Parting Thoughts

John Duchi

Stats 300b – Winter Quarter 2021

Outline of the course

- Asymptotic normality of estimators
- Uniform laws of large numbers
- Uniform central limit theorems
- Contiguity, local asymptotic normality, and optimality
- U-statistics

Reading:

- All the statistics journals (and more)

Asymptotic normality of estimators

basic approach:

- show consistency (e.g., by growth of objective near optimum)
- Taylor expansions
- a bit more advanced: look at deviations of local process

Estimators today

- idea of *growth* of loss central to modern analyses
- high-dimensional and structural statistics (Candès and Tao, 2008; Recht, 2011; Negahban et al., 2012; Cai and Zhang, 2015; Wainwright, 2019)
- learning without concentration (Mendelson, 2014, 2015)
- non-convex optimization (Candès et al., 2015; Ma et al., 2020; Duchi and Ruan, 2018; Davis et al., 2020)

Uniform laws of large numbers

basic approach:

- show pointwise limits $P_n f \stackrel{a.s.}{\to} Pf$
- cover space $\mathcal{F}$ (either bracketing or metric entropies)

alternative (related):

- use bounded differences on $\sup_{f \in \mathcal{F}} |P_n f - Pf|$
- chaining or other integral bounds

Uniform laws of large numbers today

- central to much of machine learning theory (Bartlett and Mendelson, 2002; Bartlett et al., 2005; Boucheron et al., 2005; Koltchinskii, 2006)
- often a good first approach when full understanding of "structural" properties not known yet
  - nonconvex problems (e.g. Candès et al., 2015)
  - robustness and distributionally robust optimization (Duchi and Namkoong, 2020)
- still a building block in high-dimensional statistics (Negahban et al., 2012; Wainwright, 2019)

Uniform central limit theorems

basic approach:

- finite dimensional convergence

$$\sqrt{n}(P_n h - Ph) \stackrel{d}{\to} \mathcal{N}(0, \mathrm{Cov}(h))$$

  for each h of form $h(x) = (f_1(x), \ldots, f_k(x))$ for $f_i \in \mathcal{F}$

- some type of continuity (approximation by finite maxima) over $\mathcal{F}$
- get central limit theorem

$$\sqrt{n}(P_n - P) \stackrel{d}{\to} \mathbb{G} \quad \text{in} \quad L^\infty(\mathcal{F})$$

  where $\mathbb{G} : \mathcal{F} \to \mathbb{R}$ is a Gaussian process

Some applications (including those we haven't seen)

- testing (e.g. Kolmogorov Smirnov tests)
- diffusion limits in stochastic optimization and sequential processes
  - stochastic optimization (Kushner and Yin, 2003) of minimizing $f(\theta) = \mathbb{E}[F(\theta; X)]$: interpolate process

$$\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla F(\theta^k; X_k)$$

  - sequential analysis, reinforcement learning, bandits (Siegmund, 1985; Wager and Xu, 2021)

Contiguity, local asymptotic normality, optimality

- to understand optimality, look at sequences of *local* alternatives
- local asymptotic theory: in local experiments $\{P_{h/\sqrt{n}}\}_{h\in\mathbb{R}^d}$, limits must look Gaussian

$$\sqrt{n}\log\frac{dP^n_{h/\sqrt{n}}}{dP^n_0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_0^T h - h^T P_0\dot{\ell}_0\dot{\ell}_0^T h + o_P(\|h\|)$$

- parametric local minimax theorem (optimality for the distribution *at hand*:

$$\lim_{c\to\infty}\liminf_n \inf_{\widehat{\theta}_n}\int \mathbb{E}_{\theta+h/\sqrt{n}}\left[L(\sqrt{n}(\widehat{\theta}_n-(\theta+h/\sqrt{n})))\right]d\pi_c(h)$$
$$\geq \mathbb{E}[L(\mathcal{N}(0,I_\theta^{-1}))] \quad \text{for } I_\theta = P\dot{\ell}_\theta\dot{\ell}_\theta^T$$

Current treatments of optimality

- minimax, global worst case (Yang and Barron, 1999; Arias-Castro et al., 2013; Tsybakov, 2009; Duchi, 2019; Wainwright, 2019)
  - often construct local "packing" of central point θ_0, perturbations $\theta_v = \theta_0 + v/\sqrt{n}$
  - argue cannot test (and hence estimate) which v one got
- much interest in optimality theory beyond standard parametric models, especially semiparametric efficiency
  - stochastic optimization (Duchi and Ruan, 2020a)
  - causal inference (all over econometrics)
- move toward more "instance-optimal" behavior (Cai and Low, 2015; Zhu et al., 2016; Duchi and Ruan, 2020b; Asi and Duchi, 2020; Khamaru et al., 2020; Roughgarden, 2020)

U-statistics

- kernel $h : \mathcal{X}^r \to \mathbb{R}$ and parameter $\theta = \mathbb{E}[h(X_1, \ldots, X_r)]$
- unbiased statistic

$$U_n := \binom{n}{r}^{-1} \sum_{|S|=r} h(X_S)$$

approximated well by linearization

$$\widehat{U}_n := \frac{r}{n} \sum_{i=1}^{n} h_1(X_i), \quad h_1(x) = \mathbb{E}[h(x, X_2, \ldots, X_r)] - \theta$$

- two-sample U-statistics, $h : \mathcal{X}^r \times \mathcal{Y}^s \to \mathbb{R}$,

$$U_N = \binom{m}{r}^{-1} \binom{n}{s}^{-1} \sum_{|A|=r, |B|=s} h(X_A, Y_B)$$

well approximated by

$$\widehat{U}_N = \frac{r}{m} \sum_{i=1}^{m} h_{1,0}(X_i) + \frac{s}{n} \sum_{i=1}^{n} h_{0,1}(Y_i)$$

Applications and (potential) future of U-statistics

- frequent in two-sample (and related) testing, e.g. kernel embeddings (Gretton et al., 2012a,b) (but see also Ramdas et al. (2015))
- much modern data collection *aggregates* information together (Russakovsky et al., 2015; Krishna et al., 2017; Recht et al., 2019)

References I

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.

H. Asi and J. Duchi. Near instance-optimality in differential privacy. *arXiv:2005.10630 [cs.CR]*, 2020.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9: 323–375, 2005.

T. Cai and M. Low. A framework for estimating convex functions. *Statistica Sinica*, 25:423–456, 2015.

T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *Annals of Statistics*, 43(1):102–138, 2015.

References II

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5): 2053–2080, 2008.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4): 2652–2695, 2020.

J. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference*, iay015, 2018.

J. C. Duchi. Information theory and statistics. Lecture Notes for Statistics 311/EE 377, Stanford University, 2019. URL `http://web.stanford.edu/class/stats311/lecture-notes.pdf`. Accessed May 2019.

References III

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, to appear, 2020. URL https://arXiv.org/abs/1810.08750.

J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *Annals of Statistics*, To Appear, 2020a.

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*, 2020b.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13: 723–773, 2012a.

A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, 2012b.

K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv:2003.07337 [stat.ML]*, 2020.

References IV

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition, 2003.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):461–632, 2020.

S. Mendelson. Learning without concentration. In *Proceedings of the Twenty Seventh Annual Conference on Computational Learning Theory*, 2014.

V. K. S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

References V

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

A. Ramdas, S. Reddi, B. Pòczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance-based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Thirty-Second National Conference on Artificial Intelligence*, 2015.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

T. Roughgarden, editor. *Beyond the Worst Case Analysis of Algorithms*. Cambridge University Press, 2020.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

D. O. Siegmund. *Sequential Analysis*. Springer, 1985.

References VI

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

S. Wager and K. Xu. Diffusion asymptotics for sequential experiments. *arXiv:2101.09855 [math.ST]*, 2021.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 29*, 2016.