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{2\sigma^2 \operatorname{card}(\mathcal{A}) H(A^\star)}\sqrt{n}.$$

We make a few remarks. First, the entropy $H(A^\star)$ of the optimal action is never greater than $\log \operatorname{card}(\mathcal{A})$, and so the regret essentially scales at worst as $\sqrt{\operatorname{card}(\mathcal{A}) n}$. As an example application of Theorem 18.3.6, let us revisit the Bernoulli bandit problem in Example 18.3.4.

Example 18.3.7 (Bernoulli bandits, continued)**:** Consider the Bernoulli bandit setting of Example 18.3.4. Then the mean regret for loss $\ell(a, y(a)) = -y(a)$ is

$$\overline{\mathsf{Reg}}_n = \sum_{t=1}^n \mathbb{E}_\pi[\theta_{A^\star} - \theta_{A_t}],$$

the gap between the best mean reward and the mean reward of the played arm A_t. Because there are K arms, with $\{0, 1\}$-valued rewards, each is $\frac{1}{4}$-sub-Gaussian, and so

$$\overline{\mathsf{Reg}}_n \le \sqrt{\frac{K \log K}{2}}\sqrt{n}.$$

To within the $\log K$ factor, this is minimax rate optimal, as we shall see, and removes the $\log n$ multiplicative factors (18.2.5) present in our other analyses. ◇

Exercise 18.9 suggests experiments to compare the UCB algorithms of the preceding section with Thompson sampling and mirror-descent based methods that we develop in Section 18.4. Figure 18.5 exhibits the behavior of Thompson sampling on a related linear bandit problem, and Fig. 18.7 compares the three algorithms; empirically, Thompson sampling and similar posterior sampling algorithms appear to be generally quite effective.

18.3.3 Information-based exploration

The loss/information ratio (18.3.5) around which the regret bound in Theorem 18.3.2 centers suggests strategies that iteratively choose the distribution ρ to minimize the ratio. This leads to *information-directed sampling*, which at iteration t chooses

$$\rho_t^{\text{ids}} := \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathcal{A})} \left\{ R_t(\rho) = \frac{\mathbb{E}_\rho[\ell(A, Y(A)) - \ell(A^\star, Y(A^\star)) \mid \mathcal{H}_{t-1}]^2}{I_t(\rho)} \right\}. \tag{18.3.10}$$

Here we let $\mathcal{P}(\mathcal{A})$ denote all probability distributions on $\mathcal{A}$, and $\mathbb{E}_\rho$ denotes expectation over $A \sim \rho$.

When $\mathcal{A}$ is finite and the losses are sub-Gaussian, Thompson sampling guarantees the bound $R_t(\rho_t) \le 2\sigma^2 \operatorname{card}(\mathcal{A})$ (recall inequality (18.3.9)), so similarly $R_t(\rho_t^{\text{ids}}) \le 2\sigma^2 \operatorname{card}(\mathcal{A})$. Thus information directed sampling achieves no worse regret:

Corollary 18.3.8. *Let ρ_t^{ids} be the information-directed sampling distribution (18.3.10), and assume the other conditions of Theorem 18.3.6. Then*

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{2\sigma^2 \operatorname{card}(\mathcal{A}) H(A^\star)} \sqrt{n}.$$

It may be challenging to exactly compute the information-directed sampling distribution (18.3.10), but we can relax it by replacing information with a variance proxy:

Lemma 18.3.9. *Assume that $\ell(a, Y(a))$ is σ^2-sub-Gaussian (18.3.7) for each $a \in \mathcal{A}$. Then*

$$\operatorname{Var}(\mathbb{E}[\ell(a, Y(a)) \mid A^\star]) \le 2\sigma^2 I(A^\star; Y(A), A \mid A = a).$$

Proof Let $P_{a|A^\star=a^\star}$ denote the distribution of $\ell(a, Y(a))$ conditional on $A^\star = a^\star$, and let ρ be the distribution of $A^\star$. Then

$$\begin{aligned}
2\sigma^2 \cdot I(A^\star; Y(A), A \mid A = a) &\overset{(i)}{\ge} 2\sigma^2 \cdot I(A^\star; \ell(a, Y(a)) \mid A = a) \\
&\overset{(ii)}{=} 2\sigma^2 \int_{a^\star \in \mathcal{A}} D_{\rm kl}\left(P_{a|A^\star=a^\star} \| P_a\right) d\rho(a^\star) \\
&\overset{(iii)}{\ge} \int_{a^\star \in \mathcal{A}} \left(\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a^\star]\right)^2 d\rho(a^\star) \\
&= \operatorname{Var}(\mathbb{E}[\ell(a, Y(a)) \mid A^\star]),
\end{aligned}$$

where step (i) follows by the data processing inequality, step (ii) via the representation (9.4.4) of the mutual information, and step (iii) uses the sub-Gaussianity of $\ell(a, Y(a))$ (Theorem 7.1.2), and the final equality follows because $\mathbb{E}[\mathbb{E}[\ell(a, Y(a)) \mid A^\star]] = \mathbb{E}[\ell(a, Y(a))]$. □

We consider Lemma 18.3.9 conditional on the history $\mathcal{H}_{t-1}$. Defining the shorthand notation

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_{t-1}] \quad \text{and} \quad \operatorname{Var}_t(\cdot) := \operatorname{Var}(\cdot \mid \mathcal{H}_{t-1})$$

for the conditional expectation and variance given the history, the lemma shows that for any distribution ρ on $\mathcal{A}$, the loss/information ratio $R_t(\rho) \le 2\sigma^2 V_t(\rho)$ for the loss/variance ratio

$$V_t(\rho) := \frac{(\int_{\mathcal{A}} \mathbb{E}_t[\ell(a, Y(a)) - \ell(A^\star, Y(A^\star))] d\rho(a))^2}{\int_{\mathcal{A}} \operatorname{Var}_t(\mathbb{E}_t[\ell(a, Y(a)) \mid A^\star]) d\rho(a)}. \tag{18.3.11}$$

The quantity $V_t(\rho)$ is the ratio of a quadratic function in ρ over a linear function of ρ, and so is convex in ρ. (See Exercise 18.1.) Thus, at least from a computational perspective, it is natural to use *variance-directed sampling*, which chooses

$$\rho_t^{\mathrm{vds}} := \operatorname*{argmin}_{\rho \in \Delta(\mathcal{A})} V_t(\rho).$$

Whenever we can guarantee the variance ratio $\inf_\rho V_t(\rho)$ is bounded for all t, then evidently we have an order $\sqrt{n}$ regret bound, and we do not require information calculations.

For a discrete action set, we identify ρ with a vector $\rho \in \mathbb{R}_+^k$ or $\rho \in \mathbb{R}_+^{\mathbb{N}}$. Then using the shorthands $v = [\mathrm{Var}(\mathbb{E}[\ell(a, Y(a)) \mid A^\star])]_{a \in \mathcal{A}}$ and $r = [\mathbb{E}[\ell(a, Y(a)) - \ell(A^\star, Y(A^\star))]]_{a \in \mathcal{A}}$ for the vectors of variances and instantaneous regrets, respectively, the variance ratio becomes

$$V(\rho) := \frac{(\sum_{a \in \mathcal{A}} \rho(a) \mathbb{E}[\ell(a, Y(a)) - \ell(A^\star, Y(A^\star))])^2}{\sum_{a \in \mathcal{A}} \rho(a) \mathrm{Var}(\mathbb{E}[\ell(a, Y(a)) \mid A^\star])} = \frac{\langle \rho, r \rangle^2}{\langle \rho, v \rangle}.$$

For the K-armed bandit with sub-Gaussian arms, we can modify the argument that Thompson sampling has bounded loss/information ratio (18.3.9) to obtain an identical variance ratio bound:

Lemma 18.3.10. *Let $\mathcal{A}$ be discrete. Variance-directed sampling and Thompson sampling satisfy*

$$V_t(\rho_t^{\mathrm{vds}}) = \inf_\rho V_t(\rho) \le V_t(\rho_t^{\mathrm{ts}}) \le \mathrm{card}(\mathcal{A}). \tag{18.3.12}$$

Proof By construction, the distribution ρ_t^{ts} is identical to the distribution $\rho_t^\star = \rho^\star(\cdot \mid \mathcal{H}_{t-1})$ of $A^\star$, so $\inf_{\rho \in \Delta(\mathcal{A})} V_t(\rho) \le V_t(\rho_t^{\mathrm{ts}}) = V_t(\rho_t^\star)$. We may now proceed tacitly conditioning on the history $\mathcal{H}_{t-1}$ to avoid notational overload. For $A, A^\star \overset{\mathrm{iid}}{\sim} \rho^\star$,

$$\begin{aligned}
\mathbb{E}[\ell(A, Y(A)) - \ell(A^\star, Y(A^\star))] &= \sum_{a \in \mathcal{A}} \rho^\star(a) \mathbb{E}[\ell(a, Y(a))] - \sum_{a \in \mathcal{A}} \rho^\star(a) \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a] \\
&= \sum_{a \in \mathcal{A}} \rho^\star(a) (\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a]),
\end{aligned}$$

while the vector $v = [\mathrm{Var}(\mathbb{E}[\ell(a, Y(a)) \mid A^\star])]_{a \in \mathcal{A}}$ of variances satisfies

$$\langle \rho^\star, v \rangle = \sum_{a \in \mathcal{A}} \rho^\star(a) \sum_{a^\star \in \mathcal{A}} \rho^\star(a^\star) (\mathbb{E}[\ell(a, Y(a)) \mid A^\star = a^\star] - \mathbb{E}[\ell(a, Y(a))])^2.$$

For $K = \mathrm{card}(\mathcal{A})$, the Cauchy-Schwarz inequality implies that

$$\begin{aligned}
(\mathbb{E}[\ell(A, Y(A)) - \ell(A^\star, Y(A^\star))])^2 &\le K \sum_{a \in \mathcal{A}} (\rho^\star(a))^2 (\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a])^2 \\
&\le K \sum_{a, a^\star \in \mathcal{A}} \rho^\star(a) \rho^\star(a^\star) (\mathbb{E}[\ell(a, Y(a))] - \mathbb{E}[\ell(a, Y(a)) \mid A^\star = a^\star])^2 = K \langle \rho^\star, v \rangle,
\end{aligned}$$

because we have simply added more positive terms. □

As an immediate consequence of inequality (18.3.12), Theorem 18.3.2 implies the following.

Corollary 18.3.11. *Let ρ_t^{vds} be the variance-directed sampling distribution, and assume the other conditions of Theorem 18.3.6. Then*

$$\mathsf{Reg}_n(\mathcal{A}, \ell, \pi) \le \sqrt{2\sigma^2 \, \mathrm{card}(\mathcal{A}) H(A^\star)} \sqrt{n}.$$

18.3.4 An extended example: linear bandits

In many bandit and other exploration/exploitation problems, losses $\ell(a, Y(a))$ for a particular action $a \in \mathcal{A}$ correlate with those for other actions $a' \in \mathcal{A}$. *Linear bandits* provide the cleanest model capturing this setting, making them a natural target of analysis. In this case, we identify actions $\mathcal{A}$ with a subset of $\mathbb{R}^d$, and the "true" state of the world with a parameter $\theta \in \mathbb{R}^d$. Upon taking action $a \in \mathcal{A}$, we receive feedback and loss

$$Y(a) = \langle a, \theta \rangle + \varepsilon \quad \text{and} \quad \ell(a, Y(a)) = -Y(a), \tag{18.3.13}$$

where ε is conditionally mean-zero noise, and we therefore wish to maximize the inner product

$$\langle a, \theta \rangle = -\mathbb{E}[\ell(a, Y(a)) \mid \theta].$$

Example 18.3.12 (Network routing)**:** In the source routing problem, one wishes to send a sequence of packets from a source to a given destination, and one chooses a path (sequence of links) on which to send the packets among nodes $i = 1, \ldots, d$. Representing a path via a matrix $a \in \{0, 1\}^{d \times d}$, where $a_{ij} = 1$ if the packet is sent on the link from node i to j and 0 otherwise, then the (noiseless) total transit cost is $\sum_{ij} a_{ij}\theta_{ij}$, where θ_{ij} indicates the delay on edge (i, j). The actions $\mathcal{A}$ correspond to valid paths between nodes in the network. $\diamond$

Given the setting (18.3.13), we can extend Lemma 18.3.10.

Proposition 18.3.13. *Assume the linear bandit model* (18.3.13)*. Then variance-directed sampling and Thompson sampling satisfy*

$$V_t(\rho_t^{\text{vds}}) \le V_t(\rho_t^{\text{ts}}) \le d,$$

where V_t denotes the loss/variance ratio (18.3.11)*. If additionally $Y(a) = \langle \theta, a \rangle + \varepsilon$ is σ^2-sub-Gaussian for each $a \in \mathcal{A}$, then information-directed sampling, variance-directed sampling, and Thompson sampling satisfy the loss/information ratio bound*

$$R_t(\rho_t^{\text{ids}}) \le 2\sigma^2 V_t(\rho_t^{\text{vds}}) \le 2\sigma^2 V_t(\rho_t^{\text{ts}}) \le 2\sigma^2 d.$$

Proof We tacitly ignore the history $\mathcal{H}_{t-1}$ as it is immaterial for the proof, and let $V = V_t$ be the loss/variance ratio. Let $\rho = \rho^{\text{ts}}$, so that $A^\star \sim \rho$. Then $V(\rho^{\text{vds}}) \le V(\rho)$, and defining the matrix $M = \text{Cov}(\mathbb{E}[\theta \mid A^\star]) \in \mathbb{R}^{d \times d}$, when $A \sim \rho$ the variance becomes

$$\mathbb{E}_\rho[\text{Var}(\langle \mathbb{E}[\theta \mid A^\star], A \rangle \mid A)] = \langle \text{Cov}(\mathbb{E}[\theta \mid A^\star]), \mathbb{E}_\rho[AA^\top] \rangle = \langle M, \mathbb{E}_\rho[AA^\top] \rangle.$$

Noting that $\mathbb{E}[A] = \mathbb{E}[A^\star]$ when $A, A^\star \sim \rho$, because A is independent of θ we also have

$$\begin{aligned}\mathbb{E}[\ell(A, Y(A)) - \ell(A^\star, Y(A^\star))] = \mathbb{E}[\langle \theta, A^\star - A \rangle] &= \mathbb{E}[\langle \theta, A^\star \rangle - \langle \mathbb{E}[\theta], A^\star \rangle] \\ &= \mathbb{E}[\langle \mathbb{E}[\theta \mid A^\star] - \mathbb{E}[\theta], A^\star \rangle].\end{aligned}$$

Writing M^{-1} for the inverse on the range of M (the Moore-Penrose pseudo-inverse), Cauchy-Schwarz implies

$$\begin{aligned}\mathbb{E}[\langle \mathbb{E}[\theta \mid A^\star] - \mathbb{E}[\theta], A^\star \rangle]^2 &= \mathbb{E}\left[\left\langle M^{-1/2}\left(\mathbb{E}[\theta \mid A^\star] - \mathbb{E}[\theta]\right), M^{1/2}A^\star \right\rangle\right]^2 \\ &\le \mathbb{E}\left[\left\|M^{-1/2}(\mathbb{E}[\theta \mid A^\star] - \mathbb{E}[\theta])\right\|_2^2\right] \mathbb{E}\left[\left\|M^{1/2}A^\star\right\|_2^2\right] \\ &= \langle M^{-1}, \text{Cov}(\mathbb{E}[\theta \mid A^\star]) \rangle \left\langle M, \mathbb{E}[AA^\top] \right\rangle\end{aligned}$$

because A and $A^\star$ have identical distributions. Finally, we use that $\langle M^{-1}, M\rangle = \mathrm{rank}(M) \le d$, implying $V(\rho^{\mathrm{vds}}) \le V(\rho^{\mathrm{ts}}) = V(\rho) \le d$.

For the second claim of the proposition, we simply apply Lemma 18.3.10. □

As a corollary, we can obtain the "standard" regret bounds for linear bandits, assuming we have a prior π over the parameters $\Theta \subset \mathbb{R}^d$. In this case, the loss

$$\ell(a, y) = -y \;\; \text{satisfies} \;\; \mathbb{E}[\ell(a, Y(a)) \mid \theta] = -\langle a, \theta\rangle,$$

and so minimizing the loss over $a \in \mathcal{A}$ corresponds to maximizing $\langle a, \theta\rangle$. We then obtain the following corollary.

Corollary 18.3.14. *Let $\mathcal{A}$ be countable and $H(A^\star) < \infty$. Assume the linear bandit model (18.3.13) and that for each $a \in \mathcal{A}$, $Y(a) = \langle \theta, a\rangle + \varepsilon$ is σ^2-sub-Gaussian (18.3.7). Then information-directed sampling, variance-directed sampling, and Thompson sampling each have regret bound*

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{2\sigma^2 d\, H(A^\star)} \cdot \sqrt{n}.$$

Let us work through one example here, which shows how Thompson-style sampling procedures can apply to linear bandit problems.

Example 18.3.15 (Thompson sampling for linear bandits with a Gaussian prior)**:** Consider the linear bandit setting (18.3.13), and let $\Theta = \mathbb{R}^d$ be all of d-dimensional space; put a Gaussian prior $\mathsf{N}(0, \tau^2 I_d)$ on θ, where $\tau^2 < \infty$ captures the prior variance, and assume the noise $\varepsilon \stackrel{\mathrm{iid}}{\sim} \mathsf{N}(0, \sigma^2)$. At least at the first step $t = 1$, the observed loss

$$-\ell(a, Y(a)) = Y(a) = \langle a, \theta\rangle + \varepsilon \sim \mathsf{N}\left(0, \|a\|_2^2 \tau^2 + \sigma^2\right),$$

which is sub-Gaussian so long as $\mathcal{A}$ is bounded.

We claim that for $\mathcal{H}_t = \{Y_1(A_1), A_1, \ldots, Y_t(A_t), A_t\}$, we have

$$\theta \mid \mathcal{H}_t \sim \mathsf{N}\left(\mathbb{E}[\theta \mid \mathcal{H}_t], t^{-1}\Sigma_t\right) \tag{18.3.14}$$

for

$$\Sigma_t = \left(\frac{1}{t\tau^2} I_d + \frac{1}{t\sigma^2}\sum_{i=1}^t A_i A_i^\top\right)^{-1} \;\; \text{and} \;\; \mathbb{E}[\theta \mid \mathcal{H}_t] = \frac{1}{\sigma^2}\Sigma_t\left(\frac{1}{t}\sum_{i=1}^t A_i Y_i(A_i)\right).$$

Assuming that the distributional equality (18.3.14) holds, then we have the necessary sub-Gaussianity (18.3.7) for the regret bounds to apply, as conditional on $\mathcal{H}_t$, $\ell(a, Y(a))$ is Gaussian with variance $\frac{1}{t} a^\top \Sigma_t a + \sigma^2 \le \tau^2 \|a\|_2^2 + \sigma^2$, where we used $\Sigma_t \preceq t\tau^2$. So then for any action set $\mathcal{A}$ and $\mathsf{N}(0, \tau^2 I_d)$ prior π, we have the following expected regret bound for information-directed, variance-directed, or Thompson sampling and linear bandits: whenever $\mathcal{A}$ is finite,

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \pi) \le \sqrt{2d\left(\tau^2 \max_{a \in \mathcal{A}} \|a\|_2^2 + \sigma^2\right) \log \mathrm{card}(\mathcal{A})} \cdot \sqrt{n}.$$

Let us return to develop the posterior distribution (18.3.14) of θ. We prove the result inductively. Let $\pi(\theta \mid y_1^t, a_1^t)$ be shorthand for the density of θ given $Y_i(a_i) = y_i$ and $A_i = a_i$ for $i \le t$. We show that

$$\pi(\theta \mid y_1^t, a_1^t) \propto \pi(\theta \mid y_1^{t-1}, a_1^{t-1}) \exp\left(-\frac{1}{2\sigma^2}(y_t - \langle \theta, a_t\rangle)^2\right). \tag{18.3.15}$$

To see this, note that at time t, we have the graphical structure in Figure 18.4, so that $Y_t(A_t)$ is conditionally independent of the history $\mathcal{H}_{t-1}$ given A_t, θ. Thus we can write (with some abuse of notation) that

$$\begin{aligned}\pi(\theta \mid y_1^t, a_1^t) \propto p(\theta, y_t, a_t \mid y_1^{t-1}, a_1^{t-1}) &= \pi(\theta \mid y_1^{t-1}, a_1^{t-1}) p(y_t, a_t \mid \theta, y_1^{t-1}, a_1^{t-1}) \\ &= \pi(\theta \mid y_1^{t-1}, a_1^{t-1}) p(y_t \mid a_t, \theta) p(a_t \mid y_1^{t-1}, a_1^{t-1}),\end{aligned}$$

where we use p to denote an appropriate density or p.m.f. As $p(a_t \mid y_1^{t-1}, a_1^{t-1})$ does not depend on θ, we have the claim (18.3.15). For $t = 1$, $\pi(\theta \mid y_1^0, a_1^0) = \pi(\theta)$, and so inductively, we obtain

$$\pi(\theta \mid a_1^t, y_1^t) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^t (y_i - \langle a_i, \theta \rangle)^2 - \frac{1}{2\tau^2} \|\theta\|_2^2\right).$$

Given this equality, the conditional distribution (18.3.14) follows by algebraic manipulations (see Exercise 18.2). $\diamond$

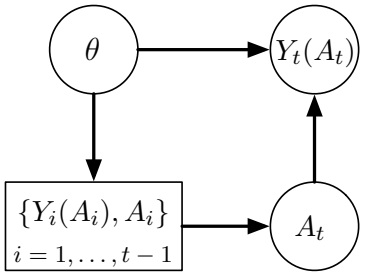


Figure 18.4. Graphical independence structure in sequential problems. Action A_t is drawn conditional on past observations and actions $\mathcal{H}_{t-1} = \{Y_i(A_i), A_i\}_{i=1}^{t-1}$, and given θ, observation $Y_t(A_t)$ is conditionally independent of $\mathcal{H}_{t-1}$.

The sampling distribution (18.3.14) is easy to use in Example 18.3.15: letting

$$\widehat{\theta}_t = \operatorname{argmin}\left\{\frac{1}{2\sigma^2} \sum_{i=1}^t (Y_i - \langle A_i, \theta \rangle)^2 + \frac{1}{2\tau^2} \|\theta\|_2^2\right\}$$

be the minimizer of the regularized squared error, we draw $\theta_{t+1} \sim \mathsf{N}(\widehat{\theta}_t, \frac{1}{t}\Sigma_t)$ for the covariance Σ_t in (18.3.14), and then choose

$$A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} \langle a, \theta_{t+1} \rangle.$$

For example, if $\mathcal{A} = \{-1, 1\}^d / \sqrt{d}$ consists of all $\{-1, 1\}$-valued vectors normalized to the unit ball, then so long as $\tau^2, \sigma^2 = O(1)$, we have expected regret $\mathsf{Reg}_n \le O(1) d\sqrt{n}$. Figure 18.5 shows example behavior with this sampling strategy, with the minor modification that we take $\mathcal{A} = \{a \in \mathbb{R}^d \mid \|a\|_2 \le 1\}$ to be the ℓ_2-ball. The figure shows that our analysis captures the typical behavior of the methods: the majority of the time, the (average) regret scales no worse than $d/\sqrt{n}$, and this appears to capture the typical behavior as well.

18.4 Online gradient descent approaches

Returning to the more basic multi-armed bandit setting of Section 18.1, we leverage the online optimization approaches from Chapter 17 to tackle these problems. We must formulate the bandit

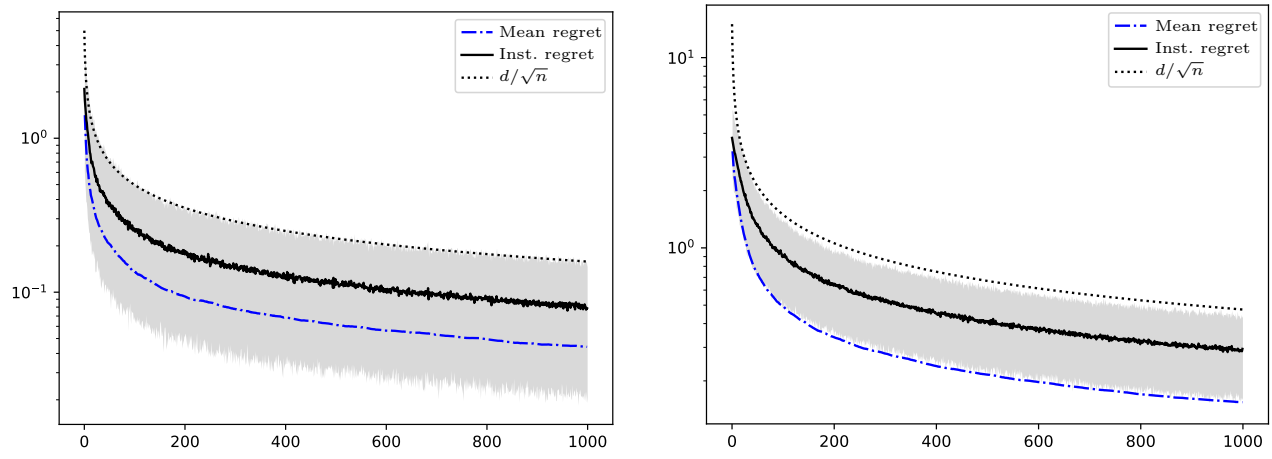


Figure 18.5. The regret behavior of Thompson sampling for the linear bandit problem (18.3.13). Each plot shows the results of 500 experiments run for 1000 iterations using the Gaussian prior of Example 18.3.15, where the action set $\mathcal{A} = \{a \in \mathbb{R}^d \mid \|a\|_2 \le 1\}$ is the ℓ_2-ball of radius 1. The solid black line (Inst. regret) in each plot shows the instantaneous regret $\langle \theta, A^\star - A_t \rangle$ at each iteration t, averaged over the 500 experiments. The dot-dashed line (Mean regret) shows the "posterior mean" regret: taking $\overline{\theta}_t = \mathbb{E}[\theta \mid \mathcal{H}_t]$ and the "mean" action $\overline{a}_t = \overline{\theta}_t / \|\overline{\theta}_t\|_2$. The dotted line shows $d/\sqrt{n}$, while the shaded region gives the 10–90% quantiles of the instantaneous regret across experiments.

problem as one of minimizing a partially observed sequence of convex losses, and the approaches in Chapter 17 require us to construct unbiased gradient estimators. Let us assume a setting analogous to the K armed bandit problem in Section 18.1, except that now we ask to find the minimal element of an unknown mean vector $\mu \in \mathbb{R}_+^K$ with nonnegative entries.

At each step t of the procedure, we play a distribution $w_t \in \Delta_K$ on the arms and select arm $A_t = a$ randomly with probability $w_{t,a}$. For a choice $A_t = a$, we have realized (and observed) loss $\ell(a, Y_t(a)) = Y_t(a)$ as in Section 18.3, but we suffer expected loss $\ell_t(w_t) = \langle w_t, \mu \rangle$. Because of its connections with probability estimation, we use the exponentiated gradient algorithm, Example 17.2.2, to play this game. We face one main difficulty: we must estimate the gradient of the expected losses, $\nabla \ell_t(w_t) = \mu$. We can construct such an estimate without too much additional variance.

Lemma 18.4.1. *Let $Y \in \mathbb{R}^K$ be a random variable with $\mathbb{E}[Y] = \mu$ and $w \in \Delta_K$ be a probability vector with positive entries. Choose coordinate $A = a \in [K]$ with probability w_a and define*

$$\widetilde{Y}(j) = \begin{cases} Y(j)/w_j & \text{if } j = a \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, K$. Then $\mathbb{E}[\widetilde{Y} \mid Y] = Y$.

Proof For each coordinate j of $\widetilde{Y}$, we have $\mathbb{E}[\widetilde{Y}(j) \mid Y] = w_j Y(j)/w_j = Y(j)$. □

Lemma 18.4.1 suggests the procedure 18.6, a variant of Auer et al.'s EXP3 (Exponentiated gradient for Exploration and Exploitation) algorithm [36]. We can provide an expected regret bound by leveraging the refined analysis of exponentiated gradients in Proposition 17.4.2.

Input: stepsize parameter η, initial vector $w_1 = [\frac{1}{K} \; \cdots \; \frac{1}{K}]^\top$
Repeat: for each iteration t,
Choose random action $A_t = a$ with probability $w_{t,a}$
Receive nonnegative loss $\ell(a, Y_t(a)) = Y_t(a)$, and define

$$g_{t,j} = \begin{cases} Y_t(j)/w_{t,j} & \text{if } A_t = j \\ 0 & \text{otherwise.} \end{cases}$$

Update for each $i = 1, \ldots, K$

$$w_{t+1,i} = \frac{w_{t,i} \exp(-\eta g_{t,i})}{\sum_j w_{t,j} \exp(-\eta g_{t,j})}.$$

Figure 18.6: Exponentiated gradient for bandit problems.

Proposition 18.4.2. *Assume that for each $j \in [K]$, $\mathbb{E}[Y(j)^2] \le \sigma^2$ and the observed loss $Y(j) \ge 0$ and let $a^\star$ minimize μ_a over $a \in [K]$. Then Algorithm 18.6 attains expected regret*

$$\overline{\mathsf{Reg}}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{a^\star}] \le \frac{\log K}{\eta} + \frac{\eta}{2}\sigma^2 K n.$$

In particular, choosing $\eta = \sqrt{\log K/(K\sigma^2 n)}$ gives

$$\overline{\mathsf{Reg}}_n = \sum_{t=1}^n \mathbb{E}[\mu_{A_t} - \mu_{a^\star}] \le \frac{3}{2}\sigma\sqrt{Kn \log K}.$$

Proof Recall the refined regret bound of Proposition 17.4.2. For $w^\star \in \Delta_K$ and any sequence of vectors $g_1, g_2, \ldots$ with $g_t \in \mathbb{R}_+^K$, the exponentiated gradient method achieves

$$\sum_{t=1}^n \langle g_t, w_t - w^\star \rangle \le \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^K w_{t,j} g_{t,j}^2.$$

We take expectations, as

$$\mathbb{E}[g_t \mid w_t] = \mathbb{E}[Y] = \mu$$

by construction, and we also have

$$\mathbb{E}\Big[\sum_{j=1}^K w_{t,j} g_{t,j}^2 \mid w_t\Big] = \sum_{j=1}^K w_{t,j}^2 \mathbb{E}[Y_t(j)^2/w_{t,j}^2 \mid w_t] = \sum_{j=1}^K \mathbb{E}[Y(j)^2] = \mathbb{E}[\|Y\|_2^2].$$

(This careful normalizing is essential, and it fails for more naive applications of online convex optimization bounds.) In particular, we have

$$\overline{\mathsf{Reg}}_n = \sum_{t=1}^n \mathbb{E}[\langle \mu, w_t - w^\star \rangle] = \sum_{t=1}^n \mathbb{E}[\langle g_t, w_t - w^\star \rangle] \le \frac{\log K}{\eta} + \frac{\eta}{2} n \mathbb{E}[\|Y\|_2^2].$$

Taking expectations gives the result. □

Proposition 18.4.2 provides a regret bound for the multi-armed bandit problem that applies when the observations $Y_t(a)$ are nonnegative with finite second moment. When $Y(a)$ is bounded or sub-Gaussian, it is possible to achieve high-probability guarantees on the regret, though this is beyond our scope.

18.4.1 Some empirical comparisons

In spite of the similarities in the regret bounds between each procedure we have developed, the methods frequently exhibit strikingly different performance. It appears that Thompson-sampling strategies and their variants—such as information-directed sampling—typically obtain the best empirical performance, even in non-Bayesian settings. We show a few results for K-armed Bernoulli bandits, where $Y(a) \sim \mathsf{Bernoulli}(\theta_a)$ for a vector $\theta \in [0,1]^K$. As a brief remark, the EXP3 algorithm (Fig. 18.6) often gets "stuck"—the weights on one arm become exponentially larger than the weights on others. Practically it appears important to use a variant that encourages additional exploration by mixing the weights with a uniform distribution via

$$w_{t,a} := \frac{\exp(-\eta \sum_{\tau<t} g_{\tau,a})}{\sum_j \exp(-\eta \sum_{\tau<t} g_{\tau,j})}, \quad \epsilon_t := \min\left\{\frac{1}{K}, \sqrt{\frac{\log K}{Kt}}\right\}, \quad \text{and} \quad \rho_t(a) = (1 - K\epsilon_t) w_{t,a} + \epsilon_t \tag{18.4.1}$$

for each $a \in \{1, \ldots, K\}$. Then we draw $A_t \sim \rho_t$ in iteration t. The strategy (18.4.1) encourages more exploration without substantial regret; see the references for pointers to its analysis.

In Figure 18.7, we plot results that show typical behavior of Upper Confidence Bound (UCB) algorithms (Section 18.2), Thompson/posterior sampling algorithms (Section 18.3.2), and ϵ-exploration variant (18.4.1) of exponentiated gradient (EXP3) algorithms. Each figure plots a summary of the instantaneous regrets $\theta_{A^\star} - \theta_{A_t}$ over 1000 iterations, collating the results of 4000 experiments. In each experiment, we draw $\theta \sim \mathsf{Uniform}([0,1]^K)$, and at each iteration t, observe $Y_t(a) \sim \mathsf{Bernoulli}(\theta_a)$ upon choosing action $A_t = a \in [K]$. The average regrets appear to decrease polynomially in t, consistent with the analyses here. Thompson sampling outperforms the other algorithms; it is a reasonable default procedure. Finally, the bottom plot in Fig. 18.7 shows that eventually, Thompson sampling appears to find the best arm, incurring 0 regret in most samples.

18.5 Minimax lower bounds

We turn to the converse problem of lower bounds for decision-making problems in the framework of Figure 18.2. We consider two variants of the problem: a pure exploration scenario, when there is no accumulated loss during the run of the procedure, so we measure only

$$\mathbb{E}[\ell(A_{n+1}, Y(A_{n+1}))] - \mathbb{E}[\ell(A^\star, Y(A^\star))] \tag{18.5.1}$$

and the regret-based scenarios that Figure 18.2 targets. Any lower bound on the former immediately implies a lower bound on the latter: indeed, define the action sampling distribution $\rho := \frac{1}{n}\sum_{t=1}^n \rho_t$ to be the average of the distributions played throughout the procedure. Let $A_{n+1} \sim \rho$, and let θ parameterize P_θ. Then

$$n\mathbb{E}_\theta[\ell(A_{n+1}, Y(A_{n+1})) - \ell(A^\star, Y(A^\star))] = \sum_{t=1}^n \mathbb{E}_\theta[\ell(A_t, Y(A_t)) - \ell(A^\star, Y(A^\star))] = \overline{\mathsf{Reg}}_n(\mathcal{A}, \ell, \theta),$$

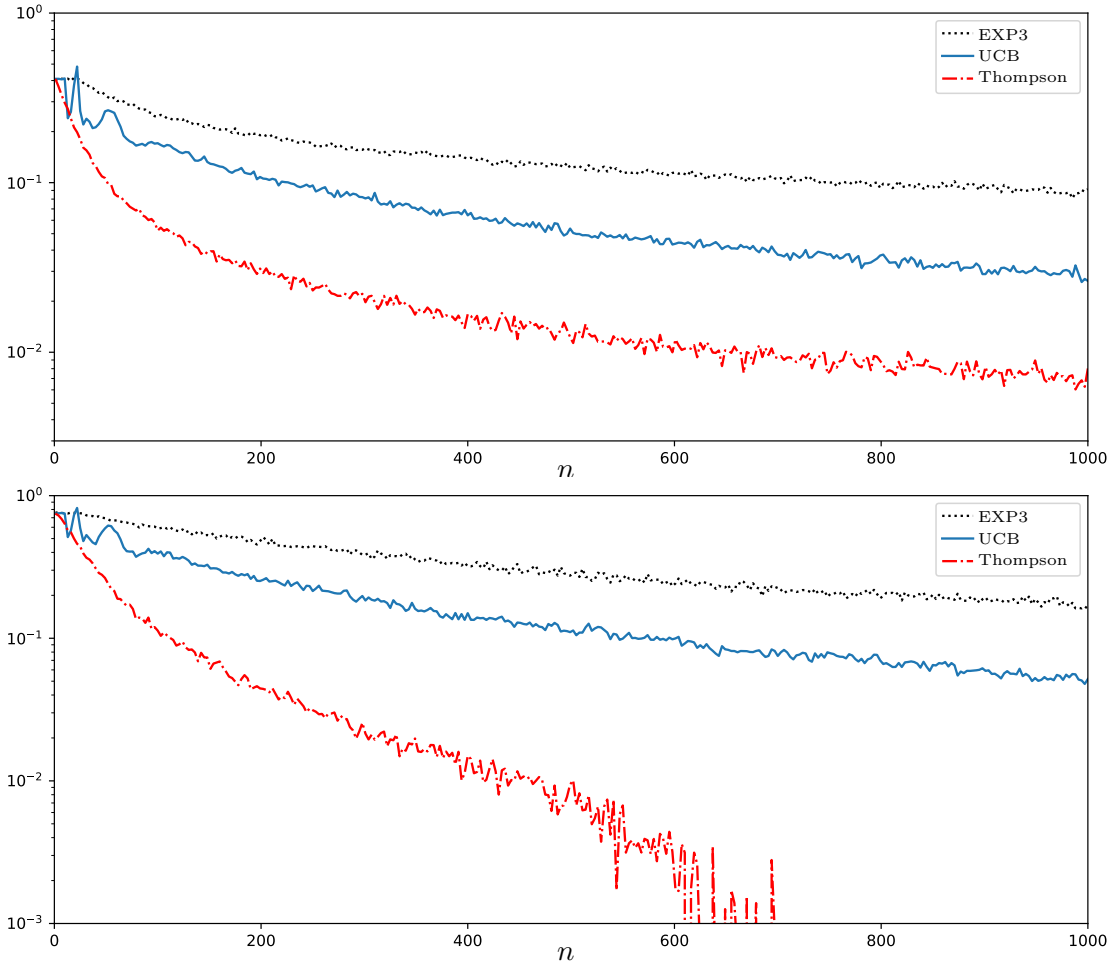


Figure 18.7. An empirical comparison of average regret for $K = 10$-armed Bernoulli bandit algorithms. Each figure represents the results of 4000 experiments. Top: mean instantaneous regret $\theta_{A^\star} - \theta_{A_t}$. Bottom: 90th percentile of instantaneous regret $\theta_{A^\star} - \theta_{A_t}$ across experiments. EXP3 uses the exploration strategy (18.4.1) with $\eta = \sqrt{\log K}/\sqrt{KT}$. UCB uses confidence $\delta = 10^{-5}$. Thompson sampling uses independent $\mathsf{Beta}(1,1)$ priors for each coordinate of $\theta \in [0,1]^K$.

and so any lower bound on the gap (18.5.1) immediately implies one on the regret.

Let us set notation to obtain cleaner statements. We say P is a *causal distribution* when it maps actions $a \in \mathcal{A}$ to a distribution over losses $L \in \mathbb{R}$ and observations $Y \in \mathcal{Y}$. In the notation of Figure 18.2, we have $L = \ell(A, Y(A))$, though we allow the losses to be any scalar and Y to denote the observed feedback under the sampling distribution $P(a)$, which for simplicity we assume includes the loss L. Thus, we observe $Y \sim P(a)$. The *gap* of an action a for a causal distribution P is

$$\mathsf{gap}_P(a) := \sup_{a^\star \in \mathcal{A}} \mathbb{E}_{P(a)}[L] - \mathbb{E}_{P(a^\star)}[L], \tag{18.5.2}$$

corresponding to $\mathbb{E}[\ell(a, Y(a)) - \ell(A^\star, Y(A^\star))]$ as in (18.5.1). We say a distribution q is an *exploration strategy* if for each $t = 1, 2, \ldots$, it defines a distribution $A_t \sim q(\cdot \mid \mathcal{H}_{t-1})$, for $\mathcal{H}_t = \{A_i, Y_i(A_i)\}_{i \le t}$.

Given the gap (18.5.2) and definition of an exploration strategy, the *minimax risk* of a decision

estimation problem is then

$$\mathfrak{M}_n(\mathcal{P}) := \inf_{\rho,q} \sup_{P\in\mathcal{P}} \mathbb{E}_{P,\rho,q}\left[\mathsf{gap}_P(A_{n+1})\right], \tag{18.5.3}$$

where the infimum is taken over all exploration strategies q and distributions ρ on actions A_{n+1}, where ρ is a measurable function of $\mathcal{H}_n = (A_1^n, Y_1^n)$. This quantity is the minimax risk for *best arm identification*, which measures only the error in identifying the optimal action $a \in \mathcal{A}$. We define the minimax regret similarly, letting

$$\mathfrak{R}_n(\mathcal{P}) := \inf_{q} \sup_{P\in\mathcal{P}} \sum_{t=1}^n \mathbb{E}_{P,q}[\mathsf{gap}_P(A_t)],$$

where we take the infimum over strategies q. The discussion above shows that

$$\mathfrak{R}_n(\mathcal{P}) \ge n \cdot \mathfrak{M}_n(\mathcal{P}),$$

so we focus essentially exclusively on $\mathfrak{M}_n(\mathcal{P})$. We will provide two strategies for lower bounds. The first derives from the modulus of continuity, or geometrizing, approach to functional estimation lower bounds we develop in Chapter 13. The second derives from Assouad's method (Section 9.5). In both, we must address the major challenge that all information gathering is interactive, which adds an additional wrinkle to the lower bound approaches that Part II of the text develops.

18.5.1 Action separation and a modulus of continuity

For our first lower bound on the minimax risk (18.5.3), we provide an analogue of the moduli of continuity that were central to lower bounds on scalar functional estimation in Chapter 13, except that now this modulus relates to decision making problems. For a collection $\mathcal{P}$ of causal distributions with associated action set $\mathcal{A}$, define the *action separation* of $\mathcal{P}$

$$\delta_{\mathcal{A}}(\mathcal{P}) := \sup_{P_1,P_2\in\mathcal{P}} \left\{\delta \ge 0 \mid \text{for all } a \in \mathcal{A},\ \mathsf{gap}_{P_1}(a) \ge \delta \text{ or } \mathsf{gap}_{P_2}(a) \ge \delta\right\}.$$

This measures the separation between distributions in that any action $a \in \mathcal{A}$ is sub-optimal for at least one of P_1 or P_2. For any causal distribution $\overline{P}$ and probability distribution q on $\mathcal{A}$, we can define ϵ-neighborhood of $\overline{P}$ indexed by $A \sim q$ by

$$\mathcal{P}_{q,\epsilon}(\overline{P}) := \left\{P \in \mathcal{P} \mid \mathbb{E}_q\left[D_{\rm kl}\left(\overline{P}(A)\|P(A)\right)\right] \le \epsilon^2\right\}.$$

To keep the notation clear, we mean $\mathbb{E}_q[D_{\rm kl}\left(\overline{P}(A)\|P(A)\right)] = \int_{\mathcal{A}} D_{\rm kl}\left(\overline{P}(a)\|P(a)\right) dq(a)$ to be the average KL-divergence between $\overline{P}(a)$ and $P(a)$ over q.

We can now define the analogue of the modulus of a parameter with respect to the Hellinger distance (13.1.1), which we shall term the *modulus of action separation* by

$$\omega_{\mathcal{A}}(\epsilon \mid \mathcal{P}, \overline{P}) := \inf_q \delta_{\mathcal{A}}\left(\mathcal{P}_{q,\epsilon}(\overline{P})\right). \tag{18.5.4}$$

Intuitively, we expect this quantity to control lower bounds: when ϵ is small, the definition of the KL-neighborhood $\mathcal{P}_{q,\epsilon}$ means that playing actions $A \sim q$ does not distinguish distributions P_1 and P_2 in $\mathcal{P}_{q,\epsilon}$, but any action a has some non-trivial loss for at least one of P_1 and P_2.

Continuing our analogy with Chapter 13, we have the following result, which shows that the action modulus (18.5.4) at radius $1/\sqrt{n}$ lower bounds the minimax risk:

Theorem 18.5.1. *For any causal distribution $\overline{P}$, not necessarily in $\mathcal{P}$, the minimax risk satisfies*

$$\mathfrak{M}_n(\mathcal{P}) \geq \frac{1}{4}\omega_{\mathcal{A}}\left(\frac{1}{\sqrt{8n}} \mid \mathcal{P}, \overline{P}\right).$$

We defer the proof of Theorem 18.5.1 to Section 18.5.3, instead giving two applications of the result.

The key in Theorem 18.5.1 is that a geometric-like quantity—the modulus (18.5.4)—guarantees lower bounds on the minimax risk for decision estimation. The next example exhibits a recipe that lower bounds the modulus to obtain a minimax lower bound.

Example 18.5.2 (Multi-armed bandits)**:** By Theorem 18.5.1, to prove a lower bound we can show that a family $\mathcal{P}$ of causal distributions, restricted to a neighborhood of some model $\overline{P}$, has large enough action separation. Consider the k-armed Gaussian bandit with rewards $P(a) = \mathsf{N}(\mu_a, \sigma^2)$ and loss $L = -Y(a)$ for means $\mu_a \in \mathbb{R}$. Then for the "null" $\overline{P}$ with reward distribution $\overline{P}(a) = \mathsf{N}(0, \sigma^2)$, for any p.m.f. q and mean reward vector $\mu \in \mathbb{R}^k$, we have

$$\sum_{i=1}^k q_i D_{\rm kl}\left(\overline{P}(i) \| P(i)\right) = \frac{1}{2\sigma^2}\sum_{i=1}^k q_i \mu_i^2.$$

Letting $\delta \geq 0$, when $\mu = \delta e_a$ for a standard basis vector e_a, $\mathbb{E}_q[D_{\rm kl}\left(\overline{P}(A) \| P(A)\right)] = \frac{q_a \delta^2}{2\sigma^2}$.

We now compute a lower bound on the action separation. For any q, there is necessarily at least one index i with $q_i \leq \frac{1}{k}$, and at least one distinct index j such that $q_j \leq \frac{1}{k-1}$. (Otherwise, we would have $\sum_i q_i > 1$.) Let P_1 and P_2 correspond to mean vectors $\mu = \delta e_i$ and δe_j for these two indices. Then at least one of $\mathsf{gap}_{P_1}(a) \geq \delta$ or $\mathsf{gap}_{P_2}(a) \geq \delta$, while

$$\mathbb{E}_q[D_{\rm kl}\left(\overline{P}(A) \| P_1(A)\right)] \leq \frac{\delta^2}{2k\sigma^2} \quad \text{and} \quad \mathbb{E}_q[D_{\rm kl}\left(\overline{P}(A) \| P_2(A)\right)] \leq \frac{\delta^2}{2(k-1)\sigma^2}.$$

So for any distribution q on A and $\epsilon > 0$, we obtain action separation

$$\delta_{\mathcal{A}}(\mathcal{P}_{q,\epsilon}(\overline{P})) \geq \sup\left\{\delta \geq 0 \mid \frac{\delta^2}{2(k-1)\sigma^2} \leq \epsilon^2\right\} = \sigma\epsilon\sqrt{2(k-1)}.$$

In particular, $\omega_{\mathcal{A}}(\epsilon \mid \mathcal{P}, \overline{P}) \geq \sigma\epsilon\sqrt{2(k-1)}$, and so the minimax risk for best-arm identification in a k-armed Gaussian bandit is at least

$$\mathfrak{M}_n(\mathcal{P}) \geq \frac{\sigma\sqrt{k-1}}{8\sqrt{n}}$$

by Theorem 18.5.1. As in our discussion of UCB algorithms at the end of Section 18.2, this is sharp. $\diamond$

Exercise 18.10 develops a second application of the technique, applying it to linear bandits, though the lower bound is loose by a factor of $\sqrt{d}$, motivating the different lower bounding technique to which we now turn.

18.5.2 Assouad's method for lower bounds

Assouad's method, which we initially develop in Chapter 9.5, extends elegantly to adaptive problems. The key insight is that when we reduce to collections of binary tests, as in the lower bound (9.5.4), we can apply chain rules for KL-divergence even with adaptive information gathering. We apply this idea to linear bandit problems, which helps both to highlight the ideas behind its application and in providing rate-optimal minimax and Bayesian lower bounds for both pure exploration problems and regret.

Recall the key steps in applying Assouad's method: we embed the problem in the hypercube $\mathcal{V} = \{\pm 1\}^d$, demonstrate a loss with a Hamming separation, and then show that distributions $P_v, P_{v'}$ for which v, v' differ in only a single coordinate are close in KL-divergence. Following this recipe, we will work in the linear bandit with Gaussian noise setting, where for a parameter $\theta \in \mathbb{R}^d$ and actions $x \in \mathbb{R}^d$, we observe rewards

$$y(x) = \langle \theta, x \rangle + \varepsilon \quad \text{and} \quad \mathsf{gap}_{P_\theta}(a) = \sup_{x \in \mathcal{A}} \langle \theta, x \rangle - \langle \theta, a \rangle.$$

The interactions between the action set $\mathcal{A}$ and the parameter space Θ are sophisticated, so we focus on two standard cases: we always take $\Theta = \mathbb{B}_2^d$, the ℓ_2-ball, and consider either the actions $\mathcal{A} = [-1, 1]^d$ or $\mathcal{A} = \mathbb{B}_2^d$. We begin by stating the main theorem of this section and a few corollaries. In the theorem, we elaborate the notation (18.5.3) for the minimax risk to include the action set $\mathcal{A}$, so $\mathfrak{M}_n(\mathcal{P}, \mathcal{A}) = \inf_{\rho, q} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,\rho,q}[\mathsf{gap}_P(A_{n+1})]$, and ρ and q draw actions in $\mathcal{A}$.

Theorem 18.5.3. *Let $\mathcal{P}$ be the collection of linear bandit models $y(x) = \langle \theta, x \rangle + \varepsilon$, $\|\theta\|_2 \le 1$, with $\varepsilon \sim \mathsf{N}(0, \sigma^2)$-noise. Then the following minimax lower bounds hold.*

(i) For any action set satisfying $\{-1, 1\}^d \subset \mathcal{A} \subset [-1, 1]^d$,

$$\mathfrak{M}_n(\mathcal{P}, \mathcal{A}) \ge \frac{1}{8} \cdot \left(\frac{d\sigma}{\sqrt{n}} \wedge \sqrt{d} \right).$$

(ii) For any action set satisfying $\{\pm 1/\sqrt{d}\}^d \subset \mathcal{A} \subset \mathbb{B}_2^d$,

$$\mathfrak{M}_n(\mathcal{P}, \mathcal{A}) \ge \frac{1}{16} \cdot \left(\frac{d\sigma}{\sqrt{n}} \wedge 1 \right).$$

Before proving the theorem, we can develop a few simple corollaries to the result and proof technique by considering Bayesian regret (18.3.2).

Corollary 18.5.4. *Let the conditions of Theorem 18.5.3 hold and assume that $n \ge d$. Let $\mathcal{A}$ either satisfy condition (i) or (ii) of the theorem and let $\ell_{\mathrm{id}}(a, y) = -y$ be the (negative) identity loss. Then there is a $\delta \in [0, 1/\sqrt{d}]$ such that for π_{Uniform} uniform over $\theta \in \{-\delta, \delta\}^d$,*

$$\mathsf{R}\overline{\mathsf{eg}}_n(\mathcal{A}, \ell_{\mathrm{id}}, \pi_{\mathsf{Uniform}}) \ge \frac{1}{16} d\sigma\sqrt{n}.$$

Of course, the worst-case regret also satisfies the lower bound in the corollary, so that

$$\sup_{\theta \in \Theta} \mathsf{R}\overline{\mathsf{eg}}_n(\mathcal{A}, \ell_{\mathrm{id}}, \theta) \ge \frac{1}{16} d\sigma\sqrt{n}$$

under the same conditions. The corollary shows that our analyses of information-directed and Thompson sampling were sharp, at least to within numerical constants; Corollary 18.3.14 shows that if $\mathcal{A} = \{\pm 1/\sqrt{d}\}^d$ or $\mathcal{A} = \{-1, 1\}^d$, then because $H(A^\star) \le d \log 2$, we have

$$\overline{\mathsf{Reg}}_n(\mathcal{A}, \ell_{\rm id}, \pi) \le \sqrt{2 \log 2} \cdot d\sigma\sqrt{n}$$

for any prior π.

Theorem 18.5.3 and Corollary 18.5.4 also highlight how the action modulus (18.5.4), while successful for K-armed bandit problems, fails to capture the correct lower bounds for linear bandit problems. Exercise 18.3 asks you to show how to use stochastic subgradient methods to achieve the lower bounds in Theorem 18.5.3 for the pure exploration case, highlighting a fairly straightforward approach. Exercise 18.4 shows that a minor variant of the problem leads to quite different convergence behavior, highlighting some of the subtleties in decision estimation problems, especially with regards to the interaction between action sets $\mathcal{A}$ and the underlying distributions.

Proof of Theorem 18.5.3. We follow Assouad's method. Beginning with the embedding, for $v \in \mathcal{V} = \{-1, 1\}^d$, let $\theta_v = \frac{\delta}{\sqrt{d}} v$, where $\delta \le 1$ is to be chosen in each construction. The next lemma then demonstrates a Hamming separation (9.5.1) for each action set.

Lemma 18.5.5. *Let $\theta \in \{\theta_v\}_{v \in \mathcal{V}} = \{\pm\delta/\sqrt{d}\}^d$ as above, and let $a^\star(\theta) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta \rangle$. Then*

(i) For any action set satisfying $\{-1, 1\}^d \subset \mathcal{A} \subset [-1, 1]^d$,

$$\langle \theta, a^\star(\theta) - a \rangle \ge \frac{\delta}{\sqrt{d}} \sum_{j=1}^d \mathbf{1}\left\{\operatorname{sign}(a_j) \ne \operatorname{sign}(\theta_j)\right\}.$$

(ii) For any action set satisfying $\{\pm 1/\sqrt{d}\}^d \subset \mathcal{A} \subset \mathbb{B}_2^d$,

$$\langle \theta, a^\star(\theta) - a \rangle \ge \frac{\delta}{2d} \sum_{j=1}^d \mathbf{1}\left\{\operatorname{sign}(a_j) \ne \operatorname{sign}(\theta_j)\right\}.$$

Proof For the first claim of the lemma, we have $a^\star(\theta) = \operatorname{sign}(\theta)$ (defined elementwise), and so

$$\langle \theta, a^\star(\theta) - a \rangle = \sum_{j=1}^d (|\theta_j| - a_j \theta_j) \ge \sum_{j=1}^d |\theta_j| \mathbf{1}\left\{\operatorname{sign}(a_j) \ne \operatorname{sign}(\theta_j)\right\},$$

where we used that $a_j \in [-1, 1]$. Noting that $|\theta_j| = \delta/\sqrt{d}$ gives the first result.

For the second result, note that whenever $\|a\|_2 \le 1$ and $v \in \{-1, 1\}^d$, we have

$$\sqrt{d} \sum_{j=1}^d \left(\frac{1}{\sqrt{d}} - a_j v_j\right)^2 = \sqrt{d}(1 - 2\langle a, v\rangle/\sqrt{d} + \|a\|_2^2) \le 2\sqrt{d}\left(1 - \langle a, v\rangle/\sqrt{d}\right).$$

Scaling terms appropriately, for $v = \sqrt{d}\theta/\delta$ we obtain

$$\begin{aligned}\frac{2}{\delta}\langle \theta, a^\star(\theta) - a\rangle = \frac{2}{\sqrt{d}}\langle v, a^\star(\theta) - a\rangle &= 2\left(1 - \langle v, a\rangle/\sqrt{d}\right) \\ &\ge \sum_{j=1}^d \left(\frac{1}{\sqrt{d}} - a_j v_j\right)^2 \ge \sum_{j=1}^d \frac{1}{d} \mathbf{1}\left\{\operatorname{sign}(a_j) \ne v_j\right\}.\end{aligned}$$

Multiplying by $\frac{\delta}{2}$ gives the lemma. □

As the last step, we must control the variation distance terms in the lower bound (9.5.4). Let $P_{v,+j}$ and $P_{v,-j}$ denote the distribution of the observations $Y_1, \ldots, Y_n$ under the model $\theta = \delta v/\sqrt{d}$ where v_j is set to (respectively) 1 or -1. Then we have the following lemma:

Lemma 18.5.6. *Let $\mathcal{A} \subset \mathbb{R}^d$ be any action set satisfying $\|a\|_2 \le D < \infty$ for all $a \in \mathcal{A}$. Then for the Gaussian linear bandit model,*

$$\frac{1}{d2^d} \sum_{j=1}^{d} \sum_{v \in \mathcal{V}} \|P_{v,+j} - P_{v,-j}\|_{\mathrm{TV}}^2 \le \frac{n\delta^2}{\sigma^2 d^2} D^2.$$

Proof Fix $v \in \{-1, 1\}^d$ temporarily, and let $P = P_{v,+j}$ and $Q = P_{v,-j}$. Use the shorthand that P_i is the distribution of the ith observation Y_i conditional on the history $\mathcal{H}_{i-1}$ and similarly for Q_i. Let $\theta_+ \in \mathbb{R}^d$ and $\theta_- \in \mathbb{R}^d$ be the parameters associated with $P_{v,+j}$ and $P_{v,-j}$, respectively. Then applying the chain rule for the KL-divergence (Proposition 2.1.13), we obtain

$$\begin{aligned} D_{\mathrm{kl}}\left(P_{v,+j} \| P_{v,-j}\right) = \sum_{i=1}^{n} \mathbb{E}_P[D_{\mathrm{kl}}\left(P_i \| Q_i\right)] &= \sum_{i=1}^{n} \mathbb{E}_P\left[D_{\mathrm{kl}}\left(\mathsf{N}(\langle A_i, \theta_+ \rangle, \sigma^2) \| \mathsf{N}(\langle A_i, \theta_- \rangle, \sigma^2) \mid A_i\right)\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_P\left[\frac{\langle A_i, \theta_+ - \theta_- \rangle^2}{2\sigma^2}\right] = \frac{2\delta^2}{d\sigma^2} \sum_{i=1}^{n} \mathbb{E}[A_{ij}^2]. \end{aligned}$$

Now averaging over all $v \in \{\pm 1\}^d$ and coordinates $j = 1, \ldots, d$, we obtain

$$\frac{1}{2^d d} \sum_{j=1}^{d} \sum_{v \in \mathcal{V}} D_{\mathrm{kl}}\left(P_{v,+j} \| P_{v,-j}\right) \le \frac{2\delta^2}{\sigma^2 d^2} \sum_{i=1}^{n} \mathbb{E}\left[\|A_i\|_2^2\right],$$

and applying Pinsker's inequality gives the result. □

Combining Lemmas 18.5.5 and 18.5.6, we can complete the proof of the theorem by considering the cases (i) and (ii) distinguishing the action sets $\mathcal{A}$. We first consider case (i), the hypercube-like action set, where $\{-1, 1\}^d \subset \mathcal{A} \subset [-1, 1]^d$. Then part (i) of Lemma 18.5.5, Assouad's method (9.5.4), and Lemma 18.5.6 (take $D = \sqrt{d}$) show that

$$\mathfrak{M}_n(\mathcal{P}, \mathcal{A}) \ge \frac{\sqrt{d}\delta}{2} \left[1 - \left(\frac{n\delta^2}{\sigma^2 d}\right)^{1/2}\right].$$

Setting $\delta = \frac{\sqrt{d\sigma^2/n}}{2} \wedge 1$ implies the first claim of the theorem.

For case (ii), the ℓ_2-type action set satisfying $\{\pm 1/\sqrt{d}\}^d \subset \mathcal{A} \subset \mathbb{B}_2^d$, the same argument (except using part (ii) of Lemma 18.5.5 and setting the action radius $D = 1$) implies

$$\mathfrak{M}_n(\mathcal{P}, \mathcal{A}) \ge \frac{\delta}{4} \left[1 - \left(\frac{n\delta^2}{\sigma^2 d^2}\right)^{1/2}\right].$$

Set $\delta = \frac{d\sigma}{2\sqrt{n}} \wedge 1$.

18.5.3 Proof of Theorem 18.5.1

We begin with a minor remark to avoid some trivialities. If $\omega_{\mathcal{A}}(\epsilon \mid \mathcal{P}, \overline{P}) > 0$, then *a fortiori* the set $\mathcal{P}_{q,\epsilon}(\overline{P})$ is non-empty for all distributions q on actions, because $\delta_{\mathcal{A}}(\emptyset) = 0$. We thus proceed as if $\mathcal{P}_{q,\epsilon}(\overline{P})$ is non-empty, as otherwise the lower bound of 0 is vacuous. We shall also assume w.l.o.g. that $\mathcal{A}$ is discrete, though this is only for notational convenience.

We begin by letting $\delta > 0$ to be chosen. For simplicity in notation, we assume the observations Y include the observed losses. For any exploration strategy q and probability distribution ρ, the latter defined conditional on the history $\mathcal{H}_n := (Y_1^n, A_1^n)$, we define the distribution $\mathbb{P}_{P,\rho,q}$ on actions by $\mathbb{P}_{P,\rho,q}(A_{n+1} = a) = \mathbb{E}_{P,q}[\rho(a \mid \mathcal{H}_n)]$, where the expectation is taken jointly over Y_1^n drawn from P under the exploration strategy q. We let

$$q_P(a) := \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P,q} \left[q(a \mid \mathcal{H}_{i-1}) \right] \tag{18.5.5}$$

be the average exploration distribution.

Markov's inequality implies the trivial inequality

$$\mathbb{E}_{P,\rho,q}[\mathsf{gap}_P(A)] \geq \delta \cdot \mathbb{P}_{P,\rho,q}(\mathsf{gap}_P(A) \geq \delta).$$

We modify the probabilistic model to an arbitrary $\overline{P}$, observing that

$$\mathbb{E}_{P,\rho,q}[\mathsf{gap}_P(A)] \geq \delta \cdot \left(\mathbb{P}_{\overline{P},\rho,q}(\mathsf{gap}_P(A) \geq \delta) - \left\| \mathbb{P}_{\overline{P},\rho,q} - \mathbb{P}_{P,\rho,q} \right\|_{\mathrm{TV}} \right).$$

As is typical in our lower bound proofs, we now show that the first probability is constant, and use that P and $\overline{P}$ are "close" to bound the variation distance. Let $\mathbb{P}_{P,q}$ be the joint distribution of the observations (Y_1^n, A_1^n) under P and q, so that the data processing inequality implies

$$\left\| \mathbb{P}_{\overline{P},\rho,q} - \mathbb{P}_{P,\rho,q} \right\|_{\mathrm{TV}}^2 \leq \left\| \mathbb{P}_{\overline{P},q} - \mathbb{P}_{P,q} \right\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}} \left(\mathbb{P}_{\overline{P},q} \| \mathbb{P}_{P,q} \right)$$

because of the Markov structure $(A_1^n, Y_1^n) \to A_{n+1}$, by Pinsker's inequality.

At step i under exploration distribution q and causal distribution P, Y_i has distribution $Y_i \mid A_i = a \sim P(a)$. By the Markovian action sampling structure that $A_i \mid \mathcal{H}_{i-1} \sim q(\cdot \mid \mathcal{H}_{i-1})$ under both $\overline{P}$ and P, the chain rule (Proposition 2.1.13) implies that

$$\begin{aligned} D_{\mathrm{kl}} \left(\mathbb{P}_{\overline{P},q} \| \mathbb{P}_{P,q} \right) & = \sum_{i=1}^n \mathbb{E}_{\overline{P},q} \left[D_{\mathrm{kl}} \left(\mathbb{P}_{\overline{P},q}(A_i \in \cdot, Y_i \in \cdot \mid \mathcal{H}_{i-1}) \| \mathbb{P}_{P,q}(A_i \in \cdot, Y_i \in \cdot \mid \mathcal{H}_{i-1}) \right) \right] \\ & = \sum_{i=1}^n \mathbb{E}_{\overline{P},q} \left[D_{\mathrm{kl}} \left(\overline{P}(A_i) \| P(A_i) \right) \right] = n \mathbb{E}_{q_{\overline{P}}} \left[D_{\mathrm{kl}} \left(\overline{P}(A) \| P(A) \right) \right], \end{aligned}$$

where the last equality follows by definition (18.5.5) of q_P.

Combining the steps, we have shown that for any pair ρ, q and any models $P, \overline{P}$, we have

$$\mathbb{E}_{P,\rho,q}[\mathsf{gap}_P(A_{n+1})] \geq \delta \left(\mathbb{P}_{\overline{P},\rho,q} \left(\mathsf{gap}_P(A_{n+1}) \geq \delta \right) - \sqrt{\frac{n}{2} \mathbb{E}_{q_{\overline{P}}} \left[D_{\mathrm{kl}} \left(\overline{P}(A) \| P(A) \right) \right]} \right).$$

Now we use the definition of the action separation: for any distribution q on actions $a \in \mathcal{A}$, we have $\inf_{q^\star} \delta_{\mathcal{A}}(\mathcal{P}_{q^\star,\epsilon}(\overline{P})) \leq \delta_{\mathcal{A}}(\mathcal{P}_{q,\epsilon}(\overline{P}))$, and in particular, this holds for $q = q_{\overline{P}}$. So taking

$\delta < \delta_{\mathcal{A}}(\mathcal{P}_{q_{\overline{P}},\epsilon}(\overline{P}))$, there must exist P_1 and P_2 so that for all $a \in \mathcal{A}$ we have at least one of $\mathsf{gap}_{P_1}(a) \ge \delta$ or $\mathsf{gap}_{P_2}(a) \ge \delta$ while $\mathbb{E}_{q_{\overline{P}}}[D_{\rm kl}\left(\overline{P}(a) \| P_i(a)\right)] \le \epsilon^2$ for $i = 1, 2$. In particular,

$$\begin{aligned}
&2 \max_{i \in \{1,2\}} \mathbb{E}_{P_i,\rho,q}[\mathsf{gap}_{P_i}(A_{n+1})] \\
&\ge \delta \left(\mathbb{P}_{\overline{P},\rho,q}\left(\mathsf{gap}_{P_1}(A_{n+1}) \ge \delta\right) + \mathbb{P}_{\overline{P},\rho,q}\left(\mathsf{gap}_{P_2}(A_{n+1}) \ge \delta\right) - 2\sqrt{\frac{n}{2}\epsilon^2} \right) \\
&\ge \delta \left(1 - \sqrt{2n\epsilon^2}\right),
\end{aligned}$$

because at least one of the events $\mathsf{gap}_{P_i}(A_{n+1}) \ge \delta$ must occur.

Finally, we substitute appropriate values. So long as the separation $\epsilon^2 \le \frac{1}{8n}$, we have $2n\epsilon^2 \le \frac{1}{4}$, and therefore

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{P,\rho,q}[\mathsf{gap}_P(A_{n+1})] \ge \frac{\delta}{2}\left(1 - \sqrt{2n\epsilon^2}\right) \ge \frac{\delta}{4}.$$

Take $\delta \uparrow \delta_{\mathcal{A}}(\mathcal{P}_{q_{\overline{P}},\epsilon}(\overline{P})) \ge \omega_{\mathcal{A}}(\epsilon \mid \mathcal{P}, \overline{P})$ with $\epsilon^2 = \frac{1}{8n}$.

18.6 A technical proof

18.6.1 Proof of Lemma 18.2.1

Without loss of generality, we can assume that any algorithm choosing arms is, at time t, a function of the past arms $A_1, \ldots, A_{t-1}$, responses $Y_1(A_1), \ldots, Y_{t-1}(A_{t-1})$, and an auxiliary random variable $U \sim \mathsf{Uniform}[0,1]$. Let $\mathcal{H}_t = \sigma(\{A_1, \ldots, A_t, Y_1(A_1), \ldots, Y_t(A_t), U\})$ be the history σ-field generated by U, the actions, and responses through time t. Then $A_t \in \mathcal{H}_{t-1}$.

To see Lemma 18.2.1, we use the standard fact [69] that the characteristic function of a random variable completely characterizes its distribution. Let $\varphi_i(\lambda) = \mathbb{E}[e^{\lambda Y(i)}] = \mathbb{E}[e^{\lambda Y'(i)}]$ for $\lambda \in \mathbb{C}$ denote the complex extension of the moment generating function of $Y(i)$, which is identical to that of $Y'(i)$ by construction. Writing the joint MGF of $N_t(i)\widehat{\mu}_t(i)$ and $N_t(i)$, we obtain by the tower property of expectation[1] that for $\lambda_1, \lambda_2 \in \mathbb{C}$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda_1 \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\} Y_\tau(i) + \lambda_2 N_t(i)\right)\right] \\
&= \mathbb{E}\left[\prod_{\tau=1}^t \mathbb{E}\left[\exp\left(\lambda_1 \mathbf{1}\left\{A_\tau = i\right\} Y_\tau(i) + \lambda_2 \mathbf{1}\left\{A_\tau = i\right\}\right) \mid \mathcal{H}_{\tau-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{\tau=1}^t \left(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2} \mathbb{E}\left[\exp(\lambda_1 Y_\tau(i)) \mid \mathcal{H}_{\tau-1}\right] + \mathbf{1}\left\{A_\tau \ne i\right\}\right)\right]
\end{aligned}$$

because $N_t(i) = \sum_{\tau=1}^t \mathbf{1}\left\{A_\tau = i\right\}$ and $A_\tau \in \mathcal{H}_{\tau-1}$, that is, it is a function of the history. Then we observe that conditional on $\mathcal{H}_{\tau-1}$, $Y_\tau(i)$ and $Y'_\tau(i)$ have identical distribution, so that

$$\mathbb{E}[\exp(\lambda_1 Y_\tau(i)) \mid \mathcal{H}_{\tau-1}] = \varphi_i(\lambda_1) = \mathbb{E}[\exp(\lambda_1 Y'_\tau(i)) \mid \mathcal{H}_{\tau-1}].$$

[1] That if X_t is $\mathcal{H}_{t-1}$ measurable, then $\mathbb{E}[X_1 \cdots X_N] = \mathbb{E}[\prod_{t=1}^N \mathbb{E}[X_t \mid \mathcal{H}_{t-1}]]$

Unrolling the above product, we therefore obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda_1 \sum_{\tau=1}^{t} \mathbf{1}\left\{A_\tau = i\right\} Y_\tau(i) + \lambda_2 N_t(i)\right)\right] \\
&= \mathbb{E}\left[\prod_{\tau=1}^{t} \left(\mathbf{1}\left\{A_\tau = i\right\} e^{\lambda_2} \mathbb{E}\left[\exp(\lambda_1 Y'_\tau(i)) \mid \mathcal{H}_{\tau-1}\right] + \mathbf{1}\left\{A_\tau \neq i\right\}\right)\right] \\
&= \mathbb{E}\left[\prod_{\tau=1}^{t} \mathbb{E}\left[\exp(\lambda_1 Y'_\tau(i)\mathbf{1}\left\{A_\tau = i\right\} + \lambda_2 \mathbf{1}\left\{A_\tau = i\right\}) \mid \mathcal{H}_{\tau-1}\right]\right] \\
&= \mathbb{E}\left[\exp\left(\lambda_1 \sum_{\tau=1}^{t} \mathbf{1}\left\{A_\tau = i\right\} Y'_\tau(i) + \lambda_2 N_t(i)\right)\right]
\end{aligned}$$

by again using the tower property of expectations. This gives the equality (18.2.1).

18.7 Bibliography and further reading

Causal estimation problems gave birth to statistics through the 1920s work of Ronald Fisher and his assistant Winifred Mackenzie on variation in crop responses to weather, fertilizer, and other treatments [186, 188]. At the same time, Neyman [325] developed his eponymous causal model through the study of plant yields. Such causal inference problems, especially in the context of the potential outcomes framework (Examples 11.2.5 and 18.1.1), remain an active area of research. The books by Imbens and Rubin [247] and Wager [418] survey and provide references for much of this literature.

Bandit problems have existed since at least Thompson's 1933 paper [395]. In the statistics literature, "design of experiments" considers choosing covariates in linear regression to best estimate a regression vector θ [349], which Robbins considered in sequential settings in the 1950s [361], yielding the linear bandit problems Section 18.3.4 outlines. Lai and Robbins [271] initiated the "modern" analysis of bandit problems, showing how confidence-based algorithms achieve regret of order $K \log n / \Delta^2$, where $\Delta = \min_{i \neq i^\star} \mu_{i^\star} - \mu_i$, and providing matching lower bounds.

More recent work in machine learning (of which there are too many references to list) has been motivated by online auctions, medical scenarios, and reinforcement learning problems. The books [105, 91, 275] provide excellent references. Auer, Cesa-Bianchi, Freund, and Schapire [36] introduced upper confidence bound (UCB) algorithms and provided finite sample bounds on their regret, while Auer, Cesa-Bianchi, and Fischer [35] introduced EXP3 and its analysis. Interest in Thompson sampling reignited when Chapelle and Li [108] showed that, in spite of its (at the time) heuristic nature, it was competitive with established procedures, and Kaufmann et al. [264] provided the first analysis of the procedure.

Russo and Van Roy [370, 368, 369] pioneer the information-ratio approach to Bayesian bandits in Section 18.3. These analyses require a certain well-specification of the procedures, so that the prior and posterior distributions on the parameter are accurate, as otherwise the regret bounds fail; extending Thompson sampling to unknown priors requires additional techniques (see, e.g. [11] or [275, Chapter 36]). That most analyses of bandit problems repose on some type of well-specification assumption—e.g., for linear bandits, that rewards $Y(a) = \langle \theta, a \rangle + \varepsilon$—poses a problem for saying anything rigorous about their behavior in real-world scenarios. Researchers have begun to investigate this issue, showing that in the case of mis-specified models, finding the best approximation

in the class can take time exponential in the dimension d of the approximator [151], while weaker approximation guarantees admit efficient algorithms [276, 195]. Many questions in this direction remain open.

The approach we take in Section 18.5 derives from a variety of sources. Section 18.5.1 presents a distilled and weaker version of Foster et al.'s *decision estimation coefficient* [193, 194]. We, however, take a perspective a little closer to the modulus of continuity of a statistic with respect to divergences, as in Chapter 13. The idea of using the empirical distribution of plays A_t to lower bound the regret (18.5.3) we learned from Bubeck and Cesa-Bianchi [91, Ch. 3.3]. The results using Assouad's method to lower bound the minimax risk of an action A_{n+1}—the pure exploration case—appear to be new, though they take inspiration from Lattimore and Szepesvári [275, Chapter 24] (which in turn builds off of the papers [376, 367]).

18.8 Exercises

Exercise 18.1 (Convexity of quadratic-over-linear functions)**:**

(a) For $x, y \in \mathbb{R}$, define $h(x, y) = \frac{x^2}{y}$. Show that h is convex over the domain $y > 0$.

(b) Let $x \in \mathbb{R}^n$ and $y > 0$. Show that $h(x, y) = \|x\|^2 / y$ is convex for any norm $\|\cdot\|$.

(c) Let X and Y be vector spaces, and let $f : X \to \mathbb{R}^n$ and $g : Y \to \mathbb{R}$ be affine functions. Show that $h(x, y) = \|f(x)\|^2 / g(y)$ is convex over the domain $\{y \in Y \mid g(y) > 0\}$.

(d) Assume f and g are, in addition to being affine, also continuous, and that $f(x) \neq 0$ for all x. Show that h is a closed convex function.

Exercise 18.2: Let $\sigma^2 > 0$ and $\tau^2 > 0$, $x_1, \ldots, x_n \in \mathbb{R}^d$, and $y_1, \ldots, y_n \in \mathbb{R}$. Let $\theta \in \mathbb{R}^d$ have density

$$\pi(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta\rangle)^2 - \frac{1}{2\tau^2}\|\theta\|_2^2\right).$$

Show that $\theta \sim \mathsf{N}(\widehat{\theta}, \Sigma)$ for

$$\Sigma = \left(\frac{1}{\tau^2} I_d + \frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i^\top\right)^{-1} \quad \text{and} \quad \widehat{\theta} = \frac{1}{\sigma^2}\Sigma\sum_{i=1}^n x_i y_i = \operatorname*{argmin}_\theta \left\{\sum_{i=1}^n \frac{(y_i - \langle x_i, \theta\rangle)^2}{2\sigma^2} + \frac{1}{2\tau^2}\|\theta\|_2^2\right\}.$$

Exercise 18.3: We describe an approach using stochastic gradient methods to solve pure-exploration linear bandit problems. Let $\theta \in \mathbb{B}_2^d$ and the action set $\mathcal{A} = \mathbb{B}_2^d$, and assume that $y(x) = \langle \theta, x\rangle + \varepsilon$ for ε an independent mean-zero variable with $\mathbb{E}[\varepsilon^2] \le \sigma^2$. We wish to find a (potentially random) action $A \in \mathcal{A}$ so that $\mathbb{E}[\langle A, \theta\rangle] \le -\|\theta\|_2 + o(1)$, that is, minimizing $\langle a, \theta\rangle$. Consider the following procedure: at iteration t, draw $x_t \sim \mathsf{Uniform}(\mathbb{S}^{d-1})$, observe $y_t = \langle \theta, x_t\rangle + \varepsilon_t$, and then set the estimated subgradient $g_t = d \cdot x_t y_t$.

(a) Show that $\mathbb{E}[g_t \mid \mathcal{H}_{t-1}] = \theta$ and $\mathbb{E}[\|g_t\|_2^2 \mid \mathcal{H}_{t-1}] \le d\|\theta\|_2^2 + d^2\sigma^2$.

(b) Let $\eta > 0$ be a stepsize (which you will determine). Define the iterates

$$\theta_t = \operatorname*{argmin}_{\theta \in \mathbb{B}_2^d}\left\{\langle g_t, \theta\rangle + \frac{1}{2\eta}\|\theta - \theta_{t-1}\|_2^2\right\}$$

and $\overline{\theta}_n = \frac{1}{n}\sum_{t=1}^n \theta_t$. Give a setting of the stepsize η so that

$$\mathbb{E}[\langle \theta, \overline{\theta}_n \rangle] \le -\left\|\theta\right\|_2 + O(1)\sqrt{\frac{d^2\sigma^2 + d}{n}}.$$

Exercise 18.4 (Exploration in linear bandits without noise [376])**:** Consider a variant of the linear bandit problem with convex action set $\mathcal{A} \subset \mathbb{R}^d$ where at time t, given action x, we observe $y(x) = \langle \theta_t, x \rangle$ for θ_t chosen i.i.d. but satisfying $\mathbb{E}[\theta_t] = \theta$. Assume the action set $\mathcal{A} \supset \alpha\{-1, 1\}^d$ for some $\alpha > 0$ and that $\mathcal{A}$ is convex.

(a) Show that the vector $g_t = \alpha^{-2}\langle \theta_t, A \rangle A$ for $A \sim \mathsf{Uniform}(\alpha\{-1, 1\}^d)$ satisfies

$$\mathbb{E}[g_t] = \theta \quad \text{and} \quad \mathbb{E}[\|g_t\|_2^2] = d \cdot \mathbb{E}[\|\theta_t\|_2^2].$$

(b) Assume that the noise in θ_t is such that $\mathbb{E}[\|\theta_t\|_2^2] \le \kappa^2$. Using the result of Exercise 18.3, show how to construct an action $\widehat{a}_n$ based on n observations such that

$$\mathbb{E}[\langle \theta, \widehat{a}_n \rangle] \le \inf_{a \in \mathcal{A}} \langle \theta, a \rangle + O(1)\kappa \operatorname{diam}(\mathcal{A})\sqrt{\frac{d}{n}}.$$

(c) Compare the result of part (b) to achievable convergence guarantees in the full information case, where we may observe θ_t, and the upper and lower bounds for cases in which $y(x) = \langle \theta, x \rangle + \varepsilon$.

Exercise 18.5: Consider a "full information" bandit problem in which the observation at time t takes the form $Y_t = \mu + \varepsilon_t \in \mathbb{R}^k$, where the noise ε_t is i.i.d., mean-zero, and has at least two moments. Take as action set $\mathcal{A} = \Delta_k = \{a \in \mathbb{R}_+^k \mid \langle a, \mathbf{1} \rangle = 1\}$, and let the losses be $\ell(a, y) = y^\top a$. Exhibit an online method achieving expected regret $O(1)\sqrt{n \log k} \cdot \mathbb{E}[\|Y_1\|_\infty^2]$.

Exercise 18.6 (A success in information-directed sampling [368])**:** This exercise shows that, when prior information is available, information-directed sampling approaches (as in Section 18.3) can achieve substantially better convergence than strategies based on confidence. Consider a non-noisy linear bandit setting in which responses and rewards $Y = \langle \theta, x \rangle$ coincide, where $\|\theta\|_0 = 1$ and $\theta \sim \mathsf{Uniform}\{e_j\}_{j=1}^d$. As action set take $\mathcal{A} = \{x / \left\|x\right\|_1 \mid x \in \{0, 1\}^d\}$, and let $\mathcal{A}^\star = \{e_j\}_{j=1}^d$ denote the collection of possible optimal actions.

(a) Show that any algorithm that chooses only actions $a \in \mathcal{A}^\star$ incurs average regret scaling at least linearly in the dimension d. (This includes the Thompson sampling and UCB procedures.)

(b) Show that information-directed sampling (Sec. 18.3.3) achieves regret scaling at most as $\log d$.

Exercise 18.7 (The number of incorrect pulls)**:** Let $\sigma \ge \delta > 0$, and let $\mathcal{P}_0$ be the family of Gaussian bandit problems with 2 arms, $a \in \{0, 1\}$, where $Y(a) \sim \mathsf{N}(\mu_a, \sigma^2)$ and $\mu_0 - \mu_1 \ge \delta$. Let $\mathcal{P}_1$ be the family of Gaussian bandits where instead $\mu_1 - \mu_0 \ge \delta$. Consider any bandit procedure pulling arms $A_t \in \{0, 1\}$ for n steps, and let $N(0) = \sum_{t=1}^n \mathbf{1}\{A_t = 0\}$ and $N(1) = \sum_{t=1}^n \mathbf{1}\{A_t = 1\}$ be the number of pulls of arms 0 and 1, respectively.

(a) Show that for a numerical constant $c > 0$,

$$\max\left\{\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[N(1)], \sup_{P_1 \in \mathcal{P}_1} \mathbb{E}_{P_1}[N(0)]\right\} \ge c \cdot \frac{\sigma^2}{\delta^2}.$$

Hint. Use the Bretagnolle-Huber inequality, Proposition 2.2.9(b).

(b) Show that the bound (18.2.4) is tight to within a logarithmic factor in that for any n,

$$\mathfrak{R}_n(\mathcal{P}_0 \cup \mathcal{P}_1) \geq c\frac{\sigma^2}{\delta}.$$

Exercise 18.8: Let $n \geq K$. Demonstrate that for UCB (Fig. 18.1), there exists a numerical constant such that executing UCB with probabilities $\delta_t = \frac{\delta}{Kt^2}$ gives regret

$$\mathsf{Reg}_n \leq C\sqrt{Kn\sigma^2 \log \frac{n}{\delta}}$$

with probability at least $1 - \delta$.

Exercise 18.9 (An empirical comparison of Bandit algorithms)**:** This question investigates three algorithms for solving the K-armed bandit problem: the Upper Confidence Bound algorithm (UCB), Thompson sampling, and exponentiated gradient. Set the rewards for the arms as follows:

i. Let $\theta_1 = \frac{1}{2}$ and $\theta_2 = \frac{1}{2} - \epsilon, \ldots, \theta_K = \frac{1}{2} - \epsilon$.

ii. When arm j is sampled, return $Y \sim \mathsf{Bernoulli}(\theta_j)$

Repeat the following experiment with the values given in each part.

(a) $K = 10$, $\epsilon = .1$, and $n = 10^6$ steps

(b) $K = 10$, $\epsilon = .02$, and $n = 10^6$ steps.

Perform Thompson sampling (Example 18.3.4) assuming prior $\theta_j \stackrel{\text{iid}}{\sim} \mathsf{Beta}(1, 1)$. Perform UCB with the confidence parameter $\delta_t = 1/\sqrt{t}$ (Algorithm 18.1) and the appropriate choice of the sub-Gaussian parameter σ^2. Perform exponentiated gradient (Algorithm 18.6) using the optimal stepsize choice η, assuming that $\sigma^2 = \frac{1}{2}$. Plot and discuss results for each of experiments (a) and (b).

Exercise 18.10 (A sub-optimal lower bound for linear bandits)**:** Consider linear bandits, where we index causal distributions P by $\theta \in \mathbb{B}_2^d$, take actions $x \in \mathcal{A} := \mathbb{B}_2^d$, and under P_θ observe reward

$$y(x) = \langle x, \theta \rangle + \varepsilon, \quad \varepsilon \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

This is a standard linear regression problem where we can design x. This exercise demonstrates a minimax lower bound using the approach in Section 18.5.1 based on the action modulus of continuity (18.5.4), but which is not quite sharp in the dimension.

(a) Show that the gap (18.5.2) satisfies $\mathsf{gap}_{P_\theta}(x) = \|\theta\|_2 - \langle x, \theta \rangle$.

(b) Show that for any distribution q on actions x and $\overline{P}(x) = \mathsf{N}(0, \sigma^2)$ and $\delta \geq 0$, there exist linear bandit models P_1 and P_2 indexed by orthogonal unit vectors $\theta = \delta u$ and $\theta = \delta v$ such that

$$\mathbb{E}_q[D_{\rm kl}\left(\overline{P}(X) \| P_1(X)\right)] \leq \frac{\delta^2}{2d\sigma^2} \quad \text{and} \quad \mathbb{E}_q[D_{\rm kl}\left(\overline{P}(X) \| P_2(X)\right)] \leq \frac{\delta^2}{2(d-1)\sigma^2}.$$

(c) Show in the same setting that $\max\{\mathsf{gap}_{P_1}(x), \mathsf{gap}_{P_2}(x)\} \geq \frac{2-\sqrt{2}}{2}\delta$.

(d) Give the minimax lower bound $\mathfrak{M}_n(\mathcal{P}) \gtrsim \frac{\sigma\sqrt{d}}{\sqrt{n}}$.

Chapter 19

Minimax games and Bayesian estimation

This final chapter connects online learning, probabilistic prediction, and Bayesian statistics via information theoretic analyses. Consider the following probabilistic game between nature, who chooses a sampling distribution over a (temporarily finite) set $\mathcal{X}$, and a player who wishes to model the sampling distribution. The player chooses a distribution Q on $\mathcal{X}$ with probability mass function q, and then nature chooses $P \in \mathcal{P}$, where $\mathcal{P}$ is a collection of distributions on $\mathcal{X}$. For $X \sim P$, the player suffers expected log loss $\mathbb{E}_P[-\log q(X)] = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}$. Playing optimally against the worst-possible choice of nature then entails solving the minimax problem

$$\operatorname*{minimize}_Q \sup_{P \in \mathcal{P}} \mathbb{E}_P[-\log q(X)]. \tag{19.0.1}$$

Instead of using the objective (19.0.1), we frequently relativize losses to $\mathcal{P}$ to compete against the best $P \in \mathcal{P}$, seeking distributions Q with little *regret* relative to the losses we would suffer if we had played the optimal P. Then we consider the objective

$$\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P[-\log q(X)] - \inf_{Q \in \mathcal{P}} \mathbb{E}_P[-\log q(X)] \right\} = \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\log \frac{p(X)}{q(X)} \right] = \sup_{P \in \mathcal{P}} D_{\mathrm{kl}}\left(P \| Q\right).$$

More generally, we can revisit Chapter 14, where for a proper loss ℓ, we suffer $\ell(Q, x)$ for predicting distribution Q on realization x; the problem (19.0.1) corresponds to the logarithmic loss. We wish to minimize the worst-case expected regret of playing Q instead of the optimal P, that is, to solve

$$\operatorname*{minimize}_Q \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)] \right\}. \tag{19.0.2}$$

We can modify the game (19.0.2) to let the minimizer play last. Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ be parameterized by θ (this is no loss of generality). Now nature chooses a prior distribution π on Θ, and with knowledge of π we choose the predictive distribution Q. This leads to the *Bayesian regret*,

$$\int_\Theta \left(\mathbb{E}_{P_\theta}[\ell(Q, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] \right) \pi(\theta) d\theta.$$

The optimal play for Q is immediate. Indeed, let P_π denote the marginal distribution over X obtained by drawing $\theta \sim \pi$ and then $X \sim P_\theta$; then, when ℓ is proper, P_π minimizes the Bayesian regret. Nature may choose a prior to make this regret as large as possible, but the Q player always

plays with knowledge of nature's choice. This leads to the main question of this chapter: when are the (worst case) Bayesian regret and the game (19.0.2) where Q plays first equivalent?

We motivate problems (19.0.1) and (19.0.2) in two ways: first, in Section 19.1, as a form of robust Bayesian procedure, which leads to classical maximum entropy estimators. In Section 19.2, we connect coding problems to the games (19.0.1) and (19.0.2) to overview results on universal and sequential prediction. We then present a fundamental duality result: for many classes $\mathcal{P}$ of distributions, it does not matter which player goes first in problem (19.0.2), and similarly in problem (19.0.1): there are robust choices of Q that nature—the P player—can, essentially, gain no advantage against. The remainder of the chapter provides Bayesian statistical procedures, which asymptotically find optimal Q distributions for the log loss $-\mathbb{E}_P[\log q(X_1^n)]$ when $X_i \stackrel{\text{iid}}{\sim} P$.

19.1 Robust Bayesian procedures and maximum entropy

Consider Bayesian statistics, where a data analyst has a prior π on the states of nature and wishes to infer the "true" state of nature using this prior. Letting π be a density on Θ for a collection $\{P_\theta\}_{\theta\in\Theta}$, the Bayesian chooses Q to maximize the log-likelihood of observed data on average over the possible states θ of nature, that is, to minimize $\int \mathbb{E}_{P_\theta}[-\log q(X)]\pi(\theta)d\theta$. A "robust" Bayesian chooses the worst prior π on Θ from a collection Π of priors on Θ. Recalling the notation $P_\pi(\cdot) = \int P_\theta(\cdot)\pi(\theta)d\theta$, we thus think of problem (19.0.1) as seeking a *robust Bayes* act against $\{P_\pi\}_{\pi\in\Pi}$: we wish to solve

$$\operatorname*{minimize}_Q \sup_{\pi\in\Pi} \mathbb{E}_{P_\pi}[-\log q(X)]. \tag{19.1.1}$$

Then Q is robust Bayes for the priors Π if it minimizes the supremum risk (19.1.1), achieving the best possible logarithmic loss measured in a uniform sense against $\mathcal{P} = \{P_\pi\}_{\pi\in\Pi}$.

Consider temporarily a discrete set $\mathcal{X}$, and let $H(P) = -\sum_x p(x)\log p(x)$ be the usual (Shannon) entropy. Because the KL-divergence is always nonnegative, for any distribution P,

$$\mathbb{E}_P[-\log q(X)] = \mathbb{E}_P[-\log p(X)] + D_{\rm kl}\left(P\|Q\right) \geq H(P),$$

and equality holds if and only if $Q = P$. If we switch the order of play in the game (19.1.1), forcing nature to play first, then

$$\sup_{P\in\mathcal{P}} \inf_Q \mathbb{E}_P[-\log q(X)] = \sup_{P\in\mathcal{P}} H(P),$$

the maximum entropy in $\mathcal{P}$. Then our introductory question—does it matter whether the Q player (statistician) or P player (nature) plays first in the game (19.1.1)?—becomes one of duality: when does

$$\sup_{P\in\mathcal{P}} \inf_Q \mathbb{E}_P[-\log q(X)] = \inf_Q \sup_{P\in\mathcal{P}} \mathbb{E}_P[-\log q(X)]? \tag{19.1.2}$$

A digression on min-max games We briefly discuss min-max duality for context. For $L : U \times V \to \mathbb{R}$, where U and V are arbitrary, the min-max game considers solving

$$\operatorname*{minimize}_{u\in U} \sup_{v\in V} L(u,v) \quad \text{and} \quad \operatorname*{maximize}_{v\in V} \inf_{u\in U} L(u,v), \tag{19.1.3}$$

where we think of a u-player and a v-player choosing u (respectively v), after which the other player chooses a best response. The game has a value if $\inf_u \sup_v L(u,v) = \sup_v \inf_u L(u,v)$. A pair $(u^\star, v^\star) \in U \times V$ is a *saddle point* for the game if

$$\sup_{v\in V} L(u^\star, v) \leq L(u^\star, v^\star) \leq \inf_{u\in U} L(u, v^\star).$$

For such a point, we have

$$\inf_{u\in U}\sup_{v\in V} L(u,v) \le \sup_{v\in V} L(u^\star,v) \le L(u^\star,v^\star) \le \inf_{u\in U} L(u,v^\star) \le \sup_{v\in V}\inf_{u\in U} L(u,v),$$

and a quick calculation shows each inequality is thus necessarily an equality. We capture this as a proposition to highlight its importance, as our later results will rely on exhibiting saddle points:

Proposition 19.1.1. *If a point $(u^\star, v^\star) \in U \times V$ is a saddle point for the game (19.1.3), then the game has value $L(u^\star, v^\star)$, independent of the order of play, and*

$$L(u^\star,v^\star) = \inf_{u\in U}\sup_{v\in V} L(u,v) = \sup_{v\in V}\inf_{u\in U} L(u,v).$$

Moreover, $u^\star$ minimizes $\sup_{v\in V} L(u,v)$ over $u \in U$, and $v^\star$ maximizes $\inf_{u\in U} L(u,v)$ over $v \in V$.

Classical results in convex optimization and game theory provide sufficient conditions for such saddles to exist (see also the min-max Exercises 17.3 and 17.4 in Chapter 17). While min-max theory is important, the particular structure of the regret minimization game (19.0.2) allows a more direct "information-theoretic" proof of duality, with the advantage that it applies to finite- and infinite-dimensional problems.

19.1.1 Saddle points for maximum entropy

Returning to the log-loss (robust Bayes) game (19.0.1) and letting q denote a density or p.m.f. of Q, under appropriate conditions we expect to have a saddle point $(Q^\star, P^\star)$ with

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[-\log q^\star(X)] \le \mathbb{E}_{P^\star}[-\log q^\star(X)] \le \inf_{Q} \mathbb{E}_{P^\star}[-\log q(X)],$$

which implies the equality (19.1.2) by Proposition 19.1.1. A saddle point $(Q^\star, P^\star)$ necessarily achieves the maximum entropy $H(P^\star) = \sup_{P\in\mathcal{P}} H(P)$, which is thus the value for the game (19.0.1). The saddle criterion means $q^\star$ uniquely minimizes $\mathbb{E}_{P^\star}[-\log q(X)] = H(P^\star) + D_{\rm kl}\left(P^\star \| Q\right)$, so $Q^\star$ maximizes entropy over $\mathcal{P}$ as well, and maximum entropy becomes robust Bayes against $\mathcal{P}$: it solves the worst-case problem (19.0.1). In summary, for the log-loss game, any saddle point takes the form

$$(P^\star, P^\star),$$

and the distribution $P^\star$ maximizes the entropy.

Section 19.6* shows that saddle points typically exist for the general expected regret game (19.0.2) when the losses ℓ are strictly proper, connecting the development of proper losses in Chapter 14 to the game formulations (19.0.1) and (19.0.2). As corollaries, we develop the existence of saddle points in the log-loss game and more general scenarios. First, however, we show that exponential family models provide a clean treatment of the log-loss game.

19.1.2 Exponential family models as robust Bayesian procedures

The minimax game (19.0.1) naturally gives rise to the exponential family models from Chapter 3, which become robust Bayes procedures against certain families $\mathcal{P}$ of distributions. To develop this, let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a statistic, and for $\alpha \in \mathbb{R}^d$ consider the mean-value constrained set

$$\mathcal{P} = \mathcal{P}_\alpha^{\rm lin} := \left\{\text{distributions } P \text{ on } \mathcal{X} \text{ s.t. } \mathbb{E}_P[\phi(X)] = \alpha\right\}.$$

Continuing to consider the case that $\mathcal{X}$ is discrete, let P_θ denote the exponential family distribution with p.m.f. $p_\theta(x) = h(x)\exp(\langle\theta, \phi(x)\rangle - A(\theta))$, where h is the carrier. We have the following.

Proposition 19.1.2. *Let the conditions above hold. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then*

$$\inf_Q \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log q(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] = \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \inf_Q \mathbb{E}_P[-\log q(X)].$$

Proof First, note that

$$\begin{aligned}\sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\log p_\theta(X)] &= \sup_{P \in \mathcal{P}_\alpha^{\rm lin}} \mathbb{E}_P[-\langle \phi(X), \theta\rangle + A(\theta)] \\ &= -\langle \alpha, \theta\rangle + A(\theta) = \mathbb{E}_{P_\theta}[-\langle \theta, \phi(X)\rangle + A(\theta)] = H(P_\theta),\end{aligned}$$

where H denotes the Shannon entropy, for any distribution $P \in \mathcal{P}_\alpha^{\rm lin}$. For $Q \neq P_\theta$,

$$\sup_P \mathbb{E}_P[-\log q(X)] \geq \mathbb{E}_{P_\theta}[-\log q(X)] > \mathbb{E}_{P_\theta}[-\log p_\theta(X)] = H(P_\theta),$$

where the inequality follows because $D_{\rm kl}\left(P_\theta \| Q\right) > 0$. This shows the first equality in the proposition. For the second equality, note that

$$\inf_Q \mathbb{E}_P[-\log q(X)] = \inf_Q \mathbb{E}_P\Big[\log \frac{p(X)}{q(X)}\Big] - \mathbb{E}_P[\log p(X)] = H(P).$$

Standard maximum entropy results (Theorem 14.4.7) show P_θ maximizes the entropy over $\mathcal{P}_\alpha^{\rm lin}$. □

The same result immediately extends beyond discrete sets $\mathcal{X}$, which we state here without proof, as the argument parallels the proof of Proposition 19.1.2. Let ν be a measure on $\mathcal{X}$, and let

$$\mathcal{P}_\alpha^{\rm lin} := \left\{\text{distributions } P \ll \nu \text{ on } \mathcal{X} \text{ s.t. } \mathbb{E}_P[\phi(X)] = \alpha\right\}.$$

Then the following generalization of Proposition 19.1.2 (see Exercise 19.9) follows.

Theorem 19.1.3. *Let $\mathcal{E} = \{P_\theta\}_{\theta \in \mathbb{R}^d}$ be the exponential family with densities $p_\theta(x) = \exp(\langle \theta, \phi(x)\rangle - A(\theta))$ relative to ν. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then P_θ is uniquely robust Bayes against $\mathcal{P}_\alpha^{\rm lin}$.*

In short: maximum entropy is equivalent to robust prediction procedures for linear families of distributions $\mathcal{P}_\alpha^{\rm lin}$. In turn, this is equivalent to maximum likelihood estimation in exponential families, which corresponds to moment-matching (3.2.5).

19.2 The coding game and sequential prediction

An additional motivation for the log-loss game (19.0.1) comes from source coding. We assume we receive n random variables $X_i \stackrel{\rm iid}{\sim} P$, where P is unknown, and suffer the sequential prediction loss

$$\mathbb{E}_P[-\log q(X_1^n)] = \sum_{i=1}^n \mathbb{E}_P\left[\log \frac{1}{q(X_i \mid X_1^{i-1})}\right],$$

which corresponds to predicting X_i given X_1^{i-1} as well as possible. The connection between instantaneously decodable codes and probability distributions underlying classical interpretations of entropy (Section 2.4) provides a concrete grounding for the minimax game (19.0.1).

Example 19.2.1 (The coding game)**:** Let the set $\mathcal{X}$ be finite or countable, and consider the problem of encoding $\mathcal{X}$ into $\{0,1\}$-valued sequences using as few bits as possible. The Kraft-McMillan inequality (recall Theorem 2.4.2) states that if $C : \mathcal{X} \to \{0,1\}^*$ is a uniquely decodable code, and $\ell_C(x)$ denotes the bit-length of the encoding for the symbol $x \in \mathcal{X}$, then

$$\sum_x 2^{-\ell_C(x)} \le 1.$$

Extending C to $\mathcal{X}^*$ by concatenation, $C(x_1^n) = C(x_1) \cdots C(x_n)$, so $\ell_C(x_1^n) = \sum_{i=1}^n \ell_C(x_i)$, define the distribution $q_C(x) = 2^{-\ell_C(x)} / \sum_x 2^{-\ell_C(x)}$ on $x \in \mathcal{X}^*$. Then for any sequence x_1^n,

$$-\log_2 q_C(x_1^n) = \sum_{i=1}^n \left[\ell_C(x_i) + \log_2 \sum_{x \in \mathcal{X}} 2^{-\ell_C(x)} \right] \le \sum_{i=1}^n \ell_C(x_i).$$

Conversely, given a p.m.f. q on $\mathcal{X}^n$, the length function $\ell(x_1^n) := \lceil -\log_2 q(x_1^n) \rceil$ satisfies $\sum_{x_1^n} 2^{-\ell(x_1^n)} \le 1$, so there is a prefix code $C : \mathcal{X}^n \to \{0,1\}^*$ with length $\ell_C(x_1^n) = \lceil -\log_2 q(x_1^n) \rceil$ by the converse of the Kraft inequality in Theorem 2.4.2. Thus for $X_1^n \sim P$,

$$\mathbb{E}_P \left[\ell_C(X_1^n) \right] = \mathbb{E}_P \left[\lceil -\log_2 q(X_1^n) \rceil \right] \le 1 + \mathbb{E}_P[-\log_2 q(X_1^n)].$$

The minimax game (19.0.1) corresponds to a coding game where we seek a sequential coding scheme that has as small an expected length as possible, uniformly over distributions P. $\diamond$

When we have sequences $x_1, \ldots, x_n$, the minimax log-loss game (19.0.1) thus is equivalent to a sequential setting in which we attempt to encode the sequence x_1^n online. Rather than measuring performance by the losses $-\log q(x)$, however, we measure against reference distributions P, and ask whether we can predict the sequence (nearly) as well as if we knew P. In more general settings, we would like to predict the data as well as all predictive distributions P from some family of distributions $\mathcal{P}$, even if *a priori* we know little about the coming sequence of data.

This leads to an online setting relating to the online convex optimization scenarios we investigated in Chapter 17. We consider two versions of the sequential prediction game: adversarial and probabilistic. For both of the following definitions of sequential prediction games, we assume that p and q are densities or probability mass functions in the case that $\mathcal{X}$ is continuous or discrete (this is no real loss of generality) for distributions P and Q.

Adversarial regret Given a sequence $x_1^n \in \mathcal{X}^n$, the *regret* of the distribution Q for the sequence x_1^n with respect to the distribution P is

$$\mathsf{Reg}(Q, P, x_1^n) := \log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} = \sum_{i=1}^n \left[\log \frac{1}{q(x_i \mid x_1^{i-1})} - \log \frac{1}{p(x_i \mid x_1^{i-1})} \right], \tag{19.2.1}$$

where the sum over $q(x_i \mid x_1^{i-1})$ emphasizes the sequential nature of the game. In the context of the coding game in Example 19.2.1, the regret (19.2.1) corresponds to the realized code length (in nats) necessary to encode x_1^n over a putative encoding through p. Associated with the regret of the sequence x_1^n is the *adversarial regret* of Q with respect to the family $\mathcal{P}$ of distributions,

$$\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) := \sup_{P \in \mathcal{P}, x_1^n \in \mathcal{X}^n} \mathsf{Reg}(Q, P, x_1^n). \tag{19.2.2}$$

Redundancy A less adversarial problem is to minimize the *expected regret* with respect to a distribution P, which for the log loss receives the special name *redundancy*. We thus define

$$\mathsf{Red}_n(Q, P) := \mathbb{E}_P \left[\log \frac{1}{q(X_1^n)} - \log \frac{1}{p(X_1^n)} \right] = D_{\rm kl} \left(P \| Q \right), \tag{19.2.3}$$

where the dependence on n is implicit in the KL-divergence. The name follows from the connections to coding games in Example 19.2.1: giving the encoding schemes C_p and $C_q : \mathcal{X} \to \{0,1\}^*$ p.m.f.s $p(x) = 2^{-\ell_{C_p}(x)}$ and $q(x) = 2^{-\ell_{C_q}(x)}$ through the Kraft-McMillan correspondence,

$$\frac{1}{\log 2} \mathsf{Red}_n(Q, P) = \sum_{i=1}^n \mathbb{E}_P \left[\ell_{C_q}(X_i) - \ell_{C_p}(X_i) \right].$$

Thus redundancy measures the additional expected length of the implicit encoding by Q relative to P, and as in Section 2.4.1, the length $\ell(x) = \lceil -\log_2 p(x) \rceil$ is essentially optimal.

19.3 Expected regret, information capacity, and redundancy

The elegant reformulation of the log-loss in terms of the KL-divergence is a special case of more general results connecting divergences and the proper losses of Chapter 14. Thus, we consider the expected regret for a proper loss ℓ as in problem (19.0.2), where $\ell(P, x)$ indicates the loss suffered by playing P on the realized point x. Then we define the *expected regret*

$$\mathsf{EReg}(Q, P) = \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)],$$

where the expectation is taken over $X \sim P$, and which is always nonnegative because ℓ is proper.

Though for simplicity we focus on the case $n = 1$ (equivalently, that we measure losses $\ell(Q, x_1^n)$ against the vector x_1^n), we can make an essentially complete analogy with the redundancy (19.2.3). Recall the Gneiting-Raftery-Savage representation of proper losses from Section 14.2. Letting $\mathbf{1}_x$ denote either a standard basis vector or point mass on x as appropriate, so $\mathbb{E}_P[\mathbf{1}_X] = P$, Corollaries 14.2.5 and 14.2.9 show ℓ is (strictly) proper if and only if there exists a (strictly) convex Ω with selection $\Omega'(Q) \in \partial\Omega(Q)$ such that

$$\ell(Q, x) = -\Omega(Q) - \langle \Omega'(Q), \mathbf{1}_x - Q \rangle$$

(we abuse notation slightly to write inner products). For this negative entropy Ω,

$$D_\Omega(P, Q) = \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)],$$

where we recall the Bregman divergence $D_\Omega(P, Q) = \Omega(P) - \Omega(Q) - \langle \Omega'(Q), P - Q \rangle$. The expected regret

$$\mathsf{EReg}(Q, P) = D_\Omega(P, Q) \tag{19.3.1}$$

is thus the divergence between P and Q, as in the case (19.2.3) of the log loss with the KL-divergence, and the representation (19.3.1) via divergence will be the key to the dualities we show.

With this setup, consider the min-max game formulation. Without loss of generality, we identify $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ for a space Θ, and letting $\Pi(\Theta)$ be the collection of probability distributions on Θ,

$$\sup_{P \in \mathcal{P}} \mathsf{EReg}(Q, P) = \sup_{\pi \in \Pi(\Theta)} \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta),$$

where the rightmost quantity is linear in π. Then we can ask for a distribution Q minimizing the worst-case expected regret, and we can provide conditions under which the duality

$$\inf_Q \sup_\pi \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta) = \sup_\pi \inf_Q \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta)$$

obtains, that is, it is immaterial whether we choose Q first and nature chooses π or vice versa.

19.3.1 Information capacity and regret duality

Recall Section 14.1.3, where we defined the *information in an experiment* as the generalized entropy reduction that observing a variable X provides for Y. Adapting our notation here, assume the Markov chain $T \to X$, where we draw $T \sim \pi$ for a prior distribution on Θ, and conditional on $T = \theta$, we draw $X \sim P_\theta$. Then $I_\ell(T; X) = H_\ell(X) - H_\ell(X \mid T)$ denotes the information that T carries about X, that is, that knowing $T = \theta$ allows improved prediction of X. To highlight the importance of the prior distribution π on T, we incorporate it into our notation and define the information T carries about X according to the loss ℓ via

$$\begin{aligned} I_\ell(\pi; X) &:= \inf_Q \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta) \\ &= \inf_Q \int \mathbb{E}_{P_\theta}[\ell(Q, X)] d\pi(\theta) - \int \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] d\pi(\theta) \\ &= H_\ell(X) - H_\ell(X \mid T), \end{aligned} \tag{19.3.2}$$

where the final equality uses the generalized entropy (14.1.4) and tacitly draws $T \sim \pi$.

Because the loss ℓ is proper, the marginal distribution

$$P_\pi(A) := \int P_\theta(A) d\pi(\theta) \tag{19.3.3}$$

attains the infimum over Q in the definition (19.3.2), that is, P_π is the marginal distribution on X after drawing $T \sim \pi$ and then $X \sim P_T$. To see this, simply note that $\int \mathbb{E}_{P_\theta}[\ell(Q, X)] d\pi(\theta) = \mathbb{E}_{P_\pi}[\ell(Q, X)]$, which P_π minimizes by propriety. Recalling the representation of expected regret as the divergence (19.3.1) for the (generalized) negative entropy $\Omega(P) = -\inf_Q \mathbb{E}_P[\ell(Q, X)]$, we thus may represent the information as the average Bregman divergence

$$I_\ell(\pi; X) = \inf_Q \int D_\Omega(P_\theta, Q) d\pi(\theta) = \int D_\Omega(P_\theta, P_\pi) d\pi(\theta).$$

This should be familiar, as it generalizes the result (9.4.4) for the Shannon mutual information (corresponding to the log loss) that $I(T; X) = \int D_{\rm kl}\left(P_\theta \| P_\pi\right) d\pi(\theta)$.

The notation (19.3.2) makes the dependence on the prior π clear, so for a collection $\Pi \subset \Pi(\Theta)$ of probability distributions on Θ, we define the *capacity* of the channel $T \to X$ by

$$C_\ell(\Pi) := \sup_{\pi \in \Pi} I_\ell(\pi; X). \tag{19.3.4}$$

This measures the most information X can contain about T when we know that T is drawn from the prior $\pi \in \Pi$. To rephrase our questions of duality as a saddle point problem, we define the expected loss gap (regret) with respect to the prior π,

$$L(Q, \pi) := \int \mathbb{E}_{P_\theta}[\ell(Q, X)] d\pi(\theta) - \int \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] d\pi(\theta) = \int D_\Omega(P_\theta, Q) d\pi(\theta). \tag{19.3.5}$$

Then the worst-case expected regret (redundancy) of a distribution Q on $\mathcal{X}$ is

$$\sup_{\pi \in \Pi} L(Q, \pi) = \sup_{\pi \in \Pi} \int D_\Omega(P_\theta, Q) d\pi(\theta),$$

while the capacity (19.3.4) of the family $\{P_\theta\}_{\theta \in \Theta}$ is

$$C_\ell(\Pi) := \sup_{\pi \in \Pi} I_\ell(\pi; X) = \sup_{\pi \in \Pi} \inf_Q L(Q, \pi) = \sup_{\pi \in \Pi} L(P_\pi, \pi),$$

for the marginal distribution P_π (Eq. (19.3.3)), so

$$\inf_Q L(Q, \pi) = L(P_\pi, \pi). \tag{19.3.6}$$

We can therefore rephrase our questions of duality as the following *regret/capacity duality* question: under what circumstances do we have min-max equality for the saddle objective (19.3.5),

$$\inf_Q \sup_{\pi \in \Pi} L(Q, \pi) = \sup_{\pi \in \Pi} \inf_Q L(Q, \pi) = \sup_{\pi \in \Pi} I_\ell(\pi; X)? \tag{19.3.7}$$

Because $\Pi \subset \Pi(\Theta)$ may be infinite-dimensional, this equality does not follow so immediately from classical minimax theorems. Nonetheless, the duality (19.3.7) does indeed typically hold. For the logarithmic loss, the regret/capacity duality equality becomes the *redundancy/capacity* duality that

$$\sup_{\pi \text{ on } T} I(T; X) = \inf_Q \sup_\pi \int D_{\rm kl}\left(P_\theta \| Q\right) d\pi(\theta),$$

which captures that abstractly—assuming the equality holds—in developing a robust encoding as in the redundancy (19.2.3), it does not matter whether nature or the encoder plays first.

Stating things informally for now, the following theorem captures the prototypical regret/capacity duality:

Theorem 19.3.1 (Informal regret/capacity duality)**.** *Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ be a collection of distributions on $\mathcal{X}$, Π a convex collection of priors on Θ, and assume the loss ℓ is "proper enough" and the capacity (19.3.4) is finite. Then there exists a unique distribution $Q^\star$ on $\mathcal{X}$ such that*

$$\sup_{\pi \in \Pi} \int \mathsf{EReg}(Q^\star, P_\theta) d\pi(\theta) = C_\ell(\Pi).$$

In particular, the duality (19.3.7) holds, and

$$\inf_Q \sup_{\pi \in \Pi} \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta) = C_\ell(\Pi) = \sup_{\pi \in \Pi} \inf_Q \int \mathsf{EReg}(Q, P_\theta) d\pi(\theta).$$

While Theorem 19.3.1 is informal, stating merely that the loss ℓ is "proper enough," we provide several formalizations of the theorem in Section 19.6*, presenting a full version in Theorem 19.6.6 in Section 19.6.5. In brief, "proper enough" means that the loss ℓ is strictly proper plus a bit more, so that convergence of distributions in the Bregman divergence representation of expected regret (19.3.1), guarantees convergence in an appropriate topology on distributions. The next subsection instantiates several versions of the theorem, and Sections 19.4 and 19.5 exhibit procedures for achieving asymptotically optimal regret, especially in sequential settings.

19.3.2 Instantiations and corollaries of regret/capacity duality

For the log-loss, the redundancy (19.2.3) matches the regret as $\mathsf{EReg}(Q, P) = D_{\rm kl}(P\|Q)$. Then the information I_ℓ is simply the mutual information, and we have the classical capacity

$$C(\Pi) := \sup_\pi \{I(X;T) \mid \pi \text{ on } T \in \Theta, \pi \in \Pi\},$$

where the supremum is over a convex collection Π of distributions on T, and conditional on $T = \theta$ we draw $X \sim P_\theta$ as usual. We have the following *redundancy/capacity* duality result, one of the foundational results in information theory.

Corollary 19.3.2 (Redundancy/capacity duality)**.** *Let $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be distributions on a set $\mathcal{X}$, and assume that the capacity $C(\Pi) < \infty$. Then there exists a distribution $Q^\star$ on $\mathcal{X}$ such that*

$$\sup_{\pi\in\Pi} \int D_{\rm kl}(P_\theta\|Q^\star)\, d\pi(\theta) = C(\Pi).$$

If $\Pi = \Pi(\Theta)$ consists of all distributions on Θ, then

$$\sup_{\theta\in\Theta} D_{\rm kl}(P_\theta\|Q^\star) = C(\Pi(\Theta)).$$

Lastly, if $\pi^\star$ achieves the supremum in the capacity, then $Q^\star = P_{\pi^\star}$ is the marginal over X when $T \sim \pi^\star$, and we have the saddle point

$$\sup_{\pi\in\Pi} \int D_{\rm kl}(P_\theta\|P_{\pi^\star})\, d\pi(\theta) \le \int D_{\rm kl}(P_\theta\|P_{\pi^\star})\, d\pi^\star(\theta) = I(X;T) \le \inf_Q \int D_{\rm kl}(P_\theta\|Q)\, d\pi^\star(\theta).$$

When $\mathcal{X}$ is finite, we more or less have duality without conditions beyond ℓ being strictly proper and lower semicontinuous, that is, that $Q \mapsto \ell(Q, x)$ is lower semicontinuous for each x.

Corollary 19.3.3 (Regret/capacity duality for finite domains)**.** *Let $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be distributions on a finite set $\mathcal{X}$, let $\Pi \subset \Pi(\Theta)$ be a convex collection of distributions on Θ, and assume the capacity (19.3.4) is finite: $C_\ell(\Pi) < \infty$. If ℓ is strictly proper and lower semicontinuous, then there exists a unique distribution $Q^\star$ on $\mathcal{X}$ such that*

$$\sup_{\pi\in\Pi} \int \left(\mathbb{E}_{P_\theta}[\ell(Q^\star, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) d\pi(\theta) = C_\ell(\Pi).$$

If additionally $\Pi = \Pi(\Theta)$ consists of all distributions on Θ, then

$$\sup_{\theta\in\Theta} \mathbb{E}_{P_\theta}[\ell(Q^\star, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] = C_\ell(\Pi(\Theta)).$$

Lastly, if $\pi^\star$ achieves the supremum in the capacity, then $Q^\star = P_{\pi^\star}$, and $(P_{\pi^\star}, \pi^\star)$ is a saddle point.

We state one more corollary of Theorem 19.6.6 to come, which shows an application of the results to the continuous ranked probability scores we discuss in Examples 14.2.6 and 14.2.10. Recall that in this case, for a CDF F, we have loss $\ell_{\rm crps}(F, y) = \int (F(t) - \mathbf{1}\{y \le t\})^2 dt$.

Corollary 19.3.4 (Regret/capacity duality for ranked scores)**.** *Let $\mathcal{P} = \{F_\theta\}_{\theta\in\Theta}$ be any collection of cumulative distributions on a compact interval $[t_0, t_1]$ and let Π be a convex collection of distributions on Θ. Then there exists a unique cumulative distribution $F^\star$ such that*

$$\sup_{\pi\in\Pi} \int \left(\mathbb{E}_{F_\theta}[\ell_{\rm crps}(F^\star, X)] - \mathbb{E}_{F_\theta}[\ell_{\rm crps}(F_\theta, X)]\right) d\pi(\theta) = C_{\ell_{\rm crps}}(\Pi).$$

If $\pi^\star$ achieves the capacity $C_{\ell_{\rm crps}}(\Pi)$, then $F^\star(t) = \int F_\theta(t) d\pi^\star(\theta)$ is the marginal CDF.

19.3.3 Maximum generalized entropy and Robust Bayesian procedures

The instantiations in Corollaries 19.3.2, 19.3.3, and 19.3.4 show the duality between expected regret (or redundancy) and capacity. We can specialize these results to show that there are robust Bayesian procedures for many of the loss-minimization games motivating Section 19.1. To demonstrate these dualities, we consider the identity channel $X \to X$. We normalize the losses so that

$$\inf_P \ell(P, x) = \ell(\mathbf{1}_x, x) = 0 \;\; \text{for all } x \in \mathcal{X}, \tag{19.3.8}$$

where $\mathbf{1}_x$ denotes the point mass on x. In this case, the (expected) regret is

$$\mathbb{E}_P[\ell(Q, X)] \geq 0,$$

giving worst case game

$$\operatorname*{minimize}_Q \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(Q, X)]. \tag{19.3.9}$$

Example 19.3.5: If $\mathcal{X}$ is discrete, then the log-loss $\ell(q, x) = -\log q(x)$ satisfies the condition (19.3.8), because $q = \mathbf{1}_x$ satisfies $\ell(q, x) = -\log 1 = 0$. $\diamond$

Example 19.3.6: If $\mathcal{X} \subset \mathbb{R}$, the continuous ranked probability score (CRPS) $\ell_{\text{crps}}(F, x) = \int (F(t) - \mathbf{1}\{x \leq t\})^2 dt$ satisfies the condition (19.3.8), where we take $F(t) = \mathbf{1}\{x \leq t\}$. $\diamond$

Let $\mathcal{P}$ be a convex collection of distributions on $\mathcal{X}$. Then for the generalized entropy (14.1.5),

$$H_\ell(P) = \inf_Q \mathbb{E}_P[\ell(Q, X)] = \mathbb{E}_P[\ell(P, X)],$$

we say that the robust game (19.3.9) has a solution if there exists $Q^\star$ such that

$$\sup_{P \in \mathcal{P}} H_\ell(P) = \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(Q^\star, X)],$$

and a saddle point $(Q^\star, P^\star)$ if

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(Q^\star, X)] \leq \mathbb{E}_{P^\star}[\ell(Q^\star, X)] \leq \inf_Q \mathbb{E}_{P^\star}[\ell(Q, X)].$$

If a saddle point exists, then whenever ℓ is strictly proper, it must be the case that $P^\star = Q^\star$, and moreover, $P^\star$ must maximize the generalized entropy $H_\ell(P)$ over $\mathcal{P}$.

We can therefore provide corollaries of Theorem 19.3.1 (more accurately, Theorem 19.6.6).

Corollary 19.3.7. *Let $\mathcal{X}$ be finite or countable, let the loss $\ell(q, x) = -\log q(x)$ be the log-loss, and let $\mathcal{P}$ be a convex collection of distributions. For $H(P) = -\sum_x p(x) \log p(x)$, assume that the maximum entropy $\sup_{P \in \mathcal{P}} H(P) < \infty$. Then*

$$\sup_{P \in \mathcal{P}} H(P) = \inf_Q \sup_{P \in \mathcal{P}} \mathbb{E}_P[-\log q(X)] < \infty,$$

and there exists a unique minimizer $Q^\star$ of $\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(Q, X)]$. If $P^\star$ attains the maximum entropy, then $Q^\star = P^\star$ and $(P^\star, P^\star)$ is a saddle point for the game (19.3.9).

So maximum entropy distributions are robust Bayes.

To avoid annoying technicalities, we do not state the most general version of the results here. We can, however, give a few more special cases, leaving generalizations to the exercises. Let $\mathcal{X}$ be finite. To give a more explicitly Bayesian flavor to the results, assume that we parameterize distributions on $\mathcal{X}$ by some $\theta \in \Theta$, so that for any collection $\{P_\theta\}_{\theta\in\Theta}$, we have

$$\mathrm{Conv}\{P_\theta\}_{\theta\in\Theta} = \{P_\pi\}_{\pi\in\Pi(\Theta)},$$

where $\Pi(\Theta)$ is the collection of probability distributions on Θ and P_π denotes the marginal over $T \sim \pi$ and $X \sim P_\theta$ conditional on $T = \theta$. (See Corollary B.1.24 in Appendix B.1.3.) Then for a π, propriety means that P_π minimizes $\int \mathbb{E}_{P_\theta}[\ell(Q, X)]d\pi(\theta) = \mathbb{E}_{P_\pi}[\ell(Q, X)]$. We call $\pi^\star$ a worst-case prior if it maximizes

$$\inf_Q \int \mathbb{E}_{P_\theta}[\ell(Q, X)]d\pi(\theta).$$

Corollary 19.3.8. *Let $\mathcal{X}$ be finite and ℓ be strictly proper. Let $\{P_\theta\}_{\theta\in\Theta}$ be any collection of distributions on $\mathcal{X}$. Assume the maximum generalized entropy $\sup_{\pi\in\Pi(\Theta)} H_\ell(P_\pi) < \infty$. Then there exists a unique minimizer $Q^\star$ of $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\ell(Q, X)]$, and*

$$\sup_{\pi\in\Pi(\Theta)} H_\ell(P_\pi) = \sup_{\pi\in\Pi(\Theta)} \mathbb{E}_{P_\pi}[\ell(Q^\star, X)] = \inf_Q \sup_{\pi\in\Pi(\Theta)} \int \mathbb{E}_{P_\theta}[\ell(Q, X)]d\pi(\theta) < \infty.$$

If the convex hull $\{P_\pi\}_{\pi\in\Pi(\Theta)}$ is closed, then there is a prior $\pi^\star$ maximizing the generalized entropy $H_\ell(P_\pi)$, $\pi^\star$ is the worst-case prior, $Q^\star = P_{\pi^\star}$, and $(P_{\pi^\star}, P_{\pi^\star})$ is a saddle point for the game (19.3.9).

Proof All that we need to argue is that there is a prior $\pi^\star$ maximizing the entropy $H_\ell(P_\pi)$. Because $\mathcal{P} = \mathrm{Conv}\{P_\theta\}_{\theta\in\Theta} \subset \Delta_\mathcal{X}$ is bounded, if it is closed it is compact. Then noting that $H_\ell(P)$ is the infimum of linear functions of P, it is concave and upper semi-continuous, so that its maximizers are attained over compact sets. So there is $P^\star \in \mathcal{P}$ maximizing $H_\ell(P)$; that $\mathcal{P}$ is closed implies there exists at least one prior $\pi^\star$ on Θ for which $P^\star = P_{\pi^\star}$. □

We have no particular need to work with finite domains, though (as we note above), this usually necessitates more care in the constructions. As a particular example, however, Corollary 19.3.4 implies the following result (see Exercise 19.1):

Corollary 19.3.9. *Let $\mathcal{F}$ be the collection of cumulative distributions on the compact interval $[t_0, t_1] \subset \mathbb{R}$ and ℓ_{crps} be the continuous ranked probability loss. Then the CDF $F^\star$ of the uniform distribution on $\{t_0, t_1\}$ uniquely minimizes $\sup_{F\in\mathcal{F}} \mathbb{E}_F[\ell_{\mathrm{crps}}(G, X)]$ over CDFs G, and*

$$\sup_{F\in\mathcal{F}} \int_{t_0}^{t_1} (1 - F(t))F(t)dt = \sup_{F\in\mathcal{F}} \mathbb{E}_F[\ell_{\mathrm{crps}}(F^\star, X)].$$

19.4 Minimax strategies for regret

With general theory about the existence of saddle points and minimax solutions in place, we turn to more explicit strategies for attaining optimal regret as the sample size $n \to \infty$. We discuss the adversarial setting only briefly, as optimal strategies are somewhat difficult to implement. We focus on the log loss, because (near-)optimal strategies for other losses typically remain open questions.

We begin by describing a notion of complexity that captures the best possible regret in the adversarial setting. Assume without loss of generality that we have a set of distributions $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ parameterized by $\theta \in \Theta$, where the distributions are supported on $\mathcal{X}^n$. We define the complexity of the set $\mathcal{P}$ (i.e. the complexity of Θ) as

$$\mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) dx_1^n \quad \text{or} \quad \mathsf{Comp}_n(\Theta) := \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) d\nu(x_1^n), \tag{19.4.1}$$

where ν is a base measure on $\mathcal{X}^n$. We may have $\mathsf{Comp}_n(\Theta) = \infty$, and this is not particularly uncommon; for example, consider the case of a normal location family with $\Theta = \mathbb{R}$.

The complexity (19.4.1) coincides with the minimax regret in the adversarial setting.

Proposition 19.4.1. *The minimax regret*

$$\inf_Q \mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n(\Theta).$$

If $\mathsf{Comp}_n(\Theta) < \infty$*, then the* normalized maximum likelihood *distribution* $\overline{Q}$ *with density*

$$\overline{q}(x_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(x_1^n)}{\int \sup_\theta p_\theta(x_1^n) dx_1^n},$$

is uniquely minimax optimal.

(The distribution $\overline{Q}$ is also known as the *Shtarkov* distribution.) The proposition completely characterizes the minimax regret in the adversarial setting, and it gives the unique distribution achieving the regret. In most cases it is challenging to compute the minimax optimal distribution $\overline{Q}$, so we must make approximations of some type.

Proof We first prove the result in the case that $\mathsf{Comp}_n < \infty$. The normalized maximum likelihood distribution $\overline{Q}$ has constant regret:

$$\begin{aligned}
\mathfrak{R}_n^{\mathcal{X}}(\overline{Q}, \mathcal{P}) &= \sup_{x_1^n \in \mathcal{X}^n} \left[\log \frac{1}{\overline{q}(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \right] \\
&= \sup_{x_1^n} \left[\log \frac{\int \sup_\theta p_\theta(x_1^n) dx_1^n}{\sup_\theta p_\theta(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \right] = \mathsf{Comp}_n(\Theta).
\end{aligned}$$

For any distribution Q on $\mathcal{X}^n$ we have

$$\begin{aligned}
\mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) &\geq \int \left[\log \frac{1}{q(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \right] \overline{q}(x_1^n) dx_1^n \\
&= \int \left[\log \frac{\overline{q}(x_1^n)}{q(x_1^n)} + \mathsf{Comp}_n(\Theta) \right] \overline{q}(x_1^n) dx_1^n \\
&= D_{\rm kl}\left(\overline{Q} \| Q\right) + \mathsf{Comp}_n(\Theta),
\end{aligned} \tag{19.4.2}$$

so that $\overline{Q}$ is uniquely minimax optimal, as $D_{\rm kl}\left(\overline{Q} \| Q\right) > 0$ unless $\overline{Q} = Q$.

To extend the lower bound (19.4.2) to the case $\mathsf{Comp}_n(\Theta) = \infty$, assume without loss of generality that $\mathcal{X}$ is countable and consists of points $x_1, x_2, \ldots$ (we can discretize $\mathcal{X}$ otherwise) and $n = 1$. Fix $\epsilon \in (0, 1)$, construct $\theta_1, \theta_2, \ldots$ so $p_{\theta_j}(x_j) \geq (1 - \epsilon) \sup_{\theta \in \Theta} p_\theta(x_j)$, and define the sets $\Theta_j =$

$\{\theta_1, \ldots, \theta_j\}$. Clearly $\mathsf{Comp}(\Theta_j) \le \log j$, and defining $\overline{q}_j(x) = \max_{\theta\in\Theta_j} p_\theta(x) / \sum_{x\in\mathcal{X}} \max_{\theta\in\Theta_j} p_\theta(x)$, we extend the reasoning yielding inequality (19.4.2) to obtain

$$\begin{aligned}
\mathfrak{R}^{\mathcal{X}}(Q, \mathcal{P}) &= \sup_{x\in\mathcal{X}} \left[\log \frac{1}{q(x)} - \log \frac{1}{\sup_{\theta\in\Theta} p_\theta(x)}\right] \\
&\ge \sum_x \overline{q}_j(x) \left[\log \frac{1}{q(x)} - \log \frac{1}{\max_{\theta\in\Theta_j} p_\theta(x)}\right] \\
&= \sum_x \overline{q}_j(x) \left[\log \frac{\overline{q}_j(x)}{q(x)} + \log \sum_{x'} \max_{\theta\in\Theta_j} p_\theta(x')\right] = D_{\rm kl}\left(\overline{Q}_j \| Q\right) + \mathsf{Comp}(\Theta_j).
\end{aligned}$$

Now note $\mathsf{Comp}(\Theta_j) \ge (1-\epsilon) \sum_{i=1}^j \sup_\theta p_\theta(x_i) + \sum_{i>j} \max_{\theta\in\Theta_j} p_\theta(x_i) \to \infty$ as $j \to \infty$. □

While typically computing the minimax (adversarial) regret is challenging, some simple cases, such as the family of i.i.d. Bernoulli distributions, admit evaluation.

Example 19.4.2: Consider the family of Bernoulli distributions with $P_\theta(x) = \theta^x (1-\theta)^{1-x}$ for $x \in \{0, 1\}$ and $0 \le \theta \le 1$. For a sequence $x_1^n \in \{0,1\}^n$ with m ones, $\widehat{\theta} = m/n$ satisfies

$$\sup_{\theta\in[0,1]} P_\theta(x_1^n) = P_{\widehat{\theta}}(x_1^n) = \widehat{\theta}^m (1-\widehat{\theta})^{n-m} = \exp(-n h_2(\widehat{\theta})),$$

where $h_2(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy. Using this representation gives complexity

$$\mathsf{Comp}_n([0,1]) = \log \sum_{m=0}^n \binom{n}{m} e^{-n h_2(\frac{m}{n})}.$$

We claim (see Exercise 19.4) that for a value c_n satisfying $\frac{1}{\sqrt{8}} \le c_n \le \frac{1}{\sqrt{\pi}}$, as $n \to \infty$,

$$\sum_{m=0}^n \binom{n}{m} e^{-n h_2(\frac{m}{n})} = (1 + o(1)) \cdot c_n \cdot n \cdot \int_0^1 \theta^{-1/2} (1-\theta)^{-1/2} d\theta.$$

This gives regret

$$\inf_Q \mathfrak{R}_n^{\mathcal{X}}(Q, \mathcal{P}) = \mathsf{Comp}_n([0,1]) = \frac{1}{2}\log n + \log \int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} d\theta + \log c_n + o(1).$$

We remark in passing that this is equal to $\frac{1}{2}\log n + \log \int_0^1 \sqrt{J_\theta} d\theta$, where J_θ denotes the Fisher information of the Bernoulli family (recall Example 3.1.1). We will see that this holds in more generality for redundancy in the sequel. ◇

19.5 Mixture (Bayesian) strategies and redundancy

In the less adversarial setting of expected regret, we compete against a random sequence X_1^n, drawn from a fixed distribution P. Thinking of this problem as a game, we choose a distribution Q according to which we make predictions (based on previous data), and nature chooses a distribution

$P_\theta \in \mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$. Simplifying the setting (19.2.3) we consider the i.i.d. case that $X_1^n \stackrel{\text{iid}}{\sim} P_\theta$, and we suffer expected regret (or redundancy)

$$\mathsf{Red}_n(Q, P_\theta) = \mathbb{E}_\theta \left[\log \frac{1}{q(X_1^n)}\right] - \mathbb{E}_\theta \left[\log \frac{1}{p_\theta(X_1^n)}\right] = D_{\rm kl}\left(P_\theta^n \| Q_n\right), \tag{19.5.1}$$

where Q_n denotes the law that Q induces on the full sample X_1^n. In this expression, q and p denote the densities of Q and P, respectively. In a slightly more general setting, we may consider the expected regret of Q with respect to a distribution P_θ under model mis-specification, meaning the data follows an alternative distribution P. The (more general) redundancy then becomes

$$\mathbb{E}_P \left[\log \frac{1}{q(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)}\right]. \tag{19.5.2}$$

In both cases (19.5.1) and (19.5.2), we wish to guarantee that the redundancy grows more slowly than n by finding distributions Q such that, for any $\theta_0 \in \Theta$, we have $\frac{1}{n} D_{\rm kl}\left(P_{\theta_0}^n \| Q_n\right) \to 0$ as $n \to \infty$. Assuming we could actually obtain such a distribution in general, this is interesting because (even in the i.i.d. case) for *any* fixed distribution $P_\theta \neq P_{\theta_0}$, we must have $D_{\rm kl}\left(P_{\theta_0}^n \| P_\theta^n\right) = nD_{\rm kl}\left(P_{\theta_0} \| P_\theta\right) = \Omega(n)$. Motivated by the redundancy/capacity duality Corollary 19.3.2, a natural approach is to consider mixtures, where we choose Q as a convex combination of all the possible source distributions P_θ for $\theta \in \Theta$. In contrast to the exact dualities in Section 19.3, we will rely on asymptotics to show approximate optimality, though we will be a bit non-rigorous.

To see how we make predictions from such mixtures, let $Q_\pi^n(A) = \int P_\theta^n(A) d\pi(\theta)$ for $A \subset \mathcal{X}^n$, noting that Q_π^n will typically not be a product distribution. At time i, this mixture plays the density

$$q_\pi(x_i \mid x_1^{i-1}) = \int_\Theta q(x_i, \theta \mid x_1^{i-1}) d\theta = \int_\Theta p_\theta(x_i) \pi(\theta \mid x_1^{i-1}) d\theta$$

by construction of the mixture Q_π^n. Here, the posterior $\pi(\theta \mid x_1^{i-1})$ takes the form

$$\pi(\theta \mid x_1^{i-1}) = \frac{\pi(\theta) p_\theta(x_1^{i-1})}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'} = \frac{\pi(\theta) \exp\left(-\log \frac{1}{p_\theta(x_1^{i-1})}\right)}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'}, \tag{19.5.3}$$

where we have emphasized that this strategy exhibits an *exponential weighting* approach, scaling distribution weights exponentially by their previous loss performance $-\log p_\theta(x_1^{i-1})$.

The weighting scheme (19.5.3) attains asymptotically strong performance. So long as the prior π puts non-zero mass over all of Θ, under appropriate smoothness conditions, the scheme Q_π is *universal*, meaning that $D_{\rm kl}\left(P_\theta^n \| Q_\pi^n\right) = o(n)$ for any $\theta \in \operatorname{int} \Theta$. Clarke and Barron [112, 113] make these statements rigorous; while the particular conditions are beyond our scope, they require roughly the following: the Fisher information J_θ for the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ exists in a compact set K interior to Θ, and the distributions P_θ are sufficiently regular that differentiation and integration can be interchanged (essentially, uniform versions of the conditions (i)–(v) necessary for the van Trees inequality, Theorem 11.2.7).

Theorem 19.5.1 (Clarke and Barron [112, 113])**.** *Let $Q_\pi^n = \int P_\theta^n \pi(\theta) d\theta$ be the marginal distribution over X_1^n obtained by drawing $T \sim \pi$ and then $X_i \stackrel{\text{iid}}{\sim} P_\theta$ conditional on $T = \theta$. Let π have compact support Θ, and $K \subset \operatorname{int} \Theta$ be any compact set. Then under appropriate smoothness conditions on the distributions P_θ and π,*

$$D_{\rm kl}\left(P_\theta^n \| Q_\pi^n\right) - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{\pi(\theta)} + \frac{1}{2} \log \det(J_\theta) \quad \text{as } n \to \infty \tag{19.5.4a}$$

uniformly in $\theta \in K$, and

$$\left| \int D_{\rm kl}\left(P_\theta^n \| Q_\pi^n\right) \pi(\theta) d\theta - \frac{d}{2} \log \frac{n}{2\pi e} - \int \pi(\theta) \log \frac{\sqrt{\det(J_\theta)}}{\pi(\theta)} d\theta \right| \to 0. \tag{19.5.4b}$$

While we do not rigorously prove the theorem, we give a sketch showing the main components of the result based on asymptotic normality arguments in Section 19.5.2.

Example 19.5.2 (Bernoulli distributions with Beta prior)**:** Consider the class of X_i drawn i.i.d. $\mathsf{Bernoulli}(\theta)$, where $\theta = P_\theta(X = 1) \in [0, 1]$. The $\mathsf{Beta}(\alpha, \beta)$-distribution has density

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}$$

on $[0, 1]$, where $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ denotes the gamma function. Under the $\mathsf{Beta}(\alpha, \beta)$ distribution, we have $\mathbb{E}_\pi[\theta] = \frac{\alpha}{\alpha + \beta}$. (Any undergraduate probability text contains such results.)

To play a Beta mixture of Bernoulli distributions, we first obtain the posterior (19.5.3). Let $S_i = \sum_{j=1}^i X_j$ be the partial sum up to iteration i. Then

$$\pi(\theta \mid x_1^i) = \frac{p_\theta(x_1^i)\pi(\theta)}{q(x_1^i)} \propto \theta^{S_i}(1 - \theta)^{i - S_i} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1} = \theta^{\alpha + S_i - 1}(1 - \theta)^{\beta + i - S_i - 1},$$

where we have ignored the denominator as we must simply normalize the quantity in θ. By inspection, $\theta \mid X_1^i$ has distribution $\mathsf{Beta}(\alpha + S_i, \beta + i - S_i)$. Thus, to compute the predictive distribution Q, we note that $\mathbb{E}_\theta[X] = \theta$, so

$$Q(X_{i+1} = 1 \mid X_1^i) = \mathbb{E}_\pi[\theta \mid X_1^i] = \frac{S_i + \alpha}{i + \alpha + \beta}.$$

Theorem 19.5.1 shows that the redundancy scales as

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_\pi^n\right) = \frac{1}{2} \log \frac{n}{2\pi e} + \log \left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \frac{1}{\theta_0^{\alpha - 1}(1 - \theta_0)^{\beta - 1}} \right] + \frac{1}{2} \log \frac{1}{\theta_0(1 - \theta_0)} + o(1)$$

for $\theta_0 \in (0, 1)$, and the prediction scheme with $\mathsf{Beta}(\alpha, \beta)$-prior is universal. $\diamond$

We can also show how to play sequentially with Gaussian models.

Example 19.5.3 (Gaussian distributions with a Gaussian prior)**:** Consider the collection of $X_i \stackrel{\rm iid}{\sim} \mathsf{N}(\theta, \sigma^2 I)$, $\theta \in \mathbb{R}^d$. Then if θ has $\pi = \mathsf{N}(0, \tau^2 I)$, a calculation yields posterior density

$$\pi(\theta \mid x_1^n) \propto \exp\left(-\frac{1}{2} \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2} \right) \left\| \theta - \frac{n\tau^2}{n\tau^2 + \sigma^2} \overline{x}_n \right\|_2^2 \right),$$

that is, if $X_i \mid T = \theta \sim \mathsf{N}(\theta, \sigma^2 I)$ and $T \sim \mathsf{N}(0, \tau^2 I)$,

$$T \mid X_1^n \sim \mathsf{N}\left(\frac{\tau^2 n}{\tau^2 n + \sigma^2} \overline{X}_n, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2 n} I \right).$$

Thus, the predictive distribution $Q_\pi(X_{n+1} \in \cdot \mid X_1^n)$ is Gaussian with mean $\mathbb{E}_Q[X_{n+1} \mid X_1^n] = \mathbb{E}_\pi[T \mid X_1^n] = \frac{\tau^2 n}{\tau^2 n + \sigma^2} \overline{X}_n$ and covariance

$$\mathrm{Cov}(X_{n+1} \mid X_1^n) = \mathrm{Cov}(T \mid X_1^n) + \sigma^2 I = \left(\frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2 n} + \sigma^2 \right) I.$$

Direct (but tedious; see Exercise 19.5) calculations show this agrees with the limit (19.5.4a). $\diamond$

While Theorem 19.5.1 addresses the case in which the data are generated i.i.d. P_θ, so that the model P_θ is well-specified, mixture models enjoy a type of robustness even under model misspecification, that is, when the true distribution generating the data does not belong to the class $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$. In this case, we look at the generalized redundancy (19.5.2), measuring loss relative to $D_{\rm kl}\left(P \| P_\theta\right)$. We now define the restricted mixture distributions on a subset $\Theta_0 \subset \Theta$ by

$$Q_{\pi, \Theta_0}(A) := \frac{1}{\pi(\Theta_0)} \int_{\Theta_0} P_\theta(A) d\pi(\theta).$$

With this, we obtain the following robustness result.

Proposition 19.5.4. *Assume that P_θ have densities p_θ over $\mathcal{X}$, let P be any distribution having density p over $\mathcal{X}$, and let q_π be the density associated with Q_π. Then for each $\theta \in \Theta$,*

$$\mathbb{E}_P\left[\log \frac{1}{q_\pi(X)} - \log \frac{1}{p_\theta(X)}\right] \le \inf_{\Theta_0 \subset \Theta} \left\{ \log \frac{1}{\pi(\Theta_0)} + D_{\rm kl}\left(P \| Q_{\pi, \Theta_0}\right) - D_{\rm kl}\left(P \| P_\theta\right) \right\}.$$

Proposition 19.5.4 shows that so long as some subset $\Theta_0 \subset \Theta$ allows mixture distributions Q_{π, Θ_0} to closely approximate P, we attain a convergence guarantee nearly as good as any in the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$. (This result is a spiritual predecessor to the mutual information bound (12.2.3), Corollary 12.2.2, and the *index of resolvability* there.)

Proof Fix $\Theta_0 \subset \Theta$. Then $q_\pi(x) = \int_\Theta p_\theta(x) d\pi(\theta) \ge \int_{\Theta_0} p_\theta(x) d\pi(\theta)$. Thus we have

$$\begin{aligned}
\mathbb{E}_P\left[\log \frac{p(X)}{q_\pi(X)}\right] &\le \mathbb{E}_P\left[\inf_{\Theta_0 \subset \Theta} \log \frac{p(X)}{\int_{\Theta_0} p_\theta(X) d\pi(\theta)}\right] \\
&= \mathbb{E}_P\left[\inf_{\Theta_0} \log \frac{p(X)\pi(\Theta_0)}{\pi(\Theta_0) \int_{\Theta_0} p_\theta(X) d\pi(\theta)}\right] = \mathbb{E}_P\left[\inf_{\Theta_0} \log \frac{p(X)}{\pi(\Theta_0) q_{\pi, \Theta_0}(X)}\right].
\end{aligned}$$

This is certainly smaller than the same quantity with the infimum outside the expectation, and noting that

$$\mathbb{E}_P\left[\log \frac{1}{q_\pi(X)} - \log \frac{1}{p_\theta(X)}\right] = \mathbb{E}_P\left[\log \frac{p(X)}{q_\pi(X)}\right] - \mathbb{E}_P\left[\log \frac{p(X)}{p_\theta(X)}\right]$$

gives the result. $\square$

19.5.1 Bayesian redundancy and objective, reference, and Jeffreys priors

Consider again the Bayesian redundancy

$$\inf_Q \int_\Theta \pi(\theta) D_{\rm kl}\left(P_\theta^n \| Q\right) d\theta = \int_\Theta \pi(\theta) D_{\rm kl}\left(P_\theta^n \| Q_\pi^n\right) d\theta = I_\pi(T; X_1^n),$$

the mutual information between a random variable $T \sim \pi$ and $X_1^n \stackrel{\text{iid}}{\sim} P_\theta$ drawn conditional on $T = \theta$. With Theorem 19.5.1 in hand, we can give a somewhat more nuanced picture of this mutual information quantity. As a first consequence of the theorem,

$$I_\pi(T; X_1^n) = \frac{d}{2} \log \frac{n}{2\pi e} + \int \pi(\theta) \log \frac{\sqrt{\det J_\theta}}{\pi(\theta)} d\theta + o(1), \tag{19.5.5}$$

where J_θ denotes the Fisher information matrix for the family $\{P_\theta\}_{\theta \in \Theta}$. *Reference analysis*, a strand of Bayesian statistics, advocates choosing the prior π that maximizes the mutual information between the unknown parameter θ and the observations X_1^n (before observing X_1^n). The asymptotic formula (19.5.5) allows us to choose such a prior. The bibliographic section contains more discussion.

In a different vein, Jeffreys [257] proposed that if the square root of the determinant of the Fisher information was integrable, then one should take the eponymous *Jeffreys prior*

$$\pi_{\text{jeffreys}}(\theta) = \frac{\sqrt{\det J_\theta}}{\int_\Theta \sqrt{\det J_\theta} d\theta}.$$

Jeffreys proposes this for invariance, as inferences on θ under the prior π_{jeffreys} are identical to those on a transformed parameter $\phi(\theta)$ under the appropriately transformed Jeffreys prior.

Proceeding non-rigorously by not worrying about integrability or uniformity in the prior π, the asymptotic expression (19.5.5) shows the Jeffreys and asymptotic reference priors coincide. Indeed, computing the integral in (19.5.5), we have

$$\begin{aligned} \int_\Theta \pi(\theta) \log \frac{\sqrt{\det J_\theta}}{\pi(\theta)} d\theta &= \int_\Theta \pi(\theta) \log \frac{\pi_{\text{jeffreys}}(\theta)}{\pi(\theta)} d\theta + \log \int \sqrt{\det J_\theta} d\theta \\ &= -D_{\rm kl}\left(\pi \| \pi_{\text{jeffreys}}\right) + \log \int \sqrt{\det J_\theta} d\theta, \end{aligned}$$

whenever the Jeffreys prior exists. Moreover, we see that in an asymptotic sense, the Jeffreys prior is the worst-case prior distribution π for nature to play, as otherwise the $-D_{\rm kl}\left(\pi \| \pi_{\text{jeffreys}}\right)$ term in the expected (Bayesian) redundancy is negative.

Example 19.5.5 (Jeffreys priors and the exponential distribution)**:** Let us now assume that our source distributions P_θ are exponential distributions, meaning that $\theta \in (0, \infty)$ and we have density $p_\theta(x) = \exp(-\theta x - \log \frac{1}{\theta})$ for $x \in [0, \infty)$. This is an exponential family model with Fisher information $J_\theta = \frac{\partial^2}{\partial \theta^2} \log \frac{1}{\theta} = 1/\theta^2$ (cf. Eq. (3.3.2)).

In this case, the Jeffreys prior $\pi_{\text{jeffreys}}(\theta) \propto \sqrt{J_\theta} = 1/\theta$ fails to integrate over $[0, \infty)$. To address this difficulty, Bernardo [64, Definition 3] (among others) advocates proceeding formally, noting that given a single observation, the "posterior" distribution $\pi(\theta \mid X)$ is well-defined. Following this idea, note that after seeing data $X_1, \ldots, X_i$ with partial sum $S_i = \sum_{j=1}^i X_j$,

$$\pi(\theta \mid x_1^i) \propto p_\theta(x_1^i) \pi_{\text{jeffreys}}(\theta) = \theta^i \exp\left(-\theta \sum_{j=1}^i x_j\right) \frac{1}{\theta} = \theta^{i-1} \exp(-\theta S_i).$$

Integrating, for $s_i = \sum_{j=1}^i x_j$ we obtain

$$q(x \mid x_1^i) = \int_0^\infty p_\theta(x) \pi(\theta \mid x_1^i) d\theta \propto \int_0^\infty \theta e^{-\theta x} \theta^{i-1} e^{-\theta s_i} d\theta = \frac{1}{(s_i + x)^{i+1}} \int_0^\infty u^i e^{-u} du,$$

using the change of variables $u = \theta(s_i + x)$. This is at least a distribution that normalizes, so often one simply assumes the existence of a piece of fake data. For example, by saying we "observe" $x_0 = 1$, we have prior proportional to $\pi(\theta) = e^{-\theta}$, which yields redundancy

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_\pi^n\right) = \frac{1}{2}\log\frac{n}{2\pi e} + \theta_0 + \log\frac{1}{\theta_0} + o(1).$$

The difference is that, in this case, the redundancy bound is no longer uniform in θ_0, as it would be for the true reference (or Jeffreys, if it exists) prior. $\diamond$

19.5.2 Heuristic calculations: normality and Theorem 19.5.1

We heuristically justify the asymptotic expression (19.5.4a), which predicts the expectation version (19.5.4b). To do this, we argue that the posterior distribution $\pi(\theta \mid X_1^n)$ should be approximately normal with appropriate variance.

Fix θ_0. We begin by expanding the log likelihood in the divergence

$$D_{\rm kl}\left(P_{\theta_0}^n \| Q_\pi^n\right) = \mathbb{E}_{P_{\theta_0}}\left[\log\frac{p_{\theta_0}(X_1^n)}{q_\pi(X_1^n)}\right] = \mathbb{E}_{P_{\theta_0}}\left[\log\frac{p_{\theta_0}(X_1^n)}{\int \pi(\theta)p_\theta(X_1^n)d\theta}\right].$$

By a Taylor expansion of θ around θ_0, we have

$$\log p_\theta(x) = \log p_{\theta_0}(x) + \nabla \log p_{\theta_0}(x)^\top(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top \nabla^2 \log p_{\theta_0}(x)(\theta - \theta_0) + o(\|\theta - \theta_0\|^2).$$

Assuming the expansion is uniform in x, then recalling the *Fisher score* $S_n := \sum_{i=1}^n \nabla \log p_{\theta_0}(X_i)$, the law of large numbers implies $\frac{1}{n}\sum_{i=1}^n \nabla^2 \log p_{\theta_0}(X_i) \to \mathbb{E}_{\theta_0}[\nabla^2 \log p_{\theta_0}(X)] = -J_{\theta_0}$, so

$$\log p_\theta(X_1^n) \approx \log p_{\theta_0}(X_1^n) + S_n^\top(\theta - \theta_0) - \frac{n}{2}(\theta - \theta_0)^\top J_{\theta_0}(\theta - \theta_0).$$

With this approximation, we complete the square so for $\overline{S}_n = \frac{1}{n}S_n$ and θ near θ_0,

$$\log p_\theta(X_1^n) \approx \log p_{\theta_0}(X_1^n) - \frac{n}{2}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n)^\top J_{\theta_0}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n) + \frac{n}{2}\overline{S}_n^\top J_{\theta_0}^{-1}\overline{S}_n. \tag{19.5.6}$$

Because $\mathbb{E}_{\theta_0}[\overline{S}_n^\top J_{\theta_0}^{-1}\overline{S}_n] = \frac{d}{n}$, so typically $J_{\theta_0}^{-1}\overline{S}_n = O(1/\sqrt{n})$, and when θ is farther than distance $O(1/\sqrt{n})$ from θ_0,

$$\frac{n}{2}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n)^\top J_{\theta_0}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n) \gg 1.$$

Assuming the approximation (19.5.6) is accurate enough, if we let Θ_0 be the points within $O(1/\sqrt{n})$ of θ_0, then for "most" samples X_1^n we have

$$\begin{aligned}
&\int \pi(\theta)p_\theta(X_1^n)d\theta \\
&\approx p_{\theta_0}(X_1^n)\pi(\theta_0)e^{\frac{n}{2}\overline{S}_n^\top J_{\theta_0}^{-1}\overline{S}_n}\int_{\Theta_0}\exp\left(-\frac{n}{2}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n)^\top J_{\theta_0}(\theta - \theta_0 - J_{\theta_0}^{-1}\overline{S}_n)\right)d\theta
\end{aligned}$$

because π is continuous. Making the *Laplace approximation* that the last integral is that of a Gaussian density, recalling $\int_{\mathbb{R}^d}\exp(-\frac{1}{2}x^\top Jx)dx = (2\pi)^{d/2}\det(J)^{-1/2}$ for $J \succ 0$, we obtain

$$\int \pi(\theta)p_\theta(X_1^n)d\theta \approx p_{\theta_0}(X_1^n)\pi(\theta_0)\exp\left(\frac{n}{2}\overline{S}_n^\top J_{\theta_0}^{-1}\overline{S}_n\right)(2\pi)^{d/2}\det(nJ_{\theta_0})^{-1/2}. \tag{19.5.7}$$

Substituting the heuristic approximation (19.5.7) into the KL-divergence yields

$$\begin{aligned} D_{\rm kl}\left(P_{\theta_0}^n \| Q_\pi^n\right) &\approx \mathbb{E}_{P_{\theta_0}}\left[\log \frac{1}{\pi(\theta_0)} - \frac{n}{2}\overline{S}_n^\top J_{\theta_0}^{-1}\overline{S}_n - \frac{d}{2}\log(2\pi) + \frac{d}{2}\log n + \frac{1}{2}\log\det(J_{\theta_0})\right] \\ &= \log\frac{1}{\pi(\theta_0)} + \frac{d}{2}\log\frac{1}{2\pi e} + \frac{d}{2}\log n + \frac{1}{2}\log\det(J_{\theta_0}), \end{aligned}$$

because $\mathbb{E}_{P_{\theta_0}}[\overline{S}_n\overline{S}_n^\top] = n^{-1}J_{\theta_0}$. This is the result (19.5.4a).

19.6* Regret/capacity dualities and Theorem 19.3.1 with rigor

To develop rigorous versions of Theorem 19.3.1 requires substantial additional mathematical care, which we present here for more advanced readers. The basic strategy of the argument is to demonstrate a sequence of priors π_n and associated marginals $Q_{\pi_n} = \int P_\theta d\pi_n(\theta)$, which (because of some type of compactness) have subsequent limits, which must be optimal. Recall that for $\Omega(P) = -\inf_Q \mathbb{E}_P[\ell(Q, X)]$, we have the saddle point objective (19.3.5)

$$L(Q, \pi) = \int D_\Omega(P_\theta, Q) d\pi(\theta),$$

where $D_\Omega(P, Q) = \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)]$ is the Bregman divergence (expected regret) associated with Ω. We first prove the theorem in the special case that $\mathcal{X}$ is finite (Section 19.6.2), which corresponds to Corollary 19.3.3, because this contains the main ideas. The extension to arbitrary spaces $\mathcal{X}$ then "just" requires more careful treatment of probabilistic convergence.

19.6.1 Lower semicontinuity and capacity-achieving priors

A key intermediate result is that when there is a prior $\pi^\star$ achieving the capacity, then we are guaranteed a saddle point, and so long as ℓ is strictly proper, we have uniqueness. Because we wish to achieve minimizers, we require the losses to have some type of lower semicontinuity. We begin with a one-dimensional notion that sometimes suffices, and say that $D_\Omega(P, Q)$ is *directionally lower semicontinuous in* Q if for any two distributions Q_0, Q_1,

$$\liminf_{\lambda \downarrow 0} D_\Omega(P, (1-\lambda)Q_0 + \lambda Q_1) \geq D_\Omega(P, Q_0). \tag{19.6.1}$$

Under this condition, achieving the maximum capacity is sufficient to guarantee the existence of saddle points, so the duality (19.3.7) obtains.

Lemma 19.6.1. *Let Π be a convex collection of distributions on Θ. Assume that the capacity $C_\ell(\Pi) := \sup_{\pi\in\Pi} I_\ell(\pi; X)$ is finite, and that $\pi^\star \in \Pi$ achieves it:*

$$I_\ell(\pi^\star; X) = \sup_{\pi\in\Pi} I_\ell(\pi; X) = C_\ell(\Pi) < \infty.$$

Let the divergence $D_\Omega(P_\theta, Q)$ be directionally lower semicontinuous in Q for each P_θ. Then the marginal $P_{\pi^\star}$ satisfies

$$\sup_{\pi\in\Pi} L(P_{\pi^\star}, \pi) \leq L(P_{\pi^\star}, \pi^\star) \leq \inf_Q L(Q, \pi^\star),$$

so $(P_{\pi^\star}, \pi^\star)$ is a saddle point for the expected regret (19.3.5), and $L(P_{\pi^\star}, \pi^\star) = C_\ell(\Pi) = I_\ell(\pi^\star; X)$. If additionally ℓ is strictly proper, then $P_{\pi^\star}$ uniquely achieves the infimum in $L(Q, \pi^\star)$.

Writing the lemma in terms of the gaps in the expected losses, if $\pi^\star$ maximizes $I_\ell(\pi; X)$ over $\pi \in \Pi$, then $Q^\star = P_{\pi^\star}$, and the regret/capacity game (19.0.2) has saddle point $(P_{\pi^\star}, \pi^\star)$:

$$\sup_{\pi \in \Pi} \int \left(\mathbb{E}_{P_\theta}[\ell(P_{\pi^\star}, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) d\pi(\theta) \le \int \left(\mathbb{E}_{P_\theta}[\ell(P_{\pi^\star}, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) d\pi^\star(\theta)$$
$$\le \inf_Q \int \left(\mathbb{E}_{P_\theta}[\ell(Q, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) d\pi^\star(\theta).$$

Proof of Lemma 19.6.1 Because ℓ is proper, for any π, the distribution $P_\pi := \int P_\theta d\pi(\theta)$ minimizes $L(Q, \pi)$ over distributions Q (recall Eq. (19.3.6)). Let $Q^\star = P_{\pi^\star}$. Then for any Q,

$$C_\ell = I_\ell(\pi^\star; X) = L(Q^\star, \pi^\star) = \inf_Q L(Q, \pi^\star) \le L(Q, \pi^\star). \tag{19.6.2}$$

Now, take any distribution $\pi \in \Pi$, and let $\lambda \in (0, 1)$, defining $\pi_\lambda = (1 - \lambda)\pi^\star + \lambda\pi$ (recall we have assumed Π is convex). Then $I_\ell(\pi_\lambda; X) \le C_\ell$ immediately. Let $P_\lambda = (1 - \lambda)P_{\pi^\star} + \lambda P_\pi$ and $T \in \Theta$ represent a draw from π or $\pi^\star$, and abuse notation so that P_T denotes the distribution on X conditional on the realization T. Then

$$\begin{aligned}
C_\ell \ge I_\ell(\pi_\lambda; X) &= \inf_Q \left\{ \int \mathbb{E}_{P_\theta}[\ell(Q, X)] d\pi_\lambda(\theta) - \lambda \mathbb{E}_\pi[H_\ell(P_T)] - (1 - \lambda)\mathbb{E}_{\pi^\star}[H_\ell(P_T)] \right\} \\
&\overset{(i)}{=} \int \mathbb{E}_{P_\theta}[\ell(P_\lambda, X)] d\pi_\lambda(\theta) - \lambda \mathbb{E}_\pi[H_\ell(P_T)] - (1 - \lambda)\mathbb{E}_{\pi^\star}[H_\ell(P_T)] \\
&= (1 - \lambda) \left[\int \mathbb{E}_{P_\theta}[\ell(P_\lambda, X)] d\pi^\star(\theta) - \mathbb{E}_{\pi^\star}[H_\ell(P_T)] \right] + \lambda \left[\int \mathbb{E}_{P_\theta}[\ell(P_\lambda, X)] d\pi(\theta) - \mathbb{E}_\pi[H_\ell(P_T)] \right] \\
&\overset{(ii)}{\ge} (1 - \lambda)C_\ell + \lambda \left(\int \mathbb{E}_{P_\theta}[\ell(P_\lambda, X)] d\pi(\theta) - \mathbb{E}_\pi[H_\ell(P_T)] \right),
\end{aligned}$$

where step (i) uses the propriety of the loss and step (ii) follows from the earlier bound (19.6.2). Rearranging and dividing through by $\lambda > 0$, we obtain by definition of L that

$$L(\lambda P_\pi + (1 - \lambda)P_{\pi^\star}, \pi) = \int \mathbb{E}_{P_\theta}[\ell(\lambda P_\pi + (1 - \lambda)P_{\pi^\star}, X)] d\pi(\theta) - \mathbb{E}_\pi[H_\ell(P_T)] \le C_\ell$$

for all $\pi \in \Pi$ and $\lambda \in (0, 1)$.

Finally, we use that the expected regret $D_\Omega(P_\theta, Q) = \mathbb{E}_{P_\theta}[\ell(Q, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]$ is directionally lower semicontinuous (19.6.1). This implies that

$$\liminf_{\lambda \downarrow 0} \mathbb{E}_{P_\theta}\left[\ell(\lambda P + (1 - \lambda)Q, X)\right] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] \ge D_\Omega(P_\theta, Q)$$

for any P, Q, P_θ. Applying Fatou's lemma (see Proposition A.1.1 in Appendix A.1) then gives

$$\begin{aligned}
L(P_{\pi^\star}, \pi) = \int D_\Omega(P_\theta, P_{\pi^\star}) d\pi(\theta) &\le \int \liminf_{\lambda \downarrow 0} \left(\mathbb{E}_{P_\theta}\left[\ell(\lambda P_\pi + (1 - \lambda)P_{\pi^\star}, X)\right] - H_\ell(P_\theta) \right) d\pi(\theta) \\
&\le \liminf_{\lambda \downarrow 0} L(\lambda P_\pi + (1 - \lambda)P_{\pi^\star}, \pi) \le C_\ell.
\end{aligned}$$

This implies the desired saddle point. □

19.6.2 Duality when the domain is finite

To show that a regret/capacity duality obtains when $\mathcal{X}$ is finite (and also when $\mathcal{X}$ is infinite in the sections to come), we leverage the Bregman divergence representation of expected regret (19.3.1). Let us recapitulate to keep the presentation self-contained. By Theorem 14.2.1, we have the Gneiting-Raftery-Savage representation (14.2.1) associated to any proper loss,

$$\ell(Q, x) = -\Omega(Q) - \langle \Omega'(Q), e_x - Q \rangle \tag{19.6.3}$$

for some convex Ω. Equation (19.6.3) abuses notation slightly, in that we are (without comment) viewing distributions Q as elements of $\Delta_{\mathcal{X}} = \{q \in \mathbb{R}_+^{\mathcal{X}} \mid \langle \mathbf{1}, q \rangle = 1\} \subset \mathbb{R}^{\mathcal{X}}$, and e_x is the "xth" standard basis vector. Then $\Omega'(Q)$ is a (particular, fixed) element of the subdifferential $\partial\Omega(Q)$, which necessarily exists because Ω is proper (Theorem 14.2.1). As $\Omega(P) = \sup_Q -\mathbb{E}_P[\ell(Q, X)]$ for proper losses, we then have the equality (19.3.1), that is,

$$D_\Omega(P, Q) := \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)] = \Omega(P) - \Omega(Q) - \langle \Omega'(Q), P - Q \rangle,$$

the Bregman divergence between P and Q. To avoid issues of non-determinacy in the selection of $\Omega'(Q)$, we always take the gap in the expected losses as the definition of the divergence.

When the space $\mathcal{X}$ is finite and the loss ℓ is strictly proper, the next lemma shows that convergence of distributions in D_Ω implies convergence in any norm. To make clear that we work with vectors $p \in \mathbb{R}^{\mathcal{X}}$, we use lower case letters. We require one additional consideration: that

$$D_\Omega(p, q) := \mathbb{E}_p[\ell(q, X)] - \mathbb{E}_p[\ell(p, X)] = \mathbb{E}_p[\ell(q, X)] - H_\ell(p) = \Omega(p) + \mathbb{E}_p[\ell(q, X)]$$

is lower semicontinuous in p, q, meaning that if $p_n \to p$ and $q_n \to q$, then $\liminf_n D_\Omega(p_n, q_n) \geq D_\Omega(p, q)$. Because $\Omega(p)$ is the supremum of linear functions of p, it is lower semicontinuous, and so because $\mathcal{X}$ is finite, it is enough that $\ell(q, x)$ be lower semicontinuous in q for each x.

Lemma 19.6.2. *Let $\mathcal{X}$ be finite, and assume the loss ℓ is strictly proper and that D_Ω is lower semicontinuous. Then there exists a function ω with $\omega(\epsilon) > 0$ for all $\epsilon > 0$ such that*

$$D_\Omega(p, q) \geq \omega(\|p - q\|_1).$$

Proof We identify $\mathcal{P} = \Delta_{\mathcal{X}} = \{p \in \mathbb{R}_+^{\mathcal{X}} \mid \langle \mathbf{1}, p \rangle = 1\}$, a compact set. Assume for the sake of contradiction that

$$\inf_{p,q \in \Delta_{\mathcal{X}}} \{D_\Omega(p, q) \mid \|p - q\|_1 \geq \epsilon\} = 0$$

for some $\epsilon > 0$. Let $p_n, q_n \in \Delta_{\mathcal{X}}$ tend to this infimum, that is, satisfy $\|q_n - p_n\|_1 \geq \epsilon$ and $D_\Omega(p_n, q_n) \to 0$. Taking subsequences as necessary, we may assume that $q_n \to q$ and $p_n \to p$. By the lower semicontinuity assumption, for any $\delta > 0$ there exists N such that $n \geq N$ implies

$$\mathbb{E}_{p_n}[\ell(q_n, X)] - \mathbb{E}_{p_n}[\ell(p_n, X)] \geq \mathbb{E}_p[\ell(q, X)] + \Omega(p) - \delta = D_\Omega(p, q) - \delta.$$

Because $\delta > 0$ was arbitrary, we in fact have

$$\liminf_n \mathbb{E}_{p_n}[\ell(q_n, X)] - \mathbb{E}_{p_n}[\ell(p_n, X)] \geq D_\Omega(p, q) > 0$$

because ℓ is strictly proper, our desired contradiction. □

As a consequence, if p_n is a Cauchy sequence for D_Ω, meaning that $\sup_{m \geq 0} D_\Omega(p_n, p_{n+m}) \to 0$ as $n \to \infty$, it is also a Cauchy sequence in $\Delta_{\mathcal{X}}$ and thus converges to some p.

19.6.3 Proof of Corollary 19.3.3

We turn to the proof of Corollary 19.3.3. Recall the convex collection Π of prior distributions π on Θ. Let Q be any distribution on $\mathcal{X}$ (meaning we may identify it with a vector in $\mathbb{R}_+^{\mathcal{X}}$), and define the saddle objective

$$\begin{aligned} L(Q, \pi) &:= \int \mathbb{E}_{P_\theta}[\ell(Q, X)] d\pi(\theta) - \int \mathbb{E}_{P_\theta}[\ell(P_\theta, X)] d\pi(\theta) \\ &= \int \left[\Omega(P_\theta) - \Omega(Q) - \langle \Omega'(Q), P_\theta - Q \rangle\right] d\pi(\theta) = \int D_\Omega(P_\theta, Q) d\pi(\theta). \end{aligned}$$

We decompose the proof into several parts. We first recall from Lemma 19.6.1 that *if* $\hat{\pi}$ attains the supremum in the definition (19.3.4) of the information in the experiment $\{P_\theta\}$ for the loss ℓ, then $(P_{\hat{\pi}}, \hat{\pi})$ is a saddle point. Using this intermediate result, we use finite-dimensional approximations of the capacity $\sup_\pi I_\ell(\pi; X)$, which in turn give a sequence of (explicit) saddle points $Q_n = P_{\pi_n}$ minimizing the worst-case redundancy. Finally, with a bit of algebraic manipulation and the careful use of the relationship between Bregman divergences D_Ω and proper losses ℓ, we can use completeness to show that limits of the sequence Q_n necessarily exist, and they are unique.

Finite-dimensional approximation and a limiting Q With Lemma 19.6.1 in hand, we proceed via finite dimensional approximations. Let $\pi_n \in \Pi$ be a sequence of priors satisfying $I_\ell(\pi_n; X) \to \sup_{\pi \in \Pi} I_\ell(\pi; X) = C_\ell$. Then the set

$$\Pi_n := \mathrm{Conv}\{\pi_1, \ldots, \pi_n\} \tag{19.6.4}$$

is isomorphic to an at most $n-1$ dimensional set, and so because the saddle objective $L(Q, \pi)$ is linear in π, $\inf_Q L(Q, \pi)$ is a concave function in π (and upper semicontinuous, as it is the infimum of linear functions). So

$$\pi_n^\star = \operatorname*{argmax}_{\pi \in \Pi_n} \left\{ \inf_Q L(Q, \pi) = I_\ell(\pi; X) \right\}$$

is attained. The sequence $\pi_n^\star$ of priors also satisfies $I_\ell(\pi_n^\star; X) \to C_\ell$ because $C_\ell \geq I_\ell(\pi_n^\star; X) \geq I_\ell(\pi_n; X) \to C_\ell$. Because the loss ℓ is proper, the mixture distribution $Q_n = P_{\pi_n^\star}$ minimizes $L(Q, \pi_n^\star)$, and Lemma 19.6.1 implies $\sup_{\pi \in \Pi_n} L(Q_n, \pi) \leq L(Q_n, \pi_n^\star) \leq \inf_Q L(Q, \pi_n^\star) \to C_\ell$.

Now for $k \in \mathbb{N}$ consider the pair $Q_n = P_{\pi_n^\star}$ and $Q_{n+k} = P_{\pi_{n+k}^\star}$. We have

$$L(Q_{n+k}, \pi_n^\star) - L(Q_n, \pi_n^\star) = -\Omega(Q_{n+k}) - \langle \Omega'(Q_{n+k}), P_{\pi_n^\star} - Q_{n+k} \rangle + \Omega(Q_n) = D_\Omega(Q_n, Q_{n+k}).$$

On the other hand, we also have

$$\begin{aligned} L(Q_{n+k}, \pi_n^\star) - L(Q_n, \pi_n^\star) &\leq L(Q_{n+k}, \pi_{n+k}^\star) - L(Q_n, \pi_n^\star) \\ &= I_\ell(\pi_{n+k}^\star; X) - I_\ell(\pi_n^\star; X) \leq C_\ell - I_\ell(\pi_n^\star; X) \to 0, \end{aligned}$$

the inequality following again from Lemma 19.6.1. As k is arbitrary, we have the limit (in n)

$$\sup_{k \geq 1} D_\Omega(Q_n, Q_{n+k}) \to 0. \tag{19.6.5}$$

Here we use that the space $\mathcal{X}$ is finite; by Lemma 19.6.2, there is necessarily some $Q^\star \in \Delta_{\mathcal{X}}$ such that $Q_n \to Q^\star$. We show this $Q^\star$ is a saddle.

Optimality of the limiting Q Using Lemma 19.6.1 again, we have

$$L(Q_n, \pi_n^\star) \le I_\ell(\pi_n^\star; X) \le C_\ell.$$

For any $\pi \in \Pi_\infty := \cup_{n=1}^\infty \Pi_n$, we have $\pi \in \Pi_n$ for all large enough n, so that similarly

$$L(Q_n, \pi) \le C_\ell$$

for all $\pi \in \Pi_\infty$. Now we use the lower semicontinuity of the regret $D_\Omega(P_\theta, Q)$ in Q, which implies

$$\begin{aligned}\int \left(\mathbb{E}_{P_\theta}[\ell(Q^\star, X)] - H_\ell(P_\theta)\right) d\pi(\theta) &\stackrel{(i)}{\le} \int \liminf_n \left(\mathbb{E}_{P_\theta}[\ell(Q_n, X)] - H_\ell(P_\theta)\right) d\pi(\theta) \\ &\stackrel{(ii)}{\le} \liminf_n \int \left(\mathbb{E}_{P_\theta}[\ell(Q_n, X)] - H_\ell(P_\theta)\right) d\pi(\theta) \le C_\ell,\end{aligned}$$

where step (i) used lower semicontinuity of ℓ and step (ii) used Fatou's lemma (Proposition A.1.1). In brief, we have demonstrated that

$$L(Q^\star, \pi) \le \sup_{\pi \in \Pi} I_\ell(\pi; X) = C_\ell \quad \text{for all } \pi \in \Pi_\infty = \bigcup_{n \ge 1} \Pi_n. \tag{19.6.6}$$

For the final step to showing that $Q^\star$ is optimal, we show that the inequality (19.6.6) holds for any $\pi \in \Pi$. To that end, let $\pi \in \Pi$ be any fixed prior, and replace the set (19.6.4) with $\widetilde{\Pi}_n = \mathrm{Conv}\{\pi_1, \ldots, \pi_n, \pi\} \subset \Pi$. This is isomorphic to an n-dimensional set, and a *mutatis mutandis* derivation yields a sequence of priors $\widetilde{\pi}_n^\star$ and distributions $\widetilde{Q}_n = P_{\widetilde{\pi}_n^\star}$. Then

$$L(\widetilde{Q}_n, \pi_n^\star) - L(Q_n, \pi_n^\star) = -\Omega(\widetilde{Q}_n) - \langle \Omega'(\widetilde{Q}_n), P_{\pi_n^\star} - \widetilde{Q}_n \rangle + \Omega(Q_n) = D_\Omega(Q_n, \widetilde{Q}_n).$$

Similarly, we have

$$\begin{aligned}L(\widetilde{Q}_n, \pi_n^\star) - L(Q_n, \pi_n^\star) &\le L(\widetilde{Q}_n, \widetilde{\pi}_n^\star) - L(Q_n, \pi_n^\star) \\ &= I_\ell(\widetilde{\pi}_n^\star; X) - I_\ell(\pi_n^\star; X) \le C_\ell - I_\ell(\pi_n^\star; X) \to 0\end{aligned}$$

by the assumption that $I_\ell(\pi_n; X) \to C_\ell = \sup_{\pi \in \Pi} I_\ell(\pi; X)$. Thus $D_\Omega(Q_n, \widetilde{Q}_n) \to 0$, and so again by Lemma 19.6.2, $\widetilde{Q}_n \to Q^\star$ as well. Thus we extend inequality (19.6.6) to

$$L(Q^\star, \pi) \le \sup_{\pi \in \Pi} I_\ell(\pi; X) = C_\ell \quad \text{for all } \pi \in \Pi. \tag{19.6.7}$$

Saddle point and its uniqueness As an immediate consequence of inequality (19.6.7), we see that $Q^\star$ is indeed a saddle point: we have

$$\sup_{\pi \in \Pi} L(Q^\star, \pi) \le C_\ell = \sup_{\pi \in \Pi} \inf_Q L(Q, \pi) \le \inf_Q \sup_{\pi \in \Pi} L(Q, \pi) \le \sup_{\pi \in \Pi} L(Q^\star, \pi).$$

by the weak min-max inequality, so equality holds in each step. Finally, we show that $Q^\star$ is unique. Suppose that $\widetilde{Q}$ is a different distribution satisfying $C_\ell = \sup_{\pi \in \Pi} L(\widetilde{Q}, \pi)$. Then letting

$$\begin{aligned}C_\ell \ge L(\widetilde{Q}, \pi_n^\star) &= L(\widetilde{Q}, \pi_n^\star) - L(Q_n, \pi_n^\star) + L(Q_n, \pi_n^\star) \\ &= -\Omega(\widetilde{Q}) - \langle \Omega'(\widetilde{Q}), P_{\pi_n^\star} - \widetilde{Q} \rangle + \Omega(Q_n) + L(Q_n, \pi_n^\star) = L(Q_n, \pi_n^\star) + D_\Omega(Q_n, \widetilde{Q}).\end{aligned}$$

But we know that $L(Q_n, \pi_n^\star) \to C_\ell$, and thus

$$D_\Omega(Q_n, \widetilde{Q}) \to 0. \tag{19.6.8}$$

Because $Q_n \to Q^\star$, it must be the case that $\widetilde{Q} = Q^\star$.

If $\pi^\star$ achieves the capacity, that is, maximizes $\inf_Q L(Q, \pi)$ over $\pi \in \Pi$, then $\inf_Q L(Q, \pi^\star) = C_\ell(\Pi)$, and by the inequality (19.6.7),

$$\sup_{\pi \in \Pi} L(Q^\star, \pi) \le \inf_Q L(Q, \pi^\star).$$

Then $L(Q^\star, \pi^\star) \le \inf_Q L(Q, \pi^\star)$, and because ℓ is strictly proper we necessarily have $Q^\star = P_{\pi^\star}$.

19.6.4 Regret/capacity duality for arbitrary domains

Our proof of Corollary 19.3.3 in Section 19.6.3 made little use of the fact that $\mathcal{X}$ was finite. An inspection of the proof shows that we used the finiteness of $\mathcal{X}$ in only a few places: the first is in equation (19.6.3) representing the expected regret via proper losses and the Gneiting-Raftery-Savage representation (14.2.1) from Theorem 14.2.1. In this case, we can rewrite the loss using the general representation in Theorem 14.2.8,

$$\ell(Q, x) = -\Omega(Q) - \langle \Omega'(Q), \mathbf{1}_x - Q \rangle \tag{19.6.9}$$

for some convex Ω. Here, we have abused notation to mimic that in (19.6.3): we take $\mathbf{1}_x$ to be the point mass at x and inner products to mean expectations, i.e.,

$$\langle \Omega'(Q), P \rangle = \mathbb{E}_P[\Omega'(Q, X)]$$

for an appropriate subgradient $\Omega'(Q, x)$, as in expression (14.2.6). We then have the equality

$$D_\Omega(P, Q) := \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)] = \Omega(P) - \Omega(Q) - \langle \Omega'(Q), P - Q \rangle,$$

exactly as in the finite-dimensional case. Therefore, as in the proof of Corollary 19.3.3, we still have the saddle point representation

$$L(Q, \pi) = \int D_\Omega(P_\theta, Q) d\pi(\theta)$$

and the infimal representation (19.3.6) that $\inf_Q L(Q, \pi) = L(P_\pi, \pi) = I_\ell(\pi; X)$, and so at least insofar as setting up the problem, finiteness of $\mathcal{X}$ is immaterial because of the proper loss representation in Theorem 14.2.8.

The other places in which we use finite dimensionality all involve continuity of divergence measures and in completeness of collections of distributions. The statement of Lemma 19.6.1 assumes that the expected regret (Bregman divergence) $D_\Omega(P_\theta, Q)$ is lower semicontinuous along one-dimensional directions in Q, and inequality (19.6.6) requires lower semicontinuity of the loss ℓ, which we may replace with lower semicontinuity of $D_\Omega(P_\theta, Q)$ in Q. The remaining three uses of finite dimensionality all regard completeness: equation (19.6.5) uses that if Q_n is Cauchy for the Bregman divergence, i.e., $\sup_{k \ge 1} D_\Omega(Q_n, Q_{n+k}) \to 0$, then $Q_n \to Q^\star$ for some distribution $Q^\star$. We use this same completeness to argue that if $Q_n \to Q^\star$ and $D_\Omega(Q_n, \widetilde{Q}_n) \to 0$, then $\widetilde{Q}_n \to Q^\star$ as well in the derivation of inequality (19.6.7). And finally, the convergence (19.6.8) uses that $D_\Omega(Q_n, \widetilde{Q}) \to 0$ and $D_\Omega(Q_n, Q) \to 0$ implies $Q = \widetilde{Q}$, which proves that $Q^\star$ is unique.

Thus, while we will require somewhat careful topological and metric space considerations, to prove a rigorous version of Theorem 19.3.1 for arbitrary domains requires demonstrating only two conditions relating to modes of convergence of probability distributions:

(C.1) The divergence D_Ω is complete for the desired mode of convergence: if P_n is Cauchy for D_Ω, there exists P such that $P_n \to P$, and if $D_\Omega(P_n, P) \to 0$ and $D_\Omega(P_n, Q) \to 0$, then $P = Q$.

(C.2) For each $\theta \in \Theta$, the divergence $D_\Omega(P_\theta, Q)$ is lower semicontinuous in Q for the convergence mode in Condition (C.1): if $Q_n \to Q$ then $\liminf_n D_\Omega(P_\theta, Q_n) \geq D_\Omega(P_\theta, Q)$.

If we can show both conditions (C.1) and (C.2), then Theorem 19.3.1 will be airtight. This requires a detour into the convergence of distributions in different topologies, all of which coincide when the space $\mathcal{X}$ is finite. (Appendix A.2 discusses different types of distributional convergence on metric spaces.) We focus on two modes of convergence: in total variation and in distribution.

Convergence in total variation Convergence in total variation provides a more straightforward way to verify the above desiderata, thus extending Corollary 19.3.3 to arbitrary spaces. Let $\mathcal{X}$ be an arbitrary space and P_n and P be probability measures on $\mathcal{X}$. Then we say that

$$P_n \to P \text{ in total variation if } \|P_n - P\|_{\rm TV} \to 0.$$

Total variation is particularly convenient for convergence of distributions, as it is complete: if P_n is a Cauchy sequence for total variation, then it necessarily has a limit distribution P for which $\|P_n - P\|_{\rm TV} \to 0$. (See Lemma A.2.4 in Appendix A.2.) Because we frequently work with divergences other than the variation distance, it is useful to allow other completeness notions.

Definition 19.1. *Let $\mathcal{P}$ be a collection of probability distributions on $\mathcal{X}$ and $D : \mathcal{P} \times \mathcal{P} \to \overline{\mathbb{R}}_+$. Then D is* complete for $\mathcal{P}$ *if*

i. Whenever $P_n \in \mathcal{P}$ is Cauchy for D, there exists $P \in \mathcal{P}$ for which $\|P_n - P\|_{\rm TV} \to 0$, and

ii. If $D(P_n, Q_n) \to 0$, then $\|P_n - Q_n\|_{\rm TV} \to 0$.

For the KL-divergence, Pinsker's inequality essentially immediately provides completeness:

Example 19.6.3: Let $D(P, Q) = D_{\rm kl}(P\|Q)$. Then Pinsker's inequality (Proposition 2.2.9) implies that $D(P, Q) \geq 2\|P - Q\|_{\rm TV}^2$. Many choices for the set $\mathcal{P}$ are possible here so long as $\mathcal{P}$ is closed for the variation distance. If $\mathcal{P}$ consists of the collection of all distributions on $\mathcal{X}$, then the KL-divergence is complete. Similarly, if $\mathcal{P}$ consists of all distributions on $\mathcal{X}$ absolutely continuous with respect to a base measure ν, Definition 19.1 is satisfied as well (again, see Lemma A.2.4 in Appendix A.2).

The KL-divergence corresponds to the logarithmic loss $\ell(Q, x) = -\log q(x)$, as we have seen several times. Example 14.2.11 treats the case of general spaces $\mathcal{X}$. $\diamond$

The divergence $D(P, Q) = D_{\rm kl}(P\|Q)$ and saddle objective $L(Q, \pi) = \int D_{\rm kl}(P_\theta\|Q)\, d\pi(\theta)$ are well-defined regardless of the domain $\mathcal{X}$, so that the redundancy/capacity theorem holds so long as we can verify the lower semicontinuity (C.2). For this, we have the following corollary of the Donsker-Varadhan representation of the KL-divergence (Theorem 6.1.1):

Corollary 19.6.4 (Semicontinuity of KL-divergence)**.** *Let $\|P_n - P\|_{\rm TV}$ and $\|Q_n - Q\|_{\rm TV} \to 0$. Then $\liminf_n D_{\rm kl}(P_n\|Q_n) \geq D_{\rm kl}(P\|Q)$.*

Proof Let $g(x) = \sum_{i=1}^k \alpha_i \mathbf{1}\{x \in A_i\}$ for scalars α_i and measurable sets A_i. Then clearly $\mathbb{E}_{P_n}[g] \to \mathbb{E}_P[g]$ and $\mathbb{E}_{Q_n}[e^g] \to \mathbb{E}_Q[e^g]$ by the convergence in total variation. So

$$D_{\rm kl}(P_n\|Q_n) \geq \mathbb{E}_{P_n}[g] - \log \mathbb{E}_{Q_n}[e^g] \to \mathbb{E}_P[g] - \log \mathbb{E}_Q[e^g].$$

Taking a supremum over simple functions g achieves the KL-divergence on the right. □

Combining this corollary with the discussion in Example 19.6.3, we see we have satisfied desiderata (C.2) and (C.1), and have thus proved the redundancy/capacity duality in Corollary 19.3.2.

Convergence in distribution Let $\mathcal{X}$ be a metric space (we leave the metric implicit). A sequence of distributions[1] P_n on $\mathcal{X}$ *converge in distribution* to a distribution P, which we denote

$$P_n \overset{d}{\rightsquigarrow} P,$$

if for every bounded continuous function $f : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_{P_n}[f(X)] \to \mathbb{E}_P[f(X)].$$

When $\mathcal{X} \subset \mathbb{R}^d$, this is equivalent to convergence of cumulative distribution functions continuity points: letting $F_n(t) = P_n(X \preceq t)$ and $F(t) = P(X \preceq t)$ be CDFs, $P_n \overset{d}{\rightsquigarrow} P$ if and only if

$$F_n(t) \to F(t)$$

at all continuity points of F. Appendix A.2 discusses these modes of convergence.

Especially in infinite-dimensional problems, it can be easier to obtain completeness for convergence in distribution rather than in total variation (Definition 19.1).

Definition 19.2. *The divergence measure $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ is* complete for $\mathcal{P}$ in distribution *if*

i. Whenever $P_n \in \mathcal{P}$ is Cauchy for D, there exists $P \in \mathcal{P}$ for which $P_n \overset{d}{\rightsquigarrow} P$

ii. If $P_n \overset{d}{\rightsquigarrow} P$ and $D(P_n, Q_n) \to 0$, then $Q_n \overset{d}{\rightsquigarrow} P$.

To distinguish this from Definition 19.1, consider predicting cumulative distributions, where natural losses include the continuous ranked probability score (CRPS) in Examples 14.2.6 and 14.2.10.

Example 19.6.5: For distributions P and Q on $\mathbb{R}^d$, let

$$D(P, Q) = \int \left(P(X \preceq t) - Q(X \preceq t)\right)^2 dt$$

be the average squared distance between their cumulative distribution functions (CDFs). Let $\mathcal{P}$ be the collection of all (Borel) probability distributions on a compact set $T \subset \mathbb{R}^d$. Then Proposition A.2.7 in Appendix A.2 shows that this divergence is complete for convergence in distribution: we have $D(P_n, P) \to 0$ if and only if $P_n \overset{d}{\rightsquigarrow} P$, and P_n is Cauchy for the Bregman divergence if and only if it has a limit $P_n \overset{d}{\rightsquigarrow} P$. That is, it satisfies Definition 19.2.

The divergence corresponds to a multivariate version of the CRPS loss $\ell_{\text{crps}}(Q, y) := \int (Q(Y \preceq t) - \mathbf{1}\{y \preceq t\})^2 dt$, slightly generalizing Examples 14.2.6 and 14.2.10. ◇

Inspecting Definition 19.2 and Example 14.2.10, we see that both of the desiderata (C.1) and (C.2) hold: the semicontinuity follows because for $Q_n \overset{d}{\rightsquigarrow} Q$,

$$\left|D(P, Q_n)^{1/2} - D(P, Q)^{1/2}\right| \le D(Q, Q_n)^{1/2} \to 0$$

by the triangle inequality and the example. The completeness we have shown.

[1] We assume all probabilities are Borel measures, meaning they are defined on the Borel sets.

19.6.5 A formal statement of regret/capacity duality

With the examples above in mind, we can provide a formal statement that makes Theorem 19.3.1 precise. All that we really need is an appropriate assumption on the losses ℓ to guarantee some type of completeness. Recall the representation (19.6.9) of a proper loss, so that for the negative generalized entropy Ω we have Bregman divergence

$$D_\Omega(P, Q) = \mathbb{E}_P[\ell(Q, X)] - \mathbb{E}_P[\ell(P, X)].$$

Definition 19.3. *Let $\{P_\theta\}_{\theta\in\Theta}$ be a collection of distributions on $\mathcal{X}$, Π be a convex collection of distributions on Θ, and let $\mathcal{P} \supset \{P_\pi\}_{\pi\in\Pi}$. The loss ℓ is* identifying for $\mathcal{P}$ *if*

i. it is strictly proper;

ii. for the negative generalized entropy Ω in its representation (19.6.9), either

a. the divergence D_Ω is complete for $\mathcal{P}$ in total variation (Definition 19.1), or

b. the divergence D_Ω is complete for $\mathcal{P}$ in distribution (Definition 19.2);

iii. for each $\theta \in \Theta$, $D_\Omega(P_\theta, Q)$ is lower semicontinuous in Q for the convergence mode in item ii.

By our discussion so far, this is precisely what is needed for a non-informal version of Theorem 19.3.1. Recalling the definition (19.3.5) of the saddle objective

$$L(Q, \pi) := \int \left(\mathbb{E}_{P_\theta}[\ell(Q, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) d\pi(\theta),$$

we have the following theorem.

Theorem 19.6.6. *Let $\{P_\theta\}_{\theta\in\Theta}$ be a collection of distributions on $X \in \mathcal{X}$ and Π be a convex collection of prior distributions on $\theta \in \Theta$. Let ℓ be identifying for $\mathcal{P} \supset \{P_\pi\}_{\pi\in\Pi}$ (Definition 19.3), and assume the capacity*

$$C_\ell(\Pi) := \sup_{\pi\in\Pi} I_\ell(\pi; X) < \infty.$$

Then there exists a unique $Q^\star \in \mathcal{P}$ such that

$$\sup_{\pi\in\Pi} L(Q^\star, \pi) = \sup_{\pi\in\Pi} I_\ell(\pi; X).$$

If additionally $\Pi = \Pi(\Theta)$ consists of all distributions on Θ, then

$$\sup_{\theta\in\Theta} \left(\mathbb{E}_{P_\theta}[\ell(Q^\star, X)] - \mathbb{E}_{P_\theta}[\ell(P_\theta, X)]\right) = C_\ell(\Pi(\Theta)) = \sup_{\pi\in\Pi(\Theta)} I_\ell(\pi; X).$$

Lastly, if $\pi^\star$ achieves the capacity $C_\ell(\Pi)$, then $Q^\star = P_{\pi^\star}$, and $(P_{\pi^\star}, \pi^\star)$ is a saddle point:

$$\sup_{\pi\in\Pi} L(P_{\pi^\star}, \pi) \le L(P_{\pi^\star}, \pi^\star) \le \inf_Q L(Q, \pi^\star).$$

Clearly, Theorem 19.6.6 shows that we have the general regret/capacity duality (19.3.7).

19.7 Bibliography and further reading

The two perspectives this chapter presents—one arising out of a robust Bayesian view and the other from coding—both have enjoyed substantial work. Within the information theory literature, Davisson [130] introduced the problem of universal coding—to obtain encodings whose lengths were asymptotically no longer than an optimal code—and Gallager [203] first proved the redundancy-capacity theorem (duality) in Corollary 19.3.2 for finitely supported distributions. Rissanen [360] extends these results to show that in a sense, the behavior of parametric families in Theorem 19.5.1 is the rule: all except for a vanishing fraction of sources P_θ in the class $\{P_\theta\}_{\theta\in\Theta}$ have $\frac{d}{2}\log n$ lower bounds on the regret. Merhav and Feder [314] and Haussler [231] present versions of the redundancy/capacity theorem for general alphabets (i.e. spaces $\mathcal{X}$).

A universal encoding scheme asymptotically achieves average length at most the entropy rate $\lim_n \frac{1}{n}H(X_1^n)$ of the process (assuming it exists). Several effective algorithms achieve these guarantees, including Shannon-Fano-Elias [118, Ch. 5.9] and Lempel-Ziv [441] coding; Cover and Thomas [118, Chapter 13] succintly overviews the area. The broader question of whether sequences are compressible typically reposes on Kolmogorov complexity or minimum description length principles; see Grünwald's book [219].

Similar questions arise from the statistics literature. Reference analysis, which we mention in Section 19.5.1, suggests performing Bayesian analyses by choosing a prior π to maximize the information in an experiment. Bernardo [64] surveys the field, and Berger and collaborators [60, 61] provide elegant treatments of "objective" Bayesian analysis. Clarke and Barron [112, 113] prove rigorous versions of the heuristics we sketch in Section 19.5, making tight connections between Bayesian statistical methods and information theory.

The development of the regret/capacity duality theorems in Section 19.3 follow a few strands of research. Grünwald and Dawid [220] develop generalized maximum entropy/robust Bayesian duality by considering the maximum entropy problems in Section 19.1, and they provide general theorems that guarantee the min-max duality (19.1.2) even beyond the logarithmic loss, by building off of work of Csiszár [123] and Topsøe [397, 398]. Their maximum entropy/robust Bayesian principles, which apply for general losses, directly lead to the material in Section 19.1. The relativized losses we adopt in Sections 19.3 and 19.6* allow us to avoid some subtleties in that (generalized) entropies $H(P) = \inf_Q \mathbb{E}_P[\ell(Q, X)]$ often require a base measure to be sensible, and they sometimes yield infinite losses, which the capacity formulation $L(Q, \pi) = \int D_\Omega(P_\theta, Q)d\pi(\theta)$ avoids. Grünwald and Dawid [220, Sec. 9] provide similar minimax theorems by appealing to infinite-dimensional duality theory. Our proof of Theorem 19.6.6 closely follows that of Polyanskiy and Wu [346, Theorem 5.9], who prove the classical redundancy capacity theorem. They also show the neat proof of Corollary 19.3.7 as a consequence of the redundancy/capacity theorem, which Corollary 19.3.8 mimics. The formulation of proper losses via the Gneiting-Raftery-Savage representation shows redundancy/capacity duality extends far beyond the log-loss and KL-divergence.

19.8 Exercises

Exercise 19.1: This exercise gives one proof of Corollary 19.3.9. Let $L(G, F) = \mathbb{E}_F[\ell_{\rm crps}(G, X)]$ and $\mathcal{F}$ be the collection of CDFs on $[t_0, t_1]$, as in the corollary.

(a) Show that $\widehat{F}$ corresponding to the uniform distribution on the endpoints $\{t_0, t_1\}$ is an equalizer rule, so that $L(\widehat{F}, F) = \frac{1}{4}$ for all F.

(b) Use Corollary 19.3.4 and saddle point duality (Proposition 19.1.1) to conclude Corollary 19.3.9.

Exercise 19.2 (Minimax redundancy and loss functions)**:** Assume that $X_i \stackrel{\text{iid}}{\sim} \mathsf{Bernoulli}(p)$. We consider three different loss functions, and the associated expected regret, for measuring the accuracy of predictions of these X_i. Redefine the expected regret (making the loss ℓ explicit)

$$\mathsf{EReg}_n(\widehat{\theta}, P, \ell) := \sum_{i=1}^n \mathbb{E}_P[\ell(\widehat{\theta}(X_1^{i-1}), X_i)] - \inf_\theta \sum_{i=1}^n \mathbb{E}_P[\ell(\theta, X_i)], \tag{19.8.1}$$

where $\widehat{\theta}$ is a predictor based on $X_1, \ldots, X_{i-1}$ at time i. Define $S_i = \sum_{j=1}^i X_j$ to be the partial sum up to time i. For each of parts (a)–(c), at time i use the predictor

$$\widehat{\theta}_i = \widehat{\theta}(X_1^{i-1}) = \frac{S_{i-1} + \frac{1}{2}}{i}.$$

(a) Loss function: $\ell(\theta, x) = \frac{1}{2}(x - \theta)^2$. Show that $\mathsf{EReg}_n(\widehat{\theta}, P, \ell) \le C \cdot \log n$ where C is a constant.

(b) Loss function: $\ell(\theta, x) = x \log \frac{1}{\theta} + (1 - x) \log \frac{1}{1-\theta}$, the usual log loss for predicting probabilities. Show that $\mathsf{EReg}_n(\widehat{\theta}, P, \ell) \le C \cdot \log n$ whenever the true probability $p \in (0, 1)$, where C is a constant. *Hint.* Give a prior π for which $\widehat{\theta}$ is a Bayes strategy.

(c) Loss function: $\ell(\theta, x) = |x - \theta|$. Show that $\mathsf{EReg}_n(\widehat{\theta}, P, \ell) \ge c \cdot n$, where $c > 0$ is a constant depending on p, whenever the true probability $p \notin \{0, \frac{1}{2}, 1\}$.

Exercise 19.3: Let the conditions of Exercise 19.2 hold. Show that there is a numerical constant $c > 0$ such that for any procedure $\widehat{\theta}$, the worst-case expected regret $\sup_{p \in [0,1]} \mathsf{EReg}_n(\widehat{\theta}, \mathsf{Bernoulli}(p), \ell) \ge c\sqrt{n}$ for the absolute loss ℓ in Exercise 19.2.(c). Give a strategy attaining this redundancy.

Exercise 19.4: Use Stirling's approximation [118, Lemma 17.5.1] that for $0 < p < 1$ with $np \in \mathbb{N}$, $\sqrt{\pi np(1-p)}\binom{n}{np} \le \exp(nh_2(p)) \le \sqrt{8np(1-p)}\binom{n}{np}$ to complete the details in the calculations for Example 19.4.2. *Hint.* Recall the Riemann sum for the integral $\int_0^1 \theta^{-1/2}(1-\theta)^{-1/2} d\theta$.

Exercise 19.5: Complete the details in the calculations for Example 19.5.3.

Exercise 19.6 (Universal prediction via the log loss)**:** Let $\ell : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ be a loss on actions $\mathcal{A}$ and outcomes $\mathcal{X}$, and consider a sequential game in which a player plays predictive distributions $Q(\cdot \mid X_1^{i-1})$ and chooses

$$\widehat{A}_i \in \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_Q[\ell(a, X_i) \mid X_1^{i-1}].$$

Let $\{P_\theta\}_{\theta \in \Theta}$ be a collection of distributions on sequences $\{X_i\}_{i \in \mathbb{N}}$ and define the expected regret $\mathsf{EReg}_n(Q, P, \ell) = \sum_{i=1}^n \mathbb{E}_P[\ell(\widehat{A}_i, X_i) - \ell(A_i^\star, X_i)]$, where $A_i^\star = \operatorname{argmin}_{a \in \mathcal{A}} \mathbb{E}_P[\ell(a, X_i) \mid X_1^{i-1}]$ is the P-optimal action. Assume the bound $\ell(a, x) - \ell(a^\star, x) \le L$ for all $a, a^\star \in \mathcal{A}$ and $x \in \mathcal{X}$.

(a) Show that

$$\mathsf{EReg}_n(Q, P, \ell) \le 2L \sum_{i=1}^n \mathbb{E}\left[\left\|P(\cdot \mid X_1^{i-1}) - Q(\cdot \mid X_1^{i-1})\right\|_{\text{TV}}\right].$$

(b) Show that if P^n and Q^n denote the (not necessarily) product distributions induced on $X_1^n \sim P$ and $X_1^n \sim Q$, then

$$\mathsf{EReg}_n(Q, P, \ell) \le L\sqrt{2n}\sqrt{D_{\text{kl}}\left(P^n \| Q^n\right)}.$$

(c) Conclude that if Q has average redundancy (19.2.3) satisfying $\frac{1}{n}\mathsf{Red}_n(Q, P) = n^{-1}D_{\rm kl}\left(P^n\|Q^n\right) \to 0$, then the expected regret satisfies $n^{-1}\mathsf{ERreg}_n(Q, P, \ell) \to 0$, that is, sublinear redundancy implies sublinear regret for bounded losses.

Exercise 19.7 (Strong redundancy [360, 314])**:** Assume that for a given $\theta \in \Theta$ we draw $X \sim P_\theta$. For a probability measure π on Θ, define the Bayes redundancy for a family $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ as

$$C^\pi := \inf_Q \int D_{\rm kl}\left(P_\theta\|Q\right) d\pi(\theta) = I_\pi(T; X),$$

where $T \sim \pi$, conditional on $T = \theta$ we draw $X \sim P_\theta$, and I_π denotes the mutual information. Define the maximin redundancy $C^* := \sup_\pi C^\pi$. This exercise shows that for "most" θ under the prior π, no distribution Q can have substantially better redundancy than the mixture $\overline{Q} = \int P_\theta d\pi(\theta)$.

Consider a distribution Q on $\mathcal{X}$ and let $\epsilon \in [0, 1]$, and define the ϵ-improvement set

$$B_\epsilon := \left\{\theta \in \Theta : D_{\rm kl}\left(P_\theta\|Q\right) \le (1 - \epsilon)C^*\right\}.$$

(a) Show that for any prior π,

$$\pi(B_\epsilon) \le \frac{\log 2 + C^* - I_\pi(T; X)}{\epsilon C^*}.$$

(b) Assume that π attains the supremum in the definition of C^*. Show that

$$\pi(B_\epsilon) \le O(1) \cdot \exp(-\epsilon\, C^*).$$

Hint. Introduce the random variable Z to be 1 if the random variable $T \in B_\epsilon$ and 0 otherwise, then use that $Z \to T \to X_1^n$ forms a Markov chain, and expand the mutual information. For part (b), the inequality $\frac{1-x}{x}\log\frac{1}{1-x} \le 1$ for all $x \in [0, 1]$ may be useful.

Exercise 19.8 (Mixtures versus point distributions)**:** Let P be a $\mathsf{Laplace}(\lambda)$ distribution on $\mathbb{R}$ with density

$$p(x) = \frac{\lambda}{2}\exp(-\lambda|x|).$$

Let $X_1, \ldots, X_n \stackrel{\rm iid}{\sim} P$ and P^n denote the n-fold product of P. We compare the predictive performance of distributions from the normal location family $\mathcal{P} = \{\mathsf{N}(\theta, \sigma^2)\}_{\theta\in\mathbb{R}}$ with the mixture Q^π over $\mathcal{P}$ defined by the normal prior $\mathsf{N}(\mu, \tau^2)$, that is, $\pi(\theta) = (2\pi\tau^2)^{-1/2}\exp(-(\theta - \mu)^2/2\tau^2)$.

(a) Let $P_{\theta,\Sigma}$ be the multivariate normal distribution with mean $\theta \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$. Give $D_{\rm kl}\left(P^n\|P_{\theta,\Sigma}\right)$.

(b) Show that $\inf_{\theta\in\mathbb{R}^n} D_{\rm kl}\left(P^n\|P_{\theta,\Sigma}\right) = D_{\rm kl}\left(P^n\|P_{0,\Sigma}\right)$, that is, the mean-zero normal distribution has the smallest KL-divergence from the Laplace distribution.

(c) Let Q_n^π be the mixture of the n-fold products in $\mathcal{P}$, that is, Q_n^π has density

$$q_n^\pi(x_1^n) = \int_{-\infty}^{\infty} \pi(\theta)p_\theta(x_1)\cdots p_\theta(x_n)d\theta,$$

where π is $\mathsf{N}(0, \tau^2)$. What is $D_{\rm kl}\left(P^n\|Q_n^\pi\right)$?

(d) Show that the redundancy of Q_n^π under the distribution P is asymptotically nearly as good as the redundancy of any $P_\theta \in \mathcal{P} = \{\mathsf{N}(\theta, \sigma^2)\}_{\theta \in \mathbb{R}}$. That is, show that

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_P \left[\log \frac{1}{q_n^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] \le \mathrm{Const}(\tau^2, \sigma^2, \mu^2) \cdot \log n$$

for any prior mean μ and variance $\tau^2 > 0$, where the Const term hides dependence on τ^2, σ^2, μ^2.

Exercise 19.9: Prove Theorem 19.1.3.

Exercise 19.10: This question shows that the maximum entropy/robust Bayesian duality of Section 19.1 frequently holds for distributions on finite domains. Let $\Delta_k = \{p \in \mathbb{R}_+^k \mid \langle \mathbf{1}, p \rangle = 1\}$ and identify $X \sim p$ as elements of $\mathbb{R}^k$, so $X = e_j$ with probability p_j. Let ℓ be a strictly proper loss and $H(p) = \inf_q \mathbb{E}_p[\ell(q, X)]$ (and $\Omega = -H$) be the associated (negative) entropy. Assume that Ω is Legendre (Definition 14.4) and $\Omega(p) > -\infty$ for $p \in \Delta_k$.

(a) Show that if $\mathcal{P} \subset \Delta_k$ is closed, then $p^\star = \operatorname{argmin}_{p \in \mathcal{P}} \Omega(p)$ exists, is unique, and satisfies $p^\star \in \operatorname{relint} \mathcal{P}$. *Hint.* See Corollary C.3.10 in Appendix C.3.3.

(b) Use the conjugate loss $\varphi(s, x) = -s^\top x + \Omega^*(s)$ to show that

$$\inf_s \sup_{p \in \Delta_k} \mathbb{E}_p[\varphi(s, X)] = \sup_{p \in \Delta_k} H(p),$$

and moreover, there exists a pair $(s^\star, p^\star)$ with $p^\star \in \operatorname{relint} \Delta_k$ and $p^\star = \nabla \Omega^*(s^\star)$ for which

$$\sup_{p \in \Delta_k} \mathbb{E}_p[\varphi(s^\star, X)] \le \mathbb{E}_{p^\star}[\varphi(s^\star, X)] \le \inf_s \mathbb{E}_{p^\star}[\varphi(s, X)],$$

and $p^\star$ maximizes the entropy $H(p)$.

(c) Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta} \subset \Delta_k$ be a collection of distributions indexed by θ, which need not be closed, and let $P_\pi = \int P_\theta d\pi(\theta)$ be the marginal that π induces. Let Π be all distributions on Θ. Show that

$$\inf_q \sup_{\pi \in \Pi} \int \mathbb{E}_{P_\theta}[\ell(q, X)] d\pi(\theta) = \sup_{\pi \in \Pi} H(P_\pi).$$

(d) Argue that if Π consists of all distributions as in part (c), there exists $\pi^\star$ maximizing $H(P_\pi)$, and that $Q = P_{\pi^\star}$ solves the generalization

$$\operatorname*{minimize}_q \sup_{\pi \in \Pi} \int \mathbb{E}_{P_\theta}[\ell(q, X)] d\pi(\theta)$$

of the robust Bayesian game (19.1.1).

Part V

Appendices

Appendix A

Miscellaneous mathematical results

This appendix collects measure-theoretic and probabilistic facts used throughout the book. We include proofs when a convenient reference is unavailable or when the argument is short and useful.

A.1 Integral convergence and approximation

We record several results on the convergence of integrals. These allow exchanging integration and differentiation and discussing completeness of probability spaces. We prove only those for which we do not know standard references. We presume knowledge of Lebesgue integration, and for vector (Banach-space)-valued functions f, "measurable" means strongly measurable, and use the Bochner integral to define $\int f(x)d\mu(x)$. (That is, we define $\int f d\mu$ for simple functions, and perform the usual Lebesgue construction of approximating $\int f(x)d\mu(x)$ via Cauchy sequences of integrals of simple functions $s_n \to f$; see, for example, Cohn [115, Appendix E].) We typically leave the σ-algebra necessary for definitions implicit.

The first two results are standard; see, for example, Royden [365].

Proposition A.1.1 (Extended Fatou's lemma)**.** *Let f_n be a sequence of $\mathbb{R}$-valued μ-measurable functions and g be measurable with $f_n \geq g$. Let $f(x) = \liminf_n f_n(x)$, and let μ be a measure, and assume $\int |g(x)|d\mu(x) < \infty$. Then f is measurable, and*

$$\int f(x)d\mu(x) \leq \liminf \int f_n(x)d\mu(x).$$

Definition A.1 (Notions of convergence)**.** *Let $f_n : X \to Y$, where (X, μ) is a measure space and Y is a Banach space, be a sequence of measurable functions, and let $f : X \to Y$ be measurable.*

i. The sequence f_n converges in measure *to f if $\mu(\{x \in X \mid \|f_n(x) - f(x)\| > \epsilon\}) \to 0$.*

ii. The sequence f_n converges μ-almost everywhere *(denoted μ-a.e.) to f, or* μ-almost-surely*, if there exists a measurable null set $N \subset X$ with $\mu(N) = 0$, and $f_n(x) \to f(x)$ for each $x \in X \setminus N$.*

If μ is finite, μ-a.e. convergence implies convergence in measure: if $A_n = \{x \mid \|f_n(x) - f(x)\| \geq \epsilon\}$, then $N_\epsilon = \cap_m \cup_{n \geq m} A_n$ consists of those points x for which $\|f_n(x) - f(x)\| \geq \epsilon$ infinitely often. If $f_n \to f$ μ-a.e., then $\mu(N_\epsilon) = 0$, and continuity of measure implies $\lim_{m\to\infty} \mu(\cup_{n\geq m} A_n) = 0$.

Any book on graduate level real analysis (implicitly) contains these next three results (see, e.g., Tao [391, Chapters 1.3 and 1.10], Royden [365], or Rudin [366]).

Lemma A.1.2 (Monotone convergence). *Let $f_n : \mathcal{X} \to \mathbb{R}_+$ be measurable and satisfy $f_n(x) \uparrow f(x)$ for every x. Then, for every measure μ,*

$$\int f \, d\mu = \lim_n \int f_n \, d\mu,$$

where either side may be infinite.

Theorem A.1.3 (Dominated convergence). *Let $f_n, f : X \to Y$ where (X, μ) is a measure space and Y is a Banach space, where f_n are measurable. Assume that $f_n \to f$ for μ-almost every x, and that there exists an integrable dominating function g such that $g(x) \geq \|f_n(x)\|$ for all x and n, where $\int g d\mu < \infty$. Then*

$$\int \|f_n(x) - f(x)\| \, d\mu(x) \to 0 \quad \text{and} \quad \int f_n(x) d\mu(x) \to \int f(x) d\mu(x)$$

If μ is finite, it is enough to assume that $f_n \to f$ in measure.

The following extension of Scheffé's lemma appears in scalar form as [407, Proposition 2.29].

Proposition A.1.4 (An extended Scheffé's lemma). *Let $f_n : X \to Y$ be measurable, where (X, μ) is a measure space and Y is a Banach space, and let $1 \leq p < \infty$. Assume $f_n \to f$ for μ-almost every x, and that $\limsup_n \int \|f_n(x)\|^p \, d\mu(x) \leq \int \|f(x)\|^p \, d\mu(x) < \infty$. Then $\int \|f_n(x) - f(x)\|^p \, d\mu(x) \to 0$. If μ is finite, it is enough to assume that $f_n \to f$ in measure.*

Proof By convexity of the norm, $\|f_n(x) - f(x)\|^p \leq 2^p \|f_n(x)\| + 2^p \|f(x)\|^p$, so

$$0 \leq 2^p \|f_n(x)\|^p + 2^p \|f(x)\|^p - \|f_n(x) - f(x)\|^p \to 2^{p+1} \|f(x)\|^p$$

μ-almost everywhere (or in μ-measure). Then Fatou's lemma that $\int h d\mu \leq \liminf_n \int h_n d\mu$ for nonnegative h_n converging to h μ-almost everywhere implies

$$\begin{aligned} 2^{p+1} \int \|f(x)\|^p \, d\mu(x) &\leq \liminf_n \int \left(2^p \|f_n(x)\|^p + 2^p \|f(x)\|^p - \|f_n(x) - f(x)\|^p\right) d\mu(x) \\ &\leq 2^{p+1} \int \|f(x)\|^p \, d\mu(x) - \limsup_n \int \|f_n(x) - f(x)\|^p \, d\mu(x). \end{aligned}$$

Rearrange to obtain the limit. If μ is finite and $f_n \to f$ in measure, every subsequence has a further subsequence converging almost everywhere. Apply the preceding argument to that further subsequence and use the subsequence criterion to prove the result. □

Corollary A.1.5. *Let Q_n and Q be probability distributions on a countable set $\mathcal{X}$ with probability mass functions q_n and q. If $q_n(x) \to q(x)$ for every $x \in \mathcal{X}$, then $\|Q_n - Q\|_{\rm TV} \to 0$.*

Proof Observe that $\sum_x q_n(x) = 1$ and $\sum_x q(x) = 1$, so Proposition A.1.4 implies $\sum_x |q_n(x) - q(x)| \to 0$. Of course, $\|Q_n - Q\|_{\rm TV} = \frac{1}{2} \sum_x |q_n(x) - q(x)|$. □

The next approximation result, which shows that compactly supported functions are dense on nice enough probability spaces, is useful in Chapter 15. The result follows as a consequence of regularity of Borel measures and Lipschitz approximations to indicators of closed sets; see, e.g., [391, Proposition 1.10.4].

Lemma A.1.6 (Density of Lipschitz functions). *Let $\mathcal{X}$ be a complete separable metric (Polish) space, let P be a Borel probability on $\mathcal{X}$, and let $1 \leq p < \infty$ and $k \in \mathbb{N}$. Let $\mathcal{C}_c^{\rm Lip}$ be the collection of compactly supported Lipschitz functions $\mathcal{X} \to \mathbb{R}^k$. Then $\mathcal{C}_c^{\rm Lip}$ is dense in $L^p(P)$: for each $\epsilon > 0$ and f with $\mathbb{E}_P[\|f(X)\|^p] < \infty$, there exists $g \in \mathcal{C}_c^{\rm Lip}$ with $\mathbb{E}_P[\|g(X) - f(X)\|^p]^{1/p} \leq \epsilon$.*

A.2 Probabilistic convergence

In this appendix, we collect some of the distributional ocnvergence results we use. Billingsley's books [70, 69] and van der Vaart and Wellner [408] provide a much wider ranging treatment and foundation.

A.2.1 Classical results on convergence in distribution

Let $(\mathcal{X}, \rho)$ be a metric space. A sequence of probability distributions P_n on $\mathcal{X}$ *converges in distribution* or *weakly* to P, written $P_n \overset{d}{\rightsquigarrow} P$, if

$$\mathbb{E}_{P_n}[f(X)] \to \mathbb{E}_P[f(X)]$$

for all bounded continuous $f : \mathcal{X} \to \mathbb{R}$. The Portmanteau theorem provides several characterizations (see, e.g. [70, Chapter 1.2] or [408, Chapter 1.3]). Because we assume P_n and P are probability distributions inducing random elements X_n and X, we do not need to address measurability issues [408, Chapter 1.3].

Theorem A.2.1 (Portmanteau)**.** *Let $(\mathcal{X}, \rho)$ be a metric space and P_n, P be probability distributions on $\mathcal{X}$. The following are all equivalent to the convergence in distribution $P_n \overset{d}{\rightsquigarrow} P$.*

(i) $\mathbb{E}_{P_n}[f(X)] \to \mathbb{E}_P[f(X)]$ *for all bounded,* 1*-Lipschitz continuous* $f : \mathcal{X} \to \mathbb{R}$.

(ii) For all upper semicontinuous $f : \mathcal{X} \to \mathbb{R}$ *bounded from above,* $\limsup_n \mathbb{E}_{P_n}[f(X)] \le \mathbb{E}_P[f(X)]$.

(iii) For all lower semicontinuous $f : \mathcal{X} \to \mathbb{R}$ *bounded from below,* $\liminf_n \mathbb{E}_{P_n}[f(X)] \ge \mathbb{E}_P[f(X)]$.

(iv) For all open sets $O \subset \mathcal{X}$, $\liminf_n P_n(O) \ge P(O)$.

(v) For all closed sets $C \subset \mathcal{X}$, $\limsup_n P_n(C) \le P(C)$.

(vi) For all continuity sets A *of* P*, meaning sets for which the boundary* $\operatorname{bd} A = \operatorname{cl} A \setminus \operatorname{int} A$ *satisfies* $P(\operatorname{bd} A) = 0$, $\lim_n P_n(A) = P(A)$.

If additionally $\mathcal{X} \subset \mathbb{R}^d$*, write* $F_n(t) = P_n(X \preceq t)$ *and* $F(t) = P(X \preceq t)$*, where* $\preceq$ *is coordinatewise. Then* $P_n \overset{d}{\rightsquigarrow} P$ *is also equivalent to*

(vii) $F_n(t) \to F(t)$ *for all continuity points* t *of* F.

Prokhorov's theorem characterizes compactness for the topology of convergence in distribution (see [70, Chapter 1.5], [408, Chapter 1.3], or [165, Chapter 11]). We require a family of Borel probability measures defined on the same σ-algebra $\mathcal{A}$. Then $\mathcal{P}$ is *tight* if for all $\epsilon > 0$, there is a compact K such that $P(K) \ge 1 - \epsilon$ for all $P \in \mathcal{P}$.

Theorem A.2.2 (Prokhorov)**.** *Let $(\mathcal{X}, \rho)$ be a complete separable metric space and $\mathcal{P}$ be a collection of Borel measures on $\mathcal{X}$. The following are equivalent.*

(i) The class $\mathcal{P}$ is tight.

(ii) Every sequence in $\mathcal{P}$ has a subsequence converging in distribution to a Borel probability on $\mathcal{X}$.

If $\mathcal{P}$ is also weakly closed, it is sequentially compact for convergence in distribution.

When $\mathcal{X}$ is separable and complete, the Lévy-Prokhorov distance metrizes weak convergence. For a (Borel) set $A \subset \mathcal{X}$, define its ϵ-enlargment $A^\epsilon := \{x \in \mathcal{X} \mid \rho(x, A) < \epsilon\}$, where $\rho(x, A) = \inf_{y \in A} \rho(x, y)$. Then for probability distributions P, Q, define

$$d_{\prob}(P, Q) := \inf \left\{\epsilon \geq 0 \mid P(A) \leq Q(A^\epsilon) + \epsilon \text{ and } Q(A) \leq P(A^\epsilon) + \epsilon\right\}.$$

This metrizes convergence in distribution on separable complete metric spaces [70, Theorem 6.8].

Theorem A.2.3. *Let $\mathcal{X}$ be a complete separable metric space. Then*

$$d_{\prob}(P_n, P) \to 0 \text{ if and only if } P_n \overset{d}{\rightsquigarrow} P.$$

The space of Borel probability measures on $\mathcal{X}$ is complete and separable in the $d_{\prob}$ metric.

A.2.2 Assorted convergence results for probability distributions

In this appendix, we collect a few results that follow from the basic definitions of convergence of probability measures. We include them for completeness, because though they are more or less standard, we do not know specific references.

First, collections of probability distributions on a set are essentially always complete for the total variation distance.

Lemma A.2.4. *Let P_n be probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$ and assume P_n is Cauchy for total variation. Then*

(i) There is a probability measure P such that $\|P_n - P\|_{\TV} \to 0$.

(ii) If every P_n is absolutely continuous with respect to a measure ν, then P is absolutely continuous with respect to ν.

Proof We show claim (i) first. For each $A \in \mathcal{F}$, $P_n(A)$ is Cauchy and has a limit we call $P(A)$, which is finitely additive with $P(\emptyset) = 0$ and $P(\mathcal{X}) = 1$. To see that that P is countably additive, let $A_1, A_2, \ldots$ be a collection of disjoint measurable sets, and let $\epsilon > 0$ be arbitrary. For $m \in \mathbb{N}$, define the tail sets $B_m = \cup_{i > m} A_i$. For all $m \in \mathbb{N}$, finite additivity implies

$$P\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^m P(A_i) + P(B_m).$$

Now, there is some N such that $n \geq N$ implies

$$\sup_k \sup_m |P_n(B_m) - P_{n+k}(B_m)| \leq \sup_k \|P_n - P_{n+k}\|_{\TV} \leq \epsilon$$

because P_n is Cauchy for the total variation. Taking $k \to \infty$, we obtain that if $n \geq N$,

$$|P_n(B_m) - P(B_m)| \leq \epsilon \text{ for all } m \in \mathbb{N}.$$

For any fixed n, the probability of the tail sets $P_n(B_m) \to 0$ as $m \to \infty$. Take m large enough that $P_n(B_m) \leq \epsilon$. Then $0 \leq P(B_m) \leq 2\epsilon$, so P satisfies

$$\sum_{i=1}^m P(A_i) \leq \sum_{i=1}^m P(A_i) + P(B_m) = P\left(\bigcup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^m P(A_i) + 2\epsilon$$

for large m. As $\epsilon > 0$ was arbitrary, taking $m \to \infty$ gives that P is indeed countably additive and so a probability measure.

To show that $\|P_n - P\|_{\rm TV} \to 0$, fix $\epsilon > 0$ and let N be large enough that $\|P_n - P_{n+k}\|_{\rm TV} \le \epsilon$ for all $n \ge N$ and $k \ge 0$. Let A_n satisfy $\|P_n - P\|_{\rm TV} = P_n(A_n) - P(A_n)$. Then

$$\begin{aligned}\|P_n - P\|_{\rm TV} = \|P_n - P_{n+k} + P_{n+k} - P\|_{\rm TV} &\le |P_n(A_n) - P_{n+k}(A_n)| + |P_{n+k}(A_n) - P(A_n)| \\ &\le \|P_n - P_{n+k}\|_{\rm TV} + \underbrace{|P_{n+k}(A_n) - P(A_n)|}_{\to 0},\end{aligned}$$

the last limit holding as $k \to \infty$. This implies $\|P_n - P\|_{\rm TV} \to 0$.

For the final claim of the lemma, take a set A such that $\nu(A) = 0$. Then $P_n(A) = 0$ for all n, and thus $P(A) = 0$. □

We finish with a Cramér-von Mises-type distance between multivariate cumulative distribution functions. We first collect two results on the sets of discontinuities of functions.

Lemma A.2.5. *Let $f : X \to Y$ be any function between metric spaces (X, ρ_X) and (Y, ρ_Y), and let $\mathbb{B}_\delta(x) = \{z \in X \mid \rho_X(x, z) < \delta\}$ be the open δ-ball around x. Define the set*

$$\mathsf{D}_\epsilon(f) := \{x \in X \mid \text{for all } \delta > 0 \text{ there exist } y, z \in \mathbb{B}_\delta(x) \text{ s.t. } d_Y(f(y), f(z)) \ge \epsilon\}$$

of ϵ-discontinuities of f. Then $\mathsf{D}_\epsilon(f)$ is closed, and the set

$$\mathsf{D}(f) := \{x \in X \mid f \text{ is discontinuous at } x\}$$

is Borel-measurable.

Proof We show that $\mathsf{D}_\epsilon(f)$ is closed. Let x_∞ be a limit point of $\mathsf{D}_\epsilon(f)$. If x_∞ is isolated, meaning that there is some $\delta_0 > 0$ such that $(\mathbb{B}_{\delta_0}(x_\infty) \setminus \{x_\infty\}) \cap \mathsf{D}_\epsilon(f) = \emptyset$, then clearly $x_\infty \in \mathsf{D}_\epsilon(f)$. Otherwise, let $\delta > 0$ be arbitrary. Then there is some $x \in \mathsf{D}_\epsilon(f)$ with $x \in \mathbb{B}_\delta(x_\infty) \setminus \{x_\infty\}$ by definition of the limit point, and because $\mathbb{B}_\delta(x_\infty) \setminus \{x_\infty\}$ is open, there exists $\delta_0 > 0$ such that $\mathbb{B}_{\delta_0}(x) \subset \mathbb{B}_\delta(x_\infty) \setminus \{x_\infty\}$. As $x \in \mathsf{D}_\epsilon(f)$, there are $y, z \in \mathbb{B}_{\delta_0}(x)$ with $\rho_Y(f(y), f(z)) \ge \epsilon$. But then $y, z \in \mathbb{B}_\delta(x_\infty)$, and as δ was arbitrary we have $x_\infty \in \mathsf{D}_\epsilon(f)$.

Because $\mathsf{D}(f) = \cup_{n \ge 1} \mathsf{D}_{1/n}(f)$ is the countable union of closed sets it is Borel. □

Lemma A.2.6. *Let F be the cumulative distribution function of $X \in \mathbb{R}^d$. The set $\mathsf{D}(F)$ of discontinuities of F is Borel measurable and has (Lebesgue) measure 0.*

Proof The set of discontinuity points of a one-dimensional CDF F is necessarily countable: for $\epsilon > 0$, the set $\mathsf{D}_\epsilon(F) := \{t \in \mathbb{R} \mid F(t) \ge \limsup_{\delta \downarrow 0} F(t - \delta) + \epsilon\}$ has cardinality at most $1/\epsilon < \infty$, so

$$\mathsf{D}(F) := \left\{t \in \mathbb{R} \mid \lim_{\delta \downarrow 0} F(t - \delta) < F(t)\right\} = \bigcup_{n \in \mathbb{N}} \mathsf{D}_{1/n}(F).$$

Let F be the cumulative distribution function of $X \in \mathbb{R}^d$ and $F_1, \ldots, F_d$ be the (marginal) cumulative distributions of $X_1, \ldots, X_d$. Let $t, t' \in \mathbb{R}^d$, and define the elementwise maximum $t \vee t' = [\max\{t_j, t'_j\}]_{j=1}^d$ and similarly the elementwise minimum $t \wedge t'$. By monotonicity of the CDFs,

$$\begin{aligned}|F(t) - F(t')| &\le F(t \vee t') - F(t \wedge t') = \mathbb{P}(t \wedge t' \prec X \preceq t \vee t') \\ &\le \sum_{j=1}^d \mathbb{P}\left(\min\{t_j, t'_j\} < X_j \le \max\{t_j, t'_j\}\right) = \sum_{j=1}^d \left(F_j(\max\{t_j, t'_j\}) - F_j(\min\{t_j, t'_j\})\right).\end{aligned}$$

so a point t is a discontinuity of F only if it is a discontinuity for at least one of the marginal CDFs F_j. That is, we have shown that the set $\mathsf{D}(F)$ of discontinuities of F satisfies

$$\mathsf{D}(F) \subset \bigcup_{j=1}^{d} \left\{ \mathbb{R}^{j-1} \times \mathsf{D}(F_j) \times \mathbb{R}^{d-j} \right\},$$

which is Borel measurable by Lemma A.2.5. If λ denotes d-dimensional Lebesgue measure,

$$\lambda(\mathsf{D}(F)) \le \sum_{j=1}^{d} \lambda\left(\mathbb{R}^{j-1} \times \mathsf{D}(F_j) \times \mathbb{R}^{d-j}\right) = \sum_{j=1}^{d} \sum_{t \in \mathsf{D}(F_j)} \lambda\left(\mathbb{R}^{j-1} \times \{t\} \times \mathbb{R}^{d-j}\right) = 0,$$

because the Lebesgue measure of any lower-dimensional subset of $\mathbb{R}^d$ is 0. □

We now can show that $D_{\rm cdf}$ metrizes convergence in distribution on compacts.

Proposition A.2.7. *Let $T \subset \mathbb{R}^d$ be compact and $\mathcal{P}_T$ be the collection of Borel probability measures supported on T. For measures $P, Q \in \mathcal{P}_T$ define*

$$D_{\rm cdf}(P, Q) = \int (P(X \preceq t) - Q(X \preceq t))^2 dt.$$

Then for $P_n, P \in \mathcal{P}_T$, $P_n \overset{d}{\rightsquigarrow} P$ if and only if $D_{\rm cdf}(P_n, P) \to 0$. Thus $D_{\rm cdf}^{1/2}$ is a complete metric for convergence in distribution on $\mathcal{P}_T$.

Proof We first show that if $P_n \overset{d}{\rightsquigarrow} P$, then $D_{\rm cdf}(P_n, P) \to 0$. Recall from Lemma A.2.6 that the points of discontinuity of any cumulative distribution function $t \mapsto P(X \preceq t)$ have measure 0 (and are also measurable). The functions $F_n(t) = P_n(X \preceq t)$ and $F(t) = P(X \preceq t)$ are bounded and thus satisfy $F_n - F \to 0$ almost everywhere by the Portmanteau theorem. Because T is compact, dominated convergence implies that $D_{\rm cdf}(P_n, P) \to 0$.

Now we show that if $D_{\rm cdf}(P_n, P) \to 0$, then $P_n \overset{d}{\rightsquigarrow} P$. By Prokhorov's theorem (Thm. A.2.2), any subsequence $P_{n(m)}$ has a further subsequence that converges weakly to some limit $Q \in \mathcal{P}_T$. Then by the triangle inequality,

$$D_{\rm cdf}(P, Q)^{1/2} \le D_{\rm cdf}(P, P_{n(m)})^{1/2} + D_{\rm cdf}(P_{n(m)}, Q)^{1/2} \to 0,$$

so $Q = P$ by right continuity of P and Q. Using the standard topological result that if for a sequence P_n, every subsequence has a further subsequence converging to P, then P_n converges to P, we have the result.

Lastly, we show completeness. Let P_n be Borel measures and Cauchy for $D_{\rm cdf}$. Then by Prokhorov's theorem (Theorem A.2.2), because T is compact, every subsequence of P_n admits a further subsequence with $P_{n(m)} \overset{d}{\rightsquigarrow} Q$ for some $Q \in \mathcal{P}_T$. The first part of the theorem gives $D_{\rm cdf}(P_{n(m)}, Q) \to 0$, while the Cauchy property implies $D_{\rm cdf}(P_n, Q) \to 0$. □

Appendix B

Convex analysis: sets and functions

This appendix reviews results in convex analysis. Convex analysis and optimization have become practically important in the last fifty years because convex optimization problems—those in which the objective and constraints are convex—are tractable, while many others typically are not. This appendix builds analytic tools. Much of our treatment follows the original treatise by Rockafellar [363] and the elegant two volume set by Hiriart-Urruty and Lemaréchal [235, 237]. We focus mostly on the finite-dimensional case (though we present results that apply in infinite dimensional cases with proofs that extend straightforwardly, and we specify the domains of functions only if necessary). We assume throughout that our functions are *proper*, meaning that $f(x) > -\infty$ for all x, to avoid pathologies. We will use $\overline{\mathbb{R}} = (-\infty, \infty]$ to mean the upper extended real line, where the range $f : C \to \mathbb{R}$ or $f : C \to \overline{\mathbb{R}}$ denotes whether f may take infinite values on C.

B.1 Convex sets

The simplest and most important object in convex analysis is a convex set.

Definition B.1. *A set C is* convex *if for all $\lambda \in [0, 1]$ and all $x, y \in C$, we have*

$$\lambda x + (1 - \lambda)y \in C.$$

We typically work with *closed* convex sets. (See Figure B.1.)

We now consider two operations that extend sets, convexifying them.

Definition B.2. *The* affine hull *of a set C is the smallest affine set containing C. That is,*

$$\operatorname{aff}(C) := \left\{ \sum_{i=1}^{k} \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}^k, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Figure B.2 shows an example of an affine hull: it is the smallest (shifted) subspace containing C. Associated with any set is its convex hull, which Figure B.3 illustrates:

Definition B.3. *The* convex hull *of a set $C \subset \mathbb{R}^d$, denoted* $\operatorname{Conv}(C)$*, is the intersection of all convex sets containing C.*

An almost immediate associated result is that the convex hull of a set is equal to the set of all convex combinations of points in the set. (This result also holds in infinite dimensions.)

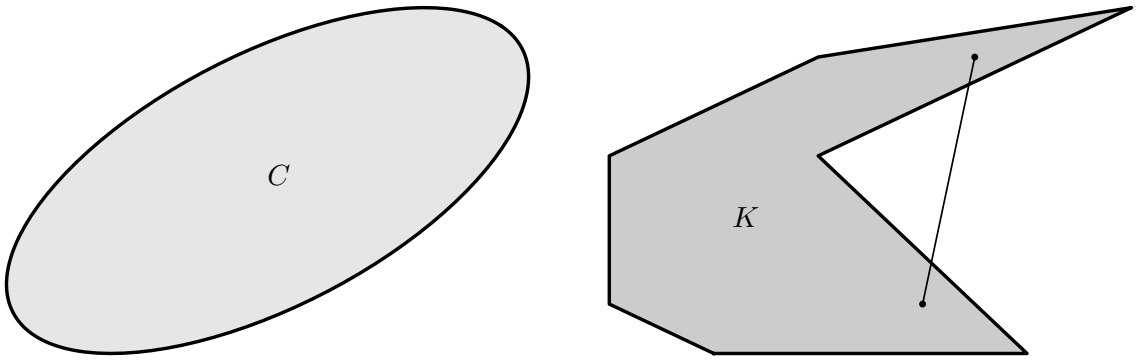


Figure B.1. Left: a convex set C. Right: a non-convex set K with a segment connecting two points in the set that does not remain within K.

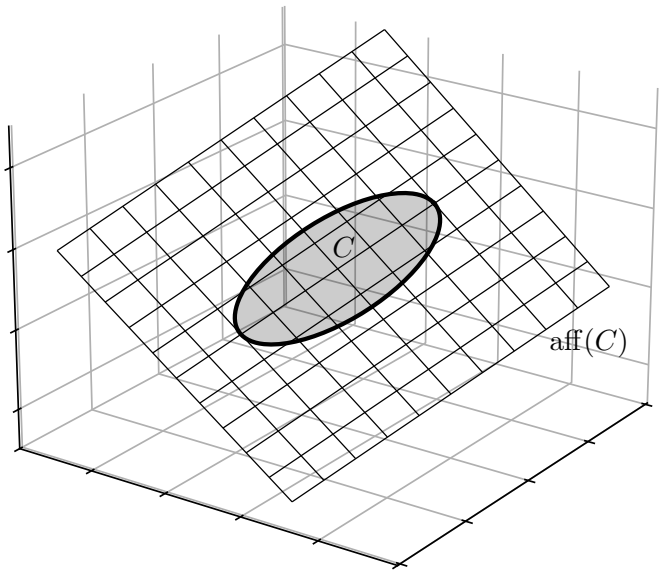


Figure B.2: The affine hull of the set C as a (shifted) subspace of $\mathbb{R}^3$.

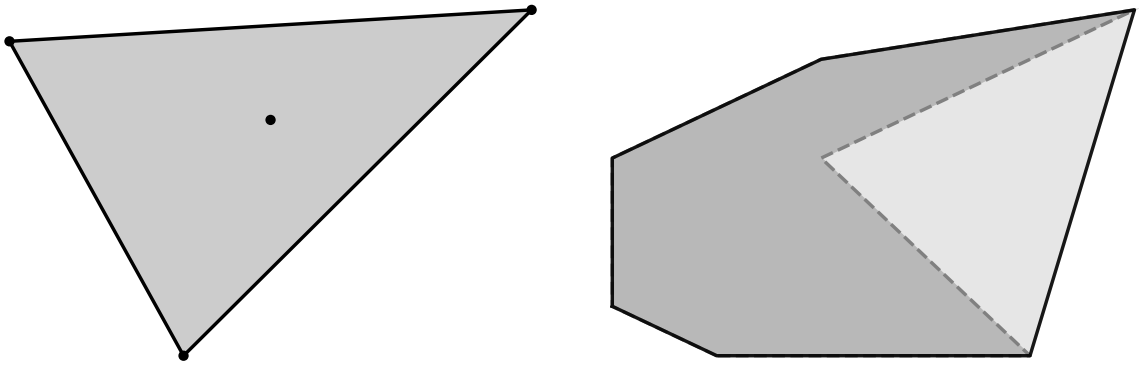


Figure B.3. Left: the convex hull of four points. Right: the convex hull of the polygon K from Figure B.1.

Proposition B.1.1. *Let C be an arbitrary set. Then*

$$\mathrm{Conv}(C) = \left\{ \sum_{i=1}^k \lambda_i x_i : k \in \mathbb{N}, x_i \in C, \lambda \in \mathbb{R}_+^k, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

Proof Call the set on the right side of the equality T. That $T \supset C$ is clear, as we may take $\lambda_1 = 1$ and vary $x \in C$. Moreover, $T \subset \mathrm{Conv}(C)$, as any convex set containing C must contain all convex combinations of its elements; similarly, any convex set $S \supset C$ must have $S \supset T$.

Thus it remains to show that T is convex. Take two points $x, y \in T$. Then $x = \sum_{i=1}^k \alpha_i x_i$ and $y = \sum_{i=1}^l \beta_i y_i$ for some $x_i, y_i \in C$. Fix $\lambda \in [0, 1]$. Then $(1 - \lambda)\beta_i \geq 0$ and $\lambda \alpha_i \geq 0$ for all i,

$$\lambda \sum_{i=1}^k \alpha_i + (1 - \lambda) \sum_{i=1}^l \beta_i = \lambda + (1 - \lambda) = 1,$$

and $\lambda x + (1 - \lambda) y$ is a convex combination of the points x_i and y_i weighted by $\lambda \alpha_i$ and $(1 - \lambda)\beta_i$, respectively. So $\lambda x + (1 - \lambda) y \in T$ and T is convex. □

In dimension $d < \infty$, *Carathéodory's theorem* states that we may restrict the definition of the convex hull of a set C to convex combinations of a at most $d + 1$ of the points in C, rather than the arbitrary convex combinations in Proposition B.1.1.

Theorem B.1.2. *Let $C \subset \mathbb{R}^d$. Then $x \in \mathrm{Conv}(C)$ if and only if there exist $x_1, \ldots, x_{d+1} \in C$ and $\lambda \in \mathbb{R}_+^{d+1}$ with $\sum_{i=1}^{d+1} \lambda_i = 1$ such that*

$$x = \sum_{i=1}^{d+1} \lambda_i x_i.$$

Proof It is clear that if x can be represented as such a sum, then $x \in \mathrm{Conv}(C)$. Conversely, Proposition B.1.1 implies that for any $x \in \mathrm{Conv}(C)$ we have

$$x = \sum_{i=1}^k \lambda_i x_i, \quad \lambda_i \geq 0, \ \sum_{i=1}^k \lambda_i = 1, \ x_i \in C$$

for some λ_i, x_i. Assume that $k > d+1$ and $\lambda_i > 0$ for each i, as otherwise, there is nothing to prove. Then the points $x_i - x_1$ are linearly dependent, and we can find (not identically zero) values $\alpha_2, \ldots, \alpha_k$ such that $\sum_{i=2}^k \alpha_i (x_i - x_1) = 0$. Let $\alpha_1 = -\sum_{i=2}^k \alpha_i$ to obtain that we have both

$$\sum_{i=1}^k \alpha_i x_i = 0 \quad \text{and} \quad \sum_{i=1}^k \alpha_i = 0. \tag{B.1.1}$$

The equalities (B.1.1) imply that at least one $\alpha_i > 0$, and if we define $\lambda^* = \min_{i:\alpha_i > 0} \frac{\lambda_i}{\alpha_i} > 0$, then setting $\lambda_i' = \lambda_i - \lambda^* \alpha_i \geq 0$, we have

$$\sum_{i=1}^k \lambda_i' = \sum_{i=1}^k \lambda_i - \lambda^* \sum_{i=1}^k \alpha_i = 1, \quad \text{and} \quad \sum_{i=1}^k \lambda_i' x_i = \sum_{i=1}^k \lambda_i x_i - \lambda^* \sum_{i=1}^k \alpha_i x_i = x.$$

But we know that at least one of the $\lambda_i' = 0$, so that we could write x as a convex combination of $k-1$ elements. Repeating this strategy until $k = d+1$ gives the theorem. □

B.1.1 Operations preserving convexity

We touch on operations that preserve convexity of convex sets.

Observation B.1.3. *Let C be a convex set. Then $C = \mathrm{Conv}(C)$.*

Observation B.1.4. *Let $\{C_\alpha\}_{\alpha \in \mathcal{A}}$ be an arbitrary collection of convex sets. Then*

$$C = \bigcap_{\alpha \in \mathcal{A}} C_\alpha$$

is convex. If C_α is closed for each α, then C is closed as well.

Observation B.1.5. *Let C be convex. Then $\mathrm{cl}(C)$ is convex.*

This is immediate: let $x_n \in C$ and $y_n \in C$ satisfy $x_n \to x$, $y_n \to y$. Then for $\lambda \in [0,1]$, we have $\lambda x_n + (1-\lambda) y_n \in C$ and $\lambda x_n + (1-\lambda) y_n \to \lambda x + (1-\lambda) y \in \mathrm{cl}(C)$.

Observation B.1.5 also implies the following result.

Observation B.1.6. *Let D be an arbitrary set. Then*

$$\bigcap \{C \text{ closed}, \ C \supset D, \ C \text{ is convex}\} = \mathrm{cl}\,\mathrm{Conv}(D).$$

Proof Let T denote the leftmost set. It is clear that $T \subset \mathrm{cl}\,\mathrm{Conv}(D)$ as $\mathrm{cl}\,\mathrm{Conv}(D)$ is closed convex (by Observation B.1.5) containing D. On the other hand, if $C \supset D$ is closed convex, then $C \supset \mathrm{Conv}(D)$, while the closedness of C implies it contains $\mathrm{cl}\,\mathrm{Conv}(D)$. Thus $T \supset \mathrm{cl}\,\mathrm{Conv}(D)$. □

Observation B.1.7. *Let $C \subset \mathbb{R}^d$ be compact. Then $\mathrm{Conv}(C)$ is compact.*

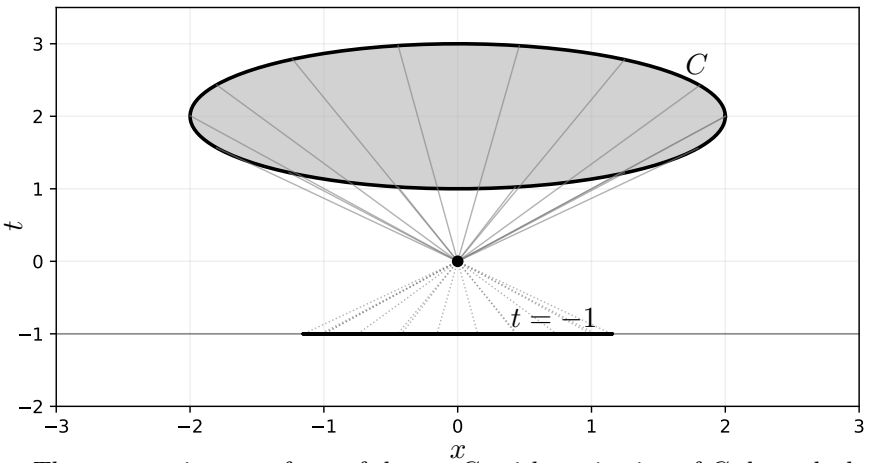


Figure B.4. The perspective transform of the set C, with projection of C through the origin $(0, 0)$. This illustrates the transformation mapping points $(x, t) \mapsto (-x/t, -1)$.

Proof Let $x^n \in \mathrm{Conv}(C)$ converge to some x. By Theorem B.1.2, for each n we can find $x_1^n, \ldots, x_{d+1}^n \in C$ and $\lambda^n \in \Delta_{d+1} = \{v \in \mathbb{R}_+^{d+1} \mid \langle \mathbf{1}, v \rangle = 1\}$ such that $x^n = \sum_{j=1}^{d+1} \lambda_j^n x_j^n$. Taking subsequences as necessary, we can assume that $x_1^n \to x_1 \in C$, as C is compact. Then taking a further subsequence, we can as well assume $x_2^n \to x_2 \in C$, and so on, so that $x_j^n \to x_j$ for each $j = 1, \ldots, d+1$. Then because Δ_{d+1} is compact, we can likewise take a (further) subsequence of λ^n so that $\lambda^n \to \lambda \in \Delta_{d+1}$. Evidently these limiting objects satisfy $x = \sum_{j=1}^{d+1} \lambda_j x_j$. □

As our last operation preserving convexity, we consider the *perspective transform* of a set. To define this, we first define the perspective function

$$\mathrm{pers} : \mathbb{R}^d \times \mathbb{R}_{>0} \to \mathbb{R}^d, \quad \mathrm{pers}(x, t) := \frac{x}{t}.$$

Then Figure B.4 illustrates the following definition.

Definition B.4. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_+$ be a set. The* perspective transform *of C is*

$$\mathrm{pers}(C) := \left\{ \frac{x}{t} : (x, t) \in C \text{ and } t > 0 \right\}.$$

This corresponds to taking all the points $z \in C$, normalizing them so their last coordinate is 1, and then removing the last coordinate. (For more on perspective functions, see Boyd and Vandenberghe [82, Chapter 2.3.3] or Hiriart-Urruty and Lemaréchal [235, Ch. IV.2.2].)

The perspective of a convex set is convex. Because convexity depends only on intervals, to see this it is nearly sufficient to consider only perspective transforms of line segments:

Lemma B.1.8. *Let $C \subset \mathbb{R}^{d+1}$ be a line segment, meaning $C = \{\lambda x + (1-\lambda)y \mid \lambda \in [0, 1]\}$, where $x_{d+1} > 0$ and $y_{d+1} > 0$. Then* $\mathrm{pers}(C) = \{\lambda \,\mathrm{pers}(x) + (1-\lambda)\,\mathrm{pers}(y) \mid \lambda \in [0, 1]\}$.

Proof Let $\lambda \in [0,1]$. Then letting $x_1^d = (x_1, \ldots, x_d)$ denote the first d components of a vector,

$$\begin{aligned}\operatorname{pers}(\lambda x + (1-\lambda)y) &= \frac{\lambda x_1^d + (1-\lambda)y_1^d}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \\ &= \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{x_1^d}{x_{d+1}} + \frac{(1-\lambda)y_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \frac{y_1^d}{y_{d+1}} \\ &= \theta \operatorname{pers}(x) + (1-\theta)\operatorname{pers}(y),\end{aligned}$$

where

$$\theta = \frac{\lambda x_{d+1}}{\lambda x_{d+1} + (1-\lambda)y_{d+1}} \in [0,1].$$

Sweeping λ from 0 to 1 sweeps $\theta \in [0,1]$, giving the result. □

Based on Lemma B.1.8, we immediately obtain the following proposition.

Proposition B.1.9. *Let $C \subset \mathbb{R}^d \times \mathbb{R}_{>0}$ be a convex set. Then* $\operatorname{pers}(C)$ *is convex.*

Proof Let $x^+, y^+ \in \operatorname{pers}(C)$, so there are $x, y \in C$ such that $x^+ = \operatorname{pers}(x)$ and $y^+ = \operatorname{pers}(y)$. Let $L = \operatorname{Conv}\{x, y\}$. Then by Lemma B.1.8, $\operatorname{pers}(L) = \operatorname{Conv}\{x^+, y^+\}$, while certainly $\operatorname{pers}(L) \subset \operatorname{pers}(C)$. Thus $\operatorname{pers}(C)$ is convex. □

B.1.2 Relative interiors

We give one more definition, which is essential in convex analysis, as it specifies the "natural" topology on convex sets and functions, allowing us to assume many sets are full-dimensional.

Definition B.5. *The* relative interior *of a set C is the interior of C relative to its affine hull, that is,*

$$\operatorname{relint}(C) := \{x \in C : B(x, \epsilon) \cap \operatorname{aff}(C) \subset C \text{ for some } \epsilon > 0\},$$

where $B(x, \epsilon) := \{y : \|y - x\| < \epsilon\}$ denotes the open ball of radius ϵ centered at x.

Two examples may make Definition B.5 clearer.

Example B.1.10 (Relative interior of a disc)**:** Consider the (convex) set

$$C = \left\{x \in \mathbb{R}^d : x_1^2 + x_2^2 \le 1, \; x_j = 0 \text{ for } j \in \{3, \ldots, d\}\right\}.$$

The affine hull $\operatorname{aff}(C) = \mathbb{R}^2 \times \{0_{d-2}\}$ is the (x_1, x_2)-plane in $\mathbb{R}^d$, while the relative interior $\operatorname{relint}(C) = \{x \in \mathbb{R}^d : x_1^2 + x_2^2 < 1\} \cap \operatorname{aff}(C)$ is the "interior" of the 2-dimensional disc in $\mathbb{R}^d$. ◇

Recall that vectors $x_1, \ldots, x_k$ are affinely independent if the augmented vectors $(1, x_i)$ are linearly independent, equivalently, no vector x_i is an affine combination of the others.

Example B.1.11 (Relative interior of simplices)**:** Let $x_1, \ldots, x_k \in \mathbb{R}^d$, $k \ge 2$, be affinely independent, $C = \operatorname{Conv}\{x_i\}_{i=1}^k$, and $X = [x_1 \; \cdots \; x_k] \in \mathbb{R}^{d \times k}$ have columns x_i. We claim that

$$\operatorname{relint} C = \left\{X\lambda \mid \lambda \in \mathbb{R}^k_{>0}, \langle \mathbf{1}, \lambda \rangle = 1\right\}.$$

To see this, let $\Delta_k = \{\lambda \in \mathbb{R}^k_+ \mid \langle \mathbf{1}, \lambda \rangle = 1\}$ be the probability simplex in $\mathbb{R}^k$, which has affine hull $\operatorname{aff}(\Delta_k) = \{\lambda \in \mathbb{R}^k \mid \langle \mathbf{1}, \lambda \rangle = 1\}$, and $A = \operatorname{aff}(C) = \{X\alpha \mid \alpha \in \operatorname{aff}(\Delta_k)\}$. For each point

$a \in A$ there is a unique $\alpha \in \mathrm{aff}(\Delta_k)$ such that $a = X\alpha$. This is a linear transformation; call $L \in \mathbb{R}^{k \times d}$ the linear mapping so that $LX\alpha = \alpha$ for $\alpha \in \mathrm{aff}(\Delta_k)$. If $x = X\lambda$ for $\lambda \in \mathbb{R}^k_{>0}$ with $\langle \mathbf{1}, \lambda \rangle = 1$, then $Lx = \lambda \succ 0$ and there is an open neighborhood U of 0 satisfying $L(x+U) \succ 0$, so $x \in \mathrm{relint}\, C$. Conversely, let $x = X\lambda$ for some $\lambda \in \Delta_k$ with an element $\lambda_j = 0$, and assume w.l.o.g. that $\lambda_1 = 0$. Let U be any open neighborhood of 0. Then there exists $\epsilon > 0$ such that $v = -\epsilon x_1 + \epsilon x_2 \in U$, while $x + v \in \mathrm{aff}(C)$. Then $L(X\lambda + \epsilon v) = \lambda + \epsilon(e_2 - e_1) \notin \Delta_k$. That is, $(x + U) \cap \mathrm{aff}(C) \not\subset C$ for any open U, and $x \notin \mathrm{relint}\, C$. $\diamond$

A specialization of this example shows that the probability simplex $\Delta_d = \mathrm{Conv}\{e_i\}_{i=1}^d$ has relative interior $\{\lambda \in \mathbb{R}^d_{>0} \mid \langle \mathbf{1}, \lambda \rangle = 1\}$. This example, coupled with a dimension counting argument, show that finite-dimensional convex sets always have non-empty relative interior:

Proposition B.1.12. *Let $C \subset \mathbb{R}^d$ be convex. Then* $\mathrm{relint}\, C \neq \varnothing$, *and* $\mathrm{aff}(C) = \mathrm{aff}(\mathrm{relint}\, C)$ *and hence* $\dim C = \dim(\mathrm{relint}\, C)$.

Proof Let $k = \dim C = \dim \mathrm{aff}(C)$. Then there exist $k+1$ affinely independent points $x_1, \ldots, x_{k+1} \in C$ with $\mathrm{aff}(C) = \mathrm{aff}\{x_1, \ldots, x_{k+1}\}$. The set $\Delta = \mathrm{Conv}\{x_1, \ldots, x_{k+1}\}$ has $\mathrm{aff}(\Delta) = \mathrm{aff}(C)$, while Example B.1.11 shows $\mathrm{relint}\, \Delta = \{\sum_{j=1}^{k+1} \lambda_j x_j \mid \langle \mathbf{1}, \lambda \rangle = 1, \lambda \succ 0\}$. $\square$

A key technical lemma involves the half-open segments $[x, y) := \{(1-t)x + ty \mid 0 \le t < 1\}$. See Figure B.5 for an illustration of the proof.

Lemma B.1.13. *Let $C \subset \mathbb{R}^d$ be a convex set, $x \in \mathrm{relint}\, C$, and $y \in \mathrm{cl}\, C$. Then $[x, y) \subset \mathrm{relint}\, C$. Moreover, there exists $\epsilon > 0$ such that $x + t(y - x) \in \mathrm{relint}\, C$ for $-\epsilon < t < 1$.*

Proof Without loss of generality we may assume $\mathrm{aff}(C) = \mathbb{R}^d$, so $\mathrm{int}\, C = \mathrm{relint}\, C$. Then there exists $\epsilon > 0$ such that $x + \epsilon\mathbb{B} \subset C$. For $0 \le t < 1$, we claim the point $x_t = (1-t)x + ty$ is in C. Indeed, let $y_n \in C$, $y_n \to y$, and let n be large enough that $\frac{t}{1-t} \|y - y_n\| < \epsilon$. Then

$$x_t = (1-t)x + ty = (1-t)\left(x + \frac{t(y - y_n)}{1-t}\right) + ty_n$$

is a convex combination of points in C. For any $x' \in x_t + (1-t)\epsilon\mathbb{B}$, there is a $u \in \mathbb{B}$ such that

$$x' = (1-t)x + (1-t)\epsilon u + t(y + 0) = (1-t)(x + \epsilon u) + ty.$$

As $x + \epsilon u \in C$, we have $x' \in C$, that is, $x_t + (1-t)\epsilon\mathbb{B} \subset C$, and so $x_t \in \mathrm{relint}\, C$.

For the second result, assume $y \neq x$ and let $v = y - x$. As $x \in \mathrm{int}\, C$, we may choose $\gamma > 0$ with $x + \gamma\mathbb{B} \subset C$. If $0 < \epsilon < \gamma/(2\|v\|)$ and $-\epsilon < t \le 0$, the point $x + tv + u \in x + \gamma\mathbb{B} \subset C$ when $\|u\| < \gamma/2$, so $x + tv \in \mathrm{int}\, C$ for $-\epsilon < t \le 0$. $\square$

Lemma B.1.13 shows that $\mathrm{relint}\, C$ is a convex set whenever C is, and Observation B.1.5 shows that $\mathrm{cl}\, C$ is convex if C is. A more sophisticated consequence of Lemma B.1.13 is the following:

Proposition B.1.14. *Let $C \subset \mathbb{R}^d$ be convex. Then the three convex sets* $\mathrm{relint}\, C$, C, *and* $\mathrm{cl}\, C$ *have the same relative interior, affine hull, and closure.*

Proof Proposition B.1.12 already shows the affine hull result. As the remaining equivalences are similar, we only show that $\mathrm{cl}\, \mathrm{relint}\, C = \mathrm{cl}\, C$. Let $y \in \mathrm{cl}\, C$ and $x \in \mathrm{relint}\, C$, which is possible in view of Proposition B.1.12. Then $[x, y) \subset \mathrm{relint}\, C$ by Lemma B.1.13, and $y = \lim_{t \to 1}(1-t)x + ty \in$

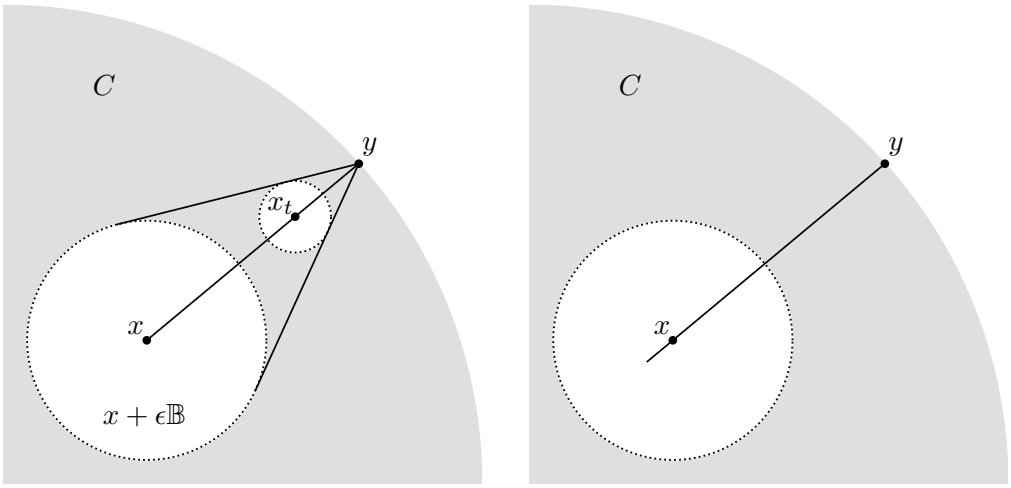


Figure B.5. Illustration of Lemma B.1.13. Left: illustration of the proof that $[x, y) \subset \operatorname{relint} C$ with point $x_t = (1-t)x + ty$. Right: stretching the line $x + t(y - x)$ a bit beyond x for $t < 0$.

$\operatorname{cl} \operatorname{relint} C$. □

Proposition B.1.14 allows us to compute the relative interior of linear transformations of convex sets, which includes as a special case a generalization of Example B.1.11.

Lemma B.1.15. *Let A be an affine mapping and C a convex set. Then* $\operatorname{relint} A(C) = A(\operatorname{relint} C)$.

Proof If C is a singleton or empty the result is trivial, so we ignore that case. Because A is continuous, $\operatorname{cl} A(S) \supset A(\operatorname{cl} S)$ for any set $S \subset \mathbb{R}^d$. So

$$A(C) \subset A(\operatorname{cl} C) = A(\operatorname{cl} \operatorname{relint} C) \subset \operatorname{cl} A(\operatorname{relint} C) \subset \operatorname{cl} A(C).$$

Thus $\operatorname{cl} A(\operatorname{relint} C) = \operatorname{cl} A(C)$, and as $A(C)$ and $A(\operatorname{relint} C)$ have the same closure, they have the same relative interior (Proposition B.1.14), and so $\operatorname{relint} A(C) = \operatorname{relint} A(\operatorname{relint} C) \subset A(\operatorname{relint} C)$.

For the converse inclusion, let $w = A(x_0)$ for some $x_0 \in \operatorname{relint} C$; we must show $w \in \operatorname{relint} A(C)$. Let $v \in \operatorname{relint} A(C)$, so that $v = A(x)$ for some $x \in C$. Then there exists $y \in C$ and $0 < t < 1$ such that $x_0 = (1-t)y + tx$ because $x_0 \in \operatorname{relint} C$, and for this y, $A(x_0) = tA(y) + (1-t)A(x) \in [A(x), A(y))$ by affinity. As $A(x) \in \operatorname{relint} A(C)$, Lemma B.1.13 implies $A(x_0) \in \operatorname{relint} A(C)$. □

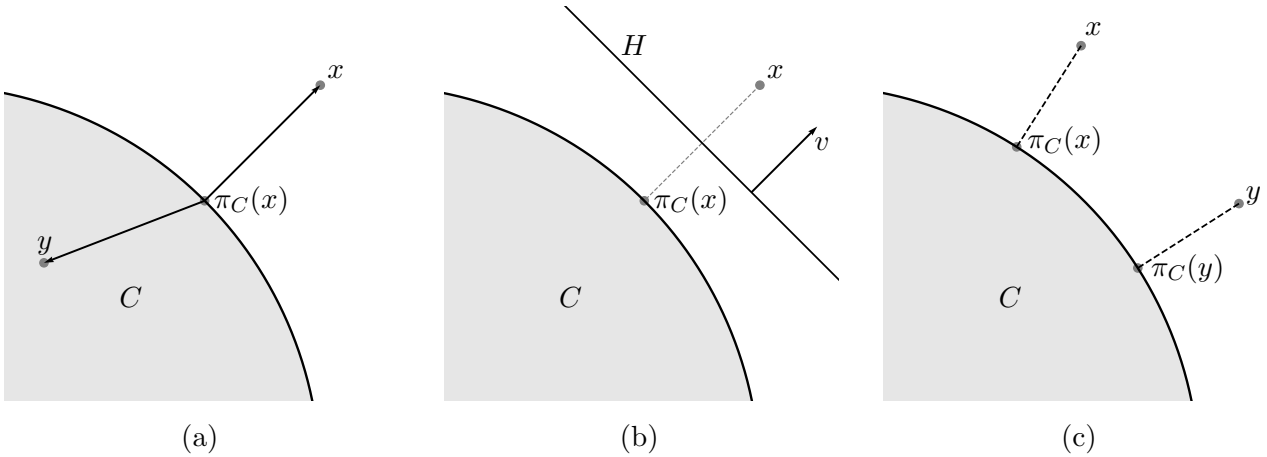
As an immediate consequence of this lemma, we can generalize Example B.1.11:

Lemma B.1.16. *Let $C = \operatorname{Conv}\{x_1, \ldots, x_k\}$ and for $\lambda \in \mathbb{R}^k_+$ with $\langle \mathbf{1}, \lambda \rangle = 1$ let $x_\lambda = \sum_{i=1}^k \lambda_i x_i$. Then* $\operatorname{relint} C = \{x_\lambda \mid \lambda \succ 0, \langle \mathbf{1}, \lambda \rangle = 1\}$.

Proof Apply Lemma B.1.15 to $\operatorname{relint} \Delta_k$. □

B.1.3 Representation and separation of convex sets

We now consider fundamental separation properties—we can separate convex sets by hyperplanes—which underpins the most important representations of convex sets. We develop the separation



(a) (b) (c)

Figure B.6. (a) The projection $\pi_C(x)$ of the point x onto C. Note that $\langle x - \pi_C(x), y - \pi_C(x)\rangle \le 0$. (b) The hyperplane H through the midpoint of the segment between x and $\pi_C(x)$ strictly separates C and x, where $H = \{y \mid \langle v, y\rangle = b\}$ for the direction $v = (x - \pi_C(x))/\left\|x - \pi_C(x)\right\|$, and $b = \langle v, x + \pi_C(x)\rangle/2 = (\left\|x\right\|^2 - \left\|\pi_C(x)\right\|^2)/2\left\|x - \pi_C(x)\right\|$. (c) Projections decrease distance: the points $\pi_C(x)$ and $\pi_C(y)$ are closer than the points x and y.

properties via the existence of projections. The projection argument is elegant in Hilbert spaces, so we state it in that setting, meaning a complete vector space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|x\| = \sqrt{\langle x, x\rangle}$. We illustrate existence in Figure B.6(a).

Theorem B.1.17. *Let C be a non-empty closed convex set. Then for any x, there exists a unique point $\pi_C(x)$ minimizing $\|y - x\|$ over $y \in C$. This point is characterized by the inequality*

$$\langle \pi_C(x) - x, y - \pi_C(x)\rangle \ge 0 \quad \text{for all } y \in C. \tag{B.1.2}$$

Proof Let $p_\star = \inf_{y\in C}\|y - x\|$ and $\{y_n\} \subset C$ be a sequence with $\|y_n - x\| \to p_\star$. We show that y_n is Cauchy. Fix $\epsilon > 0$ and let N be such that $n \ge N$ implies $\|y_n - x\|^2 \le p_\star^2 + \epsilon^2$. Let $m, n \ge N$. Then by the parallelogram identity,

$$\begin{aligned}\|y_n - y_m\|^2 = \|(x - y_n) - (x - y_m)\|^2 &= 2\left[\|x - y_n\|^2 + \|x - y_m\|^2\right] - \|(x - y_n) + (x - y_m)\|^2 \\ &= 2\left[\|x - y_n\|^2 + \|x - y_m\|^2\right] - 4\left\|x - (y_n + y_m)/2\right\|^2.\end{aligned}$$

Because $\frac{1}{2}(y_n + y_m) \in C$ by convexity, we have $\|x - (y_n + y_m)/2\|^2 \ge p_\star^2$, so

$$\|y_n - y_m\|^2 \le 2\left[p_\star^2 + \epsilon^2 + p_\star^2 + \epsilon^2\right] - 4p_\star^2 = 4\epsilon^2.$$

As $\epsilon > 0$ was arbitrary, the completeness of Hilbert space implies the first statement of the theorem, as Cauchy sequences cannot have distinct limits.

To see the second result, assume that $z \in C$ satisfies $\langle z - x, y - z\rangle \ge 0$ for all $y \in C$. If $z = x$, the result is trivial; otherwise, for any $y \in C$,

$$\|z - x\|^2 = \langle z - x, z - x\rangle = \underbrace{\langle z - x, z - y\rangle}_{\le 0} + \langle z - x, y - x\rangle \le \|z - x\|\,\|y - x\|$$

by the Cauchy-Schwarz inequality. Dividing both sides by $\|z - x\|$ yields $\|z - x\| \le \|y - x\|$ for any $y \in C$, so $z = \pi_C(x)$. Conversely, let $0 < t \le 1$ and $z = \pi_C(x)$. Then for any $y \in C$,

$$\begin{aligned}\|\pi_C(x) - x\|^2 &\le \|(1-t)\pi_C(x) + ty - x\|^2 \\ &= \|\pi_C(x) - x\|^2 + 2t\langle \pi_C(x) - x, y - \pi_C(x)\rangle + t^2 \|y - \pi_C(x)\|^2 .\end{aligned}$$

Subtracting the projection value $\|\pi_C(x) - x\|^2$ from both sides and dividing by $t > 0$, then sending $t \downarrow 0$ implies inequality (B.1.2). □

As an immediate consequence of Theorem B.1.17, we obtain several separation properties of convex sets, as well as a theorem stating that a closed convex set (not equal to the entire space in which it lies) can be represented as the intersection of all the half-spaces containing it.

Corollary B.1.18. *Let C be non-empty closed convex and $x \notin C$. Then there is a vector v strictly separating x from C, that is,*

$$\langle v, x\rangle > \sup_{y \in C} \langle v, y\rangle.$$

Moreover, $v = x - \pi_C(x)$ is one such vector.

Proof By Theorem B.1.17, taking $v = x - \pi_C(x) \neq 0$ we have

$$0 \le \langle y - \pi_C(x), \pi_C(x) - x\rangle = \langle y - \pi_C(x), -v\rangle = \langle y - x + v, -v\rangle = -\langle y, v\rangle + \langle x, v\rangle - \|v\|^2 .$$

That is, $\langle v, y\rangle \le \langle v, x\rangle - \|v\|^2$ for all $y \in C$. □

Projections also never increase distances.

Corollary B.1.19. *Let C be closed convex. Then for any x and y,*

$$\|\pi_C(x) - \pi_C(y)\| \le \|x - y\| .$$

Proof Using inequality (B.1.2) in Theorem B.1.17, observe that $\pi_C(x) \in C$ and $\pi_C(y) \in C$ so

$$0 \ge \langle \pi_C(y) - \pi_C(x), x - \pi_C(x)\rangle \quad \text{and} \quad 0 \ge \langle \pi_C(x) - \pi_C(y), y - \pi_C(y)\rangle.$$

Adding these inequalities yields

$$\langle \pi_C(x) - \pi_C(y), y - x + \pi_C(x) - \pi_C(y)\rangle \le 0,$$

or $\|\pi_C(x) - \pi_C(y)\|^2 \le \langle y - x, \pi_C(y) - \pi_C(x)\rangle \le \|y - x\| \, \|\pi_C(x) - \pi_C(y)\|$ by Cauchy-Schwarz. If $\pi_C(x) = \pi_C(y)$, the result is trivial, and otherwise, dividing by $\|\pi_C(x) - \pi_C(y)\|$ gives the result. □

Illustrations (b) and (c) in Figure B.6 show these corollaries.

In addition, we can show the existence of supporting hyperplanes, that is, hyperplanes "separating" the boundary of a convex set from itself.

Theorem B.1.20. *Let $C \subset \mathbb{R}^d$ be a convex set and $x \in \operatorname{rel}\operatorname{bd}(C) = \operatorname{cl}(C) \setminus \operatorname{relint} C$. Then there exists a non-zero $v \in \operatorname{span}(C - C)$ such that $\langle v, x\rangle \ge \sup_{y \in C}\langle v, y\rangle$. Additionally, for $x \in \operatorname{bd} C$, there exists a non-zero v such that $\langle v, x\rangle \ge \sup_{y \in C}\langle v, y\rangle$.*

Proof Let $D = \mathrm{cl}(C)$ and $x_n \in \mathrm{aff}(C) \setminus D$ with $x_n \to x$. Define the sequence of separating vectors $s_n = x_n - \pi_D(x_n)$ and the normalized version $v_n = s_n / \|s_n\|$, so $\langle v_n, x_n \rangle > \sup_{y \in C} \langle v_n, y \rangle$ for all n (Corollary B.1.18) and $v_n \in \mathrm{span}(C - C)$. By the finite-dimensionality assumption, the sequence $\{v_n\} \subset \{v \mid \|v\| = 1\}$ belongs to a compact set. Passing to a subsequence if necessary, let us assume w.l.o.g. that $v_n \to v$ with $\|v\| = 1$. Then by a standard limiting argument for the $x_n \to x$, we have

$$\langle v, x \rangle \geq \langle v, y \rangle \quad \text{for all } y \in C,$$

and as $\mathrm{span}(C - C)$ is closed, $v \in \mathrm{span}(C - C)$. Figure B.7 illustrates the construction. The general statement follows the same argument except that $x_n \in \mathbb{R}^d \setminus D$. □

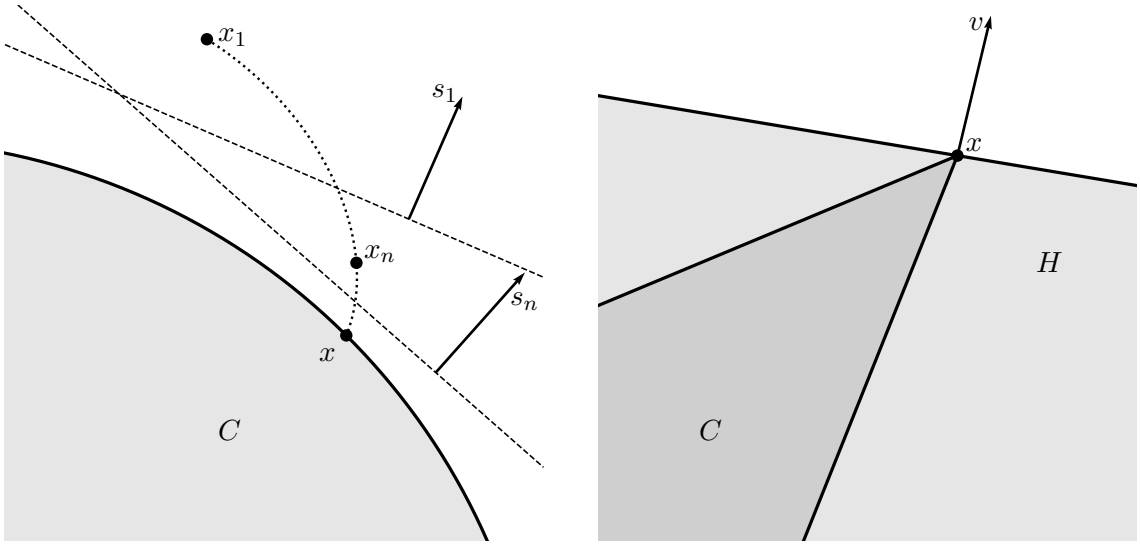


Figure B.7. Left: illustration of the proof of Theorem B.1.20. The sequence s_n converges to a vector v satisfying $\langle v, x \rangle = \sup_{y \in C} \langle v, y \rangle$. Right: the half-space H supports the set C at the point x, where $H = \{y \mid \langle v, y \rangle \leq \langle v, x \rangle\}$.

As a corollary, we have pointwise strict separation from the relative interior of convex sets:

Corollary B.1.21. *Let $C \subset \mathbb{R}^d$ be a convex set and $x \in \mathrm{rel\,bd}\, C$. Then there exists a vector v such that $\langle v, x \rangle > \langle v, y \rangle$ for all $y \in \mathrm{relint}\, C$.*

Proof By working in a subspace if necessary, we may assume $\mathrm{aff}\, C = \mathbb{R}^d$. Let v be a non-zero vector such that $\langle v, x \rangle \geq \sup_{y \in C} \langle v, y \rangle$. If $y \in \mathrm{int}\, C$, then there exists $\epsilon > 0$ such that $y + \epsilon \mathbb{B} \subset C$, and so $\langle v, x \rangle \geq \langle v, y + u \rangle$ for all $\|u\| \leq \epsilon$. Take $u = \epsilon v / \|v\|$. □

Theorem B.1.20 gives the important result that if D is an arbitrary set and $C = \mathrm{cl\,Conv}(D)$ is the closure of the convex hull of D—the smallest closed convex set containing D—then we can write C as the intersection of all the closed half-spaces containing D. To state this formally, recall that a closed half-space H is defined with respect to a non-zero vector v and real $a \in \mathbb{R}$ as

$$H := \{x \mid \langle v, x \rangle \leq a\}.$$

We say H supports C at x if $x \in C$ and H satisfies $H \supset C$ while $\langle v, x \rangle = a$. (See Figure B.7.) We have the following theorem.

Theorem B.1.22. *Let D be an arbitrary set. Then*

$$\operatorname{cl}\operatorname{Conv}(D) = \bigcap_{H \supset D} H, \tag{B.1.3}$$

where H denotes a closed half-space containing D. For a closed convex set C, let $\mathcal{H}_x$ denote the halfspaces supporting C at x. Then

$$C = \bigcap_{x \in \operatorname{bd}(C)} \bigcap_{H \in \mathcal{H}_x} H. \tag{B.1.4}$$

Proof We begin with the proof of the second result (B.1.4). By Theorem B.1.20, at each point x on the boundary of C, there exists a non-zero supporting hyperplane v, so the half-space

$$H_{x,v} := \{y \mid \langle v, y\rangle \le \langle v, x\rangle\} \supset C$$

is closed, convex, and contains C. The containment $C \subset \cap_{x \in \operatorname{bd}(C)} \cap_{H \in \mathcal{H}_x} H$ is trivial. Let $x_0 \notin C$; we show that $x_0 \notin \cap_{x \in \operatorname{bd}(C)} \cap_{H \in \mathcal{H}_x} H$. The projection $\pi_C(x_0)$ of x_0 onto C satisfies $\langle x_0 - \pi_C(x_0), x_0\rangle > \sup_{y \in C} \langle x_0 - \pi_C(x_0), y\rangle$ by Corollary B.1.18, and letting $v = x_0 - \pi_C(x_0)$, the hyperplane

$$h_{x_0,v} := \{y \mid \langle y, v\rangle = \langle \pi_C(x_0), v\rangle\}$$

supports C at the point $\pi_C(x_0)$. The half-space $\{y : \langle y, v\rangle \le \langle \pi_C(x_0), v\rangle\}$ thus contains C and does not contain x_0, implying that $x_0 \notin \cap_{x \in \operatorname{bd}(C)} \cap_{H \in \mathcal{H}_x} H$.

Now we show the result (B.1.3). Let C be the closed convex hull of D and $T = \cap_{H \supset D} H$. A trivial extension of the representation (B.1.4) yields $C = \cap_{H \supset C} H$, where H denotes any half-space containing C. As $C \supset D$, the containment $H \supset C$ implies $H \supset D$, so that

$$T = \bigcap_{H \supset D} H \subset \bigcap_{H \supset C} H = C.$$

On the other hand, Observation B.1.6 implies that any closed set containing D contains C as $C = \operatorname{cl}\operatorname{Conv}(D)$. So $H \supset D$ implies $H \supset C$, and $T = C$ as desired. □

Call a halfspace $H = \{y \mid \langle v, y\rangle \le a\}$ supporting C at x *non-trivial* if $H \supset C$, $\langle v, x\rangle = a$, and $\inf_{y \in C} \langle v, y\rangle < a$. Then we may extend the characterization (B.1.4) to include only non-trivial supporting halfspaces on the relative boundary of C; we leave the proof as Exercise B.1.

Theorem B.1.23. *Let $C \subset \mathbb{R}^d$ be a non-empty closed convex set and not an affine subspace. Let $\mathcal{H}_x$ denote the collection of non-trivial supporting halfspaces of C at x. Then*

$$C = \bigcap_{x \in \operatorname{rel}\operatorname{bd} C} \bigcap_{H \in \mathcal{H}_x} H.$$

One elegant corollary of the supporting hyperplanes for convex sets is that we can approximate convex hulls by expectations of vectors. By combining the strong law of large numbers with our descriptions of the convex hull, we have the following result.

Corollary B.1.24. *Let $X = \{x_\alpha\}_{\alpha \in \mathcal{A}} \subset \mathbb{R}^d$ be an arbitrary collection of vectors and let $\mathcal{P}$ be the collection of probability distributions on $\mathcal{A}$ for which $\mathbb{E}_P[\|x_A\|] < \infty$ when $A \sim P$. Then*

$$\operatorname{Conv}(X) = \{\mathbb{E}_P[x_A] \mid P \in \mathcal{P}\} \subset \operatorname{cl}\operatorname{Conv}(X).$$

If additionally X is compact, then $\operatorname{cl}\operatorname{Conv}(X) = \{\mathbb{E}_P[x_A] \mid P \in \mathcal{P}\}$.

Proof Let $\mathcal{M} = \{\mathbb{E}_P[x_A] \mid P \in \mathcal{P}\}$ be the collection of means. That $\operatorname{Conv}(X) \subset \mathcal{M}$ is trivial. Now fix $P \in \mathcal{P}$ and draw $A_1, A_2, \ldots, A_n \stackrel{\text{iid}}{\sim} P$. By the strong law of large numbers, $\overline{x}_n := \frac{1}{n}\sum_{i=1}^n x_{A_i} \to \mathbb{E}_P[x_A]$ with probability 1, so there exist $\overline{x}_n \in \operatorname{Conv}(X)$ satisfying $\overline{x}_n \to \mathbb{E}_P[x_A]$, and $\mathbb{E}_P[x_A] \in \operatorname{cl}\operatorname{Conv}(X)$.

Let $C = \operatorname{Conv}(X)$ for shorthand. To demonstrate that $\operatorname{Conv}(X) = \mathcal{M}$, we induct on the dimension of $\operatorname{aff} C$. When $d = 1$, C is an interval. Any random variable supported on an interval has expectation in that interval, attaining an endpoint if and only if the variable is equal to that endpoint. For the induction, let $\mu = \mathbb{E}_P[x_A]$. If $\mu \notin C$, then $\mu \in \operatorname{cl} C \setminus C$, so a supporting hyperplane to C at μ gives a nonzero vector v such that $\langle v, \mu\rangle \geq \langle v, x_\alpha\rangle$ for all $\alpha \in \mathcal{A}$. Then $\mathbb{E}_P[\langle v, x_A\rangle] = \langle v, \mu\rangle$ implies that $\langle v, x_A\rangle = \langle v, \mu\rangle$ with probability 1 over $A \sim P$, so that P has support on $X \cap H$ for $H = \{x \mid \langle v, x\rangle = \langle v, \mu\rangle\}$, whose affine hull $\operatorname{aff}(X \cap H)$ has smaller dimension. The induction hypothesis applied in H gives $\mu \in \operatorname{Conv}(X \cap H) \subset \operatorname{Conv}(X)$, a contradiction.

The final claim is simply Observation B.1.7. □

The second inclusion in Corollary B.1.24 can be strict even in one dimension: let $x_\alpha = \alpha$ for $\alpha \in (0, 1)$, so that $\operatorname{Conv}(X) = X$, and any distribution P on α yields $\mathbb{E}_P[x_A] \in (0, 1)$.

B.1.4 Cones attached to a convex set

Recall that a cone K is any set closed under positive scaling, so $x \in K$ implies $tx \in K$ for all $t > 0$. If $0 \in K$, then $tx \in K$ for all $t \geq 0$. While cones admit an extraordinary duality theory, we only require a few standard results. First, the *dual cone* K^* to a cone K is

$$K^* := \{y \mid \langle y, x\rangle \geq 0 \text{ for all } x \in K\}.$$

As K^* is the intersection of closed halfspaces, it is a closed convex set, and clearly it is a cone: if $y \in K^*$, then $ty \in K^*$ for all $t \geq 0$. Moreover, we have the following duality characterization, obtained by taking the dual to the dual cone:

Proposition B.1.25. *Let K be a set. Then $K = K^{**}$ if and only if K is a closed convex cone.*

Proof If $K = K^{**}$, then it is immediate that K is a closed convex cone. Now, assume that K is a closed convex cone. To see that $K \subset K^{**}$, observe that if $x \in K$ then $\langle x, y\rangle \geq 0$ for all $y \in K^*$ by definition of K^*, that is, $x \in K^{**}$ definitionally. If $x \notin K$, then Corollary B.1.18 guarantees there exists $v \neq 0$ such that $\langle v, x\rangle > \sup_{y \in K}\langle v, y\rangle$, and as K is a cone, it must be the case that $\langle v, y\rangle \leq 0$ for all $y \in K$. But then $-v \in K^*$, and we have $\langle -v, x\rangle < 0$, so that $x \notin K^{**}$. □

With this simple result stated, we can now define the two cones attached to any convex set. The most important for us is the *normal cone* to the set C, which is (essentially) the collection of vectors pointing out of C. Figure B.8 illustrates the concept.

Definition B.6. *Let C be a closed convex set. The* normal cone *to C at the point $x \in C$ is*

$$\mathcal{N}_C(x) := \{v \mid \langle v, y - x\rangle \leq 0 \text{ for all } y \in C\}.$$

We define $\mathcal{N}_C(x) = \emptyset$ for $x \notin C$.

So $\mathcal{N}_C(x)$ is the collection of vectors making an obtuse angle with any direction into the set C from x. It is clear that $\mathcal{N}_C(x)$ is a cone. It is closed convex, being the intersection of halfspaces. Moreover, if $x \in \operatorname{int} C$, then $\mathcal{N}_C(x) = \{0\}$, so the interesting points are $x \in \operatorname{bd} C$. We can connect supporting hyperplanes of C to its normal cones: Theorem B.1.20 gives the following corollary.

Corollary B.1.26. *Let C be non-empty closed convex. Then for any $x \in \mathrm{bd}(C)$, the normal cone $\mathcal{N}_C(x)$ is non-trivial and consists of the collection of vectors v for which the half-space $H = \{y \mid \langle v, y\rangle \le \langle v, x\rangle\} \supset C$, that is, is a supporting half-space to C at x.*

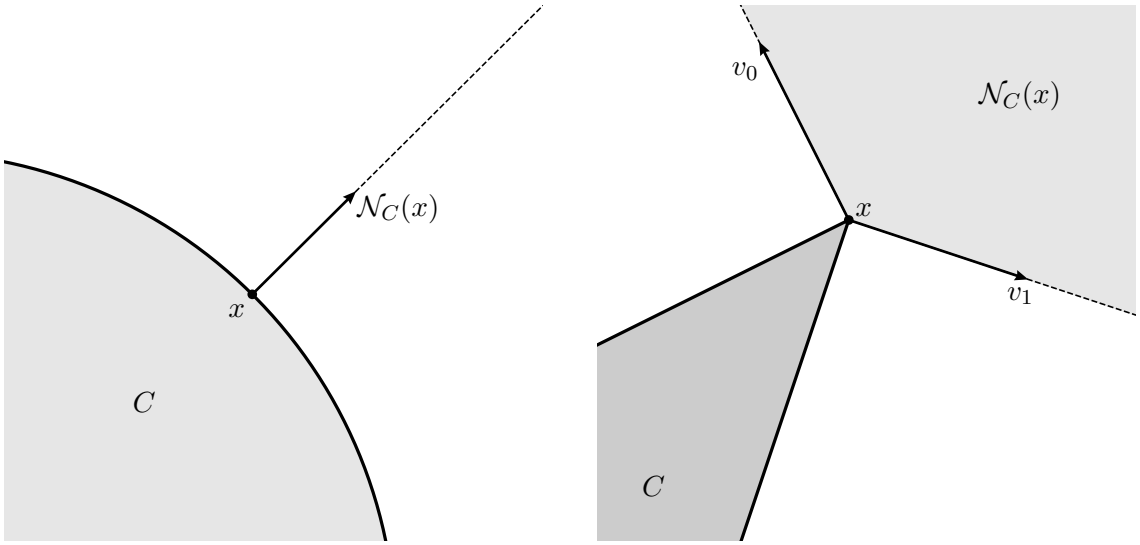


Figure B.8. Left: the normal cone to the set $C = \{y \in \mathbb{R}^2 \mid \|y\|_2 \le 1\}$ at the point x. Note that $\mathcal{N}_C(x)$ is a ray in this case. Right: the normal cone to the polyhedral set C at the point x consists of all points in the conic hull $\{t_0 v_0 + t_1 v_1 \mid t \succeq 0\}$ of the vectors v_0 and v_1 normal to C at x.

We also have the following proposition, which follows by chasing definitions:

Proposition B.1.27. *Let C be closed convex, $x \in C$, and v be a vector. Then the following properties are equivalent:*

(i) $v \in \mathcal{N}_C(x)$,

(ii) $\langle v, x\rangle = \sup_{y \in C} \langle v, y\rangle$, *that is, v supports C at x, and*

(iii) $x = \pi_C(x + v)$, *that is, x is the projection of $x + v$ onto C.*

Most texts instead begin with tangent cones, then develop the normal cone from these; we only use the normal cone, and develop the tangent cone in Exercises B.4–B.6.

B.2 Convex functions

We build off of the definitions of convex sets to define convex functions, which therefore inherit properties from the geometric (separation) properties of convex sets. First, we have

Definition B.7. *A function f is* convex *if for all $\lambda \in [0, 1]$ and $x, y \in \mathrm{dom}\, f$,*

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda) f(y). \tag{B.2.1}$$

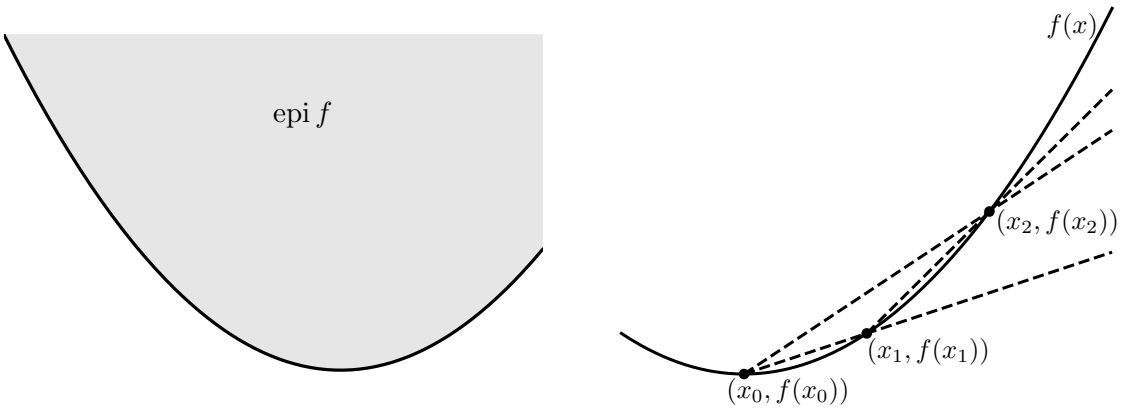


Figure B.9. Left: the epigraph of a convex function. Right: the criterion of increasing slopes. Note that the slopes of f evaluated at the points x_0, x_1, and x_2 increase.

We define the domain $\operatorname{dom} f$ of a convex function to be those points x such that $f(x) < +\infty$. Note that Definition B.7 implies that the domain of f must be convex.

An equivalent definition of convexity follows by considering a natural convex set attached to the function f, known as its epigraph.

Definition B.8. *The epigraph* $\operatorname{epi} f$ *of a function is the set*

$$\operatorname{epi} f := \{(x, t) : t \in \mathbb{R}, f(x) \le t\}.$$

The epigraph of a function f is the set of points on or above the graph of the function itself, as Figure B.9 depicts. It is immediate from the definition that f is convex if and only if $\operatorname{epi} f$ is convex. Thus, we see that any convex set $C \subset \mathbb{R}^{d+1}$ that is unbounded "above," meaning that $C = C + \{0\} \times \mathbb{R}_+$, defines a convex function, and conversely, any convex function defines such a set C. This duality in the relationship between a convex function and its epigraph is central to many of the properties we exploit.

B.2.1 Equivalent definitions of convex functions

We begin our discussion of convex functions by enumerating a few standard properties that also characterize convexity. The simplest of these relate to properties of the derivatives and second derivatives of functions. We begin by elucidating one of the most basic properties of convexity: that the slopes of convex functions are increasing. Beginning with functions on $\mathbb{R}$, let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be convex, and let $x \in \operatorname{dom} f$ and $v \in \mathbb{R}$ be otherwise arbitrary. Define the quotient function

$$q(t) := \frac{f(x + tv) - f(x)}{t}, \quad t \neq 0. \tag{B.2.2}$$

The criterion of *increasing slopes* states that q is nondecreasing in $t \neq 0$ if and only if f is convex:

Lemma B.2.1. *The function f is convex if and only if for all $x \in \operatorname{dom} f$ and v, the quotient* (B.2.2) *is nondecreasing in $t \neq 0$. (See Figure B.9.)*

Proof Define $h(z) = f(x + zv)$, and let $s < t$. For any $u \in [s, t]$, the value $\lambda = \frac{t-u}{t-s} \in [0, 1]$ satisfies $\lambda s + (1 - \lambda)t = u$, and so convexity implies

$$\begin{aligned} h(u) \le \lambda h(s) + (1 - \lambda)h(t) &= h(s) + (1 - \lambda)(h(t) - h(s)) = h(t) + \lambda(h(s) - h(t)) \\ &= h(s) + \frac{h(t) - h(s)}{t - s}(u - s) = h(t) + \frac{h(s) - h(t)}{t - s}(t - u). \end{aligned}$$

Dividing through by $u - s > 0$ and rearranging yields $\frac{1}{u-s}(h(u) - h(s)) \le \frac{1}{t-s}(h(t) - h(s))$, valid for all $s < u < t$. Similarly, dividing through by $t - u > 0$ and rearranging yields $\frac{1}{t-s}(h(t) - h(s)) \le \frac{1}{t-u}(h(t) - h(u))$ as well. That is, we have shown that if $s < u < t$, then

$$\frac{h(u) - h(s)}{u - s} \le \frac{h(t) - h(s)}{t - s} \le \frac{h(t) - h(u)}{t - u} \tag{B.2.3}$$

whenever h is convex. Considering cases and using that $s < u < t$ are arbitrary in inequality (B.2.3) gives the lemma. If $s < 0$ and $t > 0$, we may take $u = 0$ to obtain $\frac{h(s)-h(0)}{s} \le \frac{h(t)-h(0)}{t}$. If $s < u < 0$, then taking $t = 0$ yields $\frac{h(s)-h(0)}{s} \le \frac{h(u)-h(0)}{u}$. If $0 < u < t$, then setting $s = 0$ similarly implies $\frac{h(u)-h(0)}{u} \le \frac{h(t)-h(0)}{t}$.

For the converse, let $t \ge s > 0$ and define $\lambda = \frac{s}{t} \in (0, 1]$. Then

$$\begin{aligned} q(t) \ge q(s) \;\; &\text{if and only if} \;\; \lambda[f(x + tv) - f(x)] \ge f(x + \lambda tv) - f(x) \\ &\text{if and only if} \;\; \lambda f(x + tv) + (1 - \lambda)f(x) \ge f((1 - \lambda)x + \lambda(x + tv)), \end{aligned}$$

the latter holding for all λ if and only if f is convex. □

Because the quotient function (B.2.2) is nondecreasing, we can relatively straightforwardly give first-order characterizations of convexity as well. Indeed, suppose that $f : \mathbb{R} \to \mathbb{R}$ is differentiable; then convexity is equivalent to the first-order inequality that for all $x, y \in \mathbb{R}$, we have

$$f(y) \ge f(x) + f'(x)(y - x). \tag{B.2.4}$$

To see that inequality (B.2.4) implies that f is convex follows from algebraic manipulations: let $\lambda \in [0, 1]$ and $z = \lambda x + (1 - \lambda)y$, so that $y - z = \lambda(y - x)$ and $x - z = (1 - \lambda)(x - y)$. Then

$$f(y) \ge f(z) + \lambda f'(z)(y - x) \;\; \text{and} \;\; f(x) \ge f(z) + (1 - \lambda)f'(z)(x - y),$$

and multiplying the former by $(1 - \lambda)$ and the latter by λ and adding the two inequalities yields

$$\lambda f(x)+(1-\lambda)f(y) \ge \lambda f(z)+(1-\lambda)f(z)+\lambda(1-\lambda)f'(z)(y-x)+\lambda(1-\lambda)f'(z)(x-y) = f(\lambda x+(1-\lambda)y),$$

as desired. Conversely, let $v = y - x$ in the quotient (B.2.2), so that $q(t) = \frac{f(x+t(y-x))-f(x)}{t}$. If f is differentiable, then $q(0) = \lim_{t \downarrow 0} q(t) = f'(x)(y - x)$, and so

$$q(1) = f(y) - f(x) \ge q(0) = f'(x)(y - x).$$

We may also give the standard second order characterization: if $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable and $f''(x) \ge 0$ for all x, then f is convex. To see this, note that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(tx + (1 - t)y)(x - y)^2$$

for some $t \in [0,1]$ by Taylor's theorem, so that $f(y) \geq f(x) + f'(x)(y-x)$ for all x, y because $f''(tx + (1-t)y) \geq 0$. As a consequence, we obtain inequality (B.2.4), so f is convex.

As convexity is a property that depends only on properties of functions on lines—one dimensional projections—we can straightforwardly extend the preceding results to functions $f : \mathbb{R}^d \to \mathbb{R}$. Indeed, noting that if $h(t) = f(x + ty)$ then $h'(0) = \langle \nabla f(x), y \rangle$ and $h''(0) = y^\top \nabla^2 f(x) y$, then a differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \text{for all } x, y,$$

while a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x.$$

Nothing in the proof of Lemma B.2.1 that the quotient (B.2.2) is nondecreasing relies on f being a function on $\mathbb{R}$, so a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex if and only if it satisfies the *increasing slopes* criterion: for all $x \in \operatorname{dom} f$ and any vector v, the quotient

$$t \mapsto q(t) := \frac{f(x + tv) - f(x)}{t}, \tag{B.2.5}$$

is nondecreasing in $t \neq 0$ (where we leave x, v implicit). The monotonicity means that we may also define $q(0)$ to take any value in $[\sup_{t<0} q(t), \inf_{t>0} q(t)]$, which is non-empty.

We summarize each of these implications in a proposition for reference.

Proposition B.2.2 (Convexity)**.** *The following are all equivalent:*

(i) The function f is convex.

(ii) The function f satisfies the criterion of increasing slopes (B.2.5)*.*

If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable (respectively, twice differentiable), the following are also equivalent:

(iii) The function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \textit{for all } x, y.$$

(iv) The function f has positive semidefinite Hessian: $\nabla^2 f(x) \succeq 0$ for all x.

A stronger condition than convexity is *strict convexity*, which makes the secant inequality strict for distinct points. We begin with the classical definition: a function f is strictly convex if it is convex and

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

whenever $\lambda \in (0,1)$ and $x \neq y \in \operatorname{dom} f$. These are convex functions with increasing slopes—secants lie strictly above f; see Figure B.10. Tracing the one-dimensional argument for Proposition B.2.2 gives the following useful characterizations and sufficient conditions for strict convexity.

Corollary B.2.3 (Strict convexity)**.** *The following are equivalent:*

(i) The function f is strictly convex.

(ii) The function f has strictly increasing slopes (B.2.5)*.*

(iii) If f is differentiable, it satisfies

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle \quad \textit{for all } x \neq y.$$

Additionally, if f is twice differentiable and $\nabla^2 f(x) \succ 0$ for all x, then it is strictly convex.

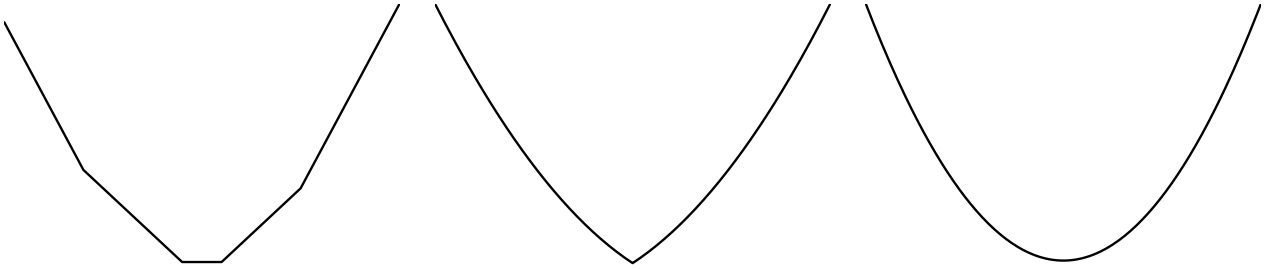


Figure B.10. Left: a convex but not strictly convex function. Middle: the strictly convex function $f(x) = \max\{(x-1)^2, (x+1)^2\}$. Right: the strictly convex function $f(x) = x^2$.

B.2.2 Continuity properties of convex functions

We now consider a few continuity properties of convex functions and a few basic relationships of the function f to its epigraph. First, we give a definition of the *subgradient* of a convex function.

Definition B.9. *A vector g is a* subgradient *of f at a point x_0 if for all x,*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle. \tag{B.2.6}$$

The subdifferential *or* subgradient set *of f at x_0 is*

$$\partial f(x_0) := \{g \mid f(x) \geq f(x_0) + \langle g, x - x_0 \rangle \text{ for all } x\}.$$

See Figure B.11 for an illustration of the affine minorizing function given by the subgradient of a convex function at a particular point.

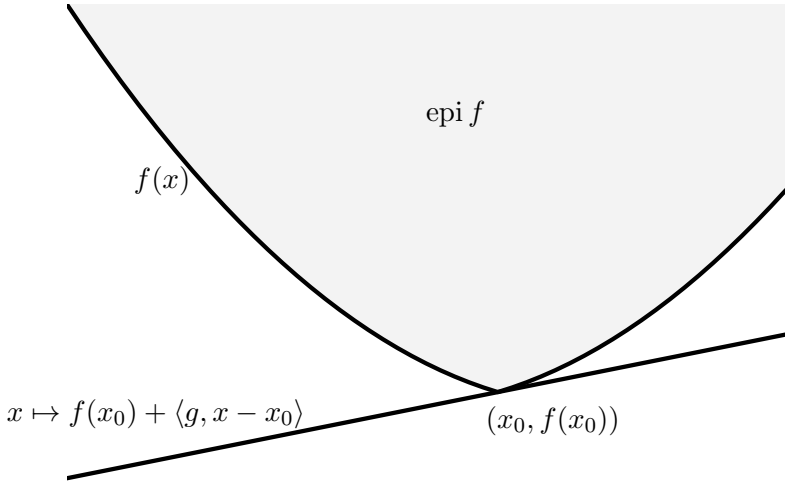


Figure B.11. The tangent (affine) function to the function f generated by a subgradient g at the point x_0. The induced hyperplane supports epi f at $(x_0, f(x_0))$.

Convex functions have subgradients (at least, nearly everywhere). This is perhaps intuitively obvious by viewing a function in conjunction with its epigraph epi f and noting that epi f has supporting hyperplanes, but here we state a result that will have further use.

Theorem B.2.4. *Let f be convex. Then there is an affine function minorizing f. More precisely, for any $x_0 \in \operatorname{relint} \operatorname{dom} f$, there exists a vector $g \in \operatorname{span}(\operatorname{dom} f - \operatorname{dom} f)$ such that*

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle.$$

Proof If $\operatorname{relint} \operatorname{dom} f = \emptyset$, then $\operatorname{dom} f = \emptyset$, making the statement trivial. Thus we assume that $\operatorname{int} \operatorname{dom} f \neq \emptyset$, as we can always change basis to work in the affine hull of $\operatorname{dom} f$.

We use Theorem B.1.20 on the existence of supporting hyperplanes to construct a subgradient. Any point of the form $(x_0, f(x_0)) \in \operatorname{bd} \operatorname{epi} f$, as for any open set O the set $(x_0, f(x_0)) + O$ contains points both inside and outside of $\operatorname{epi} f$. Thus, Theorem B.1.20 guarantees the existence of a vector v and $a \in \mathbb{R}$, not both simultaneously zero, such that

$$\langle v, x_0 \rangle + af(x_0) \leq \langle v, x \rangle + at \quad \text{for all } (x, t) \in \operatorname{epi} f. \tag{B.2.7}$$

Inequality (B.2.7) implies that $a \geq 0$, as for any x we may take $t \to +\infty$ while satisfying $(x, t) \in \operatorname{epi} f$. Now we argue that $a > 0$ strictly. To see this, note that for suitably small $\delta > 0$, we have $x = x_0 - \delta v \in \operatorname{dom} f$. Then we find by inequality (B.2.7) that

$$\langle v, x_0 \rangle + af(x_0) \leq \langle v, x_0 \rangle - \delta \left\| v \right\|^2 + af(x_0 - \delta v), \quad \text{or} \quad a \left[f(x_0) - f(x_0 - \delta v) \right] \leq -\delta \left\| v \right\|^2.$$

So if $v = 0$, then Theorem B.1.20 already guarantees $a \neq 0$, while if $v \neq 0$, then $\left\| v \right\|^2 > 0$ and we must have $a \neq 0$ and $f(x_0) \neq f(x_0 - \delta v)$. As we showed already that $a \geq 0$, we must have $a > 0$. Then by setting $t = f(x_0)$ and dividing both sides of inequality (B.2.7) by a, we obtain

$$\frac{1}{a} \langle v, x_0 - x \rangle + f(x_0) \leq f(x) \quad \text{for all } x \in \operatorname{dom} f.$$

The vector $g = -v/a$ gives the result of the theorem, as we have $f(x) = +\infty$ for $x \notin \operatorname{dom} f$. $\square$

As a corollary, for convex f, local Lipschitz continuity coincides with bounded subgradients:

Corollary B.2.5. *Let f be convex, $A = \operatorname{aff} \operatorname{dom} f$, and $U \subset A$ be a relatively open subset of $\operatorname{dom} f$. Then the following are equivalent:*

(i) f is L-Lipschitz on U.

(ii) For each $x \in U$ and $g \in \partial f(x)$, the orthogonal decomposition $g = g_A + g_\perp$ into $g_A \in A - A$ and $g_\perp \in (A - A)^\perp$ satisfies $\left\| g_A \right\| \leq L$.

We leave the proof as Exercise B.2.

Convex functions enjoy remarkable continuity properties by virtue of the defining convexity inequality (B.2.1). In particular, they are continuous on the relative interiors of their domains, and even more, they are Lipschitz continuous on compact subsets contained in these relative interiors. (See Figure B.12 for an illustration.)

Theorem B.2.6. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and $C \subset \operatorname{relint} \operatorname{dom} f$ be compact. Then there exists an $L = L(C) < \infty$ such that*

$$\left| f(x) - f(x') \right| \leq L \left\| x - x' \right\|.$$

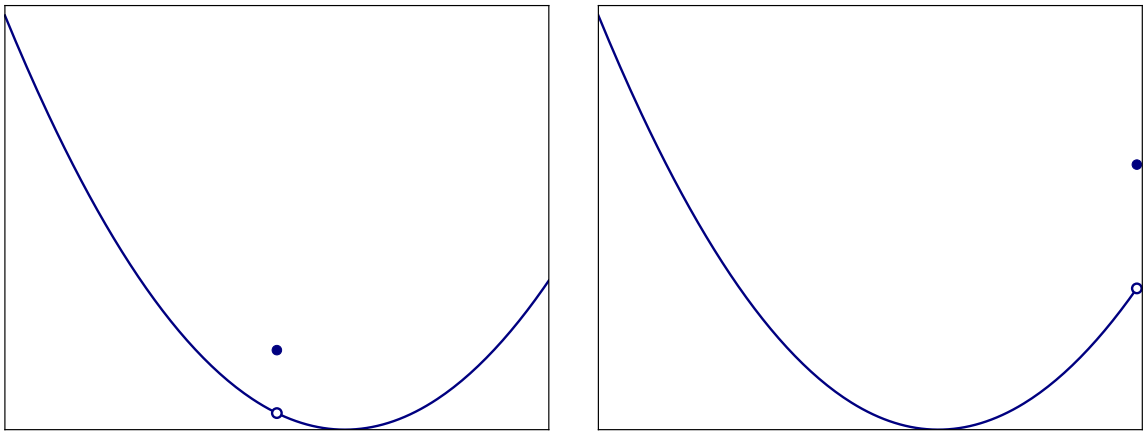


Figure B.12. Left: discontinuities in $\operatorname{int}\operatorname{dom} f$ are impossible while maintaining convexity (Theorem B.2.6). Right: At the edge of $\operatorname{dom} f$, there may be points of discontinuity.

As an immediate consequence of Theorem B.2.6, if $f : \mathbb{R}^d \to \mathbb{R}$ is convex and defined everywhere on $\mathbb{R}^d$, then it is continuous. Moreover, $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is continuous everywhere on the (relative) interior of its domain: let any $x_0 \in \operatorname{relint}\operatorname{dom} f$. Then for small enough $\epsilon > 0$, the set $\operatorname{cl}(\{x_0 + \epsilon B\} \cap \operatorname{dom} f)$, where $B = \{x : \|x\|_2 \le 1\}$, is a compact set contained in the (relative) interior of $\operatorname{dom} f$. Thus f is Lipschitz on this set, which is a neighborhood of x_0. In addition, if $f : \mathbb{R} \to \overline{\mathbb{R}}$, then f is continuous everywhere except (possibly) at the endpoints of its domain.

Proof of Theorem B.2.6 To prove the theorem, we require a technical lemma.

Lemma B.2.7. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and suppose that there are x_0, $\delta > 0$, m, and M such that*

$$m \le f(x) \le M \quad \text{for } x \in B(x_0, 2\delta) := \{x : \|x - x_0\| < 2\delta\}.$$

Then f is Lipschitz on $B(x_0, \delta)$, and moreover,

$$|f(y) - f(y')| \le \frac{M - m}{\delta} \left\|y - y'\right\| \quad \text{for } y, y' \in B(x_0, \delta).$$

Proof The case $y = y'$ is trivial. For $y \ne y'$, define $y'' = y' + \delta(y' - y)/\left\|y' - y\right\| \in B(x_0, 2\delta)$. Then

$$y' = \frac{\|y' - y\|}{\delta + \|y' - y\|} y'' + \frac{\delta}{\delta + \|y' - y\|} y$$

is a convex combination of y and y''. Thus, using the assumed bounds on f, we obtain by convexity

$$\begin{aligned} f(y') - f(y) \le \frac{\|y' - y\|}{\delta + \|y' - y\|} f(y'') + \frac{\delta}{\delta + \|y' - y\|} f(y) - f(y) &= \frac{\|y - y'\|}{\delta + \|y - y'\|} [f(y'') - f(y)] \\ &\le \frac{M - m}{\delta + \|y - y'\|} \left\|y - y'\right\|. \end{aligned}$$

Exchanging y and y' gives the reverse inequality. □

With Lemma B.2.7 in place, we proceed to the proof proper. Assume without loss of generality that $\operatorname{dom} f$ has an interior; otherwise we prove the theorem restricted to the affine hull of $\operatorname{dom} f$.

Suppose that for each $x \in C$ we construct an open ball $B_x = B(x, \delta_x)$, $\delta_x > 0$, such that

$$|f(y) - f(y')| \le L_x \left\|y - y'\right\| \quad \text{for } y, y' \in B_x. \tag{B.2.8}$$

As the B_x cover the compact set C, we can extract a finite collection $B_{x_1}, \ldots, B_{x_k}$ covering C, and then within each (overlapping) ball f has subdifferentials $\partial f(x) \subset \max_j L_{x_j} \cdot \mathbb{B}$ by Corollary B.2.5. Applying the corollary again, we obtain

$$|f(y) - f(y')| \le \max_{j \le k} L_{x_j} \left\|y - y'\right\| \quad \text{for } y, y' \in C.$$

We thus must derive inequality (B.2.8), for which we use the boundedness Lemma B.2.7. Fix $x \in \operatorname{int} \operatorname{dom} f$, and let the points $x_0, \ldots, x_d$ be affinely independent and such that

$$\Delta := \operatorname{Conv}\{x_0, \ldots, x_d\} \subset \operatorname{dom} f \quad \text{and} \quad x \in \operatorname{int} \Delta.$$

Let $\delta > 0$ be such that $B(x, 2\delta) \subset \Delta$. Each point $y \in B(x, 2\delta) \subset \Delta$, so by convexity $f(y) \le \max_{i \in \{0,\ldots,d\}} f(x_i) =: M$. Theorem B.2.4 implies that there is some affine h minorizing f; let $h(x) = a + \langle v, x \rangle$ denote this function. Then

$$m := \inf_{x \in C} f(x) \ge \inf_{x \in C} h(x) = a + \inf_{x \in C} \langle v, x \rangle > -\infty$$

exists and is finite, so that in the ball $B(x, 2\delta)$ constructed above, we have $f(y) \in [m, M]$ as required by Lemma B.2.7. This guarantees the existence of the ball B_x inequality (B.2.8) requires. □

B.2.3 Closure properties of convex functions

Our final discussion of continuity properties of convex functions revolves around the most common and analytically convenient type of convex function, the so-called *closed-convex* functions.

Definition B.10. *A function f is* closed *if its epigraph,* $\operatorname{epi} f$*, is a closed set.*

Equivalently, a function is closed if it is lower semi-continuous, meaning that

$$\liminf_{x \to x_0} f(x) \ge f(x_0) \tag{B.2.9}$$

for all x_0 and any sequence of points tending toward x_0.

In the one-dimensional case, closed convexity implies continuity. Indeed, we have the following observation (compare Figure B.12):

Observation B.2.8. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be closed convex. Then f is continuous on its domain, and for any $x_0 \in \operatorname{bd} \operatorname{dom} f$, $\lim_{x \to x_0} f(x) = f(x_0)$ whether or not $x_0 \in \operatorname{dom} f$.*

Proof By Theorem B.2.6, we need only consider the endpoints of the domain of f; let $x_0 \in \operatorname{bd} \operatorname{dom} f$. Let $y \in \operatorname{dom} f$ and define $x_\lambda = \lambda y + (1 - \lambda) x_0$. Taking $\lambda \to 0$, we have

$$f(x_\lambda) \le \lambda f(y) + (1 - \lambda) f(x_0) \to f(x_0),$$

while lower semicontinuity (B.2.9) gives $\liminf_{x \to x_0} f(x) \ge f(x_0)$. If $x_0 \notin \operatorname{dom} f$, then $f(x_0) = +\infty$ and lower semicontinuity guarantees $f(x) \to \infty$ as $x \to x_0$ from the domain. □

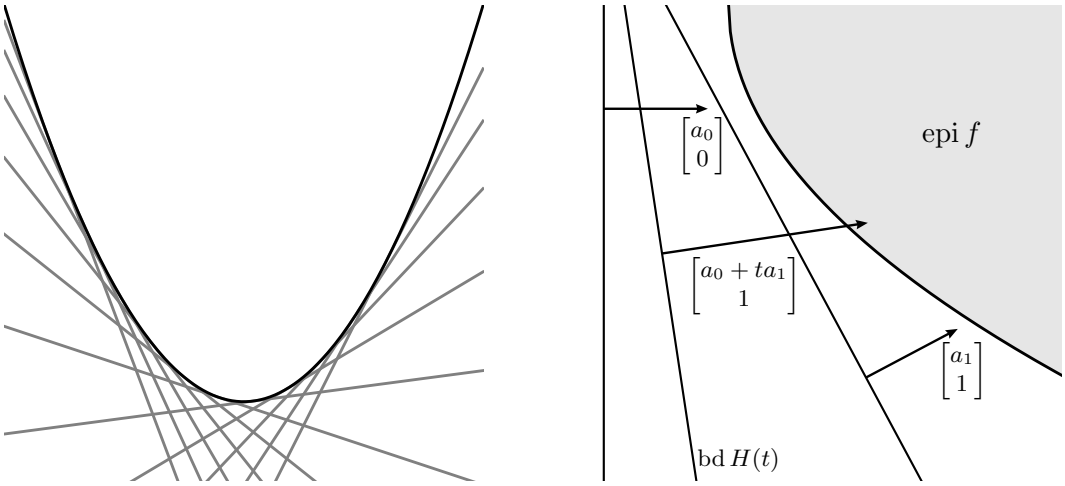


Figure B.13. Left: illustration of a convex function as the supremum of affine functions below it. Right: the construction in the proof of Theorem B.2.9. The middle plane denotes the boundary of the half space $H(t)$, which contains epi f.

In the full-dimensional case, we do not have quite the same continuity, though Theorem B.2.6 guarantees continuity on the (relative) interior of $\operatorname{dom} f$. (See Exercise B.7 for an example of a closed convex but discontinuous function.)

An important characterization of convex functions is as the supremum of all affine functionals (linear plus an offset) below them, which is one of the keys to duality relationships about functions to come. Figure B.13 illustrates this.

Theorem B.2.9. *Let f be closed convex and let $\mathcal{A}$ be the collection of affine functions h satisfying $f(x) \geq h(x)$ for all x. Then $f(x) = \sup_{h \in \mathcal{A}} h(x)$.*

Proof Let $\mathcal{H}$ be the collection of closed halfspaces $H \supset \operatorname{epi} f$, so $\operatorname{epi} f = \cap_{H \in \mathcal{H}} H$ by Theorem B.1.22. We may write any such halfspace as

$$H = \{(x, r) \in \mathbb{R}^d \times \mathbb{R} \mid \langle a, x \rangle + br \leq c\}$$

where $(a, b) \in \mathbb{R}^d \times \mathbb{R}$ is non-zero. Because $H \supset \operatorname{epi} f$ we must have $b \leq 0$, and so for any $b < 0$ we may divide through by b to rewrite H as $H = \{(x, r) \mid \langle a/b, x \rangle + r \geq c/b\}$, while if $b = 0$ then $H = \{(x, r) \mid \langle a, x \rangle \leq c\}$. That is, it is no loss of generality to set

$$\begin{aligned} \mathcal{H}_1 &:= \{\text{Halfspaces } \{(x, r) \mid \langle a, x \rangle + r \geq c\} \text{ containing } \operatorname{epi} f\} \\ \mathcal{H}_0 &:= \{\text{Halfspaces } \{(x, r) \mid \langle a, x \rangle \geq c\} \text{ containing } \operatorname{epi} f\}, \end{aligned}$$

which (respectively) correspond to the non-vertical halfspaces containing epi f and the halfspaces containing $\operatorname{dom} f \subset \mathbb{R}^d$. We have $\operatorname{epi} f = \bigcap_{H \in \mathcal{H}_1} H \cap \bigcap_{H \in \mathcal{H}_0} H$.

Identify the halfspaces $H \in \mathcal{H}_0$ or $\mathcal{H}_1$ with the associated triple $(a, 0, c)$ or $(a, 1, c)$ and abuse notation to write $(a, i, c) \in \mathcal{H}_i$ for $i \in \{0, 1\}$. For any $(a, 1, c) \in \mathcal{H}_1$, the affine function

$$l(x) = c - \langle a, x \rangle = \inf\{r \mid \langle a, x \rangle + r \geq c\} \text{ satisfies } \langle a, x \rangle + l(x) \geq c \text{ for all } x,$$

so necessarily $l(x) \le f(x)$ for all x, while for $h(x) = \sup_{(a,1,c)\in\mathcal{H}_1}\{c - \langle a, x\rangle\}$ we have

$$\operatorname{epi} h = \bigcap_{H\in\mathcal{H}_1} H.$$

Thus, if we can show that

$$\bigcap_{H\in\mathcal{H}_1} H \cap \bigcap_{H\in\mathcal{H}_0} H = \bigcap_{H\in\mathcal{H}_1} H \tag{B.2.10}$$

the proof will be complete.

Take vectors $v_0 = (a_0, 0, c_0) \in \mathcal{H}_0$ and $v_1 = (a_1, 1, c_1) \in \mathcal{H}_1$, and let $H_0 = \{(x, r) \mid \langle a_0, x\rangle \ge c_0\}$ and $H_1 = \{(x, r) \mid \langle a_1, x\rangle + r \ge c_1\}$ be the associated halfspaces. Consider the conic-like vector

$$v(t) := (a_1 + ta_0, 1, c_1 + tc_0) \quad \text{for } t \ge 0$$

and associated halfspace $H(t) := \{(x, r) \mid \langle a_1 + ta_0, x\rangle + r \ge c_1 + tc_0\}$, which Figure B.13 illustrates. Then as $\langle a_0, x\rangle \ge c_0$ if and only if $t\langle a_0, x\rangle \ge tc_0$ for all $t \ge 0$, any point $(x, r) \in H_0 \cap H_1$ satisfies

$$\langle a_1 + ta_0, x\rangle + r \ge c_1 + tc_0 \quad \text{for } t \ge 0,$$

that is, $H(t) \in \mathcal{H}_1$ and $(x, r) \in \cap_{t\ge0} H(t)$. Additionally, taking $t = 0$ we see that $H(0) = H_1$ and so $\cap_{t\ge0} H(t) \subset H_1$, while taking $t \uparrow \infty$ we obtain that each $(x, r) \in \cap_{t\ge0} H(t)$ satisfies $\langle a_0, x\rangle \ge c_0$. That is, we have

$$\bigcap_{t\ge0} H(t) = H_0 \cap H_1,$$

while $H(t) \in \mathcal{H}_1$ for all $t \ge 0$. This shows the equality (B.2.10). □

In spite of the continuity of closed convex functions on $\mathbb{R}$, closed convex functions on higher dimensional spaces need not be continuous. Indeed, it is immediate (see Proposition B.2.21 to follow) that $f(x) := \sup_{\alpha\in\mathcal{A}}\{f_\alpha(x)\}$ is closed convex whenever f_α are all closed convex for any index set $\mathcal{A}$. Exercise B.7 exhibits a discontinuous closed convex function.

Finally, we discuss closing a function, that is, replacing f with $\operatorname{cl} f$, the lower semicontinuous closure of f. For $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, we define $\operatorname{cl} f$ pointwise by

$$\operatorname{cl} f(x) := \liminf_{y\to x} f(y) \tag{B.2.11}$$

to be the lower semicontinuous closure of f. (Recall Definition B.10.) We demonstrate that this function is indeed the largest closed function below f.

Lemma B.2.10. *The function* $\operatorname{cl} f$ *is lower-semicontinuous,* $\operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$*, and if* g *is any closed function with* $g \le f$*, then* $g \le \operatorname{cl} f$*.*

Proof Assuming the second statement that $\operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$, the first follows trivially. We prove the second. Let $(x, r) \in \operatorname{cl}\operatorname{epi} f$. Then there exists a sequence $(x_n, r_n) \in \operatorname{epi} f$ such that $(x_n, r_n) \to (x, r)$, and $f(x_n) \le r_n$. Then evidently $\operatorname{cl} f(x) \le \lim_n r_n$, so that $(x, r) \in \operatorname{epi}\operatorname{cl} f$. Conversely, let $(x, r) \in \operatorname{epi}\operatorname{cl} f$. Let $x_n \to x$ satisfy $f(x_n) \to \operatorname{cl} f(x)$. Then because $r \ge \operatorname{cl} f(x)$, there exist $r_n \ge f(x_n)$ for which $r_n \to r$, while $(x_n, r_n) \in \operatorname{epi} f$, so $(x_n, r_n) \to (x, r) \in \operatorname{cl}\operatorname{epi} f$.

Finally, we prove the third claim of the lemma. If g is closed, then $\operatorname{epi} g$ is closed as well, and $\operatorname{epi} g \supset \operatorname{epi} f$. Then $\operatorname{epi} g \supset \operatorname{cl}\operatorname{epi} f = \operatorname{epi}\operatorname{cl} f$, that is, $\operatorname{cl} f(x) \le r$ implies $g(x) \le r$, i.e., $\operatorname{cl} f \ge g$. □

Using this lemma, we can provide additional color to Theorem B.2.9. In particular, Observation B.1.5 that closures of convex sets remain convex implies that if f is convex, then $\operatorname{cl} f$ is convex as well. We collect this and a bit more with the following immediate proposition.

Proposition B.2.11. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and $\mathcal{A}$ be the collection of all affine functions $h \le f$. Then $\operatorname{cl} f(x) = \sup_{h \in \mathcal{A}} h(x)$ for all x. In particular, if f is lower semicontinuous at x, then $f(x) = \sup_{h \in \mathcal{A}} h(x)$.*

This yields a way to write the closure of a convex function via directional limits:

Corollary B.2.12. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex. Then for any $x' \in \operatorname{relint} \operatorname{dom} f$,*

$$\operatorname{cl} f(x) = \lim_{t \downarrow 0} f(x + t(x' - x)).$$

Proof If $x \notin \operatorname{cl} \operatorname{dom} f$, then the definition (B.2.11) implies $\operatorname{cl} f(x) = +\infty$ and, for small enough $t > 0$, $x + t(x' - x) \notin \operatorname{cl} \operatorname{dom} f$, so that $f(x + t(x' - x)) = +\infty$. So assume $x \in \operatorname{cl} \operatorname{dom} f$. Let $\mathcal{A}$ be the collection of affine $h \le f$, so that $\operatorname{cl} f(x) = \sup_{h \in \mathcal{A}} h(x)$. If $x' \in \operatorname{relint} \operatorname{dom} f$, then for $0 \le t < 1$, $x + t(x' - x) \in \operatorname{relint} \operatorname{dom} f$ (Lemma B.1.13), and clearly

$$\liminf_{t \downarrow 0} f(x + t(x' - x)) \ge \liminf_{y \to x} f(y) = \operatorname{cl} f(x).$$

Conversely, applying Proposition B.2.11 to $x + t(x' - x) \in \operatorname{relint} \operatorname{dom} f$,

$$f(x + t(x' - x)) = \sup_{h \in \mathcal{A}} h(x + t(x' - x)) = \sup_{h \in \mathcal{A}} \left[(1 - t)h(x) + th(x')\right]$$

because h is affine. If $\operatorname{cl} f(x) < \infty$, then $\sup_{h \in \mathcal{A}} h(x) < \infty$ and similarly for $h(x')$, so that

$$\limsup_{t \downarrow 0} f(x + t(x' - x)) = \limsup_{t \downarrow 0} (1 - t) \sup_{h \in \mathcal{A}} h(x) = \sup_{h \in \mathcal{A}} h(x) = \operatorname{cl} f(x)$$

as desired. □

As a last example, we define the *closed convex hull* of f to be the largest closed convex function minorizing it. Thus, letting $\mathcal{A}$ denote the collection of all affine functions $h \le f$, define

$$\operatorname{cl} \operatorname{Conv} f(x) := \sup_{h \in \mathcal{A}} h(x). \tag{B.2.12}$$

We have the following corollary.

Corollary B.2.13. *Let g be a closed convex function with $g \le f$. Then $g \le \operatorname{cl} \operatorname{Conv} f$.*

Proof Let $\mathcal{A}_g$ be the affine functions $h \le g$. Clearly, as $g \le f$, if $h \in \mathcal{A}_g$ then $h \in \mathcal{A}$, so that $g(x) = \sup_{h \in \mathcal{A}_g} h(x) \le \sup_{h \in \mathcal{A}} h(x) = \operatorname{cl} \operatorname{Conv} f(x)$. □

B.2.4 Sublinear and support functions

Sublinear functions form the basis of the transition between convex sets and convex functions. Recall that f is convex if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ for all x, y.

Definition B.11. *A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is* sublinear *if it is convex and positively homogeneous, meaning*

$$f(tx) = tf(x) \quad \text{for all } x \in \mathbb{R}^d \text{ and } t > 0.$$

Such functions give some of the first dualities between convex sets and convex functions. They also allow us to describe first-order smoothness properties of convex functions (see Section B.2.6).

All sublinear functions have a dual construction.

Proposition B.2.14. *Let f be a closed sublinear function and $S := \{s \mid \langle s, x\rangle \le f(x) \text{ for all } x\}$. Then*

$$f(x) = \sup_{s \in S} \langle s, x\rangle.$$

Proof As f is closed convex, it has affine minorants $(s, t) \in \mathbb{R}^d \times \mathbb{R}$ satisfying $\langle s, x\rangle - t \le f(x)$ for $x \in \mathbb{R}^d$ (Theorem B.2.4). Because $f(0) = 0$ by sublinearity, $t \ge 0$, and $\langle s, \alpha x\rangle - t \le f(\alpha x)$ for $\alpha > 0$ by positive homogeneity, that is, $\langle s, x\rangle - t/\alpha \le f(x)$ for all x. Taking $\alpha \uparrow \infty$ we find that

$$\langle s, x\rangle \le f(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Because any closed convex function is the supremum of all affine functions minorizing it (Theorem B.2.9), we evidently have $f(x) = \sup_s\{\langle s, x\rangle \mid \langle s, \cdot\rangle \text{ minorizes } f\}$. □

To any set S we can associate the *support function* of S,

$$\sigma_S(x) := \sup_{s \in S} \langle s, x\rangle. \tag{B.2.13}$$

The support function is sublinear, so we immediately have the duality

Corollary B.2.15. *Let f be a closed sublinear function. Then it is the support function of the closed convex set*

$$S_f := \{s \mid \langle s, x\rangle \le f(x) \text{ for all } x \in \mathbb{R}^d\},$$

and hence if C is closed convex, then

$$C = \{x \mid \langle s, x\rangle \le \sigma_C(s) \text{ for all } s \in \mathbb{R}^d\}.$$

A few other consequences of the definition are immediate. We see that σ_S has $\operatorname{dom} \sigma_S = \mathbb{R}^d$ if and only if S is bounded: whenever $\|s\| \le L$ for all $s \in S$, then $\sigma_S(x) \le L\|x\|$. Conversely, if $\operatorname{dom} \sigma_S = \mathbb{R}^d$ then it is locally Lipschitz (Theorem B.2.6) and (by positive homogeneity) thus globally Lipschitz, so $\langle s, x\rangle \le \sigma_S(x) \le L\|x\|$ for some $L < \infty$, and taking $x = s/\|s\|$ gives $\|s\| \le L$. As another consequence, we see that support functions of a set S are the support functions of the closed convex hull of S:

Proposition B.2.16. *Let $S \subset \mathbb{R}^d$. Then*

$$\sigma_S(x) = \sigma_{\operatorname{cl}\operatorname{Conv} S}(x).$$

Exercise B.3 asks for the (straightforward) proof.

Coupled with Corollary B.2.15, this proposition shows that if sets S_1, S_2 have identical support functions, their closed convex hulls coincide.

Corollary B.2.17. *Let $S_1, S_2 \subset \mathbb{R}^d$. If $\sigma_{S_1} = \sigma_{S_2}$, then $\operatorname{cl}\operatorname{Conv} S_1 = \operatorname{cl}\operatorname{Conv} S_2$.*

Proof By Proposition B.2.16, we have $\sigma_{S_i} = \sigma_{\operatorname{cl}\operatorname{Conv} S_i}$ for each i, and Corollary B.2.15 shows that if $\sigma_{C_1} = \sigma_{C_2}$ for closed convex sets C_1 and C_2, then $C_1 = C_2$. □

As another corollary, we have

Corollary B.2.18. *Let σ_1 and σ_2 be the support functions of the nonempty closed convex sets S_1 and S_2. If $t_1 > 0$ and $t_2 > 0$,*

$$t_1\sigma_1 + t_2\sigma_2 = \sigma_{\operatorname{cl}(t_1 S_1 + t_2 S_2)}.$$

If either of S_1 or S_2 is compact, then $t_1\sigma_1 + t_2\sigma_2 = \sigma_{t_1 S_1 + t_2 S_2}$.

Proof Let $S = t_1 S_1 + t_2 S_2$. In first statement, we have

$$\sigma_{\operatorname{cl} S}(x) \stackrel{(\star)}{=} \sigma_S(x) = \sup\left\{\langle t_1 s_1 + t_2 s_2, x\rangle \mid s_1 \in S_1, s_2 \in S_2\right\},$$

equality $(\star)$ following from Proposition B.2.16. As the suprema run independently through their respective sets S_1, S_2, the latter quantity is evidently

$$\sigma_S(x) = t_1 \sup_{s_1 \in S_1} \langle s_1, x\rangle + t_2 \sup_{s_2 \in S_2} \langle s_2, x\rangle = t_1\sigma_{S_1}(x) + t_2\sigma_{S_2}(x).$$

The final result is an immediate consequence of the result that if C is a compact convex set and S is closed convex, then $C + S$ is closed convex by a standard converging subsequence argument. □

Linear transformations of support functions are also calculable. In the result, recall that for a matrix A and set S, the set $AS = \{As \mid s \in S\}$.

Proposition B.2.19. *Let $S \subset \mathbb{R}^d$ and $A \in \mathbb{R}^{m\times d}$. Then $\sigma_{\operatorname{cl} AS}(x) = \sigma_S(A^\top x)$.*

Proof We have $\sigma_{AS}(x) = \sup_{s\in S}\langle As, x\rangle = \sup_{s\in S}\langle s, A^\top x\rangle$. The closure operation changes nothing (Proposition B.2.16). □

Lastly, support functions characterize whether sets have interiors. Recall Definitions B.2 and B.5 of the affine hull and relative interior of sets.

Proposition B.2.20. *Let $S \subset \mathbb{R}^d$ be a non-empty closed convex set. Then*

(i) $s \in \operatorname{int} S$ *if and only if* $\langle s, x\rangle < \sigma_S(x)$ *for all* $x \neq 0$.

(ii) $s \in \operatorname{relint} S$ *if and only if* $\langle s, x\rangle < \sigma_S(x)$ *for all* x *with* $\sigma_S(x) + \sigma_S(-x) > 0$.

(iii) $\operatorname{int} S$ *is non-empty if and only if* $\sigma_S(x) + \sigma_S(-x) > 0$ *for all* $x \neq 0$.

Proof (i). By homogeneity, an equivalent statement is that $\sigma_S(x) > \langle s, x\rangle$ for all $x \in \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. If $s \in \operatorname{int} S$, there exists $\epsilon > 0$ such that $s + \epsilon x \in S$ for all $x \in \mathbb{S}^{d-1}$, so

$$\sigma_S(x) \geq \langle s + \epsilon x, x\rangle = \langle s, x\rangle + \epsilon.$$

Conversely, let s satisfy $\sigma_S(x) - \langle s, x\rangle > 0$ for all $x \in \mathbb{S}^{d-1}$. Because σ_S is lower semicontinuous, $\inf_{x\in\mathbb{S}^{d-1}}\{\sigma_S(x) - \langle s, x\rangle\}$ is attained at some $x^\star \in \mathbb{S}^{d-1}$ (look ahead to Proposition C.0.1). Then there is $\epsilon > 0$ such that $\langle s, x\rangle + \epsilon \le \sigma_S(x)$ for all $x \in \mathbb{S}^{d-1}$. Let u satisfy $\|u\|_2 \le \epsilon$. Then $\langle s + u, x\rangle = \langle s, x\rangle + \langle u, x\rangle \le \langle s, x\rangle + \epsilon \le \sigma_S(x)$, so Corollary B.2.15 implies $s + \epsilon\mathbb{B} \subset S$.

(ii). We decompose $\mathbb{R}^d$ into subspaces $V \oplus U$, where $U = V^\perp$ and V is parallel to $\mathrm{aff}(S)$. Writing $x = x_U + x_V$, where $x_U \in U$ and $x_V \in V$, the function $\langle s, x_U\rangle$ is constant for $s \in S$. Repeat the argument for part (i) in the subspace V.

(iii). Suppose $\operatorname{int} S$ is non-empty. Then $s \in \operatorname{int} S$ implies $\langle s, x\rangle < \sigma_S(x)$ for all $x \in \mathbb{S}^{d-1}$. Then $\sigma_S(x) + \sigma_S(-x) > \langle s, x - x\rangle = 0$. If $\operatorname{int} S$ is empty, there exists a hyperplane containing S (recall Proposition B.1.12), so $S \subset \{s \mid \langle v, s\rangle = b\}$ for some $v \neq 0$. For this $\sigma_S(v)+\sigma_S(-v) = b-b = 0$. □

B.2.5 Operations preserving convexity

We now turn to a description of operations on functions that preserve convexity. First, we extend the intersection properties of convex sets to operations on convex functions. (See Figure B.14 for an illustration of the proposition.)

Proposition B.2.21. *Let $\mathcal{A}$ index the arbitrary collection $\{f_\alpha\}_{\alpha\in\mathcal{A}}$ of convex functions. Then*

$$f(x) := \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

is convex. If for each $\alpha \in \mathcal{A}$, the function f_α is closed convex, f is closed convex.

Proof The proof is immediate once we consider the epigraph $\operatorname{epi} f$. We have that

$$\operatorname{epi} f = \bigcap_{\alpha\in\mathcal{A}} \operatorname{epi} f_\alpha,$$

is convex whenever $\operatorname{epi} f_\alpha$ is convex for all α and closed whenever $\operatorname{epi} f_\alpha$ is closed for all α. □

Composition of a convex function with an affine transformation preserves convexity:

Proposition B.2.22. *Let $A \in \mathbb{R}^{d\times n}$ and $b \in \mathbb{R}^d$, and let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex. Then the function $g(y) = f(Ay + b)$ is convex.*

Partial minimization of convex functions and some related transformations preserve convexity as well, and under some qualifications, they also preserve closedness.

Proposition B.2.23. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex, $h : \mathbb{R}^d \to \mathbb{R}^n$ be affine, and $Y \subset \mathbb{R}^d$ be convex. Then $g(x) = \inf\{f(y) \mid h(y) = x, y \in Y\}$ is convex. If Y is compact and f is closed, then g is closed convex.*

Proof Let $x_0, x_1 \in \mathbb{R}^n$. If $h(y) = x_i$ has no solution in $y \in Y$, then $g(x_i) = +\infty$, and we trivially have $g(\lambda x_0 + (1-\lambda)x_1) \le +\infty$ for $\lambda \in (0, 1)$. Assuming the sets $\{y \in Y \mid h(y) = x_0\}$ and $\{y \in Y \mid h(y) = x_1\}$ are non-empty, let $\epsilon > 0$ and $y_0, y_1 \in Y$ satisfy $h(y_i) = x_i$ and $f(y_i) \le g(x_i)+\epsilon$. Then $y_\lambda = \lambda y_0 + (1-\lambda)y_1$ satisfies $h(y_\lambda) = \lambda x_0 + (1-\lambda)x_1$, so

$$g(\lambda x_0 + (1-\lambda)x_1) \le f(\lambda y_0 + (1-\lambda)y_1) \le \lambda f(y_0) + (1-\lambda)f(y_1) \le \lambda g(x_0) + (1-\lambda)g(x_1) + \epsilon$$

for $\lambda \in [0, 1]$. Take $\epsilon \downarrow 0$.

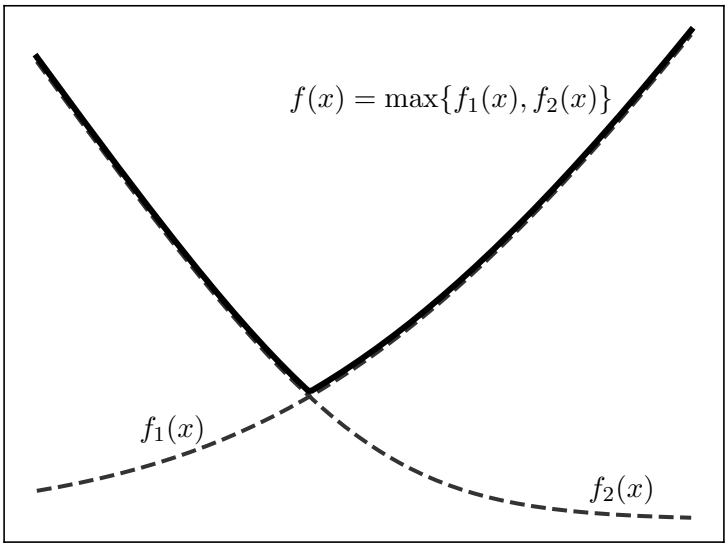


Figure B.14. The maximum of two convex functions is convex, as its epigraph is the intersection of the two epigraphs.

For the closed convexity statement, let $x_n \to x$; we wish to show that $\liminf_n g(x_n) \geq g(x)$. If $g(x_n) = +\infty$ for all x_n, we trivially have the result. Otherwise, moving to a subsequence as necessary, assume $g(x_n) \to \liminf_n g(x_n) < \infty$, let $\epsilon > 0$ be arbitrary, and let $y_n \in Y$ satisfy $h(y_n) = x_n$ and $f(y_n) \leq g(x_n) + \epsilon$. Then as Y is compact, y_n has convergent subsequences; let y be any such limit. We have $h(y) = x$, and $g(x) \leq f(y) \leq \liminf_n f(y_n) \leq \liminf_n g(x_n) + \epsilon$. □

From the proposition we immediately see that if $f(x, y)$ is jointly convex in x and y, then the partially minimized function $\inf_{y \in Y} f(x, y)$ is convex whenever Y is a convex set.

Lastly, we consider the functional analogue of the perspective transform. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *perspective transform* of f is

$$f_{\rm per}(x, t) := \begin{cases} tf\left(\frac{x}{t}\right) & \text{if } t > 0 \text{ and } \frac{x}{t} \in \operatorname{dom} f \\ +\infty & \text{otherwise.} \end{cases} \tag{B.2.14}$$

Figure B.15 illustrates this transformation. In analogue with the perspective transform of a convex set, the perspective transform of a function is (jointly) convex.

Proposition B.2.24. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex. Then $f_{\rm per} : \mathbb{R}^d \times \mathbb{R}_{>0} \to \overline{\mathbb{R}}$ is convex.*

Proof The result follows if we can show that $\operatorname{epi} f_{\rm per}$ is a convex set. Note that

$$\mathbb{R}^d \times \mathbb{R}_{>0} \times \mathbb{R} \ni (x, t, r) \in \operatorname{epi} f_{\rm per} \text{ if and only if } f\left(\frac{x}{t}\right) \leq \frac{r}{t}.$$

Rewriting this, we have

$$\begin{aligned} \operatorname{epi} f_{\rm per} &= \left\{(x, t, r) \in \mathbb{R}^d \times \mathbb{R}_{>0} \times \mathbb{R} \mid f\left(\frac{x}{t}\right) \leq \frac{r}{t}\right\} \\ &= \left\{t(x', 1, r') \mid x' \in \mathbb{R}^d, t \in \mathbb{R}_{>0}, r' \in \mathbb{R}, f(x') \leq r'\right\} \\ &= \{t(x, 1, r) \mid t > 0, (x, r) \in \operatorname{epi} f\} = \mathbb{R}_{>0} \cdot \{(x, 1, r) \mid (x, r) \in \operatorname{epi} f\}. \end{aligned}$$

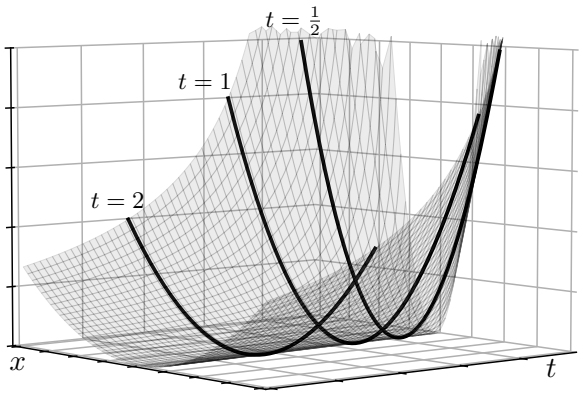


Figure B.15: Perspective transform $f_{\rm per}(x,t) = x^2/t$ of $f(x) = x^2$.

This is a convex cone because $\epsilon\text{pi}\, f$ is convex. □

Corollary B.2.12 characterizes the closure of the perspective transform of a closed convex f as well.

Corollary B.2.25. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and $f_{\rm per}$ be its perspective transform (B.2.14). Then for any $x' \in \operatorname{relint}\operatorname{dom} f$,*

$$\operatorname{cl} f_{\rm per}(x,u) = \begin{cases} uf(x/u) & \text{if } u > 0 \\ \lim_{\alpha \downarrow 0} \alpha f(x' - x + x/\alpha) & \text{if } u = 0 \\ +\infty & \text{if } u < 0. \end{cases}$$

Proof For the first equality, the definition of the closure and that f is lower semicontinuous gives $\liminf_{(y,v)\to(x,u)} vf(y/v) \geq uf(x/u)$. When $u < 0$, because $\operatorname{cl}\operatorname{dom} f_{\rm per} \subset \mathbb{R}^d \times \mathbb{R}_+$, we have $(x,u) \notin \operatorname{cl}\operatorname{dom} f_{\rm per}$ for any x, and so $\operatorname{cl} f_{\rm per}(x,u) = +\infty$ by definition of the domain. Finally, when $u = 0$, we recognize that if $x' \in \operatorname{relint}\operatorname{dom} f$, then $(x',1) \in \operatorname{relint}\operatorname{dom} f_{\rm per}$, and so using Corollary B.2.12 gives

$$\operatorname{cl} f_{\rm per}(x,0) = \lim_{\alpha \downarrow 0} f_{\rm per}(x + \alpha(x' - x), \alpha) = \lim_{\alpha \downarrow 0} \alpha f\left(x' - x + \frac{x}{\alpha}\right),$$

where the limit is consistent for any $x' \in \operatorname{relint}\operatorname{dom} f$. □

B.2.6 Smoothness properties, first-order developments for convex functions, and subdifferentiability

In addition to their continuity properties, convex functions typically enjoy strong differentiability properties. Some of these interact with the duality properties Section C.2 presents, but we give

analytic preliminaries here. Our main goal will be to show how there exist derivative-like objects for convex functions, so that for some suitably nice object $\mathsf{D}f(x; v)$ we have

$$f(x + tv) = f(x) + \mathsf{D}f(x; v)t + o(t) \tag{B.2.15}$$

for t small and any v. In the case that f is differentiable, this must coincide with the usual derivative, so that $\mathsf{D}f(x; v) = \langle \nabla f(x), v \rangle$. For convex functions, a directional derivative *always* exists, making the first-order development (B.2.15) sensible in some generality.

As one prototypical result, we leverage Rademacher's theorem on almost everywhere differentiability of Lipschitz functions to show that convex functions are almost everywhere differentiable:

Theorem B.2.26 (Rademacher)**.** *Let $U \subset \mathbb{R}^d$ be open and $f : U \to \mathbb{R}^k$ be Lipschitz continuous. Then f is differentiable almost everywhere on U.*

Proofs of this result are standard in measure-theoretic analysis texts; see, e.g., Folland [191, Section 3.5] or Villani [414, Theorem 10.8(ii)]. As any convex function is locally Lipschitz on its domain (recall Theorem B.2.6), we thus have the following result.

Corollary B.2.27. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and $\operatorname{dom} f$ have non-empty interior. Then f is differentiable except on a set of Lebesgue measure zero on its domain.*

Other differentiability properties of convex functions are also of interest. We begin by considering directional differentiability properties, after which we consider differentiability and continuous differentiability of (convex) functions. To begin, recall that the *directional derivative* of a function f in direction v at x is

$$f'(x; v) := \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} \tag{B.2.16}$$

when this quantity exists. When $f'(x; v)$ exists for all directions v and is linear in v, we call the function *Gateaux differentiable*. A (stronger in infinite dimensions) notion of differentiability is *Fréchet differentiability*: f has Fréchet differential g at x if

$$f(y) = f(x) + \langle g, y - x \rangle + o(\|y - x\|) \tag{B.2.17}$$

as $y \to x$, which is then uniform in the distance $\|y - x\|$. It is immediate that if f has Fréchet gradient g then it is Gateaux differentiable with $f'(x; v) = \langle g, v \rangle$. In finite dimensions, these coincide with the standard gradient, and $f'(x; v) = \langle \nabla f(x), v \rangle$, whenever f is locally Lipschitz.

Proposition B.2.28. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be Gateaux differentiable at x, that is, its directional derivative $f'(x; v)$ is linear in v, and locally Lipschitz, so that there exists $L < \infty$ such that $|f(x) - f(y)| \le L \|x - y\|$ for y near x. Then f is Fréchet differentiable at x with*

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_j}\right]_{j=1}^{d},$$

and $f'(x; v) = \langle \nabla f(x), v \rangle$ and $\|\nabla f(x)\| \le L$.

Proof If f is Fréchet differentiable at x with differential g, then we immediately have

$$\frac{f(x + tv) - f(x)}{t} = \frac{t\langle g, v \rangle + o(t)}{t} \to \langle g, v \rangle$$

as $t \to 0$, so that it is Gateaux differentiable.

Conversely, suppose that $f'(x; v) = \langle g, v \rangle$ for all $v \in \mathbb{R}^d$ for some $g \in \mathbb{R}^d$. Assume for the sake of contradiction that f is not Fréchet differentiable at x, so that

$$\limsup_{\|\Delta\| \downarrow 0} \frac{f(x + \Delta) - f(x) - \langle g, \Delta \rangle}{\|\Delta\|} = c > 0.$$

Take any sequence $\Delta_n \to 0$ achieving this limit supremum, and let $\Delta_n = \epsilon_n v_n$ for a sequence v_n on the sphere, that is, $\|v_n\| = 1$, so $\epsilon_n = \|\Delta_n\|$. Then by passing to a subsequence if necessary, we can assume w.l.o.g. that $v_n \to v$ with $\|v\| = 1$. Then

$$\begin{aligned}\frac{|f(x + \Delta_n) - f(x) - \langle g, \Delta_n \rangle|}{\epsilon_n} &= \frac{|f(x + \epsilon_n v + \epsilon_n (v_n - v)) - f(x) - \epsilon_n \langle g, v \rangle - \epsilon_n \langle g, v_n - v \rangle|}{\epsilon_n} \\ &\le \frac{|f(x + \epsilon_n v) - f(x) - \epsilon_n \langle g, v \rangle|}{\epsilon_n} + \frac{L \epsilon_n \|v_n - v\| + \epsilon_n \|g\| \|v_n - v\|}{\epsilon_n}.\end{aligned}$$

Both of these terms tend to zero, a contradiction, and so f is Fréchet differentiable at x, and its Fréchet derivative is g. That Fréchet differentiability implies differentiability follows by noting that the partial derivatives $f'(x; e_j) = \frac{\partial f(x)}{\partial x_j}$ for each coordinate j.

Finally, the Lipschitzian bound on $\|\nabla f(x)\|$ follows by noting that

$$L \|\Delta\| \ge |f(x + \Delta) - f(x)| = |\langle \nabla f(x), \Delta \rangle| + o(\|\Delta\|).$$

Taking $\Delta = tv$ and $t \downarrow 0$, this implies that $L \|v\| \ge \langle \nabla f(x), v \rangle$ for all v, i.e., $\|\nabla f(x)\| \le L$. □

The main consequence of convexity that is important for us is that a convex function is directionally differentiable at every point in the interior of its domain, though the directional derivative need not be linear:

Proposition B.2.29. *Let f be convex and $x \in \operatorname{int} \operatorname{dom} f$. Then $f'(x; v)$ exists and the mapping $v \mapsto f'(x; v)$ is sublinear, convex, and globally Lipschitz.*

Proof If $x \in \operatorname{int} \operatorname{dom} f$, then the criterion (B.2.5) of increasing slopes guarantees that $f'(x; v) = \lim_{t \downarrow 0} \frac{f(x+tv) - f(x)}{t}$ exists. To see that $f'(x; v)$ is convex and sublinear in v, note that positive homogeneity is immediate, as $\frac{1}{t}(f(x + \alpha t v) - f(x)) = \frac{\alpha}{\alpha t}(f(x + \alpha t v) - f(x))$ for all $\alpha > 0$, and $f'(x; 0) = 0$. That it is convex is straightforward as well: for any u, v we have

$$\frac{f(x + t(\lambda u + (1 - \lambda) v)) - f(x)}{t} \le \lambda \frac{f(x + tu) - f(x)}{t} + (1 - \lambda) \frac{f(x + tv) - f(x)}{t}$$

and take $t \downarrow 0$. For the global Lipschitz claim, note that f is already locally Lipschitz near $x \in \operatorname{int} \operatorname{dom} f$ (recall Theorem B.2.6), so that there exists some $L < \infty$ and $\epsilon > 0$ such that for all $\|v\| = 1$ and $0 \le t \le \epsilon$ we have $|f(x + tv) - f(x)| \le Lt$, whence $|f'(x; v)| \le L$ and by homogeneity $|f'(x; v)| \le L \|v\|$ for all v. □

An inspection of the proof shows that the result extends to all of $\operatorname{dom} f$ if we allow $f'(x; v) = +\infty$ whenever $x + tv \notin \operatorname{dom} f$ for all $t > 0$, though we lose that $f'(x; v)$ is finite-valued. Then $f'(x; v)$ provides a valid first-order development of f in all directions from x once we write $\infty = a(t) + o(t)$ to mean $a(t) = +\infty$ for all small $t > 0$ and take $\infty \cdot t = \infty$ for $t > 0$.

Corollary B.2.30. *Let $x \in \operatorname{dom} f$. Then*

$$f(x + tv) = f(x) + f'(x; v)t + o(t)$$

as $t \downarrow 0$ and

$$f(x + tv) \geq f(x) + f'(x; v)t \quad \textit{for all } t \geq 0.$$

Proof The first part is immediate by definition of $f'(x; v)$. The second is immediate from the criterion (B.2.5) of increasing slopes, as for convex functions $f'(x; v) = \inf_{t>0} \frac{f(x+tv)-f(x)}{t}$. □

There are strong connections between subdifferentials and directional derivatives, and hence of the local developments (B.2.15). The following result makes this clear.

Proposition B.2.31. *Let f be convex and $x \in \operatorname{relint} \operatorname{dom} f$. Then*

$$\partial f(x) = \left\{s \mid \langle s, v\rangle \leq f'(x; v) \textit{ for all } v\right\} \neq \varnothing.$$

Proof For shorthand let $S = \{s \mid \langle s, v\rangle \leq f'(x; v) \text{ all } v\}$ be the set on the right. If $s \in S$, then the criterion (B.2.5) of increasing slopes guarantees that

$$\langle s, v\rangle \leq \frac{f(x + tv) - f(x)}{t} \quad \text{for all } v \in \mathbb{R}^d, t > 0.$$

Recognizing that as v is allowed to vary over all of $\mathbb{R}^d$ and $t > 0$, then $x + tv$ describes $\mathbb{R}^d$, we see that this condition is equivalent to the definition (B.2.6) of the subgradient.

That $\partial f(x) \neq \varnothing$ is Theorem B.2.4. □

We can also extend this to $x \in \operatorname{dom} f$—not necessarily the interior—where we define

$$\partial f(x) := \left\{s \mid \langle s, v\rangle \leq f'(x; v) \text{ for all } v\right\}. \tag{B.2.18}$$

Notably, the directional derivative function $v \mapsto f'(x; v)$ always exists for $x \in \operatorname{dom} f$ and is a sublinear convex function, and thus $\partial f(x)$ above is always a closed convex set whose support function (recall (B.2.13)) is the closure of $v \mapsto f'(x; v)$. While the subdifferential $\partial f(x)$ is always a compact convex set when $x \in \operatorname{int} \operatorname{dom} f$, even when it exists it may not be compact if x is on the boundary of $\operatorname{dom} f$. To see one important example of this, consider the indicator function

$$\mathbf{I}_C(x) := \begin{cases} +\infty & \text{if } x \notin C \\ 0 & \text{if } x \in C \end{cases}$$

of a closed convex set C. For simplicity, let $C = [a, b]$ be an interval. Then we have

$$\partial \mathbf{I}_C(x) = \begin{cases} [0, \infty] & \text{if } x = b \\ \{0\} & \text{if } a < x < b \\ [-\infty, 0] & \text{if } x = a. \end{cases}$$

Whether points $\pm\infty$ are included is a matter of convenience, depending on conventions with the extended real line. (See Figure B.16.)

These representations point to a certain closure property of subgradients, namely, that the subdifferential is closed under additions of the normal cone to the domain of f:

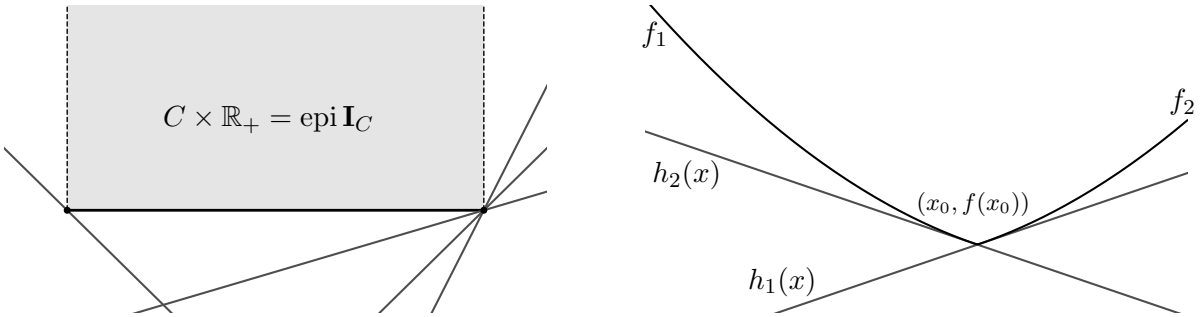


Figure B.16. (a) Subdifferentials of $\mathbf{I}_C$. (b) $\partial f(x)$ provides a first-order development of convex f. The function $f(x) = \max\{f_1(x), f_2(x)\}$ is convex, and for x near x_0, $f(x) = f(x_0) + \sup_{f_1'(x) \le g \le f_2'(x)} g(x - x_0) + o(|x - x_0|)$. The affine functions $h_1(x) = f(x_0) + f_1'(x_0)(x - x_0)$ and $h_2(x) = f(x_0) + f_2'(x_0)(x - x_0)$ capture this.

Lemma B.2.32. *Let $\mathcal{N}_{\operatorname{dom} f}(x)$ be the normal cone (Definition B.6) to $\operatorname{cl} \operatorname{dom} f$ at the point x (where $\mathcal{N}_{\operatorname{dom} f}(x) = \{0\}$ for $x \in \operatorname{int} \operatorname{dom} f$ and $\mathcal{N}_{\operatorname{dom} f}(x) = \varnothing$ for $x \notin \operatorname{dom} f$). Then*

$$\partial f(x) = \partial f(x) + \mathcal{N}_{\operatorname{dom} f}(x).$$

In particular, if x is a boundary point $x \in \operatorname{bd} \operatorname{dom} f$ of the domain of f, then either $\partial f(x) = \varnothing$ or $\partial f(x)$ is unbounded.

Proof We only need concern ourselves with points $x \in \operatorname{bd} \operatorname{dom} f$, where the normal cone $\mathcal{N} = \mathcal{N}_{\operatorname{dom} f}(x)$ is non-trivial. If $\partial f(x)$ is empty, there is nothing to prove, so assume that $\partial f(x)$ is non-empty. Then the definition (B.2.18) of the subdifferential as $\partial f(x) = \{s \mid \langle s, u \rangle \le f'(x; u)\}$ allows us to prove the result. First, consider vectors u for which $f'(x; u) = +\infty$. Then certainly, for any $s \in \partial f(x)$, we have $\langle s + v, u \rangle \le f'(x; u)$ for all $v \in \mathcal{N}$. If $f'(x; u) < \infty$, then for small enough $t > 0$ we necessarily have $x + tu \in \operatorname{dom} f$. In particular, the definition of the normal cone gives that $v \in \mathcal{N}$ satisfies $0 \ge \langle v, x + tu - x \rangle = t\langle v, u \rangle$, or that $\langle v, u \rangle \le 0$. Thus $\langle s + v, u \rangle \le \langle s, u \rangle \le f'(x; u)$, and so $s + v \in \partial f(x)$ once again.

The claim about boundedness is immediate, because $\mathcal{N}_{\operatorname{dom} f}(x)$ is a cone. □

A more compelling case for the importance of the subgradient set with respect to first-order developments and differentiability properties of convex functions is that subgradients give an accurate first-order approximation to f even when f is not differentiable. (See Figure B.16(b).)

Proposition B.2.33. *Let f be convex and $x \in \operatorname{int} \operatorname{dom} f$. Then*

$$\begin{aligned} f(y) &= f(x) + \sup_{s \in \partial f(x)} \langle s, y - x \rangle + o(\|y - x\|) \\ &= f(x) + f'(x; y - x) + o(\|y - x\|). \end{aligned}$$

Proof Theorem B.2.9 and Proposition B.2.14 immediately imply $\sup_{s \in \partial f(x)} \langle s, v \rangle = f'(x; v)$, because $f'(x; v)$ is sublinear and closed convex in v when $x \in \operatorname{int} \operatorname{dom} f$.

We thus prove the equality $f(y) = f(x) + f'(x; y - x) + o(\|y - x\|)$, where the argument is similar to that for Proposition B.2.28. Let $y_n \to x$ be any sequence and let $\Delta_n = y_n - x$, so that $\|\Delta_n\| \to 0$;

as $x \in \operatorname{int}\operatorname{dom} f$, there exists a local Lipschitz constant L such that $|f(x+\Delta) - f(x)| \le L\|\Delta\|$ for small Δ. Similarly, because $v \mapsto f'(x;v)$ is convex and sublinear, it has a Lipschitz constant, and we take this to be L as well. Now, write $\Delta_n = \epsilon_n v_n$ where $\|v_n\| = 1$ and $\epsilon_n \to 0$, and moving to a subsequence if necessary let $v_n \to v$. Then

$$\begin{aligned} f(x+\Delta_n) - f(x) - f'(x;\Delta_n) &= f(x+\epsilon_n v + \epsilon_n(v_n - v)) - f(x) - f'(x;\epsilon_n v + \epsilon_n(v_n - v)) \\ &= f(x+\epsilon_n v) - f(x) - f'(x;\epsilon_n v) \pm 2L\epsilon_n \|v_n - v\| \\ &= o(\epsilon_n) \end{aligned}$$

because $\|v_n - v\| \to 0$ and $f(x+\epsilon_n v) - f(x) = f'(x;\epsilon_n v) + o(\epsilon_n)$ by definition of the directional derivative. □

Convexity only played the role of establishing the local Lipschitz property of f in the proof of Proposition B.2.33; any locally Lipschitz function with directional derivatives will enjoy a similar first-order expansion.

As our final result on smoothness properties of convex functions, we connect subdifferentials to differentiability properties of convex f. First, we give a lemma showing that the subdifferential set ∂f is outer semicontinuous.

Lemma B.2.34 (Closure of the graph of the subdifferential)**.** *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then the graph $\{(x,s) \mid x \in \mathbb{R}^d, s \in \partial f(x)\}$ of its subdifferential is closed. Equivalently, whenever $x_n \to x$ with $s_n \in \partial f(x_n)$ and $s_n \to s$, f has non-empty subdifferential at x with $s \in \partial f(x)$.*

Proof We prove the second statement, whose equivalence to the first is definitional. Fix any $y \in \mathbb{R}^d$. Then $f(y) \ge f(x_n) + \langle s_n, y - x_n\rangle$, and because f is closed (i.e., lower semicontinuous), we have $\liminf f(x_n) \ge f(x)$. Let $\epsilon > 0$ be arbitrary. Then for all large enough n, we have $f(x_n) \ge f(x) - \epsilon$, and similarly, $\|s_n - s\| \le \epsilon$, $\|x_n - x\| \le \epsilon$, and $\|y - x_n\| \le \|y - x\| + \epsilon$. Then

$$\begin{aligned} f(y) \ge f(x_n) + \langle s_n, y - x_n\rangle &\ge f(x) + \langle s, y - x_n\rangle - \epsilon - \|s - s_n\| \|y - x_n\| \\ &\ge f(x) + \langle s, y - x\rangle - \epsilon - \epsilon\|y - x_n\| - \|s\| \|x - x_n\| \\ &\ge f(x) + \langle s, y - x\rangle - \epsilon - \epsilon(1+\epsilon)\|y - x\| - \|s\|\epsilon. \end{aligned}$$

As ϵ was arbitrary we have $f(y) \ge f(x) + \langle s, y - x\rangle$ as desired. □

Given the somewhat technical Lemma B.2.34, we can show that if f is convex and differentiable at a point, it is in fact continuously differentiable at the point.

Proposition B.2.35. *Let f be convex and $x \in \operatorname{int}\operatorname{dom} f$. Then $\partial f(x)$ is a singleton if and only if f is differentiable at x. If additionally f is differentiable on an open set U, then f is continuously differentiable on U.*

Proof Because $x \in \operatorname{int}\operatorname{dom} f$, there exists $L < \infty$ such that f is L-Lipschitz near x by Theorem B.2.6. Suppose that $\partial f(x) = \{s\}$. Then the directional derivative $f'(x;v) = \langle s, v\rangle$ for all v, and Proposition B.2.33 gives

$$f(y) = f(x) + \langle s, y - x\rangle + o(\|y - x\|)$$

as $y \to x$, that is, f is differentiable. Conversely, assume that f is differentiable at x. Then taking any vector v, $f'(x;v) = \langle \nabla f(x), v\rangle$ and Proposition B.2.31 gives that $\partial f(x) = \{\nabla f(x)\}$.

To see that f is continuously differentiable on U, let $x \in U$ and f be L-Lipschitz on a compact set $C \subset U$ containing x in its interior. Let $x_k \in \operatorname{int} C$ satisfy $x_k \to x$ and let $s_k = \nabla f(x_k) \in \partial f(x_k)$. Then $\|s_k\| \le L$, and each subsequence has a further convergent subsequence. Lemma B.2.34 implies that any convergent subsequence $s_{k(m)} \to s \in \partial f(x)$. As $\partial f(x) = \{\nabla f(x)\}$, we have $\nabla f(x_{k(m)}) \to \nabla f(x)$ and so $\nabla f(x)$ is continuous in x. □

B.2.7 Calculus rules of subgradients

We close this section with a few calculus results on subdifferentials of convex functions. These calculus rules exhibit ways that the subdifferential plays a similar role to the gradient for differentiable functions. Additionally, they allow us to take derivatives of various extremal functions.

Our first result shows that subdifferentials of sums are sums of subdifferentials, which relies on both the representation of sublinear functions as support functions for convex sets and the characterization of the subdifferential in terms of directional derivatives:

Proposition B.2.36. *Let f and g be closed convex functions, and let $x \in \operatorname{int} \operatorname{dom} f$ and g be subdifferentiable at x, meaning that $\partial g(x) \neq \varnothing$. Then*

$$\partial(f + g)(x) = \partial f(x) + \partial g(x).$$

Proof By Proposition B.2.31, the set $\partial f(x)$ is a compact convex set, and the general definition (B.2.18) of the subdifferential gives that $\partial g(x)$ is closed convex. Let $S_1 = \partial f(x)$ and $S_2 = \partial g(x)$. Then immediately $S_1 + S_2 \subset \partial(f + g)(x)$, so that

$$S := \partial(f + g)(x) = \left\{ s \mid \langle s, v \rangle \le f'(x; v) + g'(x; v) \text{ for all } v \in \mathbb{R}^d \right\}$$

is non-empty. Because of the support function equality $f'(x; v) = \sigma_{S_1}(v)$ and $g'(x; v) = \sigma_{S_2}(v)$, Corollary B.2.18 gives

$$\sigma_S(v) = \sigma_{S_1}(v) + \sigma_{S_2}(v) = \sigma_{S_1 + S_2}(v).$$

Thus (Corollary B.2.17) $S_1 + S_2 = S$. □

The subdifferential sum rule, which we only state, holds when the domains of functions overlap.

Proposition B.2.37 (Rockafellar [363], Thm. 23.8)**.** *Let f and g be convex with* $\operatorname{relint} \operatorname{dom} f \cap \operatorname{relint} \operatorname{dom} g \neq \varnothing$. *Then $\partial(f + g)(x) = \partial f(x) + \partial g(x)$ for all x.*

Other situations that arise frequently are composition with affine mappings, so that finding a calculus for these is also important.

Corollary B.2.38. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and for $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$, let $g(x) = f(Ax + b)$. Then*

$$\partial g(x) = A^\top \partial f(Ax + b).$$

Proof Using the directional derivative, we have $g'(x; v) = f'(Ax + b; Av)$ for all $v \in \mathbb{R}^d$, and applying Proposition B.2.19 gives that the latter is the support function of the convex compact set $A^\top \partial f(Ax + b)$. □

It is also useful to be able to compute subdifferentials of maxima and suprema (recall Proposition B.2.21). Consider a collection $\{f_\alpha\}_{\alpha \in \mathcal{A}}$ of convex functions, and define

$$f(x) := \sup_{\alpha \in \mathcal{A}} f_\alpha(x). \tag{B.2.19}$$

The function f is certainly convex. For a given x let

$$\mathcal{A}(x) := \{\alpha \in \mathcal{A} \mid f_\alpha(x) = f(x)\}$$

be the indices attaining the suprema, that is, the active index set (though this may be empty). Then there is an "easy" direction:

Lemma B.2.39. *For the function* (B.2.19)*,*

$$\partial f(x) \supset \operatorname{cl} \operatorname{Conv} \left\{ \bigcup \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x) \right\} = \operatorname{cl} \operatorname{Conv} \left\{ g \mid g \in \partial f_\alpha(x) \text{ for some } \alpha \in \mathcal{A}(x) \right\}.$$

Proof Let $\alpha \in \mathcal{A}(x)$ and $g \in \partial f_\alpha(x)$. Then

$$f(y) \geq f_\alpha(y) \geq f_\alpha(x) + \langle g, y - x \rangle = f(x) + \langle g, y - x \rangle.$$

Thus $g \in \partial f(x)$, which as a closed convex set must include its closed convex hull. □

A more challenging argument shows when the active index set $\mathcal{A}(x)$ characterizes the subdifferential of f at x; we state a typical result as a proposition. (Recall Figure B.16(b).)

Proposition B.2.40. *Let $\mathcal{A}$ be a compact set (for some metric space), and assume that for each x, the mapping $\alpha \mapsto f_\alpha(x)$ is upper semi-continuous. Assume there is a neighborhood U of x such that each f_α is finite on U. Then*

$$\partial f(x) = \operatorname{Conv} \left\{ \bigcup \partial f_\alpha(x) \mid \alpha \in \mathcal{A}(x) \right\} = \operatorname{Conv} \left\{ g \mid g \in \partial f_\alpha(x) \text{ for some } \alpha \in \mathcal{A}(x) \right\}.$$

For a proof, see Hiriart-Urruty and Lemaréchal [236, Theorem VI.4.4.2].

B.3 Exercises

Exercise B.1: This question outlines a proof of Theorem B.1.23. Let $C \subset \mathbb{R}^d$ be closed convex, not an affine subspace, $A = \operatorname{aff}(C)$, and $L = A - A = \operatorname{span}(C - C)$ be the subspace parallel to A.

(a) Let $y \in A \setminus C$. Let $v = y - \pi_C(y)$, and show that the halfspace $H = \{x \mid \langle v, x \rangle \leq \langle v, \pi_C(y) \rangle\}$ is a non-trivial supporting halfspace at $\pi_C(y)$.

(b) Let $y \notin A$. Decompose $y = y_A + y_\perp$, where $y_A \in A$ and $y_\perp \in L^\perp$, $y_\perp \neq 0$. Fix $x \in \operatorname{rel\,bd} C$ and use Theorem B.1.20 to exhibit a vector $u_t = v + t y_\perp$, where v gives a non-trivial supporting halfspace to C at x and $t \in \mathbb{R}$, strictly separating y from C.

(c) Conclude Theorem B.1.23.

Hint. Corollary B.1.21 may help with non-triviality.

Exercise B.2: Prove Corollary B.2.5

Exercise B.3: Prove Proposition B.2.16.

Exercise B.4: This exercise defines the tangent cone to a closed set S. A direction v is *tangent* to S at $x \in S$ if there exist sequences $t_k > 0$ and $x_k \in S$, $x_k \to x$, for which

$$t_k \downarrow 0, \quad x_k \to x, \quad \frac{1}{t_k}(x_k - x) \to v. \tag{B.3.1}$$

(a) Show that v is tangent if and only if there exist $v_k \to v$ and $t_k \downarrow 0$ for which $x + t_k v_k \in S$.

(b) Let K be the set of all tangent directions v at $x \in S$. Show that K is a closed cone.

(c) Show that the set K of all tangent directions v at $x \in S$ satisfies

$$K = \limsup_{t \downarrow 0} \frac{1}{t}(S - x) = \limsup_{t \downarrow 0} \left\{ \frac{1}{t}(s - x) \mid s \in S \right\},$$

where the limit supremum $\limsup_k A_k$ of sets consists of subsequent limits of points in A_k, that is, v for which there exists a subsequence $k_m \subset \mathbb{N}$ and $v_{k_m} \in A_{k_m}$ with $v_{k_m} \to v$.

(d) Show that v is tangent (B.3.1) if and only if for the distance function $\operatorname{dist}(x, S) = \inf_{y \in S} \|y - x\|$, $\liminf_{t \downarrow 0} \frac{1}{t} \operatorname{dist}(x + tv, S) = 0$.

Exercise B.5: Let C be a closed convex set and K be the collection of tangent directions (B.3.1) at $x \in C$. Recall that $\mathbb{R}_+ C = \{tx \mid t \geq 0, x \in C\}$.

(a) Show that $C - x \subset K$.

(b) Show that $\operatorname{cl} \mathbb{R}_+(C - x) \subset K$.

(c) Show that

$$K = \operatorname{cl} \mathbb{R}_+(C - x) = \operatorname{cl} \left\{ v \mid v = \alpha(y - x), \; y \in C, \alpha \geq 0 \right\}.$$

Exercise B.6 ([235], Sec. III.5)**:** Let C be a closed convex set and $x \in C$. Let $K = \operatorname{cl} \mathbb{R}_+(C - x)$, that is, let K be the collection of tangent directions (B.3.1) (by the result of Exercise B.5).

(a) Show that $\mathcal{N}_C(x) = -K^*$, that is, the normal cone to C at x is the polar of K.

(b) Define the set $\mathcal{T}_C(x) = -\mathcal{N}_C(x)^*$. Show that $\mathcal{T}_C(x) = K$.

Exercise B.7 (A discontinuous closed convex function)**:** Define the function $f : \mathbb{R}^2 \to \overline{\mathbb{R}}$ by

$$f(x) := \sup \left\{ \alpha x_1 + \beta x_2 \mid \frac{1}{2}\alpha^2 \leq \beta \right\}.$$

(a) Why is f closed convex?

(b) Give an explicit formula for f and show that f is discontinuous at $\mathbf{0}$.

Exercise B.8: Show that the first part of Theorem B.1.22, that if C is a closed convex set, then $C = \cap_{H \supset C} H$ is the intersection of all closed halfspaces containing it, holds in Hilbert space.

Exercise B.9: Let $C \subset \mathcal{H}$ be a closed convex subset of a Hilbert space $\mathcal{H}$ with non-empty interior, and let $x \in \mathrm{bd}(C) = C \setminus \operatorname{int} C$. Use the Banach-Alaoglu theorem[1] to show that Theorem B.1.20 extends to this case, that is, there exists a non-zero vector $v \in \mathcal{H}$ for which $\langle v, x \rangle \geq \sup_{y \in C} \langle v, y \rangle$.

Exercise B.10: Let $\mathcal{H} = \ell^2(\mathbb{N})$, the set of square summable sequences, and $C = \mathcal{H} \cap \mathbb{R}^{\mathbb{N}}_+$, the subset of nonnegative sequences. Let $x = (1, 1/2, 1/3, 1/4, \ldots) \in C$ have jth coordinate $x_j = 1/j$.

(a) Show that $x \in \operatorname{bd} C = \{y \in C \mid \operatorname{dist}(y, C^c) = 0\}$.

(b) Show that if $v \in \mathcal{N}_C(x)$, that is, $\langle v, y - x \rangle \leq 0$ for all $y \in C$, then $v = 0$. That is, $\mathcal{N}_C(x) = \{0\}$.

(c) Conclude that some type of qualification—such as that C has an interior—is necessary for Theorem B.1.20 to extend to Hilbert space.

Exercise B.11: Show that Theorem B.2.9 holds in Hilbert space.

Exercise B.12: Let $\mathcal{H}$ be a Hilbert space. We say a cone $K \subset \mathcal{H}$ is non-trivial if $K \neq \{0\}$ and $K \neq \mathcal{H}$. Let K be a non-trivial closed convex cone. Show that the dual cone K^* is non-trivial.

Exercise B.13: Let $C \subset \mathcal{H}$ be a closed convex subset of a Hilbert space $\mathcal{H}$ with an interior. Let $x \in \operatorname{bd} C$ and $\mathcal{T}_C(x)$ be the set of tangent directions (B.3.1). Show that $\mathcal{T}_C(x)$ is a non-trivial closed convex cone. *Hint.* Exercises B.4–B.6 do not use that C is finite dimensional.

Exercise B.14: Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper closed convex function with $\operatorname{int} \operatorname{dom} f \neq \varnothing$. Show that Theorem B.2.4 extends to this case. *Hint.* You may use Bauschke and Combettes [54, Corollary 8.30], which guarantees that f is continuous on $\operatorname{int} \operatorname{dom} f$.

[1] In Hilbert spaces, the Banach-Alaoglu theorem states that if v_n is a sequence of vectors with $\|v_n\| \leq 1$, then there exists a weakly convergent subsequence, that is, a vector v with $\|v\| \leq 1$ and a subsequence $n(k)$ such that $\langle v_{n(k)}, x \rangle \to \langle v, x \rangle$ for each $x \in \mathcal{H}$. For example, the sequence $e_1, e_2, e_3, \ldots$ in $\ell^2(\mathbb{N})$ has weak limit $\mathbf{0}$.

Appendix C

Convex analysis: optimality, stability, and duality

The existence and continuity properties of minimizers of (convex) optimization problems play a central role in much of statistical theory. They are especially essential in our understanding of loss functions and the associated optimality properties, as in Chapters 14, 16, and 19. This appendix records several results along these lines, and it also builds the duality theory of convex conjugates.

For an objective f and closed convex set C, we consider the generic optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C, \end{array} \tag{C.0.1}$$

where we assume $C \cap \operatorname{dom} f \neq \emptyset$. We assume that f is proper, so that $f(x) > -\infty$ for each x, and as usual $\operatorname{dom} f = \{x \mid f(x) < \infty\}$. The most basic question is when minimizers exist in the problem (C.0.1), prompting the standard result that minimizers are attained whenever C is compact and f is lower semicontinuous (B.2.9), that is, its epigraph is closed:

Proposition C.0.1. *Let C be compact and $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be lower semi-continuous (B.2.9) over C. Then $\inf_{x \in C} f(x) > -\infty$ and the infimum is attained.*

Proof Let $f^\star = \inf_{x \in C} f(x)$. Let $x_n \in C$ be a sequence of points satisfying $f(x_n) \to f^\star$. Proceeding to a subsequence, we can assume that $x_n \to x^\star \in C$ by the compactness of C. Then lower semicontinuity guarantees that $f^\star = \lim_n f(x_n) \geq f(x^\star) \geq f^\star$, and so $f(x^\star) = f^\star > -\infty$. □

When the domain C may be noncompact, alternative conditions guarantee the existence of minimizers. The most frequent is that f is *coercive*, meaning that

$$f(x) \to \infty \quad \text{whenever} \quad \|x\| \to \infty.$$

Proposition C.0.2. *Let C be closed and $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be lower semi-continuous over C and coercive. Then $\inf_{x \in C} f(x) > -\infty$ and the infimum is attained.*

Proof Let $b > \inf_{x \in C} f(x)$. Apply Proposition C.0.1 on the compact set $C \cap \{x \mid f(x) \leq b\}$. □

Finally, norms and their duals will be important for quantitative smoothness guarantees to come. For a norm $\|\cdot\|$,

$$\|y\|_* := \sup_x \{\langle x, y \rangle \mid \|x\| \leq 1\}$$

defines the *dual norm*. This is a norm as it is positively homogeneous, $\|y\|_* = 0$ if and only if $y = 0$, and it satisfies the triangle inequality. Exercise C.1 identifies several dual norms.

C.1 Optimality conditions and stability properties

With the basic results on existence of minimizers in place, we turn to convex optimization problems, where f is closed convex and C is a closed convex set, and we assume essentially without loss of generality that $\dom f \supset \inte C$ (as otherwise, we may replace C with $C \cap \cl \dom f$). The benefits of convexity appear immediately: f has no local but non-global minimizers, and moreover, if f is strictly convex, then any minimizers (if they exist) are unique.

Proposition C.1.1. *Let f be convex. Then if x is a local minimizer of f over C, it is a global minimizer of f over C. If f is strictly convex, this minimizer is unique.*

Proof To say that x is a local minimizer of f over C is to say that $f(x) \le f(x')$ for all $x' \in C$ with $\|x' - x\| \le \epsilon$ for some $\epsilon > 0$. Consider $y \in C$. By taking $\lambda > 0$ small enough, we have both $(1-\lambda)x + \lambda y \in C$ and $\|(1-\lambda)x + \lambda y - x\| \le \epsilon$, and so

$$f(x) \le f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y),$$

which rearranged yields $f(y) \ge f(x)$. If f is additionally strictly convex (recall Corollary B.2.3), the preceding inequality is strict whenever $y \neq x$. □

C.1.1 Subgradient characterizations for optimality

First-order stationary conditions are sufficient for global optimality in convex problems. We can say more once we consider subgradients:

Observation C.1.2. *Let f be convex and subdifferentiable at x. Then x minimizes f if and only if $0 \in \partial f(x)$.*

Proof If $0 \in \partial f(x)$, then $f(y) \ge f(x) + \langle 0, y - x\rangle = f(x)$ for all y. Conversely, if x minimizes f, then we have $f(y) \ge f(x)$ for all y, and in particular, $0 \in \partial f(x)$. □

Things become more complicated when we consider the constraints in the problem (C.0.1), so that the point x may be restricted. Recalling the normal cone (Def. B.6)

$$\mathcal{N}_C(x) = \{v \mid \langle v, y - x\rangle \le 0 \text{ for all } y \in C\} \quad \text{for } x \in C$$

(where $\mathcal{N}_C(x) = \varnothing$ for $x \notin C$), we may rewrite optimality conditions through subgradient calculus. If C is a closed convex set, the convex indicator function $\mathbf{I}_C(x)$ has subdifferential

$$\partial \mathbf{I}_C(x) = \mathcal{N}_C(x).$$

This subdifferential yields the following optimality characterization for problem (C.0.1), as Figure C.1 illustrates.

Proposition C.1.3. *In the problem (C.0.1), let $x \in \inte \dom f$. Then x minimizes f over C if and only if*

$$0 \in \partial f(x) + \mathcal{N}_C(x). \tag{C.1.1}$$

Proof The minimization problem (C.0.1) is equivalent to the problem

$$\operatorname*{minimize}_x \; f(x) + \mathbf{I}_C(x).$$

As $x \in \operatorname{int} \operatorname{dom} f$, f has nonempty compact convex subdifferential $\partial f(x)$, and so $\partial(f + \mathbf{I}_C)(x) = \partial f(x) + \partial \mathbf{I}_C(x) = \partial f(x) + \mathcal{N}_C(x)$ by Proposition B.2.36. Apply Observation C.1.2. □

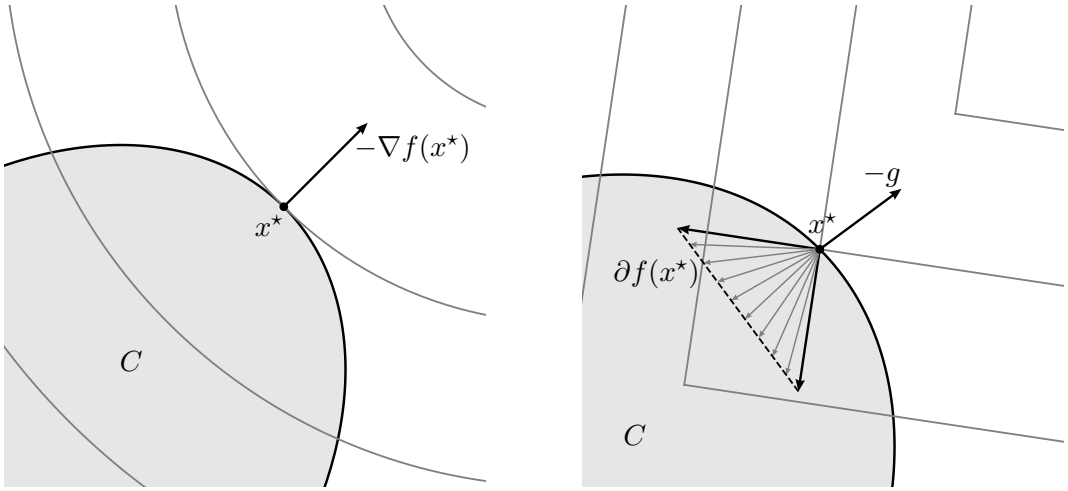


Figure C.1. Level curves of two convex functions f and a convex body C. Left: the point $x^\star$ minimizes the differentiable function f over C at $x^\star$, and $-\nabla f(x^\star) \in \mathcal{N}_C(x^\star) = \operatorname{span}\{-\nabla f(x^\star)\}$. Right: the point $x^\star$ minimizes f over C and $g \in \partial f(x^\star)$, illustrated as the cone-shaped collection of vectors. Here, $g \in \partial f(x^\star)$ satisfies $\langle g, y - x^\star \rangle \geq 0$ for all $y \in C$.

Because Proposition C.1.3 provides a general characterization of optimality, it is useful to have equivalent reformulations to use as circumstances dictate. Several equivalent versions of Proposition C.1.3 are possible. The first is that

$$-\partial f(x) \cap \mathcal{N}_C(x) \neq \varnothing,$$

that is, there is a subgradient vector $g \in \partial f(x)$ such that $-g \in \mathcal{N}_C(x)$, so that $-g$ points outside the set C. Another variant is that $x \in C$ solves problem (C.0.1) if and only if there exists $g \in \partial f(x)$ such that

$$\langle g, y - x \rangle \geq 0 \;\; \text{for all } y \in C. \tag{C.1.2}$$

Indeed, taking $g \in \partial f(x)$ to be the element satisfying $-g \in \mathcal{N}_C(x)$, then $\langle -g, y - x \rangle \leq 0$ for all $y \in C$ by definition of the normal cone, which is (C.1.2). (Again, see Figure C.1.)

C.1.2 Stability properties of minimizers

The characterizations (C.1.1) and (C.1.2) of optimality for convex optimization problems allow us to develop stability properties of the minimizers of convex problems. These, in turn, will relate to smoothness properties of various dual functions (as we explore in the sequel), which again become important in the study of loss functions. Typical results in this vein elucidate the equivalences between solution stability, meaning that small tilts of the function f do not change solutions significantly, and that the function f exhibits some local growth around the solution.

For our starting point, we consider strongly convex functions, where a function is λ-*strongly convex* with respect to the norm $\|\cdot\|$ if for all $t \in [0,1]$ and $x, y \in \dom f$,

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{\lambda}{2}t(1-t)\|x-y\|^2. \tag{C.1.3}$$

The definition (C.1.3) makes strict convexity quantitative and has several equivalent characterizations.

Proposition C.1.4 (Equivalent characterizations of strong convexity)**.** *Let f be a convex function, subdifferentiable on its domain. Then the following are equivalent.*

(i) f is λ-strongly convex (Eq. (C.1.3)).

(ii) For all $x \in \dom f$, y, and $g \in \partial f(x)$,

$$f(y) \ge f(x) + \langle g, y - x\rangle + \frac{\lambda}{2}\|x-y\|^2.$$

(iii) For all $x, y \in \dom f$ and $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y\rangle \ge \lambda\|x-y\|^2.$$

Proof We first demonstrate (ii) if and only if (iii). Let $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$ and assume (ii), so $f(y) \ge f(x) + \langle g_x, y-x\rangle + \frac{\lambda}{2}\|y-x\|^2$. Adding the symmetric inequality interchanging x and y gives

$$0 \ge \langle g_x - g_y, y - x\rangle + \lambda\|x-y\|^2.$$

Conversely, assume (iii), and for $t \in [0,1]$ let $x_t = (1-t)x + ty$ and $h(t) = f(x_t)$. Then h is convex, locally Lipschitz, and a.e. differentiable, so $h(1) = h(0) + \int_0^1 h'(t)dt$. The subgradient characterization of Proposition B.2.31 implies $h'(t) = \langle g_t, y - x\rangle$ for some $g_t \in \partial f(x_t)$, so

$$h'(t) = \langle g_t, y-x\rangle = \langle g_t - g_x, y-x\rangle + \langle g_x, y-x\rangle = \frac{1}{t}\langle g_t - g_x, (1-t)x + ty - x\rangle + \langle g_x, y-x\rangle.$$

As $h(1) = f(y)$ and $h(0) = f(x)$,

$$\begin{aligned} f(y) &= h(0) + \int_0^1 \frac{\langle g_t - g_x, (1-t)x + ty - x\rangle}{t}dt + \langle g_x, y-x\rangle \\ &\ge f(x) + \int_0^1 \frac{\lambda\|(1-t)x + ty - x\|^2}{t}dt + \langle g_x, y-x\rangle = f(x) + \langle g_x, y-x\rangle + \frac{\lambda}{2}\|y-x\|^2. \end{aligned}$$

That (ii) implies (i) is straightforward: for $g_t \in \partial f(tx + (1-t)y)$ we have

$$\begin{aligned} f(y) &\ge f(tx + (1-t)y) + t\langle g_t, y-x\rangle + \frac{\lambda}{2}t^2\|x-y\|^2 \\ f(x) &\ge f(tx + (1-t)y) + (1-t)\langle g_t, x-y\rangle + \frac{\lambda}{2}(1-t)^2\|x-y\|^2. \end{aligned}$$

Multiply the first inequality by $(1-t)$ and the second by t, then add them to obtain

$$tf(x) + (1-t)f(y) \ge f(tx + (1-t)y) + \frac{\lambda}{2}\left[(1-t)t^2 + t(1-t)^2\right]\|x-y\|^2,$$

and note that $(1-t)t^2+t(1-t)^2 = t(1-t)$. Finally, assume (i), which is equivalent to the condition

$$\frac{f((1-t)x+ty)-f(x)}{t} + \frac{\lambda}{2}(1-t)\left\|x-y\right\|^2 \le f(y)-f(x)$$

for $t \in (0,1)$. Taking $t \downarrow 0$ gives $f'(x;y-x) + \frac{\lambda}{2}\left\|x-y\right\|^2 \le f(y)-f(x)$, and because $f'(x;y-x) = \sup_{s\in\partial f(x)}\langle s, y-x\rangle$ we obtain (ii). □

As a first example application of strong convexity, consider minimizers of the tilted functions

$$f_u(x) := f(x) - \langle u, x\rangle$$

as u varies. Minimizers necessarily exist: the function $f_u(x) \to \infty$ whenever $\|x\| \to \infty$ by condition (ii) in Proposition C.1.4, and so we can restrict to minimizing f_u over compacta. Moreover, the minimizers $x_u := \operatorname{argmin}_x f_u(x)$ are unique, as the functions f_u are strongly (and hence strictly) convex. We can say more. Let C be any closed convex set and let

$$x_u = \operatorname*{argmin}_{x\in C} f_u(x). \tag{C.1.4}$$

Proposition C.1.5. *Let f be λ-strongly convex with respect to the norm $\|\cdot\|$ and subdifferentiable on C. Then the mapping $u \mapsto x_u$ is $\frac{1}{\lambda}$-Lipschitz continuous with respect to the dual norm $\|\cdot\|_*$, that is, $\|x_u - x_v\| \le \frac{1}{\lambda}\|u-v\|_*$.*

Proof We use the optimality condition (C.1.2). We have $\partial f_u(x) = \partial f(x) - u$, and thus for any u, v there exist $g_u \in \partial f(x_u)$ and $g_v \in \partial f(x_v)$ for which

$$\langle g_u - u, y - x_u\rangle \ge 0 \quad \text{and} \quad \langle g_v - v, y - x_v\rangle \ge 0$$

for all $y \in C$. Set $y = x_v$ in the first inequality and $y = x_u$ in the second and add them to obtain

$$\langle g_u - g_v + v - u, x_v - x_u\rangle \ge 0 \;\; \text{or} \;\; \langle v-u, x_v - x_u\rangle \ge \langle g_v - g_u, x_v - x_u\rangle.$$

By strong convexity the last term satisfies $\langle g_v - g_u, x_v - x_u\rangle \ge \lambda\left\|x_u - x_v\right\|^2$. By definition of the dual norm, $\|v-u\|_*\|x_v - x_u\| \ge \langle v-u, x_v-x_u\rangle$, so $\|u-v\|_*\|x_v-x_u\| \ge \lambda\|x_u-x_v\|^2$, which is the desired result. □

Uniform convexity conditions generalize strong convexity but provide similar quantitative stability properties. In analogy with the strong convexity condition (C.1.3), we say that f is (λ,κ)-uniformly convex, where $\kappa \ge 2$, over C if it is closed and for all $t \in [0,1]$ and $x, y \in C$,

$$f(tx+(1-t)y) \le tf(x) + (1-t)f(y) - \frac{\lambda}{2}t(1-t)\left\|x-y\right\|^\kappa\left[(1-t)^{\kappa-1} + t^{\kappa-1}\right]. \tag{C.1.5}$$

The $\kappa = 2$ case is simply the familiar strong convexity property. Taking $\kappa > 2$ weakens strong convexity, yielding correspondingly weaker guarantees of stability and optimality. We can nonetheless provide analogies to Propositions C.1.4 and C.1.5.

Proposition C.1.6 (Equivalent characterizations of uniform convexity)**.** *Let f be a convex function, subdifferentiable on its domain. Then the following are equivalent:*

(i) f is (λ,κ)-uniformly convex with respect to the norm $\|\cdot\|$.

(ii) For all $x \in \dom f$, $y \in \mathbb{R}^d$, and $g \in \partial f(x)$,

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\lambda}{2} \|x - y\|^\kappa .$$

Additionally, either of (i) or (ii) imply that

(iii) For all $x, y \in \dom f$ and $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y \rangle \geq \lambda \|x - y\|^\kappa ,$$

which in turn implies that f is $(\frac{2}{\kappa}\lambda, \kappa)$-uniformly convex with respect to the norm $\|\cdot\|$.

We leave the proof of the proposition as Exercise C.2, as it follows via the same arguments as those we use to prove the strong convexity version (Proposition C.1.4). An analog of Proposition C.1.5 also holds with an alternative smoothness condition.

Proposition C.1.7. *Let f be (λ, κ)-uniformly convex with respect to the norm $\|\cdot\|$ and subdifferentiable on C. Then the mapping $u \mapsto x_u$ is $\frac{1}{\kappa-1}$-Hölder, and in particular,*

$$\|x_u - x_v\| \leq \lambda^{-\frac{1}{\kappa-1}} \|u - v\|_*^{\frac{1}{\kappa-1}}$$

for all u, v.

Proposition C.1.7 follows as an immediate consequence of Proposition C.2.7 to come, where we connect smoothness of the minimizers x_u with differentiability properties of the conjugate function.

The weakest growth property is strict convexity, for which a careful reading of the proof of Proposition C.1.4 (replace all λ with 0 and inequalities with strict inequalities) gives the following characterization of equivalent definitions of strict convexity (recall also Corollary B.2.3).

Corollary C.1.8. *Let f be a convex function subdifferentiable on C. The following are equivalent.*

(i) f is strictly convex on C.

(ii) For all distinct $x, y \in C$, and $g \in \partial f(x)$, $f(y) > f(x) + \langle g, y - x \rangle$.

(iii) For all distinct $x, y \in C$ and $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y \rangle > 0.$$

Using Corollary C.1.8, we can obtain qualitative smoothness properties of the tilted minimizers x_u of the minimization (C.1.4). We state a sequence of lemmas, which guarantee growth of convex functions over their first-order approximations, though we leave their proofs as exercises, because their main consequence (Proposition C.1.11) also follows essentially as a corollary of Proposition C.2.8 below. We begin by recalling the definition (14.2.2) of the Bregman divergence. For a fixed selection $g \in \partial f(x_0)$, write

$$D_f(x, x_0) := f(x) - f(x_0) - \langle g, x - x_0 \rangle.$$

Lemma C.1.9. *Let f be convex and subdifferentiable on the closed convex set C. Let $0 < \epsilon \leq \epsilon'$ be otherwise arbitrary. Then if $\delta(\epsilon) := \inf\{D_f(x, x_0) \mid x \in C, \|x - x_0\| \geq \epsilon\}$ is finite, it is attained, nonnegative, and $\delta(\epsilon') \geq \frac{\epsilon'}{\epsilon}\delta(\epsilon)$.*

Exercise C.4 provides a hint to the proof of this lemma. Whenever f is strictly convex, because the infimum in $\delta(\epsilon)$ is attained in Lemma C.1.9, we have the following guarantee.

Lemma C.1.10. *Let the conditions of Lemma C.1.9 hold and additionally let f be strictly convex. Then $\delta(\epsilon) > 0$ for all $\epsilon > 0$.*

Combining these results yields the following qualitative version of Proposition C.1.5:

Proposition C.1.11. *Let f be strictly convex and subdifferentiable on the closed convex set C, and assume that the minimum $x_0 = \operatorname{argmin}_{x \in C} f(x)$ is attained. Then the mapping $u \mapsto x_u$ is continuous in a neighborhood of $u = 0$.*

Exercise C.5 outlines a proof using Lemma C.1.10.

C.2 Conjugacy and duality properties

Attached to any function is its *convex conjugate*, sometimes called the *Fenchel* or *Fenchel-Legendre* conjugate function, defined by

$$f^*(s) := \sup_x \left\{ \langle s, x \rangle - f(x) \right\}. \tag{C.2.1}$$

For any f, the conjugate f^* is a closed convex function, as it is the supremum of linear functions of s. This function exhibits a duality for convex functions analogous to the halfspace representation of closed convex sets (Theorem B.1.22 and equalities (B.1.3)–(B.1.4)). Figure C.2 illustrates the conjugate function.

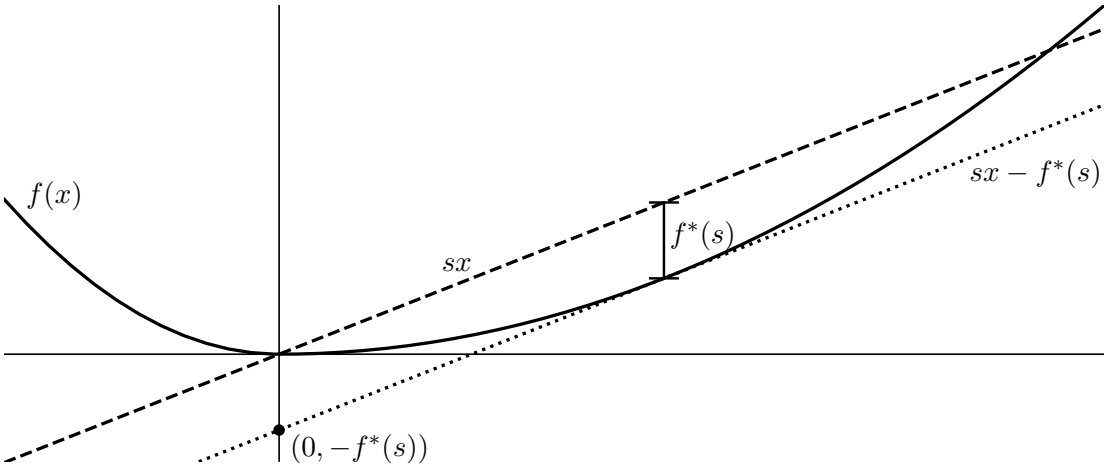


Figure C.2. Illustration of the conjugate of $f(x)$ with linear function $l(x) = \langle s, x \rangle$ (dashed line) and affine function $h(x) = \langle s, x \rangle - f^*(s)$ (dotted line). The largest gap between $\langle s, x \rangle$ and $f(x)$ equals $f^*(s)$, and $h(x)$ meets the graph of $f(x)$ at the point maximizing the difference $\langle s, x \rangle - f(x)$, where $s \in \partial f(x)$. Sweeping the functions $h_s(x) = \langle s, x \rangle - f^*(s) \le f(x)$ over $s \in \mathbb{R}^d$ traces the affine lower bounds on f, providing $\operatorname{cl}\operatorname{Conv} f(x) = \sup_s h_s(x)$ (as in the definition (B.2.12)).

The conjugate function is the largest gap between the linear functional $x \mapsto \langle s, x \rangle$ and the function f itself. The remarkable property of conjugates is that their biconjugates describe the the largest closed convex function below f (which is f itself when it is closed convex).

Theorem C.2.1. *Let f be closed convex and f^* be its conjugate (C.2.1). Then*

$$f^{**}(x) = f(x) \quad \text{for all } x.$$

Proof By definition,

$$f^{**}(x) = \sup_s \left\{\langle x, s\rangle - f^*(s)\right\},$$

and we always have $\langle x, s\rangle - f^*(s) \le f(x)$ by definition of $f^*(s) = \sup_y\{\langle s, y\rangle - f(y)\}$. So immediately we see that $f^{**}(x) \le f(x)$.

We show that the linear functions $h_s(x) := \langle x, s\rangle - f^*(s)$ describe (enough) of the global linear underestimators of f so that $f(x) = \sup_s h_s(x)$, allowing us to apply Theorem B.2.9. Indeed, let $l(x) = \langle s, x\rangle + b$ be any global underestimator of f. Then we must have $b \le f(x) - \langle s, x\rangle$ for all x, that is, $b \le \inf_x\{f(x) - \langle s, x\rangle\} = -\sup_x\{\langle s, x\rangle - f(x)\} = -f^*(s)$, so $l(x) \le \langle s, x\rangle - f^*(s) = h_s(x)$. Apply Theorem B.2.9. □

We may visualize f^{**} as pulling a string up below a function f, yielding the largest closed convex underestimator of f. Recall the definition (B.2.12) of the closed convex hull of a function f. Then combining Theorem C.2.1 with Proposition B.2.11 and Corollary B.2.13, we obtain the following corollary. (Note that we require $f(x) > -\infty$ for all x.)

Corollary C.2.2. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ and f^* be its convex conjugate (C.2.1). Then*

$$f^{**} = \operatorname{cl}\operatorname{Conv} f.$$

If f is convex, then f is lower semicontinuous at x if and only if $f(x) = \operatorname{cl} f(x) = f^{**}(x)$.

C.2.1 Gradient dualities and the Fenchel-Young inequality

It is immediate from the definition that for any pair s, x we have the *Fenchel-Young inequality*

$$\langle s, x\rangle \le f^*(s) + f(x). \tag{C.2.2}$$

Combining Theorem C.2.1 with this observation, we can exhibit a duality between subgradients of f and f^* with this inequality.

Proposition C.2.3. *Let f be closed convex. Then*

$$\langle s, x\rangle = f^*(s) + f(x) \quad \text{if and only if} \quad s \in \partial f(x) \quad \text{if and only if} \quad x \in \partial f^*(s).$$

Proof If $\langle s, x\rangle = f^*(s) + f(x)$, then $-f(x) + \langle s, x\rangle = f^*(s) \ge \langle s, y\rangle - f(y)$ for all y, and rearranging, we have $f(y) \ge f(x) + \langle s, y - x\rangle$, that is, $s \in \partial f(x)$. Conversely, if $s \in \partial f(x)$ then $0 \in \partial f(x) - s$, so that x minimizes $f(x) - \langle s, x\rangle$, or equivalently, x maximizes $\langle s, \cdot\rangle - f(\cdot)$ and so $\langle s, x\rangle - f(x) = \sup_y\{\langle s, y\rangle - f(y)\}$ as desired. The final statement is immediate from a parallel argument and the duality in Theorem C.2.1. □

Writing Proposition C.2.3 differently, we see that ∂f and ∂f^* are inverses of one another as set-valued mappings, where

$$(\partial f)^{-1}(s) := \{x \mid s \in \partial f(x)\}.$$

We have the following corollary.

Corollary C.2.4. *Let f be closed convex. Then*

$$\partial f^* = (\partial f)^{-1} \quad \text{and} \quad \partial f = (\partial f^*)^{-1}$$

and

$$\partial f^*(s) = \operatorname*{argmax}_x \left\{\langle s, x\rangle - f(x)\right\} \quad \text{and} \quad \partial f(x) = \operatorname*{argmax}_s \left\{\langle s, x\rangle - f^*(s)\right\}.$$

If f and f^* are differentiable, then $\nabla f = (\nabla f^*)^{-1}$.

Additionally, we see that the domains and images of ∂f and ∂f^* are also related.

Corollary C.2.5. *Let f be closed convex. Then*

$$\operatorname{dom} \partial f = \operatorname{Im} \partial f^* \quad \text{and} \quad \operatorname{dom} \partial f^* = \operatorname{Im} \partial f.$$

Proof Let $x \in \operatorname{dom} \partial f$, so that $\partial f(x)$ is nonempty. Then $s \in \partial f(x)$ implies that $\langle s, x\rangle = f(x) + f^*(s)$ and $x \in \partial f^*(s)$ by Proposition C.2.3. Similarly, if $x \in \operatorname{Im} \partial f^*$, then there is some s for which $x \in \partial f^*(s)$ and so $\langle s, x\rangle = f(x) + f^*(s)$ and $s \in \partial f(x)$. □

C.2.2 Smoothness and strict convexity of conjugates

The dualities in derivative mappings extend to several useful smoothness dualities. These types of results build from the stability properties of solution mappings, as in those for tilted minimizers (C.1.4) in Propositions C.1.5 and C.1.11. They also relate different smoothness properties of f and f^*, as well as their domains of definition, to the existence and continuity of minimizers of $f(x) - \langle s, x\rangle$.

When we assume that f has quantitative strong convexity or smoothness properties, we can give similar quantitative guarantees for the smoothness and strong convexity of f^*:

Proposition C.2.6. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be λ-strongly convex with respect to the norm $\|\cdot\|$ (see Eq. (C.1.3)) on its domain. Then $\operatorname{dom} f^* = \mathbb{R}^d$ and ∇f^* is $\frac{1}{\lambda}$-Lipschitz continuous with respect to the dual norm $\|\cdot\|_*$, that is, for all u, v,*

$$\|\nabla f^*(u) - \nabla f^*(v)\| \le \frac{1}{\lambda} \|u - v\|_*.$$

Conversely, let $f : \mathbb{R}^d \to \mathbb{R}$ be convex with L-Lipschitz gradient with respect to $\|\cdot\|$. Then f^* is $\frac{1}{L}$-strongly convex with respect to the dual norm $\|\cdot\|_*$ on convex subsets $C \subset \operatorname{dom} \partial f^*$, and

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \ge \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2. \tag{C.2.3}$$

Proof For the first claim, let $C = \operatorname{dom} f$. Proposition C.1.5 shows that the unique values $x_1 = \operatorname{argmin}_x\{f(x) - \langle s_1, x\rangle\}$ and $x_2 = \operatorname{argmin}_x\{f(x) - \langle s_2, x\rangle\}$ satisfy $\|x_1 - x_2\| \le \frac{1}{\lambda} \|s_1 - s_2\|_*$. Then Proposition C.2.3 shows that $x_i \in \partial f^*(s_i)$ for $i = 1, 2$, and hence $\partial f^*(s_i)$ is necessarily single-valued and $(1/\lambda)$-Lipschitz continuous.

The converse is trickier. Let x and y be arbitrary and $s_x = \nabla f(x)$ and $s_y = \nabla f(y)$; we prove inequality (C.2.3), known as *co-coercivity*. By the L-Lipschitz continuity of ∇f, we have

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\le f(x) + \langle s_x, y - x \rangle + \int_0^1 Lt \left\| y - x \right\|^2 dt = f(x) + \langle s_x, y - x \rangle + \frac{L}{2} \left\| y - x \right\|^2, \end{aligned}$$

which is valid for any x, y. Note that $f(x) - \langle s_x, x \rangle = -f^*(s_x)$, so that rearranging we have

$$\begin{aligned} f^*(s_x) \le \langle s_x, y \rangle - f(y) + \frac{L}{2} \left\| y - x \right\|^2 &= \langle s, y \rangle - f(y) + \langle s_x - s, y \rangle + \frac{L}{2} \left\| y - x \right\|^2 \\ &\le f^*(s) + \langle s_x - s, y \rangle + \frac{L}{2} \left\| y - x \right\|^2, \end{aligned}$$

valid for any vector s and any y. We may take an infimum over y on the right hand side, where

$$\begin{aligned} \inf_y \langle s_x - s, y \rangle + \frac{L}{2} \left\| y - x \right\|^2 &= \inf_z \langle s_x - s, z \rangle + \frac{L}{2} \left\| z \right\|^2 + \langle s_x - s, x \rangle \\ &= -\frac{1}{2L} \left\| s_x - s \right\|_*^2 + \langle s_x - s, x \rangle, \end{aligned}$$

by definition of the dual norm. Thus

$$f^*(s) \ge f^*(s_x) + \langle x, s - s_x \rangle + \frac{1}{2L} \left\| s - s_x \right\|_*^2$$

for all s. As $x \in \partial f^*(s_x)$, Proposition C.1.4(ii) gives the strong convexity result. The rest is algebraic manipulations with $s_y = \nabla f(y)$ and an application of Proposition C.1.4, part (iii). □

We can extend Proposition C.2.6 to the uniformly convex case (recall Proposition C.1.6), though for this we require a definition of smoothness beyond Lipschitz continuity of the gradients.

Definition C.1. *A function f is $(L, \beta, \|\cdot\|)$-smooth if it has β-Hölder continuous gradient, $0 < \beta \le 1$, with respect to the norm $\|\cdot\|$, that is,*

$$\left\| \nabla f(x) - \nabla f(y) \right\|_* \le L \left\| x - y \right\|^\beta \quad \text{for} \quad x, y \in \operatorname{dom} f.$$

We then have the following analog of Proposition C.2.6.

Proposition C.2.7. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be (λ, κ)-uniformly convex (C.1.5) with respect to the norm $\|\cdot\|$. Then $\operatorname{dom} f^* = \mathbb{R}^d$ and ∇f^* is $(\lambda^{-\frac{1}{\kappa - 1}}, \frac{1}{\kappa - 1}, \|\cdot\|_*)$-smooth. Conversely, let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $(L, \beta, \|\cdot\|)$-smooth. Then f^* is uniformly convex with respect to the dual norm $\|\cdot\|_*$ on any convex subset of $\operatorname{dom} \partial f^*$, and in particular, for $x_0 \in \partial f^*(s_0)$,*

$$f^*(s_1) \ge f^*(s_0) + \langle x_0, s_1 - s_0 \rangle + \frac{\beta}{\beta + 1} L^{-\frac{1}{\beta}} \left\| s_0 - s_1 \right\|_*^{\frac{\beta + 1}{\beta}}.$$

Exercise C.3 asks for a proof of this proposition, which more or less follows from a similar technique as that we use to prove Proposition C.2.6. In short, uniform convexity and smoothness are dual to one another (up to constants): for $\kappa \geq 2$, f is uniformly convex with power κ if and only if ∇f^* is Hölder continuous with exponent $1/(\kappa - 1)$.

There are more qualitative versions of Proposition C.2.6 that allow us to give a duality between strict convexity and continuous differentiability of f. We give one typical result.

Proposition C.2.8. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be strictly convex and closed. Then $\operatorname{int}\operatorname{dom} f^* \neq \varnothing$ and f^* is continuously differentiable on $\operatorname{int}\operatorname{dom} f^*$. Conversely, let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be differentiable on $\Omega := \operatorname{int}\operatorname{dom} f$. Then f^* is strictly convex on each convex $C \subset \nabla f(\Omega)$.*

These results follow naturally from the subgradient duality $\partial f = (\partial f^*)^{-1}$ and that $\partial f^*(s) = \operatorname{argmax}_x\{\langle s, x\rangle - f(x)\}$, because strict convexity makes this argmax unique, whence ∂f^* a singleton.

Proof To see that $\operatorname{int}\operatorname{dom} f^*$ is nonempty, we use the identification $f'_\infty(v) = \sigma_{\operatorname{dom} f^*}(v)$ in Proposition C.3.3 and the interior identification in Proposition B.2.20. Because f is strictly convex, for any $x \in \operatorname{dom} f$ we have

$$0 < \frac{f(x - tv) - f(x)}{t} + \frac{f(x + tv) - f(x)}{t} \quad \text{for } t > 0,$$

and taking $t \to \infty$ gives $0 < f'_\infty(-v) + f'_\infty(v)$. Proposition B.2.20 then shows that $\operatorname{int}\operatorname{dom} f^* \neq \varnothing$.

For the claim that f^* is continuously differentiable, take $s \in \operatorname{int}\operatorname{dom} f^*$, and suppose for the sake of contradiction that $\partial f^*(s)$ contains distinct points x_1, x_2. Then Corollary C.2.4 gives that x_1 and x_2 both minimize $f(x) - \langle s, x\rangle$ over x, while Proposition C.1.1 guarantees $x_1 = x_2$, so that $\partial f^*(s) = \{\nabla f^*(s)\}$ is a singleton, and hence f^* is continuously differentiable at s (Proposition B.2.35).

For the converse claim, let $C \subset \nabla f(\Omega)$ be a convex set. Suppose for the sake of contradiction that f^* is not strictly convex on C, so that there are distinct points $s_1, s_2 \in C$ for which f^* is affine on the line segment $[s_1, s_2] = \{ts_1 + (1 - t)s_2 \mid t \in [0, 1]\}$. As C is convex, the midpoint $s = \frac{1}{2}(s_1 + s_2) \in C$ and there exists x satisfying $\nabla f(x) = s$, or $x \in \partial f^*(s)$. Then because f^* is assumed affine on $[s_1, s_2]$, we have $f^*(s) = \frac{1}{2}f^*(s_1) + \frac{1}{2}f^*(s_2)$ and $\langle s, x\rangle = \frac{1}{2}\langle s_1 + s_2, x\rangle$, so

$$\begin{aligned} 0 &= f(x) + f^*(s) - \langle s, x\rangle \\ &= \frac{1}{2}\left[(f(x) + f^*(s_1) - \langle s_1, x\rangle) + (f(x) + f^*(s_2) - \langle s_2, x\rangle)\right]. \end{aligned}$$

Each of the terms in parenthesis is 0 if and only if $s_i \in \partial f(x)$, but by assumption $\partial f(x) = \{\nabla f(x)\}$ is a singleton, and we must have $s_1 = s_2$. □

We will later develop a class of functions for which both f and f^* are strictly convex, smooth, and whose domains we can explicitly control, but this will require an understanding of limits $f(x)$ as $\|x\| \to \infty$, so we return to this later.

C.2.3 Jensen's inequality

Jensen's inequality, which we first introduce in Chapter 2, connects naturally to conjugate duality. We provide two proofs of the inequality. Recall first Proposition 2.1.6, which we prove here.

Proposition C.2.9. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and X a random vector with $\mathbb{E}[\|X\|] < \infty$ and $\mathbb{E}[X] \in \operatorname{int}\operatorname{dom} f$. Then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)],$$

and moreover, if f is strictly convex, then $f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$ unless X is constant.

Proof Because $\mathbb{E}[X] \in \operatorname{int} \operatorname{dom} f$, $\partial f(\mathbb{E}[X])$ is non-empty (Theorem B.2.4). Letting $g \in \partial f(\mathbb{E}[X])$, observe that

$$\mathbb{E}[f(X)] - f(\mathbb{E}[X]) = \mathbb{E}[f(X) - f(\mathbb{E}[X]) - \langle g, X - \mathbb{E}[X] \rangle] \geq 0.$$

Using that $\mathbb{E}[\|X\|] < \infty$ gives the first claim of the proposition. For the strict inequality, recall from Corollary C.1.8 that $f(x) > f(\mathbb{E}[X]) + \langle g, x - \mathbb{E}[X] \rangle$ for $x \neq \mathbb{E}[X]$. Then equality with 0 holds above only if $f(X) - \langle g, X - \mathbb{E}[X] \rangle = f(\mathbb{E}[X])$ with probability 1, i.e., $X = \mathbb{E}[X]$. □

When f is closed convex, we need only assume X is integrable. This gives rise to the following variant, which extends to Hilbert spaces (cf. Exercise C.7).

Proposition C.2.10. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and X a random vector with $\mathbb{E}[\|X\|] < \infty$. Then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)],$$

where if the left side is infinite, then so is the right.

Proof Let X have distribution P. Then because f is closed convex, Theorem C.2.1 shows that

$$\begin{aligned} f(\mathbb{E}[X]) = \sup_y \left\{ \langle \mathbb{E}[X], y \rangle - f^*(y) \right\} &= \sup_y \int \left(\langle x, y \rangle - f^*(y) \right) dP(x) \\ &\leq \int \sup_y \left\{ \langle x, y \rangle - f^*(y) \right\} dP(x) = \mathbb{E}[f(X)]. \end{aligned}$$

If $\mathbb{E}[X] \notin \operatorname{dom} f$, then $f(\mathbb{E}[X]) = +\infty$, and the inequalities hold presuming that $+\infty \leq +\infty$. □

C.3 Limits at infinity of convex functions and sets

Section C.2 showed the links between conjugate functions and smoothness properties of gradients, making some connections with the existence of minimizers of convex functions. Here, we take this a step further by providing more abstract conditions for the existence of minimizers of convex functions. Part of this requires a description of different set limits—we wish to understand when we need not take points $\|x\| \to \infty$ to approach $\inf_x f(x)$—so we begin there.

C.3.1 Boundedness and closedness of convex sets

The boundedness and closedness of convex sets is central to both the growth properties of convex functions—via their epigraphs—as well as several min-max duality theorems in the coming section. The former will relate to the existence of minimizers via *recession cones* and *recession functions* associated with convex sets and functions. We develop some of the relevant theory here.

Determining the boundedness and closure of convex sets often reduces to checking the existence of (infinite) rays remaining within the set. To reduce notational complexity, throughout this section we will assume that $C \subset \mathbb{R}^d$ is a closed convex set, making note when results extend beyond this case. Our starting point is a characterization of various limiting directions that lie within convex sets. For a point $x \in C$, we define the *recession cone*

$$C_\infty(x) := \left\{ v \in \mathbb{R}^d \mid x + tv \in C \text{ for all } t \geq 0 \right\} \tag{C.3.1}$$

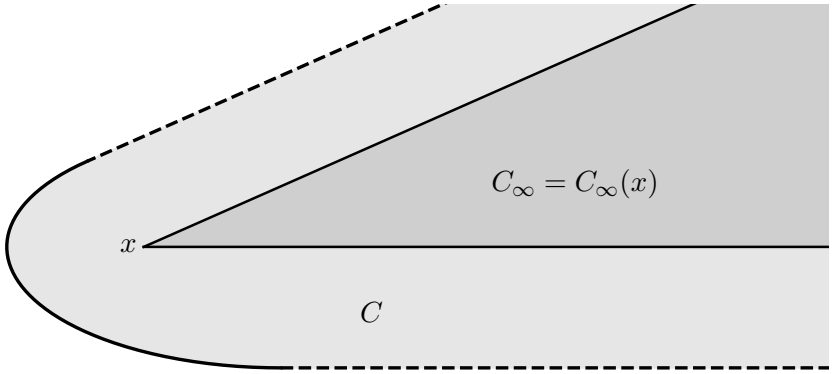


Figure C.3: The recession cone C_∞ associated with the set C.

to be the set of directions at which C extends off to ∞. (See Figure C.3.) It is immediate that $C_\infty(x)$ is a cone, because $v \in C_\infty(x)$ implies $tv \in C_\infty(x)$ for any $t \geq 0$. This definition is in fact independent of the choice of x:

Proposition C.3.1. *Let C be closed convex. The set $C_\infty(x)$ is a closed convex cone, independent of x, and*

$$C_\infty(x) = \bigcap_{t>0} \frac{1}{t}(C - x).$$

Proof Let $x_0, x_1 \in C$. For the claim that $C_\infty(x)$ is independent of x, it is enough to show that $C_\infty(x_0) \subset C_\infty(x_1)$. So let $v \in C_\infty(x_0)$ and $t \geq 0$, so that $x_0 + tv \in C$, and for $\epsilon \in [0, 1]$ let

$$y_\epsilon = (1 - \epsilon)x_0 + \epsilon x_1 + tv = (1 - \epsilon)\left(x_0 + \frac{t}{1 - \epsilon}v\right) + \epsilon x_1.$$

Then $y_\epsilon \in C$ for each $\epsilon \in (0, 1)$ because $x_0 + \frac{t}{1-\epsilon}v \in C$, and $y_1 = \lim_{\epsilon \uparrow 1} y_\epsilon = x_1 + tv \in C$ as C is closed.

To see the definition of $C_\infty(x)$ in terms of the intersection, note that $v \in C_\infty(x)$ if and only if $tv \in C - x$ for all $t > 0$, that is, $v \in \frac{1}{t}(C - x)$ for all $t > 0$. That $C_\infty(x)$ is thus the intersection of closed convex sets implies it is closed convex. □

In view of Proposition C.3.1, we can more accurately write C_∞ for the recession cone of a convex set C. The next lemma enumerates a few properties of such recession cones.

Lemma C.3.2. *Let C and $\{C_\alpha\}_{\alpha \in \mathcal{A}}$ be a (respectively, consist of) closed convex set(s). Then*

(i) $C_\infty = \{v \mid C + \mathbb{R}_+ v \subset C\}$ *and* $C = C + C_\infty$.

(ii) If $\{C_\alpha\}_{\alpha \in \mathcal{A}}$ has nonempty intersection, then $(\cap_\alpha C_\alpha)_\infty = \cap_\alpha (C_\alpha)_\infty$.

(iii) Let u be a unit vector. Then $u \in C_\infty$ if and only if there exists a sequence $x_n \in C$ with $\|x_n\| \to \infty$ while $x_n / \|x_n\| \to u$.

(iv) C is bounded if and only if $C_\infty = \{0\}$, that is, it has no directions of recession.

Proof The first claim is immediate from the definition of the recession cone and Proposition C.3.1, so we consider the remainder.

(ii) We have $v \in (C_\alpha)_\infty$ for each $\alpha \in \mathcal{A}$ if and only if for some $x \in \cap_\alpha C_\alpha$, $x + tv \in C_\alpha$ for each α and $t \geq 0$, that is, $x + tv \in \cap_\alpha C_\alpha$ for all $t \geq 0$ and $v \in (\cap_\alpha C_\alpha)_\infty$.

(iii) Let $u \in C_\infty$ be a unit vector. Let $x \in C$, and take $x_n = x + nu$ for $n \in \mathbb{N}$. Then clearly $x_n / \|x_n\| \to u$, while $x_n \in C$ for each n. Conversely, assume that $x_n / \|x_n\| \to u$ while $\|x_n\| \to \infty$ and $x_n \in C$. Let $t \geq 0$ and N be large enough that $\|x_n\| \geq t$ for $n \geq N$. Then

$$x + tu = \lim_n \underbrace{\left(1 - \frac{t}{\|x_n\|}\right) x + \frac{t}{\|x_n\|} x_n}_{\text{in } C \text{ for } n \geq N}.$$

That is, $x + tu$ is the limit of points in C, so that the closedness of C implies $x + tu \in C$.

(iv) We certainly have $C_\infty = \{0\}$ when C is bounded. If C is unbounded, then we may take a sequence $x_n \in C$ with $\|x_n\| \to \infty$, and by considering convergent subsequences, may assume w.l.o.g. that $x_n / \|x_n\| \to u$ for some $u \in \mathbb{S}^{d-1}$. Then $C_\infty \supset \{u\}$ is nonempty by part (iii).

□

C.3.2 Asymptotic growth and existence of minimizers

We can use the identification between the domains of ∂f and the images of ∂f^* to give additional characterizations of the domains of convex functions and their conjugates; the domain of f^* is intimately tied with the growth properties of f. As one example of this identification, note that if f^* is defined everywhere, that is, $\operatorname{dom} f^* = \mathbb{R}^d$, then similarly $\operatorname{dom} \partial f^* = \mathbb{R}^d$, and so the (sub)gradients of f must have image $\mathbb{R}^d$. Even more, as we shall see, this implies certain growth conditions on f.

To be rigorous, we require functions capturing the asymptotic growth of f, analogizing the recession cones (C.3.1). The following proposition defines the *recession function*, essentially an asymptotic directional derivative of f, to do this.

Proposition C.3.3. *Let f be a closed convex function and f^* its convex conjugate. Then for any $x \in \operatorname{dom} f$, we may define*

$$f'_\infty(v) := \sup_{t > 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t \to \infty} \frac{f(x + tv) - f(x)}{t} \tag{C.3.2}$$

independently of x, and moreover,

$$f'_\infty(v) = \sigma_{\operatorname{dom} f^*}(v) = \sigma_{\operatorname{cl} \operatorname{dom} f^*}(v),$$

where $\sigma_{\operatorname{dom} f^*}$ is the support function (B.2.13) of $\operatorname{dom} f^*$.

Proof That for any fixed $x \in \operatorname{dom} f$ the limit (C.3.2) exists and is equal to the supremum follows from the criterion of increasing slopes (B.2.5), giving the equality with the supremum. That $f'_\infty(v)$ is independent of x will follow once we show the equality $f'_\infty(v) = \sigma_{\operatorname{dom} f^*}(v)$.

Recall from Theorem C.2.1 that as f is closed convex,

$$\operatorname{dom} f^* = \left\{ s \mid \sup_x \{\langle s, x\rangle - f(x)\} < \infty \right\} \quad \text{and} \quad f(x) = \sup_s \left\{\langle s, x\rangle - f^*(s)\right\}.$$

Fix $x \in \operatorname{dom} f$. Then for any $s \in \operatorname{dom} f^*$, we evidently have

$$\frac{f(x+tv) - f(x)}{t} \geq \frac{\langle s, x+tv\rangle - f^*(s) - f(x)}{t} \to \langle s, v\rangle$$

as $t \uparrow \infty$. Taking a supremum over $s \in \operatorname{dom} f^*$ gives that $f'_\infty(v) \geq \sigma_{\operatorname{dom} f^*}(v)$. For the opposite direction, note that

$$\begin{aligned}\frac{f(x+tv) - f(x)}{t} &= \frac{1}{t}\left[\sup_{s \in \operatorname{dom} f^*} \{\langle s, x+tv\rangle - f^*(s)\} - \sup_{s \in \operatorname{dom} f^*} \{\langle s, x\rangle - f^*(s)\}\right] \\ &\leq \frac{1}{t} \sup_{s \in \operatorname{dom} f^*} \{\langle s, x+tv\rangle - f^*(s) - (\langle s, x\rangle - f^*(s))\} = \frac{1}{t} \sup_{s \in \operatorname{dom} f^*} t\langle s, v\rangle.\end{aligned}$$

Thus $f'_\infty(v) \leq \sigma_{\operatorname{dom} f^*}(v)$, and we have the result. □

A simple corollary of this result shows that the perspective transform of a 1-dimensional closed convex function can be written in terms of its recession function:

Corollary C.3.4. *Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be closed convex. Then*

$$\operatorname{cl} f_{\text{per}}(x, u) = \begin{cases} uf(x/u) & \text{if } u > 0 \\ f'_\infty(x) & \text{if } u = 0 \\ +\infty & \text{if } u < 0. \end{cases}$$

Proof By Corollary B.2.25, we need only consider $u = 0$. Proposition C.3.3 shows that

$$\sigma_{\operatorname{dom} f^*}(x) \leq \operatorname{cl} f_{\text{per}}(x, 0) \leq \liminf_{t \to \infty} \sigma_{\operatorname{dom} f^*}\left(x + t^{-1}(x_0 - x)\right)$$

for any $x_0 \in \operatorname{relint} \operatorname{dom} f$. For one-dimensional functions, we know that $\operatorname{cl} \operatorname{dom} f^* = [a, b]$, where $-\infty \leq a \leq b \leq \infty$, and so $\lim_{t\to\infty} \sigma_{\operatorname{dom} f^*}(x + t^{-1}v) = \sigma_{\operatorname{dom} f^*}(x)$ whenever $x \neq 0$. If $x = 0$, then Corollary B.2.25 shows that $\operatorname{cl} f_{\text{per}}(0, 0) = \lim_{\alpha \downarrow 0} \alpha f(x_0) = 0$ for any $x_0 \in \operatorname{relint} \operatorname{dom} f$. □

It is particularly interesting to understand the conditions under which $\operatorname{dom} f^* = \mathbb{R}^d$, that is, f^* is finite everywhere, and relatedly, under which the function $x \mapsto f(x) - \langle s, x\rangle$ has a minimizer. Recall that $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is *coercive* if $f(x) \to \infty$ whenever $\|x\| \to \infty$. For convex functions, the recession function of f characterizes coercivity.

Proposition C.3.5. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then f is coercive if and only if $f'_\infty(v) > 0$ for all $v \neq 0$, that is, f is coercive on each line.*

Proof If f is coercive, then for any $v \neq 0$, $f(x + tv) \to \infty$ as $t \to \infty$, so that for any $x \in \operatorname{dom} f$, $f(x + tv) - f(x) > 0$ for large enough t. In particular,

$$f'_\infty(v) = \sup_{t > 0} \frac{f(x+tv) - f(x)}{t} > 0$$

by Proposition C.3.3. Conversely, assume that $f'_\infty(v) > 0$ for all $v \neq 0$. Then if f is not coercive, it must have an unbounded level set $S_\alpha := \{x \mid f(x) \le \alpha\}$. As S_α is a convex set, if it is unbounded, then there exists a unit vector $v \in (S_\alpha)_\infty$, the recession cone of S_α by Lemma C.3.2. But then $x + tv \in S_\alpha$ for all $t \ge 0$ implies that $f(x + tv) \le \alpha$ for all $t \ge 0$, a contradiction to $f'_\infty(v) > 0$. $\square$

We call f *super-coercive* if $f(x)/\|x\| \to \infty$ whenever $\|x\| \to \infty$, so that f grows more than linearly. These concepts are central to the existence of minimizers. A priori, any function with compact domain is super-coercive, because $f(x) = +\infty$ for $x \notin \operatorname{dom} f$. The recession function (C.3.2) characterizes these functions as well. Particularly important are those f satisfying

$$f'_\infty(v) = +\infty \quad \text{for all } v \neq 0,$$

a class Rockafellar [363] calls *copositive* functions, as these exhibit superlinear growth on all rays toward infinity. We can relate this condition to the domains of f^* as well: using Proposition B.2.20, Propositions C.3.3 and C.3.5 give

Corollary C.3.6. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then $s \in \operatorname{cl}\operatorname{dom} f^*$ if and only if $\langle s, v\rangle \le f'_\infty(v)$ for all $v \neq 0$, and $s \in \operatorname{int}\operatorname{dom} f^*$ if and only if $\langle s, v\rangle < f'_\infty(v)$ for all $v \neq 0$. In particular,*

(i) If f is coercive, or, equivalently, if $f'_\infty(v) > 0$ for all $v \neq 0$, then $0 \in \operatorname{int}\operatorname{dom} f^*$ and f has a minimizer.

(ii) We have $f'_\infty(v) = +\infty$ for all $v \neq 0$ if and only if

$$\operatorname{dom} f^* = \mathbb{R}^d.$$

A sufficient condition for this is that f be super-coercive.

Proof Combine Propositions B.2.20, C.3.3, and C.3.5: for part (i), note that if $f'_\infty(v) > 0$ for all $v \neq 0$, then $0 \in \operatorname{int}\operatorname{dom} f^*$, and so f^* has a non-trivial subdifferential $\partial f^*(0)$ at 0; letting $x \in \partial f^*(0)$ we have $x \in \operatorname{argmin} f$. Part (ii) is similarly immediate. $\square$

As an example consequence, if f is closed convex and $c = \inf_{\|v\|=1} f'_\infty(v) > 0$ is the minimal value of the recession function, then the tilted function $f(\cdot) - \langle s, \cdot\rangle$ has a minimizer if $\|s\|_* < c$.

C.3.3 Essentially smooth convex functions

We close this section by investigating particularly nice classes of functions f, where f and its conjugate f^* are strictly convex and continuously differentiable. These results are central to the conjugate linkage dualities in Chapter 14.3. We therefore make the following definition:

Definition C.2. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex. Then f is of* Legendre type *if*

(i) $\operatorname{int}\operatorname{dom} f \neq \varnothing$

(ii) f is continuously differentiable on $\operatorname{int}\operatorname{dom} f$

(iii) f is strictly convex

(iv) f satisfies the gradient boundary conditions

$$\|\nabla f(x)\|_2 \to \infty \quad \text{as} \quad x \to \operatorname{bd}\operatorname{dom} f \quad \text{or} \quad \|x\|_2 \to \infty. \tag{C.3.3}$$

Thus, at the boundaries of their domains or as their argument tends off to infinity, functions of Legendre type have slopes tending to ∞. This does not guarantee that $f(x) \to \infty$ as $x \to \bd \dom f$, though it does provide guarantees of regularity that the next theorem highlights.

Theorem C.3.7. *Let f be a convex function of Legendre type (Def. C.2). Then f^* is strictly convex, continuously differentiable, and $\dom f^* = \mathbb{R}^d$.*

The theorem implies a number of results on continuity of minimizers and tilted minimizers (C.1.4), clarifying some of our earlier results. For example, we have the following corollary.

Corollary C.3.8. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a convex function of Legendre type. Then the tilted minimizer*

$$x_u := \argmin\{f(x) - \langle u, x\rangle\}$$

exists for all u, is continuous in u and unique, and $x_u \in \int \dom f$.

We turn to the proof of the theorem.

Proof of Theorem C.3.7 We state an intermediate lemma, whose proof we defer.

Lemma C.3.9. *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and satisfy the gradient boundary condition that $\|s_n\|_2 \to \infty$ for any sequence $x_n \to \bd \dom f$ and $s_n \in \partial f(x_n)$. Then the recession function*

$$f'_\infty(v) = \infty \text{ for all } v \neq 0$$

if and only if

$$\|s_n\|_2 \to \infty \quad \text{whenever} \quad \|x_n\|_2 \to \infty \text{ and } s_n \in \partial f(x_n).$$

The theorem follows straightforwardly from Lemma C.3.9. The boundary conditions (C.3.3) imply that $f'_\infty(v) = \infty$ for all $v \neq 0$, so $\dom f^* = \mathbb{R}^d$ by Corollary C.3.6. With this, Proposition C.2.8 shows that f^* is continuously differentiable and strictly convex on its domain. $\square$

Proof of Lemma C.3.9 As $\dom f^* = \mathbb{R}^d$ if and only if $f'_\infty(v) = \infty$ for all $v \neq 0$ (Corollary C.3.6), it suffices to show the result that $\int \dom f^* \neq \mathbb{R}^d$ if and only if there exists an unbounded sequence x_n with $s_n \in \partial f(x_n)$ and for which s_n is convergent.

Let us begin with the unbounded sequence x_n for which $s_n \to s \in \mathbb{R}^d$; assume for the sake of contradiction that $s \in \int \dom f^*$. Because $s_n \in \partial f(x_n)$, we have $x_n \in \partial f^*(s_n)$ by Proposition C.2.3. The assumption that $s \in \int \dom f^*$ means that there exists an $\epsilon > 0$ such that $s + \epsilon\mathbb{B} \subset \int \dom f^*$ and f^* is Lipschitz on $s + \epsilon\mathbb{B}$ (Theorem B.2.6). But then $\partial f^*(s + \epsilon\mathbb{B})$ is bounded, and $x_n \in \partial f^*(s_n) \subset \partial f^*(s + \epsilon\mathbb{B})$ for large enough n, contradicting that $\|x_n\|_2 \to \infty$, and so $s \notin \int \dom f^*$ and $\int \dom f^* \neq \mathbb{R}^d$.

Now let us assume that $\int \dom f^* \neq \mathbb{R}^d$. Let $s \in \bd \dom f^*$. Then either $\partial f^*(s) = \varnothing$ or $\partial f^*(s)$ is unbounded (Lemma B.2.32). If $\partial f^*(s) = \varnothing$, take $s_n \to s$ with $s_n \in \relint \dom f^*$, and let $x_n \in \partial f^*(s_n)$. We show that x_n must be unbounded. If x_n is bounded, then by passing to a subsequence if necessary we may assume $x_n \to x$, and the outer semicontinuity of the subdifferential (Lemma B.2.34) gives $x \in \partial f^*(s)$, contradicting that $\partial f^*(s) = \varnothing$. Thus we must have x_n unbounded, which is thus the desired unbounded sequence. On the other hand, if $\partial f^*(s)$ is unbounded, we can simply take $x_n \in \partial f^*(s)$ with $s \in \partial f(x_n)$ for each n, which is the desired convergent sequence. $\square$

As a last application of these ideas, in some cases we wish to allow constraints on the functions f to be minimized, returning to the original convex optimization problem (C.0.1) with f a function of Legendre type and C a closed convex set. We then have the following corollary.

Corollary C.3.10. *Let f be of Legendre type (Definition C.2) and $C \subset \mathbb{R}^d$ a closed convex set with $\operatorname{int}\operatorname{dom} f \cap C \neq \varnothing$. Define $f_C(x) = f(x) + \mathbf{I}_C(x)$. Then*

(i) f_C^* is continuously differentiable,

(ii) $\operatorname{dom} f_C^* = \mathbb{R}^d$, and

(iii) the constrained tilted minimizers

$$x_u = \operatorname*{argmin}_{x \in C} \{f(x) - \langle u, x \rangle\}$$

are unique, continuous in u, belong to $\operatorname{int}\operatorname{dom} f$, and for some $v \in \mathcal{N}_C(x_u)$ satisfy

$$x_u = \nabla f^*(u - v) \quad \text{and} \quad \nabla f(x_u) = u - v.$$

Proof The function $f_C := f + \mathbf{I}_C$ is closed convex. To show that $\operatorname{dom} f_C^* = \mathbb{R}^d$, we can equivalently show that $(f_C)'_\infty(v) = \infty$ for all non-zero v. Because f is Legendre type, Lemma C.3.9 guarantees that if $x \in \operatorname{int}\operatorname{dom} f \cap C$, then

$$(f_C)'_\infty(v) = \lim_{t \uparrow \infty} \frac{f(x + tv) + \mathbf{I}_C(x + tv) - f(x)}{t} \geq \lim_{t \uparrow \infty} \frac{f(x + tv) - f(x)}{t} = f'_\infty(v) = \infty.$$

So $\operatorname{dom} f_C^* = \mathbb{R}^d$, and thus $x_u = \operatorname{argmin}_{x \in C}\{f(x) - \langle u, x \rangle\} = \nabla f_C^*(u)$ exists and is unique and continuous, as f is strictly convex.

The boundary condition in Definition C.2 rules out boundary minimizers, so $x_u \in \operatorname{int}\operatorname{dom} f$. The standard subgradient conditions for optimality (Proposition C.1.3) then give

$$0 \in \nabla f(x_u) - u + \mathcal{N}_C(x_u).$$

Choose $v \in \mathcal{N}_C(x_u)$ for which $\nabla f(x_u) = u - v$, whence Corollary C.2.4 implies $x_u = \nabla f^*(u - v)$. □

C.4 Bibliography and further reading

There are many references on convex analysis and optimization, beginning with the foundational book by Rockafellar [363], which initiated the study of convex functions and optimization in earnest. Rockafellar in particular made the deep connections between convex conjugates, closures of functions, and the existence of solutions via recession cones and support functions. Many of the more nuanced results we present are essentially his; for example, our proof of Proposition C.3.3 directly follows [363, Theorem 13.3]. Since then, a variety of authors have written books on convexity that, by virtue of including figures, can be more approachable. Hiriart-Urruty and Lemaréchal [235, 237] provide two volumes explaining finite-dimensional convex analysis and some optimization methods. The abridged version [238] containing the central pieces of convex analysis is a book any statistician, information theorist, or optimizer should have at ready. Borwein and Lewis [79] and Luenberger

[295] give general treatments that include infinite-dimensional convex analysis. Bertsekas [66] provides a view of convex analysis with an eye toward the theory of optimization, and accordingly, develops many duality and saddle-point results as well.

There are books that combine theoretical treatment with questions of convex modeling and procedures for solving convex optimization problems (problems for which the objective and constraint sets are all convex). Boyd and Vandenberghe [82] give a very readable treatment for those who wish to use convex optimization techniques and modeling. Ben-Tal and Nemirovski [56], as well as Nemirovski's lecture notes [320, 321], give a theory of the tractability of convex optimization problems. Recent interest in first-order optimization methods, motivated by machine learning applications, has also underpinned several texts, including Nesterov's and Bertsekas's books [324, 67].

C.5 Exercises

Exercise C.1: Show each of the following:

(a) The ℓ_2-norm $\|x\|_2 = \sqrt{\langle x, x\rangle}$ is self-dual, i.e., its dual is $\|\cdot\|_2$.

(b) The ℓ_1 and ℓ_∞ norms are dual: $\|x\|_\infty = \sup_{\|y\|_1 \le 1} \langle x, y\rangle$ and $\|y\|_1 = \sup_{\|x\|_\infty \le 1} \langle x, y\rangle$.

(c) For all $1 < p < \infty$, the dual to the ℓ_p norm $\|x\|_p = (\sum_{j=1}^d |x_j|^p)^{1/p}$ is the ℓ_q norm with $q = \frac{p}{p-1}$.

Exercise C.2: Prove Proposition C.1.6. *Hint:* See the proof of Proposition C.1.4.

Exercise C.3: Prove Proposition C.1.7. *Hint:* See the proof of Proposition C.1.5.

Exercise C.4*: Prove Lemma C.1.9. *Hint.* Define $h(t) = D_f(x_0 + t(x - x_0), x_0)$ and use the criterion of increasing slopes (B.2.5) to consider only the compact set $\{x \mid \|x - x_0\| = \epsilon\}$.

Exercise C.5: This exercise outlines a proof of Proposition C.1.11. Let the conditions of the proposition hold, and define $\delta(\epsilon)$ as in Lemmas C.1.9 and C.1.10.

(a) Show that $f(x) - \langle u, x\rangle \ge f(x_0) - \langle u, x_0\rangle + \delta(\|x - x_0\|) - \langle u, x - x_0\rangle$ for any u, x, and x_0.

(b) Show that if $\|u\|_* < \frac{\delta(\epsilon)}{\epsilon}$ and $\|x - x_0\| > \epsilon$, then $f(x) - \langle u, x\rangle > f(x_0) - \langle u, x_0\rangle$.

(c) Prove Proposition C.1.11.

Exercise C.6: Show that Theorem C.2.1, that $f^{**} = f$ if f is closed convex, holds in Hilbert space. *Hint.* Use Exercise B.11.

Exercise C.7 (Jensen's inequality in Hilbert space)**:** Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a closed convex function, where $\mathcal{H}$ is a Hilbert space, and let $X \in \mathcal{H}$ be a random vector with $\mathbb{E}[\|X\|] < \infty$. Show that $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$ (where we use the Bochner integral for $\mathbb{E}[X]$). *Hint.* Exercise C.6.

Bibliography

[1] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

[2] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34(2):584–653, 2006.

[3] J. Acharya, C. Cannone, and H. Tyagi. Inference under information constraints I: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12): 7835–7855, 2020.

[4] J. Acharya, C. Cannone, and H. Tyagi. Inference under information constraints II: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66 (12):7856–7877, 2020.

[5] J. Acharya, C. Cannone, C. Freitag, Z. Sun, and H. Tyagi. Inference under information constraints III: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1):253–267, 2021.

[6] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687, 2003.

[7] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *Annals of Statistics*, 38(5):3063–3092, 2010.

[8] A. Agarwal and J. C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

[9] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[10] S. Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(49):1653–1674, 2014.

[11] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[12] R. Ahlswede and P. Gács. Spreading of sets in product space and hypercontraction of the Markov operator. *Annals of Probability*, 4(6):925–939, 1976.

[13] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, March 2002.

[14] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34(2):157–177, 1976.

[15] N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[16] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[17] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6 (2):170–176, 1955.

[18] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing using stable distributions. In T. Darrell, P. Indyk, and G. Shakhnarovich, editors, *Nearest Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, 2006.

[19] A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 2023.

[20] A. N. Angelopoulos, R. F. Barber, and S. Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2026+. Forthcoming.

[21] E. Arias-Castro. Detecting a vector based on linear measurements. *Electronic Journal of Statistics*, 6:547–558, 2012.

[22] E. Arias-Castro, E. J. Candès, and A. Durand. Detection of an anomalous cluster in a network. *Annals of Statistics*, pages 278–304, 2011.

[23] E. Arias-Castro, E. Candés, and M. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[24] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming, Series A*, 199: 165–214, 2023.

[25] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[26] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004.

[27] H. Asi and J. Duchi. Near instance-optimality in differential privacy. *arXiv:2005.10630 [cs.CR]*, 2020.

[28] H. Asi and J. C. Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019.

[29] H. Asi and J. C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[30] H. Asi and J. C. Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In *Advances in Neural Information Processing Systems 33*, 2020.

[31] H. Asi, J. Ullman, and L. Zakynthinou. From robustness to privacy and back. *arXiv:2302.01855 [cs.LG]*, 2023.

[32] P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences, Série I*, 296(23):1021–1024, 1983.

[33] I. Attias, G. K. Dziugaite, M. Haghifam, R. Livni, and D. M. Roy. Information complexity of stochastic convex optimization: Applications to generalization and memorization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[34] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. In *Journal of Machine Learning Research*, pages 2635–2686, 2010.

[35] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[36] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[37] S. Balakrishnan. Advanced statistical theory I. Lecture notes for Carnegie Mellon Course 36-709, 2020. URL https://www.stat.cmu.edu/~siva/teaching/709/.

[38] S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Annals of Statistics*, 47(4):1893–1927, 2019.

[39] B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems 31*, pages 6277–6287, 2018.

[40] B. Balle, J. Bell, A. Gascon, and K. Nissim. The privacy blanket of the shuffle model. In *39th Annual International Cryptology Conference (CRYPTO)*, 2019.

[41] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[42] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

[43] R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv:1412.4451 [math.ST]*, 2014.

[44] A. Barron. Entropy and the central limit theorem. *Annals of Probability*, 14(1):336–342, 1986.

[45] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic, 1991.

[46] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[47] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[48] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[49] P. L. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[50] R. Bassily, A. Smith, T. Steinke, and J. Ullman. More general queries and less generalization error in adaptive data analysis. *arXiv:1503.04843 [cs.LG]*, 2015.

[51] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, pages 1046–1059, 2016.

[52] O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems 35*, 2022.

[53] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science*, 2000.

[54] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[55] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[56] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, 2001.

[57] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

[58] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[59] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:1–7, 2013.

[60] J. O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[61] J. O. Berger, J. Bernardo, and D. Sun. The formal definition of reference priors. *Annals of Statistics*, 37(2):905–938, 2009.

[62] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

[63] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[64] J. M. Bernardo. Reference analysis. In D. Day and C. R. Rao, editors, *Bayesian Thinking, Modeling and Computation*, volume 25 of *Handbook of Statistics*, chapter 2, pages 17–90. Elsevier, 2005.

[65] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(1):1780–1815, 2013.

[66] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[67] D. P. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

[68] P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, 1998.

[69] P. Billingsley. *Probability and Measure*. Wiley, Second edition, 1986.

[70] P. Billingsley. *Convergence of Probability Measures*. Wiley, Second edition, 1999.

[71] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet*, 65:181–238, 1983.

[72] L. Birgé and P. Massart. Estimation of integral functionals of a density. *Annals of Statistics*, 23(1):11–29, 1995.

[73] D. Blackwell. Comparison of experiments. In *Proceedings of the 2nd Berkeley Symposium on Probability and Statistics*, pages 93–102. University of California Press, 1951.

[74] R. Blahut. Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4):405–417, 1974.

[75] J. Blasiok, P. Gopalan, L. Hu, and P. Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the Fifty-Fifth Annual ACM Symposium on the Theory of Computing*, 2023. URL https://arxiv.org/abs/2211.16886.

[76] J. Blasiok, P. Gopalan, L. Hu, and P. Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems 36*, 2023. URL https://arxiv.org/abs/2305.18764.

[77] V. Blondel, A. Martins, and V. Niculae. Learning with Fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

[78] S. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163:1–28, 1999.

[79] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

[80] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[81] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[82] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[83] G. Braun, C. Guzmán, and S. Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63 (7), 2017.

[84] M. Braverman. Interactive information complexity. In *Proceedings of the Forty-Fourth Annual ACM Symposium on the Theory of Computing*, pages 505–524, 2012.

[85] M. Braverman and A. Rao. Information equals amortized communication. *IEEE Transactions on Information Theory*, 60(10):6058–6069, 2014.

[86] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1506.07216.

[87] M. Brennan, G. Bresler, and W. Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018.

[88] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[89] G. Brown, M. Bun, V. Feldman, A. Smith, and K. Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the Fifty-Third Annual ACM Symposium on the Theory of Computing*, pages 123–132, 2021.

[90] L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, California, 1986.

[91] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[92] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[93] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, pages 635–658, 2016.

[94] M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

[95] T. Cai and M. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Annals of Statistics*, 39(2):1012–1041, 2011.

[96] T. Cai and M. Low. A framework for estimating convex functions. *Statistica Sinica*, 25: 423–456, 2015.

[97] T. T. Cai and Y. Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232, 2014.

[98] T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy: optimal rates of convergence for parameter estimation with differential privacy. *Annals of Statistics*, 49(5):2825–2850, 2021.

[99] T. T. Cai, Y. Wang, and L. Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv:2303.07152 [math.ST]*, 2023.

[100] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector. *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[101] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming, Series A*, 2019. doi: s10107-019-01406-y.

[102] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points II: First order methods. *Mathematical Programming, Series A*, 185:315–355, 2021. doi: s10107-019-01431-x.

[103] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes and Monographs*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. URL https://arxiv.org/abs/0712.0248.

[104] O. Catoni and I. Giulini. Dimension-free PAC-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv:1712.02747 [math.ST]*, 2017.

[105] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[106] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

[107] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the Association for Computing Machinery*, 22(3):329–340, 1975.

[108] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, 2011.

[109] R. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, 1(1):2–20, 2012.

[110] C. Cheng, J. Duchi, and D. Levy. Geometry, computation, and optimality in stochastic optimization. *arXiv:1909.10455v4 [math.OC]*, 2025.

[111] J. Cherian, I. Gibbs, and E. J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems 37*, 2024.

[112] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[113] B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[114] J. E. Cohen, Y. Iwasa, G. Rautu, M. B. Ruskai, E. Seneta, and G. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993.

[115] D. L. Cohn. *Measure Theory*. Birkhäuser, Second edition, 2013.

[116] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, A. S. Wein, and I. Zadik. Statistical and computational phase transitions in group testing. In *Proceedings of the Thirty Fifth Annual Conference on Computational Learning Theory*, 2022.

[117] J. Couzin. Whole-genome data not anonymous, challenging assumptions. *Science*, 321(5894): 1278, 2008.

[118] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[119] H. Cramér. *Mathematical Methods in Statistics*. Princeton University Press, 1946.

[120] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[121] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[122] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[123] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[124] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, second edition, 2011.

[125] A. Cutkosky and K. Boahen. Online learning without prior information. In *Proceedings of the Thirtieth Annual Conference on Computational Learning Theory*, 2017.

[126] A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018.

[127] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15: 429–444, 1977.

[128] S. Dasgupta and A. Gupta. An elementray proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2002.

[129] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[130] L. D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19 (6):783–795, 1973.

[131] A. Dawid and V. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5: 125–162, 1999.

[132] A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:278–292, 1984.

[133] A. P. Dawid. Calibration-based empirical probability. *Annals of Statistics*, 13(4):1251–1273, 1985.

[134] V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer, 1999.

[135] A. Defazio, X. Yang, H. Mehta, K. Mishchenko, A. Khaled, and A. Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems 37*, 2024.

[136] M. H. DeGroot. Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics*, 33(2):404–419, 1962.

[137] M. H. DeGroot. *Optimal Statistical Decisions*. Mcgraw-Hill College, 1970.

[138] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probability Theory and Related Fields*, 126:395–420, 2003.

[139] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 25 (2):927–939, 1997.

[140] A. Dembo and O. Zeitouni. Transportation approach to some concentration inequalities in product spaces. *Electronic Communications in Probability*, 1:83–90, 1996.

[141] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and Its Applications*, 1(1):65–80, 1956.

[142] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Theory of Probability and Its Applications*, 1(4):329–383, 1956.

[143] R. L. Dobrushin. A statistical problem arising in the theory of detection of signals in the presence of noise in a multi-channel system and leading to stable distribution laws. *Theory of Probability and Its Applications*, 3(2):161–173, 1958.

[144] J. Dong, A. Roth, and W. J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society, Series B*, 84(1):3–37, 2022.

[145] D. Donoho. One-sided inference about functionals of a density. *Annals of Statistics*, 16(4): 1390–1420, 1988.

[146] D. L. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[147] D. L. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3), 2004.

[148] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence I. Technical Report 137, University of California, Berkeley, Department of Statistics, 1987.

[149] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence II. *Annals of Statistics*, 19 (2):633–667, 1991.

[150] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence III. *Annals of Statistics*, 19 (2):688–701, 1991.

[151] S. Du, S. Kakade, R. Wang, and L. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[152] J. C. Duchi. Sample-conditional coverage in conformal prediction. In *Advances in Neural Information Processing Systems 38*, 2025.

[153] J. C. Duchi and R. Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the Thirty Second Annual Conference on Computational Learning Theory*, 2019.

[154] J. C. Duchi and F. Ruan. A constrained risk inequality for general losses. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics*, 2020.

[155] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, 49(1):21–48, 2021.

[156] J. C. Duchi and F. Ruan. The right complexity measure in locally private estimation: It is not the Fisher information. *Annals of Statistics*, 52(1):1–51, 2024.

[157] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[158] J. C. Duchi, M. I. Jordan, and H. B. McMahan. Estimation, optimization, and parallelism when data is sparse. In *Advances in Neural Information Processing Systems 26*, 2013.

[159] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

[160] J. C. Duchi, L. Mackey, and M. I. Jordan. The asymptotics of ranking algorithms. *Annals of Statistics*, 41(5):2292–2323, 2013.

[161] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order optimization: the power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[162] J. C. Duchi, K. Khosravi, and F. Ruan. Multiclass classification, information, divergence, and surrogate risk. *Annals of Statistics*, 46(6b):3246–3275, 2018.

[163] J. C. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664, 2022.

[164] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[165] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, second edition, 2002.

[166] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on the Theory of Computing*, pages 371–380, 2009.

[167] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3 & 4):211–407, 2014.

[168] C. Dwork and G. Rothblum. Concentrated differential privacy. *arXiv:1603.01887 [cs.DS]*, 2016.

[169] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, 2006.

[170] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[171] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

[172] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. *arXiv:1411.2664v2 [cs.LG]*, 2014.

[173] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on the Theory of Computing*, 2015.

[174] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving statistical validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[175] C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. Robust traceability from trace amounts. In *56th Annual Symposium on Foundations of Computer Science*, 2015. Long version available at https://jonathan-ullman.github.io/assets/tracing.pdf.

[176] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[177] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Insitute of Mathematical Statistics Monographs. Cambridge University Press, 2012.

[178] B. Efron. *Exponential Families in Theory and Practice*. Cambridge University Press, 2022.

[179] Èlizbar Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9 (1):141–142, 1964.

[180] U. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS)*, 2014.

[181] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

[182] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.

[183] K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

[184] V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the Thirty First Annual Conference on Computational Learning Theory*, 2018.

[185] V. Feldman, A. McMillan, and K. Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *62nd Annual Symposium on Foundations of Computer Science*, 2021.

[186] R. A. Fisher. Studies in crop variation i: an examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science*, 11(2):107–135, 1921.

[187] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368, 1922.

[188] R. A. Fisher and W. Mackenzie. Studies in crop variation. ii. the manurial response of different potato varieties. *Journal of Agricultural Science*, 13(3):311–320, 1923.

[189] R. A. Fisher, A. S. Corbett, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of the British Ecological Society*, 12(1):42–58, 1943.

[190] T. Fissler and J. Ziegel. Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707, 2016.

[191] G. Folland. *Real Analysis: Modern Techniques and their Applications*. Pure and Applied Mathematics. John Wiley & Sons, second edition, 1999.

[192] D. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[193] D. Foster, S. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making. *arXiv:2112.13487v3 [cs.LG]*, 2021.

[194] D. Foster, N. Golowich, and Y. Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *Proceedings of the Thirty Sixth Annual Conference on Computational Learning Theory*, 2023.

[195] D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems 33*, 2020.

[196] D. P. Foster and S. Hart. "Calibeating": Beating forecasters at their own game. *Theoretical Economics*, 18(4):1441–1474, 2023. doi: 10.3982/TE5330.

[197] D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society, Series B*, 70(2):429–444, 2008.

[198] A. Franco, N. Malhotra, and G. Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014.

[199] D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, Feb. 1975.

[200] M. Friedman and L. Savage. The utility analysis of choices involving risk. *Journal of Political Economy*, 56(4):279–304, 1948.

[201] R. Frongillo and I. A. Kash. Vector-valued property elicitation. In *Proceedings of the Twenty Eighth Annual Conference on Computational Learning Theory*, pages 710–727, 2015.

[202] R. Frongillo and B. Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems 34*, 2021.

[203] R. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, MIT Laboratory for Information and Decision Systems, 1979.

[204] D. Gamarnik and I. Zadik. Sparse high-dimensional linear regression. estimating squared error and a phase transition. *Annals of Statistics*, 50(2):880–903, 2022.

[205] D. García-García and R. C. Williamson. Divergences and risks for multiclass experiments. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[206] A. Garg, T. Ma, and H. L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, 2014.

[207] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Columbia University, 2013.

[208] C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

[209] R. P. Gilbert. *Function Theoretic Methods in Partial Differential Equations*. Academic Press, 1969.

[210] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

[211] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2021.

[212] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[213] T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1, 2014.

[214] T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[215] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B*, 69(2):243–268, 2007.

[216] I. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 1953.

[217] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.

[218] R. M. Gray. *Entropy and Information Theory*. Springer, second edition, 2009.

[219] P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[220] P. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

[221] A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

[222] C. Gupta and A. Ramdas. Top-label calibration and multiclass-to-binary reductions. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022.

[223] V. Gupta, C. Jung, G. Noarov, M. M. Pai, and A. Roth. Online multivalid learning: Means, moments, and prediction intervals. In *Proceedings of the Thirteenth Innovations in Theoretical Computer Science (ITCS)*, 2022.

[224] L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 30:105–113, 1978.

[225] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[226] J. Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 175–194, 1972.

[227] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing*, pages 705–714, 2010.

[228] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Applications*, 23:794–798, 1978.

[229] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall, 1995.

[230] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[231] D. Haussler. A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280, 1997.

[232] E. Hazan. The convex optimization approach to regret minimization. In *Optimization for Machine Learning*, chapter 10. MIT Press, 2012.

[233] E. Hazan and S. Kale. An optimal algorithm for stochastic strongly convex optimization. In *Proceedings of the Twenty Fourth Annual Conference on Computational Learning Theory*, 2011. URL http://arxiv.org/abs/1006.2425.

[234] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[235] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[236] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[237] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II*. Springer, New York, 1993.

[238] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[239] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[240] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[241] S. B. Hopkins, G. Kamath, and M. Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the Fifty-Fourth Annual ACM Symposium on the Theory of Computing*, pages 1406–1417, 2022.

[242] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[243] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[244] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

[245] K. Hung and W. Fithian. Statistical methods for replicability assessment. *Annals of Applied Statistics*, 14(3):1063–1087, 2020.

[246] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, 1981.

[247] G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[248] P. Indyk. Nearest neighbors in high-dimensional spaces. In *Handbook of Discrete and Computational Geometry*. CRC Press, 2004.

[249] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing*, 1998.

[250] Y. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[251] Y. Ingster and I. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Number 169 in Lecture Notes in Statistics. Springer, 2003.

[252] J. P. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005. doi: 10.1371/journal.pmed.0020124.

[253] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379, 1961.

[254] A. Javanmard and A. Montanari. Online rules for control of false discovery rate and false discovery exceedance. *Annals of Statistics*, 46(2):526–554, 2018.

[255] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70 (9):939–952, Sept. 1982.

[256] T. S. Jayram. Hellinger strikes back: a note on the multi-party information complexity of AND. In *Proceedings of APPROX and RANDOM 2009*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer, 2009.

[257] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 186:453–461, 1946.

[258] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[259] C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Batch multivalid conformal prediction. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[260] S. M. Kakade and D. P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

[261] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 801–808, 2009.

[262] G. Kamath, J. Li, V. Singhal, and J. R. Ullman. Privately learning high-dimensional distributions. In *Proceedings of the Thirty Second Annual Conference on Computational Learning Theory*, 2019.

[263] G. Kamath, A. Mouzakis, and V. Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems 35*, 2022.

[264] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, 2012.

[265] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2564–2572, 2018.

[266] M. J. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 311–319, 1998.

[267] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

[268] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer-Verlag, 2011.

[269] A. Kumar, P. Liang, and T. Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems 32*, 2019.

[270] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[271] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[272] N. Lambert, D. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the Ninth Annual ACM Conference on Electronic Commerce (EC)*, pages 129–138, 2008.

[273] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming, Series A*, 133(1–2):365–397, 2012.

[274] J. Langford and R. Caruana. (Not) bounding the true error. In *Advances in Neural Information Processing Systems 14*, 2001.

[275] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[276] T. Lattimore, C. Szepesvári, and G. Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[277] A. N. D. A. D. Lauger, P. E. Singer, D. Kifer, J. P. Reiter, A. Machanavajjhala, S. L. Garfinkel, S. A. Dahl, M. Graham, V. Karwa, H. Kim, P. Leclerc, I. M. Schmutte, W. N. Sexton, L. Vilhuber, and J. M. Abowd. The modernization of statistical disclosure limitation at the U.S. Census Bureau. Available online at https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf, 2017.

[278] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 3:37–98, 1953.

[279] L. Le Cam. On the asymptotic theory of estimation and hypothesis testing. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 129–156, 1956.

[280] L. Le Cam. Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 245–261, 1972.

[281] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):38–53, 1973.

[282] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[283] M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[284] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[285] D. Lee, X. Huang, H. Hassani, and E. Dobriban. T-Cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023.

[286] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

[287] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer, 2005.

[288] C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24:757–781, 2015.

[289] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[290] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[291] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[292] J. Liu, R. van Handel, and S. Verdú. Second-order converses via reverse hypercontractivity. *Mathematical Statistics and Learning*, 2(2):103–163, 2019.

[293] Y. Liu. Fisher consistency of multicategory support vector machines. In *Processing of 11th International Conference on Artificial Intelligence and Statistics*, pages 291–298, 2007.

[294] P. M. Long and R. A. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[295] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[296] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32(1):30–55, 2004.

[297] X. Lyu and K. Talwar. Fingerprinting codes meet geometry: Improved lower bounds for private query release and adaptive data analysis. In *Proceedings of the Fifty-Seventh Annual ACM Symposium on the Theory of Computing*, pages 2374–2385, 2025.

[298] L. W. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability*, 42(3):906–945, 2014.

[299] M. Madiman and A. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 53(7):2317–2329, 2007.

[300] A. Mao, M. Mohri, and Y. Zhong. Cross-entropy loss functions: theoretical analysis and applications. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[301] G. A. Margulis. Probabilistic characteristics of graphs with large connectivity. *Problems of Information Transmission*, 10(2):101–108, 1974.

[302] K. Marton. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Annals of Probability*, 24(2):857–866, 1996.

[303] K. Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110:427–439, 1998.

[304] P. Massart. Concentration inequalities and model selection. In J. Picard, editor, *Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003 Series*. Springer, 2003.

[305] J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

[306] D. A. McAllester. Some PAC-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

[307] D. A. McAllester. Simplified PAC-bayesian margin bounds. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

[308] D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[309] D. A. McAllester. A PAC-Bayestian tutorial with a dropout bound. *arXiv:1307.2118 [cs.LG]*, 2013.

[310] J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42:654–655, 1956.

[311] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[312] B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[313] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual Symposium on Foundations of Computer Science*, 2007.

[314] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, 41(3):714–722, 1995.

[315] I. Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[316] B. S. Mityagin. Approximate dimension and bases in nuclear spaces. *Russian Mathematical Surveys*, 16:59–127, 1961.

[317] C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[318] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[319] A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12:109–138, 2001.

[320] A. Nemirovski. Efficient methods in convex programming. Technion: The Israel Institute of Technology, 1994.

[321] A. Nemirovski. Lectures on modern convex optimization. Georgia Institute of Technology, 2005.

[322] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[323] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[324] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[325] J. S. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1990. Translated and edited from the 1923 original by D. M. Dabrowska and T. P. Speed.

[326] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

[327] A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: the sparse and approximate case. In *Proceedings of the Forty-Fifth Annual ACM Symposium on the Theory of Computing*, 2013.

[328] J. Niles-Weed and I. Zadik. It was 'all' for 'nothing': sharp phase transitions for noiseless discrete channels. In *Proceedings of the Thirty Fourth Annual Conference on Computational Learning Theory*, 2021.

[329] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing*, 2007.

[330] K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. The limits of post-selection generalization. In *Advances in Neural Information Processing Systems 31*, 2018.

[331] A. Nowak-Vila, F. Bach, and A. Rudi. Consistent structured prediction with max-min margin markov networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[332] S. Nowozin and C. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4), 2011.

[333] E. Ordentlich and M. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, 2005.

[334] A. Orlitsky and A. Suresh. Competitive distribution estimation: Why is Good-Turing good? In *Advances in Neural Information Processing Systems 28*, 2015.

[335] A. Orlitsky, N. Santhanam, and J. Zhang. Always Good-turing: asymptotically optimal probability estimation. In *44th Annual Symposium on Foundations of Computer Science*, 2003.

[336] K. Osband. *Providing incentives for better cost forecasting*. PhD thesis, University of California, Berkeley, 1985. URL https://doi.org/10.5281/zenodo.4355667.

[337] A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[338] D. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15:326–391, 2021.

[339] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003.

[340] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

[341] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[342] D. Petz. A survey of certain trace inequalities. *Banach Center Publications*, 30:287–298, 1994.

[343] K. Pillutla, J. Upadhyay, C. A. Choquette-Choo, K. Dvijotham, A. Ganesh, M. Henzinger, J. Katz, R. McKenna, H. B. McMahan, K. Rush, T. Steinke, and A. Thakurta. Correlated noise mechanisms for differentially private learning: A tutorial on mathematical foundations and practical aspects. *arXiv:2506.08201*, 2025.

[344] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[345] Y. Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*, volume 161 of *The IMA Volumes in Mathematics and its Applications*. Springer, 2017.

[346] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

[347] V. S. Portella and N. Harvey. Lower bounds for private estimation of Gaussian covariance matrices under all reasonable parameter regimes. *arXiv:2404.17714 [cs.DS]*, 2024.

[348] L. Pressly. *The Right to Oblivion: Privacy and the Good Life*. Harvard University Press, 2024.

[349] F. Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. SIAM, 1993.

[350] M. Raginsky. Strong data processing inequalities and ϕ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

[351] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1–2):1–250, 2014.

[352] A. Ramdas and R. Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025.

[353] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[354] A. Rao and A. Yehudayoff. *Communication Complexity and Applications*. Cambridge University Press, 2020.

[355] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

[356] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976—6994, 2011.

[357] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[358] P. Ravikumar, A. Tewari, and E. Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

[359] M. Reid and R. Williamson. Information, divergence, and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

[360] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.

[361] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[362] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[363] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[364] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958. (Reprinted in *Neurocomputing* (MIT Press, 1988).).

[365] H. Royden. *Real Analysis*. Pearson, third edition, 1988.

[366] W. Rudin. *Functional Analysis, Second Edition*. McGraw Hill, 1991.

[367] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[368] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems 27*, 2014.

[369] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[370] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[371] L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[372] T. Schramm and A. Wein. Computational barriers to estimation from low-degree polynomials. *Annals of Statistics*, 50(3):1833–1858, 2022.

[373] G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, 2019.

[374] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[375] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[376] O. Shamir. On the complexity of bandit linear optimization. In *Proceedings of the Twenty Eighth Annual Conference on Computational Learning Theory*, 2015.

[377] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[378] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Optimization Society, second edition, 2014.

[379] G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*, volume 59 of *Classics in Applied Mathematics*. SIAM, 1986.

[380] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[381] A. Slavkovic and F. Yu. Genomics and privacy. *Chance*, 28(2):37–39, 2015.

[382] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206, 1956.

[383] C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 187–195, 1956.

[384] T. Steinke and J. Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the Twenty Eighth Annual Conference on Computational Learning Theory*, pages 1588–1628, 2015.

[385] T. Steinke and J. Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2):3–22, 2017.

[386] I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

[387] S. M. Stigler. A Galtonian perspective on shrinkage estimators. *Statistical Science*, 5(1): 147–155, 1990.

[388] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B*, 66(1):187–205, 2004.

[389] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

[390] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

[391] T. Tao. *An Epsilon of Room, I: Real Analysis (pages from year three of a mathematical blog)*, volume 117 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

[392] T. Tao. The golden-thompson inequality, 2010. Accessed: 2025-09-01.

[393] B. Taskar. *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2005.

[394] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[395] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[396] R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514): 600–620, 2016.

[397] F. Topsøe. Information-theoretical optimization techniques. *Kybernetica*, 15:8–27, 1979.

[398] F. Topsøe. Maximum entropy versus minimum risk and applications to some classical discrete distributions. *IEEE Transactions on Information Theory*, 48(8):2368–2376, 2002.

[399] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[400] A. Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Annals of Statistics*, 26(6):2420–2469, 1998.

[401] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[402] J. W. Tukey. *Exploratory Data Analysis*. Pearson, 1997.

[403] G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems 26*, 2013.

[404] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Forty-Fifth Annual ACM Symposium on the Theory of Computing*, 2013.

[405] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46:429–455, 2017.

[406] A. W. van der Vaart. Superefficiency. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, chapter 27. Springer, 1997.

[407] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[408] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[409] A. W. van der Vaart and J. A. Wellner. Stein 1956: Efficient nonparametric testing and estimation. *Annals of Statistics*, 49(4):1836–1849, 2021. Memorial issue for Charles Stein.

[410] H. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley and Sons, 1968.

[411] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[412] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[413] R. Vershynin. *High Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

[414] C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

[415] J. von Neumann. Zur theorie der gesellschaftsspiele (on the theory of parlor games). *Math. Ann.*, 100:295—-320, 1928.

[416] V. Vovk. Superefficiency from the vantage point of computability. *Statistical Science*, 24(1): 73–86, 2009.

[417] V. Vovk, A. Grammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[418] S. Wager. *Causal Inference: A Statistical Learning Approach*. Lecture notes for Statistics 361. Stanford University, 2024. URL https://web.stanford.edu/~swager/causal_inf_book.pdf.

[419] K. Waghmare and J. Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and its Application*, to appear, 2025. URL https://arxiv.org/abs/2504.01781.

[420] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[421] M. J. Wainwright. Constrained forms of statistical minimax: Computation, communication, and privacy. In *Proceedings of the International Congress of Mathematicians, Volume IV*, pages 273–290, 2014.

[422] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[423] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[424] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

[425] A. Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945.

[426] A. Wald. Statistical decision functions. *Annals of Mathematical Statistics*, 20(2):165–205, 1949.

[427] S. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[428] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[429] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[430] R. C. Williamson, E. Vernet, and M. D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016.

[431] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[432] Z. Xu, Y. Zhang, G. Andrew, C. Choquette, P. Kairouz, B. Mcmahan, J. Rosenstock, and Y. Zhang. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Industry Track)*, pages 629–639, 2023.

[433] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[434] A. C.-C. Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, pages 209–213. ACM, 1979.

[435] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[436] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

[437] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85, 2004.

[438] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, 2013.

[439] N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28, 2024.

[440] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[441] J. Ziv and A. Lempel. Compression of individual sequences via variable rate coding. *IEEE Transactions on Information Theory*, 24:530–536, 1978.

Index

action separation, 519
adaptive
    composition, 144, 145, 178, 191, 206
    data analysis, 140
    estimation, 234, 236
    statistical query, 141
affine hull, 568
affinity, *see* Hellinger distance
algebra, 27
Ali-Silvey divergence, *see* f-divergence
Assouad
    lemma, 228
    method, 228, 305, 486, 497, 521
    method in communication, 258, 261
Azuma inequality, *see* Hoeffding

bandit problem, 498
Bayes risk, 436
Bayesian Cramér-Rao, *see* van Trees
Bayesian regret, *see* regret
Benjamini-Hochberg procedure, 313, 333
Bennett's inequality, 77, 80, 98
Bernstein
    condition, 79, 135, 151
    inequality, 78–80, 98, 113, 137
    inequality, matrices, 91, 95, 100, 101
best arm identification, 519
Beta distribution, 544
biconjugate, 378, 612
binary entropy, 19, 220
blowing up lemma, 167
Bonferroni correction, 312
bounded differences, 86, 163
Bregman divergence, 56, 383, 408, 470, 474, 535, 536, 611
Bretagnolle-Huber inequality, 33, 272
Brier score, 377

calibeating, 419
calibration, 407
    complete, 412
    continuous measure, 421
    distance to, 412
    error relative to a class, 412
    expected error, 409
    Lipschitz measure, 421
    partitioned error, 413
    penalized distance, 422
    sound, 412
Carathéodory's theorem, 570
chain rule, 23
    entropy, 23
    KL-divergence, 24, 242
    mutual information, 24
Chebyshev's inequality, 70
Chernoff bound, 71, 82, 173
χ^2-affinity, 351
χ^2-divergence, 32, 240
χ^2-mixture bound, 344
classification calibration, 438
    zero-one loss, 449
closed function, 588
closure
    of a function, 48, 590
co-coercivity, 614
code
    block, 49
    instantaneous, 40
    nonsingular, 40
    prefix-free, 40
    uniquely decodable, 40
coercive, 606, 620
    super-, 621
communication
    complexity, 245
    complexity of AND, 253
    complexity with public randomness, 246
    complexity, distributional, 246

complexity, randomized, 246
cost, 245
protocol, 245
transcript, 254
comparison inequality, 440
composition, *see* adaptive composition
conditional entropy, 19, 380
conjugate, *see* convex conjugate
conjugate link, 390
continuous ranked probability score, 383, 406, 538, 540
convergence in distribution, 555, 564
convex
cone, 406
function, 20, 378, 581
hull, 568
online optimization, *see* online convex optimization
set, 568
stochastic optimization, *see* stochastic convex optimization
strictly, 20, 54, 378, 584
strongly, 475, 609
uniformly, 449, 610
convex conjugate, 157, 378, 612
convex hull method, *see* Le Cam
coupling, 160, 161
covering
number, 107
set, 107
covering number, 218
for KL-divergence, 308
Cramér-Rao, 275
critical threshold, 321
cumulant generating function, *see* log partition function

data processing, 183
data processing inequality, 26, 35, 144, 232
constant, 239
strong, 239, 258
differentiable parameter, *see* influence function
differential entropy, 22
differential privacy, 178
approximate, 183, 188
central, 178
Gaussian, 203
local, 178, 241, 262
Rényi, 184
sequentially interactive, 241, 271
direct sum, 250, 251, 256
directional derivative, 597
distance generating function, 474
divergence, *see* f-divergence
Dobrushin coefficient, 240, 270
dominated convergence, 563
Donsker-Varadhan variational representation, 132, 149, 153, 554
Doob martingale, 84
doubling trick, 490, 493
dual cone, 580
dual norm, 475, 606

effective rank, 157
empirical risk minimization, 110
entropy, 19, 42
generalized, 379
rate, 43
Shannon, 397
epigraph, 582
(ε, δ)-differential privacy, *see* differential privacy
(ε, δ)-close distributions, 189
exponential family, 51, 394, 532
estimation in, 116, 122
exponential tilting, 73
exponentiated gradients, *see* mirror descent

f-divergence, 29, 35
false discovery, 312, 333
proportion, 312
rate, 313
family-wise error rate, 312
Fano
global method, 309
inequality, 37, 220, 225, 323
inequality, generalized, 234
inequality, refined, 324
local method, 222, 484
Fatou's lemma, 562
Fenchel-Legendre transform, *see* convex conjugate
Fenchel-Young inequality, 378, 613
Fisher consistent, *see* surrogate consistency
Fisher information, 57, 61, 274, 289

of a function, 283
optimality of the, 280
Fisher score, 274, 547
fooling
distribution, 250, 251
set, 248
Fréchet differentiability, 597
functional estimation, 337

Gateaux derivative, 386
Gateaux differentiability, 597
Gaussian metric concentration, 166
generalized linear model, 58
estimation in, 116, 122
Gilbert-Varshamov bound, 218, 223
GLM, *see* generalized linear model
global sensitivity, *see* sensitivity
Gneiting-Raftery-Savage, *see* proper loss
Golden-Thompson, 89
goodness of fit, 362

Hölder continuity, 611, 615
Hamming separation, 228, 306
Hellinger affinity, *see* Hellinger distance
Hellinger distance, 32, 217
Hellinger divergence, *see* Hellinger distance
Hoeffding
inequality, 74, 85, 90, 108, 206, 489
inequality, matrices, 91, 95, 100
lemma, 72, 154
hypothesis test
canonical problem, 215, 225, 228, 315
composite, 358
multiple, 347

implicit function theorem, 287
increasing slopes, 582, 584, 598
influence function, 238, 285
information
conditional mutual, 21
in an experiment, 380, 536
mutual, 21, 28
information complexity, 246, *see* communication complexity
information-directed sampling, 510
instance optimality, 288, 350
interactive, *see* adaptive
isoperimetric inequalities, 165

James-Stein, 277
Jeffreys prior, 546
Jensen's inequality, 20, 616
Johnson-Lindenstrauss, 83, 95

Kantorovich duality, *see* transport cost
kernel function, 302
kernel smoother, *see* Nadaraya-Watson
KL-divergence, *see* Kullback-Leibler divergence
Kraft-McMillan inequality, 41, 111, 534
Kullback-Leibler
divergence, 20, 28, 31, 217
divergence, conditional, 25
stability, 144, 146, 151, 262
variational representation, 132

Laplace approximation, 547
large deviations, 173
large margin, *see* margin
Le Cam
convex hull method, 343, 359
inequality, 36, 219
method, 219, 338, 484
Ledoux-Talagrand, *see* Rademacher contraction inequality
Legendre
entropy, 390
function, 496, 621
Lévy-Prokhorov distance, 565
Lie product formula, 94, 99
linear bandit, 512
Lipschitz continuity, 476, 495
local asymptotic normality, 299
log partition function, 51, 53
loss calibration, 445
loss model, 471
loss, proper, *see* proper loss
lower semicontinuous, 588
lower semicontinuous hull, *see* closure

M estimation, 116, 214
Marcinkiewicz-Zygmund, 97, 130
margin, 150, 437, 465
loss, 102, 114, 137, 476
marginal polytope, 387, 394, 450
Markov's inequality, 70, 96
martingale, 84
difference, 84

Marton's transportation inequality, *see* transportation cost
maximum entropy, 397, 398, 532, 539
McDiarmid's inequality, *see* bounded differences
mean mapping, 387, 450
mean parameters, *see* marginal polytope
mechanism, 178
  exponential, 193
  Gaussian, 184, 186
  inverse sensitivity, 196
  Laplace, 180, 194, 244
  matrix, 209
  median, 197
  subsampling, 205
metric entropy, *see* covering number
Mills ratio, 96, 151
min-max game, 494, 531
minimal exponential family, 53
minimax optimization risk, 480
minimax risk, 214, 519
  constrained, 239, 243
  excess, 214
  testing, 358, 410
mirror descent, 474
  regret, 476
modulus
  of continuity, global, 338
  action separation, 519
  of continuity, local, 337, 352
  of continuity, mixture, 355
multiclass logistic regression, 59
multiplicative weights, *see* mirror descent

Nadaraya-Watson, 303
Neyman-Pearson, 169, 318
Neyman-Rubin causal model, 499
nonparametric, 302
  rate, 305
normal cone, 580, 604
normal random variable, 22
normalized maximum likelihood, 541

online convex optimization, 470
online gradient descent, *see* projected gradient method
online-to-batch conversion, 470, 489
optimization distance, 482
order equivalence, 459, 460
Orlicz norm, 81, 155
orthosymmetric set, 486
outer semicontinuity of the subdifferential, 601

p-value, 312, 347, 360
PAC-Bayes inequality, 133, 134, 149
  information bound, 139, 141
packing
  local, 222
  number, 107
  set, 107
packing number, 218
partition, *see* quantizer
perspective transform, 48, 81, 572, 595
phase transition, *see* testing
φ-risk, *see* surrogate risk
Pinsker's inequality, 33, 155, 166, 217, 344
Poisson
  distribution, 52
  regression, 59
Poissonization, 362
Portmanteau theorem, 564
post-processing, 178, 182
post-processing gap, 415
  surrogate, 415, 416
posterior sampling, *see* Thompson sampling
potential outcomes, 281, 499
projected gradient method, 472
Prokhorov theorem, 564
proper loss, 377
proper loss, 376, 407, 535, 550, 553
  regular, 381, 388
  representation, 381
  strictly, 377
proper scoring, *see* proper loss
property testing, *see* goodness of fit
ψ_p-norm, *see* Orlicz norm
ψ-transform, 440

quadratic information bound, 222, 325
quantile loss, 384
quantizer, 27
  rectangular, 29
  risk, 455

Rényi-divergence, 183
Rademacher

complexity, 87, 105, 113
contraction inequality, 106, 114
variable, 72, 87
randomized response, 179
recession
cone, 617
function, 402, 619
reduction to testing, 215, 482
redundancy, *see* regret, expected
redundancy/capacity, *see* regret/capacity duality
reference analysis, 546
regret, 470, 476, 530, 534
adversarial, 534
Bayesian, 504, 530
expected, 499, 535, 543
regret/capacity duality, 537, 538
relative error condition, 136
relative interior, 573
risk gap, 481
risk separation, 481
risk-transfer, 350
robust Bayes, 531, 539

saddle point, 494, 531
Scheffé's lemma, 563
score, *see* Fisher score
attack, 291
sensitivity
global, 180
inverse, 196
local, 196
sequential compactness, 564
Shtarkov distribution, *see* normalized maximum likelihood
source coding, 40, 42
stability, 140, 143, *see* Kullback-Leibler
statistical query, 140
stochastic convex optimization, 470
strong converse, 171, 322
sub-exponential, 76
sub-Gaussian, 72, 96, 154
martingale, 85
vector, 109, 155
subderivative, *see* subgradient
subgradient, 378, 385
sublinear function, 592
super-efficiency, 288, 290, 350
support function, 592
surrogate
loss, 390
risk, 436
risk consistency, 436
symmetrization, 73, 90, 104

tangent cone, 581, 604
tensorization, 74, 78, 90, 217, 345
transport cost, 162
testing
critical radius, 358
Thompson sampling, 507
three term identity, 479
tight family, 564
tilt stability, 610, 614, 621, 622
tilting, 227, 237, 284, 366, 371
total variation, 31, 45, 160, 217, 240
T_p-inequality, *see* transportation cost
transportation cost, 160, 161
inequality, 155, 158, 159, 163, 166

uniform calibration, 444
uniform law, 102
universal coding, 543
universal equivalence, 455
upper confidence bound, 500

Van Trees
multivariate, 282
Van Trees inequality, 277
multivariate, 279
variance-directed sampling, 511
variation distance, *see* total variation
variational representation, *see* Donsker-Varadhan variational representation

Wasserstein distance, *see* transport cost
weak converse, 170