Geometry, Computation, and Optimality in Stochastic Optimization

Chen Cheng* John C. Duchi* Daniel Levy*
Stanford University

Abstract

We study computational and statistical consequences of problem geometry in stochastic and online optimization. By focusing on constraint set and gradient geometry, we characterize the problem families for which stochastic- and adaptive-gradient methods are (minimax) optimal and, conversely, when nonlinear updates—such as those mirror descent employs—are necessary for optimal convergence. When the constraint set is quadratically convex, diagonally pre-conditioned stochastic gradient methods are minimax optimal. We provide quantitative converses showing that the "distance" of the underlying constraints from quadratic convexity determines the sub-optimality of subgradient methods. These results apply, for example, to any ℓ_p-ball for $p < 2$, and the computation/accuracy tradeoffs they demonstrate exhibit a striking analogy to those in Gaussian sequence models.

1 Introduction

The default procedures for solving the stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \;\; f_P(x) := \mathbf{E}_P\left[F(x, S)\right] = \int F(x, s) dP(s), \tag{\textsc{Opt}}$$

where $\{F(\cdot, s), s \in \mathcal{S}\}$ are convex functions $F(\cdot, s) : \mathbf{R}^n \to \mathbf{R}$, P is a distribution on $\mathcal{S}$, and $X \subset \mathbf{R}^n$ is a closed convex set, are variants of the stochastic subgradient method, where one iteratively draws $S_k \overset{\text{iid}}{\sim} P$ and updates

$$x_{k+1} := x_k - \alpha_k g_k, \quad \text{where} \quad g_k \in \partial F(x_k, S_k). \tag{1}$$

The simplicity and scalability of this update make stochastic subgradient methods the *de facto* choice for large-scale optimization [30, 26, 7]. The geometry of the underlying underlying constraint set X and subgradients $\partial F(\cdot, s)$ impact the performance of algorithms for problem (Opt), and so a question arises: are such linear updates (1) enough to obtain (minimax rate) optimal convergence guarantees for the problem (Opt), or does does the structure of the problem *necessitate* nonlinearity to achieve optimization efficiency? Convergence guarantees for stochastic gradient methods depend on the ℓ_2-diameter of X and $\partial F(\cdot, s)$, while for non-Euclidean geometries (e.g. when X is an ℓ_1- or ℓ_∞-ball) mirror descent, dual averaging and adaptive gradient methods provide better convergence guarantees [25, 26, 5, 27, 17, 13]. We investigate these gaps by precisely quantifying convergence

[*] JCD, CC, and DL partially supported by NSF-CAREER Award 1553086, ONR-YIP N00014-19-1-2288, and the Stanford DAWN Project. JCD and CC additionally supported by NSF Award IIS-2006777 and ONR Award N00014-22-1-2669

for different method families, highlighting a particular way to trade between computational power—which we treat as whether purely linear operations suffice to optimally solve problem (Opt), or nonlinear updates are necessary—and optimization and statistical efficiency.

To set the stage, let us revisit Donoho, Liu, and MacGibbon's study of optimal estimation in Gaussian sequence models [15]. One observes a vector $x \in X$ corrupted by Gaussian noise,

$$y = x + \mathsf{N}(0, \sigma^2 I),$$

and seeks to estimate x. For such problems, one can consider linear estimators—$\widehat{x} = AY$ for a $A \in \mathbf{R}^{n \times n}$—or potentially non-linear estimators

$$\widehat{x} = \Phi(y)$$

where $\Phi : \mathbf{R}^n \to X$ is otherwise arbitrary. When X is quadratically convex, meaning the set $X^2 := \{(x_j^2) \mid x \in X\}$ is convex, Donoho et al. show there exist minimax rate-optimal linear estimators; conversely, there are non-quadratically convex X for which rate-optimal estimators $\widehat{x}$ *must* be nonlinear in y. In particular, as we discuss in Section 5, this gap depends on the difference between the Kolmogorov (linear) n-width of X and its "nonlinear" n-width, that is,

$$w^2(n) := \sup_{v \in \mathrm{Conv}(X^2)} \sum_{j>n} v_{(j)} \quad \text{versus} \quad w_{\mathrm{nl}}^2(n) := \sup_{v \in X^2} \sum_{j>n} v_{(j)}, \tag{2}$$

where $|v_{(1)}| \ge |v_{(2)}| \ge \cdots$ denote the elements of v sorted by magnitude. We show how these results follow from convex duality, and the difference between $w^2(n)$ and $w_{\mathrm{nl}}^2(n)$ allows a quantitative characterization of how far X is from being quadratically convex and the impact this distance has on the (sub)optimality of linear estimators.

Our main results show how stochastic and online convex optimization analogize these sequence models. To build the analogy, consider dual averaging [27], where for a strongly convex $h : X \to \mathbf{R}$, one iteratively receives $S_k \in \mathcal{S}$, chooses $g_k \in \partial F(x_k, S_k)$, and for a stepsize $\alpha_k > 0$ updates

$$x_{k+1} := \operatorname*{argmin}_{x \in X} \left\{ \sum_{i \le k} g_i^\top x + \frac{1}{\alpha_k} h(x) \right\}. \tag{3}$$

When $X = \mathbf{R}^n$ and h is Euclidean, that is, $h(x) = \frac{1}{2} x^\top A x$ for some $A \succ 0$, the updates are linear in the observed gradients g_i, as

$$x_k = -\alpha_k A^{-1} \sum_{i \le k} g_i.$$

Drawing a parallel between Φ in the Gaussian sequence model and h in dual averaging (3), we show that because of duality gaps in certain min-max problems, a dichotomy holds for stochastic and online convex optimization similar to that holding for the Gaussian sequence model: if X is quadratically convex, there is a Euclidean h (yielding "linear" updates (3)) that is minimax rate optimal for problem (Opt), while there exist non-quadratically convex X for which Euclidean distance-generating h are arbitrarily suboptimal. Taking a computational perspective, this means that for some problems one *must* use more sophisticated methods than "linear" updates. We show that this analogy holds, though the measurement of a set's deviance from quadratic convexity, and hence the gap in attainable performance between linear and nonlinear methods, differs between Gaussian sequence models and stochastic optimization: there are constraint sets X for which linear estimators are (rate) optimal in the Gaussian sequence model but not for stochastic optimization, and vice versa. Nonetheless, we fully characterize minimax rates when the subgradients $g \in \partial F$

lie in a quadratically convex set or a weighted ℓ_r ball, $r \geq 1$. (This issue does not arise for the Gaussian sequence model, as the observations Y come from a fixed distribution, so there is no notion of alternative norms on Y.)

More precisely, we prove that for orthosymmetric quadratically convex bodies X, subgradient methods with a fixed diagonal re-scaling are minimax rate optimal. This guarantees that for a large collection of constraints (e.g. ℓ_2-balls, weighted ℓ_p-bodies for $p \geq 2$, or hyperrectangles) a diagonal re-scaling suffices. This is important in, e.g., machine learning problems of appropriate geometry, such as in linear classification problems where the data (features) are sparse, so using a dense predictor x is natural [17, 18]. Conversely, we show that if the constraint set X is a (scaled) ℓ_p-ball, $1 \leq p < 2$, then, considering unconstrained updates (3), the regret of the best method of linear type (i.e. h quadratic) can be $\sqrt{n/\log n}$ times larger than the minimax rate in online convex optimization. As part of this, we provide new information-theoretic lower bounds on optimization for general convex constraints X. In contrast to the frequent (but illogical) practice of comparing convergence upper bounds, we demonstrate the gap between linear and non-linear methods must hold. Sections 4.2 and 5 also show how the departure from quadratic convexity affects convergence guarantees: comparing the ℓ_1 diameters of X and its second-order lifts via

$$\sup_{x \in X} \|x\|_1 \quad \text{versus} \quad \sup_{v} \left\{ \|v\|_1 \mid v^2 \in \text{Conv}\{\text{diag}(xx^\top), x \in X\} \right\},$$

the gap between the left and right quantities (essentially) characterizes the gap in performance between linear and nonlinear methods for stochastic optimization, while Kolmogorov n-widths (2) capture that in the Gaussian sequence model.

We extend our results to an additional computational consideration: whether an algorithm must be adaptive, that is, it must change its update rules over time based on observations. We demonstrate that non-adaptive linear methods necessarily suffer slower convergence rates than adaptive methods in online problems. One perspective on our results is thus computational, though with a different angle than most current work on tradeoffs between statistics and computational complexity. Much of this literature takes as inspiration the classical perspective that the gap between polynomial and non-polynomial time algorithms forms the great watershed in computational complexity, thus necessitating a class of "hard" problems while allowing essentialy arbitrary algorithms [6, 8, 9]. We take an alternative perspective that allows more nuance in the types of convergence rates we can achieve—differentiating between various polynomials—by restricting the algorithms we consider to those in families common in optimization.

Our conclusions relate to the growing literature in adaptive algorithms [4, 17, 28, 29, 14]. Our results effectively prescribe that these adaptive algorithms are useful when the constraint set is quadratically convex, as this guarantees a minimax optimal diagonal pre-conditioner. More, different sets suggest different regularizers. For example, when the constraint set is a hyperrectangle, AdaGrad has regret at most $\sqrt{2}$ times that of the best post-hoc pre-conditioner, which we show is minimax optimal, while (non-adaptive) standard gradient methods can be $\sqrt{n}$ suboptimal on such problems. Conversely, our results strongly recommend against those methods for non-quadratically convex constraint sets. Our results thus clarify existing convergence guarantees [27, 26, 17, 39]: when the geometry of X and ∂F is appropriate for adaptive gradient methods or Euclidean algorithms, one should use them; when it is not—the constraints X are not quadratically convex—one should not.

Notation We use n to refer to the dimension of problems, and we use k to denote either a sample size or number of iterations. We let $\mathbf{R}^{\mathbf{N}} = \{(x_j)_{j=1}^{\infty}\}$ denote sequence space. For a norm γ, the set $\mathbf{B}_\gamma(x_0, r) := \{x \mid \gamma(x - x_0) \leq r\}$ denotes the ball of radius r around x_0 in the γ norm. For $p \in [1, \infty]$ we use the shorthand $\mathbf{B}_p(x_0, r) := \mathbf{B}_{\|\cdot\|_p}(x_0, r)$. The dual norm of γ is

$\gamma^*(z) = \sup_{\gamma(x)\le 1} \langle x, z\rangle$. For $x, \tau \in \mathbf{R}^n$ or $\mathbf{R}^{\mathbf{N}}$, we abuse notation and define $x^2 := (x_j^2)_{j\ge 1}$, $|x| := (|x_j|)_{j\ge 1}$, $\frac{x}{\tau} := (x_j/\tau_j)_{j\ge 1}$ and $x \odot \tau := (x_j \tau_j)_{j\ge 1}$, and similarly for sets X, if $f : \mathbf{R} \to \mathbf{R}$ then we let $f(X) = \{(f(x_j))_{j\ge 1} \mid x \in X\}$ be the elementwise application of f to elements of X. The function h denotes a *distance generating function*, i.e. a function strongly convex with respect to a norm $\|\cdot\|$; $\mathrm{D}_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y\rangle$ denotes the Bregman divergence, where recall that h is strongly convex with respect to $\|\cdot\|$ if and only if $\mathrm{D}_h(x, y) \ge \frac{1}{2}\|x - y\|^2$. The subdifferential of $F(\cdot, s)$ at x is $\partial_x F(x, s)$. $\mathsf{I}(X; Y)$ is the (Shannon) mutual information between random variables X and Y. For a set Ω and $f, g : \Omega \to \mathbf{R}$, we write $f \lesssim g$ if there exists a finite numerical constant C such that $f(t) \le Cg(t)$ for $t \in \Omega$, and $f \asymp g$ if $g \lesssim f \lesssim g$.

2 Preliminaries and Background

We begin by reviewing the classical results in Gaussian sequence models and presenting and defining the minimax framework in which we analyze procedures. We also review standard stochastic subgradient methods and introduce the relevant geometric notions of convexity we require. As part of this, we give a new argument showing the optimality of linear estimators for Gaussian sequence models when the underlying constraint set is quadratically convex (which we define presently).

Quadratic convexity and orthosymmetry

A few geometric quantities are central to our development. For a set X, let $X^2 := \{x^2, x \in X\}$ denote its (elementwise) square. The set X is *quadratically convex* if X^2 is convex; typical examples of quadratically convex sets are weighted ℓ_p bodies for $p \ge 2$ or hyperrectangles. We let $\mathsf{QHull}(X)$ be the quadratic convex hull of X, meaning the smallest convex and quadratically convex set containing X. The set $X \subset \mathbf{R}^n$ or $X \subset \mathbf{R}^{\mathbf{N}}$ is *orthosymmetric* if it is invariant to flipping the signs of any coordinate: if $x \in X$ then $\sigma_j \in \{\pm 1\}$ implies $(\sigma_j x_j)_{j\ge 1} \in X$. Similarly, a norm γ is orthosymmetric if $\gamma(g) = \gamma(|g|)$ for all g, and γ is quadratically convex if it induces a quadratically convex unit ball $\mathbf{B}_\gamma(0, 1)$. For any set X, we define the squared convex hull and square root

$$\mathsf{SqHull}(X) := \mathrm{Conv}\left\{(x_j^2) \mid x \in X\right\} \quad \text{and} \quad \sqrt{\mathsf{SqHull}(X)} = \{(\sqrt{y_j}) \mid y \in \mathsf{SqHull}(X)\},$$

the latter of which is always convex by the concavity of the square root. For orthosymmetric X,

$$\mathsf{QHull}(X) = \left\{s \odot x \mid x \in \sqrt{\mathsf{SqHull}(X)}, s_j \in \{\pm 1\} \text{ for all } j\right\}.$$

2.1 The Gaussian sequence model

Gaussian sequences provide a model for analyzing parametric and nonparametric statistical procedures, and tools developed in their analysis form a bedrock of modern statistical estimation [36, 22]; we provide some perspective on estimation in the sequence model. In the Gaussian sequence model, we begin with a (typically convex and compact) set $X \subset \mathbf{R}^n$ or in sequence space $\mathbf{R}^{\mathbf{N}}$, and for an unknown $x \in X$ observe $y = x + \xi$, where $\xi \sim \mathsf{N}(0, \sigma^2 I)$. The goal is to estimate x in some sense optimally, and frequently one considers sequences with $X \subset \mathbf{R}^n$ and σ^2 scaling as $1/n$, which analogizes estimation with n observations, so that rates of convergence as $\sigma^2 \downarrow 0$ become the main focus [22]. An interesting point of contrast is when linear estimators are sufficient to achieve (near) optimal performance or nonlinear estimators are necessary. An estimator $\widehat{x} = \widehat{x}(y)$ is *linear* if it has the form $\widehat{x} = Ay$ for a linear operator A and nonlinear otherwise. We consider the *risk*

$$R(\widehat{x}, x) := \mathbf{E}\left[\|\widehat{x}(y) - x\|_2^2\right],$$

where for linear estimators of the form $\widehat{x} = Ay$, we use the shorthand

$$R(A, x) = \mathbf{E}\left[\|Ay - x\|_2^2\right] = \mathbf{E}\left[\|(A - I)x + A\xi\|_2^2\right] = \|(A - I)x\|_2^2 + \sigma^2 \|A\|_{\mathrm{Fr}}^2 . \tag{4}$$

The *maximum risk* of an estimator over the set X is

$$R^*(\widehat{x}, X) := \sup_{x \in X} R(\widehat{x}, x),$$

while the minimax risk and linear minimax risk are

$$R^*(X) := \inf_{\widehat{x}} R^*(\widehat{x}, X) \quad \text{and} \quad R^*_{\mathrm{lin}}(X) := \inf_A R^*(A, X).$$

Donoho et al. [15] proved fundamental results relating the minimax risk and linear minimax risk for the Gaussian sequence model, and among their main results is that if the set X is an orthosymmetric quadratically convex body, then $R^*_{\mathrm{lin}}(X) \le 1.25 R^*(X)$, and moreover, $R^*_{\mathrm{lin}}(X) = \sup_{H \subset X} R^*_{\mathrm{lin}}(H)$, where H is a (hyper)rectangle. We begin by giving an alternative approach to some of these arguments here via convex duality, which allows us to put these arguments and the rest of our development on similar intellectual footing. In brief, the Sion and Fan minimax theorems [34, 19], coupled with quadratic lifts of the set X, play an essential role in all of our results.

Proposition 2.1. *Assume $\sigma^2 > 0$ and $X \subset \mathbf{R}^n$ is an orthosymmetric convex body. Then the matrix A minimizing $R^*(A, X)$ is diagonal and unique. Moreover,*

$$\inf_A R^*(A, X) = \inf_{d \in \mathbf{R}^n} \sup_{x \in X} \left\{ \sum_{j=1}^n (d_j - 1)^2 x_j^2 + \sigma^2 d_j^2 \right\}.$$

We provide a proof of the proposition in Appendix B.1, noting the following corollary of our proof technique. In the corollary, we say that a measure ν is *orthosymmetric* if for all diagonal sign matrices $\Sigma \in \mathrm{diag}(\{\pm 1\}^n)$ and sets $C \subset \mathbf{R}^n$, we have $\nu(C) = \nu(\Sigma C)$.

Corollary 2.1. *There exists an orthosymmetric probability measure ν on X such that the A minimizing $R^*(A, X)$ minimizes $\int_X \mathbf{E}[\|Ay - x\|_2^2] d\nu(x)$, and it coincides with the diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$ minimizing $R^*(D, X)$, which similarly minimizes*

$$\sum_{j=1}^n \int_X \left[d_j^2 x_j^2 - 2 d_j x_j^2 + \sigma^2 d_j^2\right] d\nu(x).$$

With these results in place, we can provide alternative proofs characterizing the linear minimimax risk for the Gaussian sequence model. We first evaluate the risk.

Corollary 2.2. *Let $X \subset \mathbf{R}^n$ be an orthosymmetric quadratically convex body. Then*

$$\inf_A R^*(A, X) = \sup_{q \in \mathsf{QHull}(X)} \sum_{j=1}^n \frac{\sigma^2 q_j^2}{q_j^2 + \sigma^2}.$$

Proof. Proposition 2.1 implies that

$$\begin{aligned}
\inf_A \sup_{x \in X} \mathbf{E}_x \left[\|Ay - x\|_2^2\right] &= \inf_{d \in \mathbf{R}^n} \sup_{x \in X} \sum_{j=1}^n \left((d_j - 1)^2 x_j^2 + \sigma^2 d_j^2\right) \\
&= \inf_d \sup_{v \in \mathsf{SqHull}(X)} \sum_{j=1}^n \left((d_j - 1)^2 v_j + \sigma^2 d_j^2\right) \\
&= \sup_{v \in \mathsf{SqHull}(X)} \inf_d \left\{ \sum_{j=1}^n \left((d_j - 1)^2 v_j + \sigma^2 d_j^2\right) \right\}, \qquad (5)
\end{aligned}$$

where equality (5) is a standard convex/concave saddle-point result (we may without loss of generality restrict d to the set $[0,1]^n$). Continuing the equalities, we have

$$\inf_d \left\{(d_j-1)^2 v_j + \sigma^2 d_j^2\right\} = \frac{\sigma^2 v_j}{v_j+\sigma^2},$$

which implies the result. □

An approximation argument extends Corollary 2.2 to sequence space (see Appendix B.2).

Corollary 2.3. *Let $X \subset \mathbf{R}^{\mathbf{N}}$ be orthosymmetric and compact for $\ell_2(\mathbf{N})$. Then*

$$\inf_A R^*(A,X) = \sup_{q\in\mathsf{QHull}(X)} \sum_{j=1}^{\infty} \frac{q_j^2\sigma^2}{q_j^2+\sigma^2}.$$

The two preceding corollaries—particularly via the swap of the min/max in equality (5)—highlight that if $\mathsf{QHull}(X)=X$, then linear estimators can be essentially chosen assuming knowledge that coordinate j of x lies in a $[-|x_j|,|x_j|]$.

With these results in place, we can provide an alternative proof that if X is quadratically convex, then the linear minimax risk for the Gaussian sequence model is equal to linear minimax risk over all rectangular subsets of X (recovering [15, Theorem 7]). Let $\mathcal{R}(X)$ denote the collection of orthosymmetric rectangular subsets of X, that is, sets of the form $[-x_j,x_j]_{j\geq1}\subset X$.

Corollary 2.4. *Let X be quadratically convex, compact, and orthosymmetric. Then*

$$\inf_{\widehat{x}=Ay}\sup_{x\in X}\mathbf{E}[\|\widehat{x}-x\|_2^2] = \inf_A R^*(A,X) = \sup_{H\in\mathcal{R}(X)}\inf_A R^*(A,H) = \sigma^2\sup_{x\in X}\sum_{j\geq1}\frac{x_j^2}{x_j^2+\sigma^2}.$$

Proof. Because $X=\mathsf{QHull}(X)$, Corollary 2.3 implies $\inf_A R^*(A,X)=\sup_{x\in X}\sum_{j\geq1}\frac{\sigma^2x_j^2}{x_j^2+\sigma^2}$. Now note that given any hyperrectangle $H=\bigotimes_{j\geq1}[-x_j,x_j]$ for $x\in X$, we have again by Corollary 2.3 that $\inf_A R^*(A,H)=\sum_{j\geq1}\frac{\sigma^2x_j^2}{x_j^2+\sigma^2}$. Take a supremum. □

2.1.1 Fundamental limits for the Gaussian sequence model

To introduce the ideas for lower bounds we employ in the remainder of the paper and highlight quadratic convexity, we also review some of the fundamental lower and upper bounds in Gaussian sequence models for general orthosymmetric sets X. The following result, which for completeness we prove in Appendix B.3, is typical; it provides worse constants than those available by more careful constructions (cf. [22, Chapter 4]), but it introduces some of our main types of arguments.

Proposition 2.2. *Let X be an orthosymmetric convex set. Then for any $x\in X$,*

$$R^*(X)\geq\frac{1}{10}\sum_{j\geq1}x_j^2\wedge\sigma^2.$$

As an immediate consequence to Proposition 2.2, we see that linear estimators are minimax rate optimal whenever X is quadratically convex.

Corollary 2.5. *Let X be quadratically convex, orthosymmetric, and compact. Then*

$$R^*(X)\leq\inf_A R^*(A,X)\leq10R^*(X).$$

Proof. Observe that $x_j^2 \wedge \sigma^2 \geq \frac{x_j^2\sigma^2}{x_j^2+\sigma^2}$, and then apply Corollary 2.4 and Proposition 2.2. $\square$

Given that whenever X is quadratically convex, linear estimators are (nearly) minimax optimal, the fundamental question then becomes when we indeed require nonlinearity and which nonlinear methods are rate optimal.

2.1.2 Soft-thresholding and nonlinear estimators

Soft-thresholding, which estimates x by elementwise applying the soft-thresholding operator

$$\mathsf{S}_\lambda(y) := \mathrm{sign}(y) \cdot (|y| - \lambda)_+ ,$$

provides a nearly optimal procedure for estimation in Gaussian sequence models. Johnstone [22, Corollary 8.4] gives a paradigmatic bound for the risk of soft thresholding on a single coordinate:

Corollary 2.6. *Let $y \sim \mathsf{N}(x, \sigma^2)$ and $\delta \in (0, 1]$. Then for $\lambda = \sqrt{2\sigma^2 \log \delta^{-1}}$,*

$$\mathbf{E}[(\mathsf{S}_\lambda(y) - x)^2] \leq \delta\sigma^2 + \left(1 + 2\log\delta^{-1}\right) \cdot x^2 \wedge \sigma^2.$$

In particular, if $X \subset \mathbf{R}^n$ and $\delta = \frac{1}{n}$, then defining the estimator $\widehat{x}$ coordinatewise by $\widehat{x}_j = \mathsf{S}_\lambda(y_j)$ for $\lambda = \sqrt{2\sigma^2 \log n}$, we obtain

$$\mathbf{E}[\|\widehat{x} - x\|_2^2] \leq \sigma^2 + (2\log n + 1) \sum_{j=1}^n x_j^2 \wedge \sigma^2,$$

whose risk is within a factor $O(1) \log n$ of the minimax risk lower bound in Proposition 2.2. In fact, the factor $2 \log n$ is sharp [22, Proposition 8.8] as $n \to \infty$.

One typically considers the risk in Gaussian sequence models as $\sigma^2 \to 0$—for example, with the scaling $\sigma^2 \propto 1/n$ in an n-dimensional model—so that we wish to understand the "right" scaling in the model. For this, we adapt Donoho et al.'s results [15, Thm. 12] by combining projecting coordinates to zero and soft-thresholding. Define

$$N(\sigma, X) := \inf\left\{n \mid \sup_{x \in X} |x_j| \leq \sigma \text{ for all } j \geq n\right\},$$

and for $\lambda = \sqrt{2\sigma^2 \log N(\sigma)}$, define the truncated soft-thresholding estimator

$$\widehat{x}_j = \begin{cases} \mathsf{S}_\lambda(y_j) & \text{if } j \leq N(\sigma, X) \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

An adaptation of their arguments [15, pg. 1430] yields the following corollary.

Corollary 2.7. *The truncated soft-thresholding estimator (6) satisfies*

$$\mathbf{E}[\|\widehat{x} - x\|_2^2] \leq \sigma^2 + (2\log N(\sigma, X) + 1) \sum_{j \geq 1} x_j^2 \wedge \sigma^2.$$

If $N(\sigma, X)$ is polynomial in $1/\sigma$ as $\sigma \to 0$, there exists $C(\sigma) \leq O(1) \log \frac{1}{\sigma}$ such that

$$R(\widehat{x}, X) \leq C(\sigma) \cdot R^*(X).$$

Proof. For $N = N(\sigma, X)$ we have

$$\mathbf{E}[\|\widehat{x} - x\|_2^2] = \sum_{j=1}^{N} \mathbf{E}[(\mathsf{S}_\lambda(y_j) - x_j)^2] + \sum_{j>N} x_j^2 \le \sigma^2 + (1 + 2\log N) \sum_{j=1}^{N} x_j^2 \wedge \sigma^2 + \sum_{j>N} x_j^2 \wedge \sigma^2,$$

as $x_j^2 \le \sigma^2$ for $j > N(\sigma, X)$, which implies the first result. Proposition 2.2 implies the second. □

The assumption that $N(\sigma, X)$ is polynomial in $1/\sigma$ as $\sigma \to 0$ is typically lenient; any $X \subset \mathbf{R}^n$ evidently satisfies it, and so do sets contained in ℓ_p bodies $\{x \in \mathbf{R}^{\mathbf{N}} \mid \sum_{j=1}^\infty a_j |x_j|^p \le 1\}$ so long as $a_j \to \infty$ polynomially quickly in the index j. In brief, the (nonlinear) truncated soft-thresholding estimator (6) is nearly minimax rate-optimal: to within a logarithmic factor it achieves the minimax optimal rate for any "sufficiently compact" set X. Moreover, because $\frac{1}{2}(a \wedge b) \le \frac{ab}{a+b} \le a \wedge b$ for $a, b \ge 0$, the difference between the quantities

$$\sup_{x \in X} \sum_j x_j^2 \wedge \sigma^2 \quad \text{and} \quad \sup_{x \in \mathsf{QHull}(X)} \sum_j x_j^2 \wedge \sigma^2$$

evidently determines whether nonlinear estimators are necessary.

2.2 Minimax rate for convex stochastic optimization

We turn to stochastic optimization. We measure the complexity of problem families in two familiar ways: stochastic minimax complexity and regret [25, 1, 10, 16]. Let $X \subset \mathbf{R}^n$ be a closed convex set, $\mathcal{S}$ a sample space, and $\mathcal{F}$ a collection of functions $F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R}$. For a collection $\mathcal{P}$ of distributions over $\mathcal{S}$, recall (Opt) that $f_P(x) := \int F(x, s) dP(s)$ is the expected loss of the point x. Then the *minimax stochastic risk* is

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}, \mathcal{P}) := \inf_{\widehat{x}_k} \sup_{F \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbf{E}\left[f_P(\widehat{x}_k(S_1^k)) - \inf_{x \in X} f_P(x) \right],$$

where the expectation is taken over $S_1^k \overset{\text{iid}}{\sim} P$ and the infimum ranges over all measurable functions $\widehat{x}_k$ of $\mathcal{S}^k$. A related notion is the average *minimax regret*, which instead takes a supremum over samples $s_1^k \in \mathcal{S}^k$ and measures losses instantaneously. In this case, an algorithm consists of a sequence of decisions $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_k$, where $\widehat{x}_i$ is chosen conditional on samples s_1^{i-1}, so that

$$\mathfrak{M}_k^{\mathsf{R}}(X, \mathcal{F}, \mathcal{S}) := \inf_{\widehat{x}_{1:k}} \sup_{F \in \mathcal{F}, s_1^k \in \mathcal{S}^k, x \in X} \frac{1}{k} \sum_{i=1}^{k} \left[F\left(\widehat{x}_i\left(s_1^{i-1}\right), s_i\right) - F(x, s_i) \right].$$

In the regret case we may of course identify s_i with individual functions F. In both definitions, we do not constrain the point estimates $\widehat{x}$ to lie in the constraint sets—in language of learning theory, improper predictions—but in our cases, this does not change regret by more than a constant factor. As online-to-batch conversions make clear [11], we always have $\mathfrak{M}_k^{\mathsf{S}} \le \mathfrak{M}_k^{\mathsf{R}}$; thus we typically provide lower bounds on $\mathfrak{M}_k^{\mathsf{S}}$ and upper bounds on $\mathfrak{M}_k^{\mathsf{R}}$. In many of the cases we consider, these quantities are essentially equivalent [e.g. 33], and in cases where we wish to provide explicit lower bounds on algorithms we typically use regret.

Lipschitz continuity properties form a central lever for demonstrating convergence in general (potentially non-smooth) stochastic convex optimization [25, 1, 16], and consequently, we study functions for which a norm γ on $\mathbf{R}^n$ (γ as a mnemonic for gradient) specifies these:

$$\mathcal{F}^{\gamma, r} := \{F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R} \mid \text{for all } x \in \mathbf{R}^n,\ g \in \partial_x F(x, s),\ \gamma(g) \le r\}, \tag{7}$$

The gradient bound condition $\gamma(g) \le r$ is equivalent to the Lipschitz condition $|F(x,s) - F(x',s)| \le r\gamma^*(x - x')$, where γ^* is the dual norm to γ. We use the shorthands

$$\mathfrak{M}_k^{\mathsf{R}}(X,\gamma) := \sup_{\mathcal{S}} \mathfrak{M}_k^{\mathsf{R}}(X, \mathcal{F}^{\gamma,1}, \mathcal{S}) \ \text{ and } \ \mathfrak{M}_k^{\mathsf{S}}(X,\gamma) := \sup_{\mathcal{S}} \sup_{\mathcal{P} \subset \mathcal{P}(\mathcal{S})} \mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}^{\gamma,1}, \mathcal{P})$$

as the Lipschitzian properties of $\mathcal{F}$ in relation to X determine the minimax regret and risk.

2.3 Stochastic gradient methods, mirror descent, and regret

Let us briefly review the canonical algorithms for solving the problem (Opt) and their associated convergence guarantees. For an algorithm outputing points $x_1, \ldots, x_k$, the *regret* on the sequence $F(\cdot, s_i)$ with respect to a point x is

$$\mathsf{Regret}_k(x) := \sum_{i=1}^{k} [F(x_i, s_i) - F(x, s_i)].$$

Recalling the definition $\mathrm{D}_h(x, x_0) = h(x) - h(x_0) - \langle \nabla h(x_0), x - x_0 \rangle$ of the Bregman divergence, the mirror descent algorithm [25, 5] iteratively sets

$$g_i \in \partial_x F(x_i, s_i) \ \text{ and updates } \ x_{i+1}^{\mathsf{MD}} := \operatorname*{argmin}_{x \in X} \left\{ g_i^\top x + \frac{1}{\alpha} \mathrm{D}_h(x, x_i) \right\} \tag{8}$$

where $\alpha > 0$ is a stepsize. When the function h is 1-strongly convex with respect to a norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, the iterates (8) and the iterates (3) of dual averaging satisfy (cf. [5, 10, 27])

$$\mathsf{Regret}_k(x) \le \frac{\mathrm{D}_h(x, x_0)}{\alpha} + \frac{\alpha}{2} \sum_{i \le k} \|g_i\|_*^2 \ \text{ for all } x \in X. \tag{9}$$

The choice $h(x) = \frac{1}{2} \|x\|_2^2$ recovers the classical stochastic gradient method, while the p-norm algorithms [20, 31, 26, 16], defined for $1 < p \le 2$, use $h(x) = \frac{1}{2(p-1)} \|x\|_p^2$; each is strongly convex with respect to the ℓ_p-norm $\|\cdot\|_p$. If $G = \{g \in \partial_x F(x,s) \mid x \in X, s \in \mathcal{S}\}$ denotes the set of possible subgradients, the regret guarantee (9) becomes

$$\mathsf{Regret}_k(x) \le \frac{\|x\|_p^2}{2(p-1)\alpha} + \frac{k\alpha}{2} \sup_{g \in G} \|g\|_q^2$$

if $x_0 = 0$ and $q = \frac{p}{p-1}$ is conjugate to p. The choice $\alpha = \frac{1}{\sqrt{k}} \sup_{x \in X} \|x\|_p / \sup_{\gamma(g) \le 1} \|g\|_q$ gives the following now standard minimax regret bound [cf. 32, Corollary 2.18].

Proposition 2.3. *Let X be closed convex, γ a norm, and $1 < p \le 2$, $q = \frac{p}{p-1}$. Mirror descent with distance generating function $h(x) := \frac{1}{2(q-1)} \|x\|_p^2$ and stepsize $\alpha = \frac{\sup_{x \in X} \|x - x_0\|_p}{\sqrt{k} \sup_{\gamma(g) \le 1} \|g\|_q}$ achieves regret*

$$\mathfrak{M}_k^{\mathsf{R}}(X,\gamma) \le \frac{\sup_{x \in X} \|x\|_p \sup_{g \in \mathbf{B}_\gamma(0,1)} \|g\|_q}{\sqrt{k(p-1)}}.$$

As we previously stated in our definitions of minimax risk and regret, we do not constrain the point estimates to lie in the constraint set X, which is equivalent to taking $X = \mathbf{R}^n$ in the updates (8) or (3). The regret bound (9) still holds whenever $x \in X$. Even with unconstrained updates, the form (9) still captures the regret for all common constraint sets X [31]. To give one example of this phenomenon, take $X = \mathbf{R}^n$ in the updates (8) or (3) while choosing $h(x) = \frac{1}{2(p-1)} \|x\|_p^2$ and $p = 1 + \frac{1}{\log(2n)}$. Then with dual $q = \frac{p}{p-1} = 1 + \log(2n)$, and $x_0 = 0$, we see that for *any* set $X' \subset \mathbf{R}^n$,

$$\sup_{x \in X'} \mathsf{Regret}_k(x) \le \frac{2}{\alpha} \sup_{x \in X'} h(x) + \frac{\alpha}{2} \sum_{i \le k} \|g_i\|_q^2 \le \frac{2}{\alpha} \sup_{x \in X'} h(x) + \frac{e^2 \alpha}{2} \sum_{i \le k} \|g_i\|_\infty^2 .$$

When we take the comparator class X' to be the ℓ_1 ball $X = \{x \mid \|x\|_1 \le 1\}$, then we see the familiar logarithmic scaling

$$\sup_{\|x\|_1 \le 1} \mathsf{Regret}_k(x) \le \frac{2\log(2n)}{\alpha} + \frac{\alpha}{2} \sum_{i \le k} \|g_i\|_\infty^2 ,$$

and if $\|g_i\|_\infty \le 1$ for all i then taking $\alpha = \frac{2}{e}\sqrt{\log(2n)/k}$ gives $O(1) \cdot \sqrt{k \log n}$ regret.

We frequently focus on distance generating functions of the form $h(x) = \frac{1}{2}\langle x, Ax \rangle$ for a fixed positive semi-definite matrix A. For an arbitrary A, we will refer to these methods as **Euclidean gradient methods** and for a diagonal A as **diagonally scaled gradient methods**. In this case, the mirror descent update is the stochastic gradient update with $A^{-1}g$, where g is a stochastic subgradient. We refer to all such methods as **methods of linear type**, as their update sequence guarantees the linearity

$$x_k = -A^{-1} \sum_{i=1}^{k-1} g_i.$$

The remainder of the paper develops the analogy of linear and nonlinear updates in stochastic optimization problems with those in Gaussian sequence models, highlighting when methods of linear type are minimax rate optimal, and when more computational power—we *require* nonlinearity—is necessary.

3 Minimax optimality and quadratically convex constraint sets

We begin by providing lower bounds on the minimax risk and matching upper bounds on the minimax regret of convex optimization over quadratically convex constraint sets, where diagonally scaled gradient methods achieve the regret bounds. While the analogy with the Gaussian sequence model is nearly complete, in distinction to the work of Donoho et al. (where results depend solely on the constraints X, as in Corollary 2.4), our results necessarily depend on the geometry of the subdifferential. Consequently, we distinguish throughout this section between quadratically and non-quadratically convex geometry of the gradients. To set the stage our contributions, we begin with the classical case of $X = \mathbf{B}_p(0,1)$ with $p \in [2, \infty]$ (so that X is quadratically convex) and norm $\gamma = \|\cdot\|_r$ with $r \ge 1$. We then turn to arbitrary quadratically convex constraint sets and first show results in the case of general quadratically convex norms on the subgradients. We conclude the section by proving that, when the subgradients do not lie in a quadratically convex set but lie in a weighted ℓ_r ball (for $r \in [1, 2]$), diagonally scaled gradient methods are still minimax rate optimal.

3.1 A warm-up: p-norm constraint sets for $p \ge 2$

Though the results for the basic case that X is an ℓ_p-ball while the gradients belong to a different ℓ_r-ball are special cases of the theorems to come, the proofs (appendicized) are simpler and provide intuition for the later results. We distinguish between two cases depending on the value of r in the gradient norm. The case that $r \in [1, 2]$ corresponds roughly to "sparse" gradients, while the case

$r \geq 2$ corresponds to harder problems with dense gradients. We provide information theoretic proofs of the following two results in Appendices C.1 and C.2, respectively.

Proposition 3.1 (Sparse gradients)**.** *Let $X = \mathbf{B}_p(0,1)$ with $p \geq 2$ and $\gamma(\cdot) = \|\cdot\|_r$ where $r \in [1,2]$. Then*

$$1 \wedge \frac{n^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{k}} \lesssim \mathfrak{M}_k^{\mathsf{S}}(X,\gamma) \leq \mathfrak{M}_k^{\mathsf{R}}(X,\gamma) \lesssim 1 \wedge \frac{n^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{k}}.$$

Proposition 3.2 (Dense gradients)**.** *Let $X = \mathbf{B}_p(0,1)$ with $p \geq 2$ and $\gamma(\cdot) = \|\cdot\|_r$ with $r \geq 2$. Then*

$$1 \wedge \frac{n^{\frac{1}{2}-\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{r}}}{\sqrt{k}} \lesssim \mathfrak{M}_k^{\mathsf{S}}(X,\gamma) \leq \mathfrak{M}_k^{\mathsf{R}}(X,\gamma) \lesssim 1 \wedge \frac{n^{\frac{1}{2}-\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{r}}}{\sqrt{k}}.$$

In both cases, the stochastic gradient method achieves the regret upper bound via a straightforward optimization of the regret bounds (9) with $h(x) = \frac{1}{2}\|x\|_2^2$; a method of linear type is optimal.

3.2 General quadratically convex constraints

We now turn to the more general case that X is an arbitrary orthosymmetric quadratically convex body. We combine two techniques to develop the results. The first builds out of the ideas of Donoho et al. [15] in Gaussian sequence estimation, where, as in Section 2.1 the largest hyperrectangle in X governs the performance of linear estimators; this gives us a lower bound. The key second technique is in the upper bound, where a strong duality result holds because of the quadratic convexity of X, allowing us to prove minimax optimality of diagonally scaled Euclidean procedures. As in the previous section, we divide our analysis into cases depending on whether the gradient norm γ is quadratically convex or not (the analogs of $r \lesseqgtr 2$ in Propositions 3.1 and 3.2).

We begin with the lower bound, which relies on rectangular structures in the primal X and dual gradient spaces. For the proposition, we use a specialization of the function families (7) to rectangular sets, where for $M \in \mathbf{R}_+^n$ we define

$$\mathcal{F}^M := \Big\{F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R} \mid \text{for all } x \in \mathbf{R}^n \text{ and } g \in \partial_x f(x,s),\ \max_{j \leq n} \frac{|g_j|}{M_j} \leq 1\Big\}.$$

Proposition 3.3 (Duchi et al. [18], Proposition 1)**.** *Let $M \in \mathbf{R}_+^n$ and $\mathcal{F}^M$ be as above. Let $a \in \mathbf{R}_+^n$ and assume the hyperrectangular containment $\prod_{j=1}^n [-a_j, a_j] \subset X$. Then*

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}^M) \geq \frac{1}{8\sqrt{k \log 3}} \sum_{j=1}^n M_j a_j.$$

We begin the analysis of the general case by studying the rates of diagonally scaled gradient methods.

3.2.1 Diagonal re-scaling in gradient methods

diagonally scaled gradient methods (componentwise re-scaling of the subgradients) are equivalent to using $h_\Lambda(x) := \frac{1}{2} x^\top \Lambda x$ for $\Lambda = \mathrm{diag}(\lambda) \succeq 0$ in the mirror descent update (8). In this case, for any norm γ on the gradients, the minimax regret bound (9) becomes

$$\sup_{x \in X} \mathsf{Regret}_{k,\Lambda}(x) \leq \frac{1}{2k}\left[\sup_{x \in X} x^\top \Lambda x + \sum_{i \leq k} g_i^\top \Lambda^{-1} g_i\right] \leq \frac{1}{2k}\left[\sup_{x \in X} x^\top \Lambda x + k \sup_{g \in \mathbf{B}_\gamma(0,1)} g^\top \Lambda^{-1} g\right].$$

The rightmost term upper bounds the minimax regret, so we may take an infimum over Λ, yielding

$$\mathfrak{M}_k^{\mathsf{R}}(X, \gamma) \le \frac{1}{2k} \inf_{\lambda \succeq 0} \sup_{x \in X} \sup_{g \in \mathbf{B}_\gamma(0,1)} \left[\sum_{j \le n} \lambda_j x_j^2 + k \sum_{j \le n} \frac{1}{\lambda_j} g_j^2 \right] \tag{10}$$

The regret bound (10) holds without assumptions on X or γ. However, in the case when X is quadratically convex, strong duality allows us to simplify this quantity:

Proposition 3.4. *Let $V, X \subset \mathbf{R}^n$ be convex, quadratically convex and compact sets. Then*

$$\inf_{\lambda \succ 0} \sup_{x \in X, v \in V} \left\{ \lambda^\top x^2 + \left(\frac{1}{\lambda}\right)^\top v^2 \right\} = \sup_{x \in X, v \in V} \inf_{\lambda \succ 0} \left\{ \lambda^\top x^2 + \left(\frac{1}{\lambda}\right)^\top v^2 \right\}.$$

Proof. The quadratic convexity of the sets X and V implies that a (weighted) squared 2-norm becomes a linear functional when lifted to the squared sets $X^2 := \{x^2 \mid x \in X\}$ and V^2. Indeed, defining $J : \mathbf{R}_+^{2n} \times \mathbf{R}_+^n \to \mathbf{R}$, $J(\tau, w, \lambda) := \lambda^\top \tau + (\frac{1}{\lambda})^\top w$, the function J is concave-convex: it is linear (a fortiori concave) in (τ, w) and convex in λ. Thus, using that the set $\{\lambda \in \mathbf{R}_+^n\}$ is convex and $X^2 \times V^2$ is convex compact (because X and V are quadratically convex compact), Sion's minimax theorem [34] implies

$$\begin{aligned} \inf_{\lambda \succ 0} \sup_{x \in X, v \in V} \left\{ \lambda^\top x^2 + \left(\frac{1}{\lambda}\right)^\top v^2 \right\} &= \inf_{\lambda \succ 0} \sup_{\tau \in X^2, w \in V^2} \left\{ \lambda^\top \tau + \left(\frac{1}{\lambda}\right)^\top w \right\} \\ &= \sup_{\tau \in X^2, w \in V^2} \inf_{\lambda \succ 0} \left\{ \lambda^\top \tau + \left(\frac{1}{\lambda}\right)^\top w \right\}. \end{aligned}$$

Replacing τ with x^2 and w with v^2 gives the result. □

Proposition 3.4 provides a powerful hammer for diagonally scaled Euclidean algorithms, as we can choose an optimal scaling for any *fixed* pair x, g, taking a worst case over such pairs:

Corollary 3.1. *Let X be a convex, quadratically convex, compact set. Then*

$$\mathfrak{M}_k^{\mathsf{R}}(x, \gamma) \le \frac{1}{\sqrt{k}} \sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1)), x \in X} x^\top g,$$

and diagonally scaled gradient methods achieve this regret.

Proof. We upper bound the minimax regret (10) by taking a supremum over the quadratic hull $g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1))$, which contains $\mathbf{B}_\gamma(0,1)$. Using that for $a, b > 0$, $\inf_{\lambda > 0} a\lambda + b/\lambda = 2\sqrt{ab}$ and applying Proposition 3.4 gives the proof. □

The corollary allows us to provide concrete upper and lower bounds on minimax risk and regret, with the results differing slightly based on whether the gradient norms are quadratically convex.

3.2.2 Orthosymmetric and quadratically convex gradient norms

We now provide lower bounds on minimax risk complementary to Corollary 3.1, focusing first on the case that the gradient norm γ is quadratically convex.

Assumption A1. *The norm γ is orthosymmetric and quadratically convex, meaning $\gamma(\sigma \odot v) = \gamma(v)$ for all $\sigma \in \{\pm 1\}^n$ and $\mathbf{B}_\gamma(0,1)$ is quadratically convex.*

With this, we have the following theorem, which shows that diagonally-scaled gradient methods are minimax rate optimal, and that the constants are sharp up to a factor of 9, whenever the gradient norms are quadratically convex. While the constant 9 is looser than that Donoho et al. [15] provide for Gaussian sequence models, this theorem highlights the essential structural similarity between the sequence model case and stochastic optimization methods.

Theorem 1. *Let Assumption A1 hold and let X be quadratically convex, orthosymmetric, and compact. Then*

$$\frac{1}{8\sqrt{\log 3}} \frac{1}{\sqrt{k}} \sup_{x \in X} \gamma^*(x) \le \mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \le \mathfrak{M}_k^{\mathsf{R}}(X, \gamma) \le \frac{1}{\sqrt{k}} \sup_{x \in X} \gamma^*(x).$$

There exists $\lambda^* \in \mathbf{R}_+^n$ such that diagonally scaled gradient methods with λ^* achieve this rate.

Proof. For the upper bound, we use Corollary 3.1. Because $\mathbf{B}_\gamma(0,1)$ is quadratically convex, we have $\mathsf{QHull}(\mathbf{B}_\gamma(0,1)) = \mathbf{B}_\gamma(0,1)$, so that $\sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1))} x^\top g = \gamma^*(x)$, giving the upper bound. The lower bound uses Proposition 3.3. Define the hyperrectangle $\mathsf{Rec}(x) := \prod_{j \le n} [-|x_j|, |x_j|]$, so that, by orthosymmetry of X, $X \supset \mathsf{Rec}(x)$ for all $x \in X$. Additionally, recalling the notation (7) of $\mathcal{F}^{\gamma,1}$ and $\mathcal{F}^M$, if $M \in \mathbf{R}_+^n$ satisfies $\gamma(M) \le 1$ then, by orthosymmetry of γ, $\mathcal{F}^{\gamma,1} \supset \mathcal{F}^M$. Thus

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \mathfrak{M}_k^{\mathsf{S}}(\mathsf{Rec}(x), \gamma) \ge \mathfrak{M}_k^{\mathsf{S}}(\mathsf{Rec}(x), \mathcal{F}^M) \ge \frac{1}{8\sqrt{k \log 3}} \sum_{j \le n} |x_j| M_j$$

for all $M \in \mathbf{B}_\gamma(0,1) \cap \mathbf{R}_+^n$ and $x \in X$. Taking a supremum over $M \in \mathbf{B}_\gamma(0,1)$ and $x \in X$, we have

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \frac{1}{8\sqrt{k \log 3}} \sup_{x \in X} \sup_{\gamma(M) \le 1} x^\top M = \frac{1}{8\sqrt{k \log 3}} \sup_{x \in X} \gamma^*(x). \qquad \square$$

3.2.3 Arbitrary gradient norms

When the norm γ on the gradients defines a non-quadratically convex norm ball $\mathbf{B}_\gamma(0,1)$—for example, when the gradients belong to an ℓ_r-norm ball for $r \in [1,2]$—our results become less general. Nonetheless, when γ is a weighted ℓ_r-norm ball (for $r \in [1,2]$), diagonally scaled gradient methods remain minimax rate optimal, as Corollary 3.2 will show. When the norms γ are arbitrary we have a that uses the rescaled vector $\text{res}(x, \gamma) := (x_j / \gamma(e_j))_{j=1}^n$, where e_j are the standard basis vectors:

Theorem 2. *Let X be an orthosymmetric, quadratically convex, convex and compact set and γ an arbitrary norm. For any $d \in \mathbf{N}$,*

$$\begin{aligned} \frac{1}{8\sqrt{k \log 3}} \left(1 - \frac{d}{k \log 3}\right) \sup_{x \in X, \|x\|_0 \le d} \|\text{res}(x, \gamma)\|_2 &\le \mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \\ &\le \mathfrak{M}_k^{\mathsf{R}}(X, \gamma) \le \frac{1}{\sqrt{k}} \sup_{x \in X} \sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1))} x^\top g. \end{aligned} \tag{11}$$

Corollary 3.1 gives the upper bound in the theorem. The lower bound consists of an application of Assouad's method [2], but, in parallel to the warm-up examples, we construct well-separated functions with "sparse" gradients. See Appendix D.1 for a proof.

We can develop a corollary of this result when the norm γ is a weighted-ℓ_r norm for $r \in [1,2]$. While these do not induce quadratically convex norm balls, the previous theorem still guarantees that diagonally scaled gradient methods are minimax rate optimal.

Corollary 3.2. *Let the conditions of Theorem 2 hold and assume that $\gamma(g) = \|\beta \odot g\|_r$ with $r \in [1,2]$, $\beta_j > 0$ and $(\beta \odot g)_j = \beta_j g_j$. Then for $k \geq 2n$,*

$$\frac{1}{16}\frac{1}{\sqrt{k}} \sup_{x \in X} \|\mathrm{res}(x,\gamma)\|_2 \leq \mathfrak{M}_k^{\mathsf{S}}(X,\gamma) \leq \mathfrak{M}_k^{\mathsf{R}}(X,\gamma) \leq \frac{1}{\sqrt{k}} \sup_{x \in X} \|\mathrm{res}(x,\gamma)\|_2 .$$

There exists $\lambda^* \in \mathbf{R}_+^n$ such that diagonally scaled gradient methods with λ^* achieve this rate.

A minor modification of Theorem 2 gives the lower bound, while we obtain the upper bound by noting that the quadratic hull of a weighted-ℓ_r norm ball for $r \in [1,2]$ is the weighted-ℓ_2 norm ball. The dual norm of $\gamma(g) = \|\beta \odot g\|_2$ being $\gamma^*(g) = \|g/\beta\|_2$, the upper bound holds by duality. See Appendix D.2 for the (short) proof.

Theorem 1 and Corollary 3.2 show that for a large collection of norms γ on the gradients, diagonally scaled gradient methods is minimax rate optimal. Arguing that diagonally scaled gradient methods are minimax rate optimal when γ is neither a weighted-ℓ_r norm nor induces a quadratically convex unit ball remains an open question, though weighted-ℓ_r norms for $r \in [1,\infty]$ cover the majority of practical applications of stochastic gradient methods.

We conclude this section by generalizing our results to sets X that are rotations of orthosymmetric and quadratically convex sets. This is for example the case when features are sparse in an appropriate basis (e.g. wavelets [24]). Unsurprisingly, methods of linear type retain their optimality properties.

Corollary 3.3. *Let X_0 be a compact, orthosymmetric, convex and quadratically convex set. Let $U \in \mathcal{O}_n(\mathbf{R})$ be a rotation matrix and $X := UX_0 = \{Ux \mid x \in X_0\}$. Consider the collection*

$$\mathcal{F} := \{F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R} \mid \text{ for all } s \in \mathcal{S}, x \in \mathbf{R}^n, \text{ and } g \in \partial_x F(x,s), \gamma(U^T g) \leq 1\}.$$

A method of linear type is minimax rate optimal for the pair $(X, \mathcal{F})$.

Proof. There is a bijective mapping between $\mathcal{F}$ and $\mathcal{F}^{\gamma,1}$: for $F \in \mathcal{F}$, $x_0 \in X_0$, and $s \in \mathcal{S}$, we define $\widetilde{F}(x_0, s) := F(Ux_0, s)$. Then $\mathrm{dom}\, \widetilde{F}(\cdot, s) \supset X_0$, and its subdifferential [21, Thm. 4.2.1] is

$$\partial_x \widetilde{F}(x_0, s) = U^\top \partial_x F(Ux_0, s).$$

As $\widetilde{F}$ falls within the scope of Theorems 1 or Corollary 3.2, there exists a diagonal re-scaling Λ^* that achieves the optimal rate. We conclude the proof by observing that a diagonally re-scaled stochastic gradient update on $\widetilde{F}$ corresponds to the update $x_{i+1} = x_i - U\Lambda^* U^\top g_i$ where $g_i \in \partial_x F(x_i, S_i)$. □

4 Beyond quadratic convexity: the necessity of non-linear methods

For $X \subset \mathbf{R}^n$ quadratically convex, the results in Section 3 show that methods of linear type achieve optimal rates of convergence. When the constraint set is not quadratically convex, it is unclear whether methods of linear type are sufficient to achieve optimal rates. As we now show, they are not: we exhibit collections of problem instances in which the constraint sets are orthosymmetric convex bodies but not quadratically convex, and where methods of linear type must have regret at least a factor $\sqrt{n/\log n}$ worse than the minimax optimal rate, which (non-linear) mirror descent with appropriate distance generating function achieves. We also develop more general results to highlight the way in which the quadratic hull of the underlying constraint set X necessarily characterizes the regret of Euclidean gradient methods, which allows for a more explicity delineation of those sets X for which nonlinear methods are necessary: when $\sup_{x \in X} \|x\|_1$ is much smaller than $\sup_{x \in \mathsf{QHull}(X)} \|x\|_1$.

To construct these problem instances, we first turn to simple non-quadratically convex constraint sets: ℓ_p balls for $p \in [1,2]$. We measure subgradient norms in the dual ℓ_q norm, $q = \frac{p-1}{p}$. Our analysis consists of two steps: we first prove sharp minimax rates on these problem instances and show that mirror descent with the right (non-linear) distance generating function is minimax rate optimal. These results extend those of Agarwal et al. [1], who provide matching lower and upper bounds for $p \geq 1 + c$ for a fixed numerical constant $c > 0$. In contrast, we prove sharp minimax rates for all $p \geq 1$. To precisely characterize the gap between linear and non-linear methods, we show that for any linear pre-conditioner, we can exhibit functions for which the regret of Euclidean gradient methods is nearly the simple upper regret bound of standard gradient methods, Eq. (9) with $h(x) = \frac{1}{2}\left\|x\right\|_2^2$. Thus, when p is very close to 2 (nearly quadratically convex), the gap remains within a constant factor, whereas when p is close to 1, the gap can be as large as $\sqrt{n/\log n}$.

4.1 Minimax rates for p-norm constraint sets and general convex bodies

For $p \in [1, 2]$, we consider the constraint set $X = \mathbf{B}_p(0, 1) \subset \mathbf{R}^n$ and bound gradients with norm $\gamma = \|\cdot\|_{p^*}$. We begin by proving sharp minimax rates on this collection of problems and show that, in these cases, non-linear mirror descent is minimax optimal.

Theorem 3. *Let $p \in [1, 2]$, $X = \mathbf{B}_p(0, 1) \subset \mathbf{R}^n$ and $\gamma = \|\cdot\|_{p^*}$.*

(i) If $1 \leq p \leq 1 + 1/\log(2n)$, then

$$1 \wedge \sqrt{\frac{\log(2n)}{k}} \lesssim \mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \leq \mathfrak{M}_k^{\mathsf{R}}(X, \gamma) \lesssim 1 \wedge \sqrt{\frac{\log(2n)}{k}}.$$

(ii) If $1 + 1/\log(2n) < p \leq 2$, then

$$1 \wedge \sqrt{\frac{1}{k(p-1)}} \lesssim \mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \leq \mathfrak{M}_k^{\mathsf{R}}(X, \gamma) \lesssim 1 \wedge \sqrt{\frac{1}{k(p-1)}}$$

In both cases, mirror descent (8) with distance generating function $h(x) := \frac{1}{2(a-1)}\left\|x\right\|_a^2$ for $a = \max\{1 + \frac{1}{\log(2n)}, p\}$ achieves the optimal rate.

The upper bound essentially follows from Proposition 2.3, and the lower bound uses either reductions from estimation to testing or Assouad's method (see Appendix A.2). See Appendix E.1 for a proof.

We can also consider the minimax risk and regret for general optimization methods. To reduce complexity, we focus on the case that the gradients are bounded in ℓ_∞-norm—the "most" quadratically convex set—to make dependence on X the clearest. In the sequel it will sometimes be useful to consider sequence space $\mathbf{R}^{\mathbf{N}}$, so we give a result that allows the infinite dimensional containment $X \subset \mathbf{R}^{\mathbf{N}}$; in this case we consider domains X that are appropriately compact for the ℓ_1-norm. A sufficient condition for the results is that X have finite *effective dimension*, which we define by

$$\mathrm{effdim}(X) := \inf_{\beta \geq 0} \left\{ e^{1/\beta} \mid \sup_{x \in X} \sum_{j=1}^{\infty} j^{\beta} |x_j| \leq e \sup_{x \in X} \left\|x\right\|_1 \right\}, \tag{12}$$

where $\mathrm{effdim}(X) = +\infty$ if only $\beta = 0$ satisfies the inequality. It is immediate that if $X \subset \mathbf{R}^n$, then $\mathrm{effdim}(X) \leq n$ because $n^{1/\log n} = e$. The next observation summarizes a few sufficient conditions for the effective dimension (12) to be finite. We include the proof in Appendix E.3 for completeness.

Observation 4.1. *The following conditions are sufficient to guarantee* $\mathrm{effdim}(X) < \infty$*.*

i. If $X \subset \mathbf{R}^n$, then $\mathrm{effdim}(X) \leq n$*.*

ii. If $\lim_{\beta \downarrow 0} \sup_{x \in X} \sum_{j \ge 1} j^\beta |x_j| < \infty$*, then* $\mathrm{effdim}(X) < \infty$.

iii. Let $N < \infty$ *and* $\gamma > 0$ *satisfy the tail condition that* $\sup_{x \in X} \sum_{j=1}^N |x_j| \ge n^{2\gamma} \sup_{x \in X} \sum_{j > n} |x_j|$ *for all* $n > N$*. Then* $\mathrm{effdim}(X) \le O(1) \cdot (N^2 + \exp(\frac{3}{\gamma} \log \frac{1}{\gamma}))$.

The next proposition then shows that the ℓ_1-diameter of X always provides lower bounds on the (stochastic) minimax risk, while p-norm-based mirror descent algorithms achieve regret at most logarithmic in the effective-dimension (12), making the ℓ_1-diameter of X the central quantity governing minimax risk. See Appendix E.2 for the proof.

Proposition 4.1. *Let* X *be an orthosymmetric convex set. Then*

$$\frac{1}{8 \log 3} \frac{1}{\sqrt{k}} \sup_{x \in X} \|x\|_1 \le \mathfrak{M}_k^{\mathsf{S}}(X, \|\cdot\|_\infty).$$

If additionally X *has finite effective dimension* $N = \mathrm{effdim}(X)$ *as in* (12)*, then*

$$\mathfrak{M}_k^{\mathsf{R}}(X, \|\cdot\|_\infty) \le O(1) \cdot \sqrt{\log N} \frac{1}{\sqrt{k}} \sup_{x \in X} \|x\|_1 .$$

Letting $\beta = \frac{1}{\log N}$ *and defining the operator* A *by* $Ax = (j^\beta x_j)_{j \ge 1}$*, the mirror descent method with distance generating function* $h(x) = \frac{1}{2(p-1)} \|Ax\|_p^2$ *for* $p = \frac{2}{2-\beta}$ *achieves this regret.*

To preview our discussion to come, note the similarities between Proposition 4.1 and Corollary 2.7 in the Gaussian sequence model case: so long as X is appropriately regular, there is a "standard" nonlinear method—in the case of Gaussian sequence models, soft-thresholding with truncation (6), in the case of stochastic and online optimization, mirror descent with a p-norm-based distance-generating function—that achieves nearly rate-optimal minimax risk.

4.2 Hard problems for Euclidean gradient methods and quadratic hulls

Theorem 3 shows that (non-linear) mirror descent methods are minimax rate-optimal for ℓ_p-ball constraint sets, $p \in [1, 2]$, with gradients contained in the corresponding dual ℓ_q-norm ball ($q = \frac{p}{p-1}$). For such problems and p, standard subgradient methods achieve worst-case regret $O(n^{1/2-1/q}/\sqrt{k})$. This is an upper bound, but in fact is sharp: in the next theorem, we show that for any method of linear type, we can construct a sequence of (linear) functions such that the method's regret achieves the worst-case upper bound for standard subgradient methods, precisely quantifying the gap between linear and non-linear methods for this problem class. To that end, we let

$$\mathsf{Regret}_{k,A}(x) := \sum_{i=1}^k g_i^\top (x_i - x)$$

denote the regret of the Euclidean online mirror descent method with distance generating function $h_A(x) = \frac{1}{2} x^\top A x$ for functions F_i with subgradients g_i In the lower bounds to come, we take $F_i(x) = g_i^\top x$ to be linear, so that $\nabla F_i(x) = g_i$ is independent of x.

Theorem 4. *For any* $A \succeq 0$ *and* $p \in [1, 2]$ *with* $q = \frac{p}{p-1}$*, there exists a sequence of vectors* $g_i \in \mathbf{R}^n$*,* $\|g_i\|_q \le 1$*, and point* $x \in \mathbf{R}^n$ *with* $\|x\|_p \le 1$ *such that*

$$\mathsf{Regret}_{k,A}(x) \ge \frac{1}{2} \min \left\{ k/2, \sqrt{2k} \cdot n^{1/2-1/q} \right\} .$$

Scaled identity matrices $A = c \cdot I_n$ *achieve these bounds to within a factor of* $\sqrt{2}$ *for* $k \ge 2n^{1-2/q}$.

We provide the proof in Appendix F.1. These results explicitly exhibit a gap between methods of linear type and non-linear mirror descent methods for this problem class. In contrast to the frequent practice in literature of simply comparing regret upper bounds—prima facie illogical—we demonstrate the gap indeed must hold.

In combination with Theorem 4, Proposition 2.3 precisely characterizes the gap between linear and non-linear mirror descent on these problems for all values of $p \in [1, 2]$. Indeed, when $p = 1$, for any pre-conditioner A, there exists a problem on which Euclidean gradient methods have regret at least $\Omega(1)\sqrt{n/k}$. On the same problem, non-linear mirror descent has regret at most $O(1)\sqrt{\log n/k}$, showing the advertised $\sqrt{n/\log n}$ gap. When $p \geq 2 - 1/\log n$ (so X is nearly quadratically convex), the gap reduces to at most a constant factor.

To highlight the role of quadratically convex hulls, we provide an alternative version of Theorem 4 that allows us to focus more precisely on the constraint set X itself. In this case, we focus on scenarios where the gradients have bounded ℓ_∞-norm, $\gamma(g) = \|g\|_\infty$, which also more neatly analogizes with Gaussian sequence models $y = x + \xi$ for $\xi \sim \mathsf{N}(0, I_n)$, as we expect the scale of noise ξ_j on different coordinates to be similar. Recall also the notation (7) that $\mathcal{F}^{\|\cdot\|_\infty,1}$ is the collection of functions with subgradients $g \in \partial F(x, s)$ satisfying $\|g\|_\infty \leq 1$.

Theorem 5. *Let $X \subset \mathbf{R}^n$ be an orthosymmetric convex body. For any sequence $A(k) \succeq 0$, there exist sequences of vectors $g_i = g_i(k) \in \mathbf{R}^n$, $i = 1, \ldots, k$, with $\|g_i\|_\infty \leq 1$, such that*

$$\liminf_k \frac{1}{\sqrt{k}} \sup_{x \in X} \mathsf{Regret}_{k,A(k)}(x) \geq \sup_{x \in \mathsf{QHull}(X)} \|x\|_1 .$$

Additionally, for each k there exists a diagonal matrix D such that for any sequence of convex functions $F_i \in \mathcal{F}^{\gamma,1}$,

$$\sup_{x \in X} \mathsf{Regret}_{k,D}(x) \leq \sqrt{k} \sup_{x \in \mathsf{QHull}(X)} \|x\|_1 .$$

See Appendix F.2 for the proof of the result, which parallels that of Theorem 4.

In brief, the regret of Euclidean gradient methods *necessarily* scales with the size of the quadratic hull $\mathsf{QHull}(X)$. Contrasting this result with Proposition 4.1, we see with nonlinear methods, the regret need scale only as $\sup_{x \in X} \|x\|_1$ rather than as $\sup_{x \in \mathsf{QHull}(X)} \|x\|_1$, so that the gap between convergence achievable by linear and nonlinear methods is large precisely when

$$\frac{\sup_{x \in \mathsf{QHull}(X)} \|x\|_1}{\sup_{x \in X} \|x\|_1} \gg 1.$$

5 The role of quadratic convexity in sequence models and first-order methods

The results in Section 2.1 highlight and recapitulate some of the ways that quadratic convexity distinguishes linear and nonlinear methods in Gaussian sequence models. Theorems 4 and 5, along with the complimentary results in Proposition 4.1, address these differences for stochastic optimization problems. So in both sequence models and convex optimization, geometric aspects of the underlying set X determine nonlinear methods' necessity. The analogy between sequence models and stochastic optimization methods is not perfect, however, as there are sets X for which linear methods are minimax rate optimal for stochastic optimization problems and not for sequence models and vice versa. In both problem families, a particular "distance" of a set X from quadratic convexity delineates determines when nonlinear methods are necessary; we show that these can be different.

We begin by translating the results in Section 2.1 on Gaussian sequence models into a more geometric form; Donoho et al. [15] more or less give this translation but we make a few minor modifications for convenience in exposition. The measure of size most natural for Gaussian sequence models turns out to be (duality-based variants of) the Kolmogorov n-width of the underlying set:

Definition 5.1. *The* n-width *of a set X is*

$$w^2(n) := \inf_{\mathbf{0} \preceq d \preceq \mathbf{1}, \langle \mathbf{1}, d \rangle = n} \sup_{x \in X} \sum_j (1 - d_j) x_j^2.$$

The nonlinear n-width *of X is*

$$w_{\rm nl}^2(n) := \sup_{x \in X} \inf_{\mathbf{0} \preceq d \preceq \mathbf{1}, \langle \mathbf{1}, d \rangle = n} \sum_j (1 - d_j) x_j^2.$$

Recalling Corollaries 2.3 and 2.7, the gap between the linear minimax and nonlinear minimax risk is large for (compact) convex sets X whenever the difference between

$$\sup_{x \in X} \sum_{j=1}^{\infty} \frac{x_j^2 \sigma^2}{x_j^2 + \sigma^2} \quad \text{and} \quad \sup_{q \in \mathsf{QHull}(X)} \sum_{j=1}^{\infty} \frac{q_j^2 \sigma^2}{q_j^2 + \sigma^2}.$$

The next proposition, parts of which are present in Donoho et al. [15, Section 6], connects the n-widths to the linear and nonlinear minimax risks, where for a vector $x \in \mathbf{R}^{\mathbf{N}}$, we let $|x_{(1)}| \ge |x_{(2)}| \ge \cdots$ denote the entries of x sorted by magnitude.

Proposition 5.1. *For any compact orthosymmetric convex set X,*

$$w^2(n) = \sup_{q \in \mathsf{QHull}(X)} \sum_{j \ge n+1} q_{(j)}^2 \quad \textit{and} \quad w_{\rm nl}^2(n) = \sup_{x \in X} \sum_{j \ge n+1} x_{(j)}^2.$$

Additionally, for any σ^2,

$$\sup_{x \in X} \sum_j x_j^2 \wedge \sigma^2 = \inf_n \left\{ w_{\rm nl}^2(n) + n\sigma^2 \right\},$$

and

$$\sup_{q \in \mathsf{QHull}(X)} \sum_j q_j^2 \wedge \sigma^2 = \inf_n \left\{ w^2(n) + n\sigma^2 \right\}.$$

Proof. For the characterizations, we use duality to see that

$$\begin{aligned} w^2(n) &= \inf_{\mathbf{0} \preceq d \preceq \mathbf{1}, \langle \mathbf{1}, d \rangle = n} \sup_{v \in \mathsf{SqHull}(X)} \langle \mathbf{1} - d, v \rangle \\ &= \sup_{v \in \mathsf{SqHull}(X)} \inf_{\mathbf{0} \preceq d \preceq \mathbf{1}, \langle \mathbf{1}, d \rangle = n} \langle \mathbf{1} - d, v \rangle = \sup_{v \in \mathsf{SqHull}(X)} \sum_{j \ge n+1} v_{(j)}, \end{aligned}$$

while it is immediate that $w_{\rm nl}^2(n) = \sup_{x \in X} \sum_{j \ge n+1} x_{(j)}^2$. Then we recognize that for any sorted nonnegative vector $a_1 \ge a_2 \ge \cdots$,

$$\inf_n \left\{ \sum_{j \ge n+1} a_j + n\sigma^2 \right\} = \inf_n \left\{ \sum_{j \ge n+1} a_j + \sum_{j=1}^{n} \sigma^2 \right\} = \sum_j a_j \wedge \sigma^2$$

by choosing any n such that $a_j \le \sigma^2$ for all $j \ge n$, while $a_j \ge \sigma^2$ for $j < n$. □

Using Proposition 5.1, we see that under the conditions of Corollary 2.7, because for any $a, b \geq 0$ we have $\frac{1}{2}\min\{a,b\} \leq \frac{ab}{a+b} \leq \min\{a,b\}$, the linear sequence model risk satisfies

$$\frac{1}{2}\inf_n \left\{w^2(n) + n\sigma^2\right\} \leq R^*_{\text{lin}}(X) \leq \inf_n \left\{w^2(n) + n\sigma^2\right\},$$

while the (nonlinear) minimax risk satisfies

$$\frac{1}{4}\inf_n \left\{w^2_{\text{nl}}(n) + n\sigma^2\right\} \leq R^*(X) \lesssim \log\frac{1}{\sigma^2} \cdot \inf_n \left\{w^2_{\text{nl}}(n) + n\sigma^2\right\}.$$

The linear and nonlinear n-widths of X therefore (up to a logarithmic factor in $\frac{1}{\sigma}$) determine the risk in sequence models, so that when they are similar the linear and nonlinear minimax risks coincide. In stochastic optimization, in contrast, Section 4 shows that convergence guarantees for methods of linear type coincide with those for arbitrary methods (again, up to logarithmic factors) if and only if

$$\sup_{x \in \mathsf{QHull}(X)} \|x\|_1 \asymp \sup_{x \in X} \|x\|_1.$$

The typical scenario in sequence models one considers is the risk as $\sigma \downarrow 0$, and the following essentially trivial observation shows that for regular enough sets X, when the linear and nonlinear n-widths differ, the rates at which R^*_{lin} and R^* converge to zero differ.

Observation 5.1. *Let $\alpha > 0$ be such that $n^\alpha w^2_{\text{nl}}(n) \asymp w^* > 0$ as $n \to \infty$, and assume for some $\beta > 0$ that $w^2(n) \geq n^\beta w^2_{\text{nl}}(n)$. Then*

$$\frac{R^*(X)}{R^*_{\text{lin}}(X)} \lesssim \log\frac{1}{\sigma^2} \cdot \sigma^{\frac{2\beta}{(1+\alpha)(1+\alpha-\beta)}} \to 0 \quad \textit{as} \quad \sigma \downarrow 0.$$

Proof. The assumption that $w^2_{\text{nl}}(n) \lesssim w^* n^{-\alpha}$ guarantees the assumptions of Corollary 2.7 apply, and we necessarily have $\beta \leq \alpha$ (as otherwise we would have $w^2_{\text{nl}}(n) \to \infty$). Thus there exists a (numerical) constant $C < \infty$ such that for all small enough $\sigma^2 > 0$,

$$R^*(X) \leq C \log\frac{1}{\sigma^2} \cdot \inf_n \left\{w^* n^{-\alpha} + n\sigma^2\right\} \asymp \log\frac{1}{\sigma^2} \cdot \sigma^{\frac{2\alpha}{1+\alpha}},$$

which follows by setting $n = \sigma^{-2/(1+\alpha)}$. In contrast, the linear risk satisfies

$$R^*_{\text{lin}}(X) \gtrsim \inf_n \left\{n^\beta w^2_{\text{nl}}(n) + n\sigma^2\right\} \asymp \sigma^{\frac{2(\alpha-\beta)}{1+\alpha-\beta}}$$

as $\sigma \downarrow 0$. Then observe that $\frac{\alpha}{1+\alpha} - \frac{\delta}{1+\delta} = \frac{\alpha-\delta}{1+\alpha+\delta+\alpha\delta}$, and set $\delta = \alpha - \beta$. □

The role of quadratic convexity of X differs between Gaussian sequence models and stochastic optimization problems, however, and the remainder of this section explores two extended examples highlighting this. We focus on sets $X \subset \mathbf{R}^{\mathbf{N}}$ in sequence space.

5.1 A constraint set requiring nonlinearity only in Gaussian sequence models

We show that for a large family of ℓ_1-bodies X, minimax (rate) optimal estimation requires nonlinearity for Gaussian sequence models but not in stochastic optimization. Take

$$X := \left\{x \in \mathbf{R}^{\mathbf{N}} \mid \sum_{j=1}^{\infty} a_j |x_j| \leq 1\right\} \text{ so } \mathsf{QHull}(X) = \left\{x \in \mathbf{R}^{\mathbf{N}} \mid \sum_{j=1}^{\infty} a_j^2 x_j^2 \leq 1\right\},$$

where (a_j) is a nondecreasing positive sequence where (w.l.o.g.) we take $a_1 = 1$. Computing the ℓ_1-diameters of X and its quadratic hull, we then have

$$\sup_{x \in X} \|x\|_1 = 1 \quad \text{and} \quad \sup_{x \in \mathsf{QHull}(X)} \|x\|_1 = \left(\sum_{j=1}^{\infty} a_j^{-2} \right)^{1/2} \tag{13}$$

by the Cauchy-Schwarz inequality. On the other hand, because the coefficients a_j are increasing, we can compute both the linear and nonlinear widths via

$$w^2(n) = \sup \left\{ \sum_{j=n+1}^{\infty} x_j^2 \mid \sum_{j=1}^{\infty} a_j^2 x_j^2 \le 1,\ x_1 \ge x_2 \ge \cdots \ge 0 \right\}$$

and

$$w_{\rm nl}^2(n) = \sup \left\{ \sum_{j=n+1}^{\infty} x_j^2 \mid \sum_{j=1}^{\infty} a_j x_j \le 1,\ x_1 \ge x_2 \ge \cdots \ge 0 \right\}.$$

By convexity (we maximize a convex function over a convex set in each case), for each width the maximizing point takes the form $(t, t, \ldots, t, 0, \ldots)$, with m repeated values t. Then we obtain

$$\begin{aligned} w^2(n) &= \sup_{m \ge n, t \ge 0} \left\{ (m-n)t^2 \mid t^2 \sum_{j=1}^{m} a_j^2 \le 1 \right\} = \sup_{m \ge n} \frac{m-n}{\sum_{j=1}^m a_j^2} \\ w_{\rm nl}^2(n) &= \sup_{m \ge n, t \ge 0} \left\{ (m-n)t^2 \mid t \sum_{j=1}^{m} a_j \le 1 \right\} = \sup_{m \ge n} \frac{m-n}{(\sum_{j=1}^m a_j)^2}. \end{aligned} \tag{14}$$

Comparing the ℓ_1-diameters (13) and the widths (14), we see that whenever $\sum_j a_j^{-2}$ is summable Euclidean gradient methods are minimax (rate) optimal for stochastic optimization. If $a_j = j^{\alpha/2}$ for some $\alpha > 1$, however, we have

$$\begin{aligned} w^2(n) &= \sup_{m \ge n} \frac{m-n}{\sum_{j=1}^m j^\alpha} \asymp \sup_{m \ge n} \frac{m-n}{m^{1+\alpha}} \asymp \frac{1}{n^\alpha} \quad \text{while} \\ w_{\rm nl}^2(n) &= \sup_{m \ge n} \frac{m-n}{(\sum_{j=1}^m j^{\alpha/2})^2} \asymp \sup_{m \ge n} \frac{m-n}{m^{2+\alpha}} \asymp \frac{1}{n^{1+\alpha}} \end{aligned}$$

because $\sum_{j=1}^m j^\beta \asymp \int_0^m t^\beta dt = \frac{1}{\beta+1} m^{1+\beta}$ for $\beta > 0$. Summarizing, we have the following corollary.

Corollary 5.1. *Let $X = \{x \mid \sum_{j=1}^\infty a_j |x_j| \le 1\}$, where $a_j = j^{\alpha/2}$. Then minimax rate optimal estimation for Gaussian sequence models requires that the estimator $\widehat{x}$ be nonlinear, while Euclidean gradient methods are minimax rate optimal for stochastic optimization.*

5.2 A constraint set requiring nonlinearity only in stochastic optimization

To contrast with Corollary 5.1, we can also give families of underlying constraint sets X for which only nonlinear methods can be rate-optimal for stochastic optimization problems, while linear estimators $\widehat{x} = Ay$ can be rate-optimal in Gaussian sequence models. At the grossest level, we construct sets X for which $\sup_{x \in X} \|x\|_1 \lesssim 1$ while $\sup_{x \in \mathsf{QHull}(X)} \|x\|_1 = +\infty$, while $w_{\rm nl}^2(n)$ and $w^2(n)$ are comparable. To give a slightly more nuanced picture, we consider the rates at which the two ℓ_1-diameters approach $+\infty$, comparing

$$\sup_{x \in X} \langle \mathbf{1}_n, x \rangle \quad \text{and} \quad \sup_{x \in \mathsf{QHull}(X)} \langle \mathbf{1}_n, x \rangle,$$

where $\mathbf{1}_n \in \mathbf{R}^{\mathbf{N}}$ denotes the sequence with 1 in its first n positions and 0 elsewhere.

Here, we elaborate on the ℓ_1 bodies yielding Corollary 5.1 to consider smaller axis-aligned polyhedra. Letting e_j be the basis vectors (i.e. sequences with a 1 in position j and 0 elsewhere), let a_j be a nondecreasing sequence with $a_1 = 1$ and b_j be arbitrary (for now), define the two sets

$$C_0 := \{\sigma_j a_j e_j\}_{j \in \mathbf{N}} \quad \text{and} \quad C_1 := \left\{\frac{1}{Z(n)} \sum_{j=1}^{n} \sigma_j b_j e_j\right\}_{n \in \mathbf{N}}, \quad \text{where } \sigma_j \in \{\pm 1\}.$$

We choose the normalizing constants $Z(n)$ so that the points in C_1 all lie in

$$\mathsf{QHull}(C_0) = \mathsf{QHull}\left\{x \mid \sum_{j=1}^{\infty} a_j |x_j| \le 1\right\} = \left\{x \mid \sum_{j=1}^{\infty} a_j^2 x_j^2 \le 1\right\},$$

i.e. $Z(n) = (\sum_{j=1}^{n} a_j^2 b_j^2)^{1/2}$ so that $Z(n)^{-2} \sum_{j=1}^{n} a_j^2 b_j^2 = 1$. Then the set

$$X := \mathrm{Conv}\,(C_0 \cup C_1) \quad \text{satisfies} \quad \mathsf{QHull}(X) = \mathsf{QHull}(C_0). \tag{15}$$

We obtain the following corollary of the our convergence guarantees in Section 4 and the technical lemmas we provide in Appendix G.

Corollary 5.2. *Let $a_j = j^{\alpha/2}$ for some $0 < \alpha < 1$ and $b_j^2 = 2^j$ in the construction of X above. Then*

$$n^{-\alpha} \le w_{\mathrm{nl}}^2(n) \le w_{\mathrm{nl}}^2(n) \lesssim n^{-\alpha},$$

while

$$\frac{\sup_{x \in \mathsf{QHull}(X)} \langle \mathbf{1}_n, x\rangle}{\sup_{x \in X} \langle \mathbf{1}_n, x\rangle} \gtrsim n^{\frac{1-\alpha}{2}}.$$

In particular, linear methods are rate optimal for estimation in Gaussian sequence models, while stochastic optimization over X requires nonlinear methods.

Summarizing the conclusions of the corollary, as $\alpha \downarrow 0$, the ratio of the ℓ_1-diameters of $\mathsf{QHull}(X)$ and X grows as $\sqrt{n}$, which by Cauchy-Schwarz is as large as possible: there is a large gap between the achievable performance by nonlinear methods, as Proposition 4.1 demonstrates, and linear methods, whose regret necessarily scales as the ℓ_1-diameter of $\mathsf{QHull}(X)$ (Theorem 5). Yet the nonlinear widths are comparable for all n, so linear methods are minimax rate optimal as $\sigma^2 \downarrow 0$.

6 The need for adaptive methods

We have so far demonstrated that diagonal re-scaling is sufficient to achieve minimax optimal rates for problems over quadratically convex constraint sets. In practice, however, it is often the case that we do not know the geometry of the problem in advance, precluding selection of the optimal linear pre-conditioner. To address this problem, adaptive gradient methods choose, at each step, a (usually diagonal) matrix Λ_i conditional on the subgradients observed thus far, $\{g_l\}_{l \le i}$. The algorithm then updates the iterate based on the distance generating function $h_i(x) := \frac{1}{2} x^\top \Lambda_i x$. In this section, we present a problem instance showing that when the "scale" of the subgradients varies across dimensions, adaptive gradient methods are crucial to achieve low regret. While there exists an optimal pre-conditioner, if we do not assume knowledge of the geometry in advance, AdaGrad [17] achieves the minimax optimal regret while standard (non-adaptive) subgradient methods can be $\sqrt{n}$ suboptimal on the same problem.

We consider the following setting: $X = \mathbf{B}_\infty(0,1)$ and $\gamma_\beta(g) = \|\beta \odot g\|_1$, for an arbitrary $\beta \in \mathbf{R}^n, \beta \succ 0$. Intuitively, β_j corresponds to the "scale" of the j-th dimension. On this problem, a straightforward optimization of the regret bound (9) guarantees that stochastic gradient methods achieve regret $\sqrt{nk}/\min_j \beta_j$. We exhibit a problem instance (in Appendix F.3) such that, for any stepsize α, online gradient descent attains this worst-case regret.

Theorem 6. *Let $\mathsf{Regret}_{k,\alpha}(x) = \sum_{i\le k} g_i^\top (x_i - x)$ denote the regret of the online gradient descent method with stepsize $\alpha \ge 0$ for linear functions $F_i(x) = g_i^\top x$. For any choice of $\alpha \ge 0$ and $\beta \succ 0$, there exists a sequence of vectors $\{g_i\}_{i\le k} \subset \mathbf{R}^n$, $\gamma_\beta(g_i) \le 1$ and point $x \in X$ such that*

$$\mathsf{Regret}_{k,\alpha}(x) \ge \frac{1}{2}\min\left\{\frac{nk}{2\|\beta\|_1}, \frac{\sqrt{2nk}}{\min_{j\le n}\beta_j}\right\}.$$

In contrast, AdaGrad [17] achieves regret $\sqrt{k}\,\|1/\beta\|_2$, demonstrating suboptimality gap as large as $\sqrt{n}$ for some choices of β. Indeed, let $\mathsf{Regret}_{k,\mathsf{AdaGrad}}(x)$ be the regret of AdaGrad. Then

$$\mathsf{Regret}_{k,\mathsf{AdaGrad}}(x) \le 2\sqrt{2}\sum_{j\le n}\sqrt{\sum_{i\le k} g_{i,j}^2}.$$

(see Duchi et al. [17, Corollary 6]), and by Cauchy-Schwarz,

$$\sum_{j\le n}\sqrt{\sum_{i\le k} g_{i,j}^2} = \sum_{j\le n}\frac{1}{\beta_j}\sqrt{\sum_{i\le k}\beta_j^2 g_{i,j}^2} \le \|1/\beta\|_2\sqrt{\sum_{i\le k}\|\beta\odot g_i\|_2^2} \le \sqrt{k}\,\|1/\beta\|_2\,.$$

To concretely consider different scales across dimensions, take $\beta_j = j$. Theorem 6 guarantees that there exists a collection of linear functions such that stochastic gradient methods suffer regret $\Omega(1)\sqrt{nk}$. Given that $\|1/\beta\|_2 \le \sqrt{\zeta(2)} \le \pi/\sqrt{6}$, AdaGrad achieves regret $O(1)\sqrt{k}$—amounting to a suboptimality gap of order $\sqrt{n}$—exhibiting the need for adaptivity. This $\sqrt{n}$ gap is also the largest possible over subgradient methods, which achieve regret $\sqrt{n\sum_{i\le k}\|g_i\|_2^2} \le \sqrt{n}\sum_{j\le n}\sqrt{\sum_{i\le k} g_{i,j}^2}$ for $X = \mathbf{B}_\infty(0,1)$. Finally, we note in passing that AdaGrad is minimax optimal on this class of problems via a straightforward application of Theorem 1.

7 Discussion

We provide concrete foundations to compare adaptive, mirror, or standard gradient methods, showing how problem geometry necessarily impacts convergence. This paper puts a particular computational spin on optimization by connecting to Gaussian sequence models and linear versus nonlinear updates, which we advocate for its ability to paint a different picture than pure polynomial versus non-polynomial computational complexity. This perspective draws from information-based models in optimization [25, 1] and models in scientific computing where one uses certain families of operations—e.g., matrix vector multiplies—to build up optimal algorithms under these constraints [35, 3]. We hope to see more exploration in these directions. While Section 6 emphasizes the importance of adaptivity, the picture is not fully complete: for example, in the case of quadratically convex constraint sets, while the best diagonal pre-conditioner achieves optimal rates, the extent to which adaptive gradient algorithms find this optimal pre-conditioner remains an open question. Another avenue to explore involves the many flavors of adaptivity—while the minimax framework assumes knowledge of the problem setting (e.g. a bound on the domain or the gradient norms), it is often the case that such parameters are unknown to the practitioner. To what extent can adaptivity mitigate this and achieve optimal rates, and is minimax (i.e. worst-case) optimality truly the right measure of performance?

Acknowledgments

We thank Aditya Grover, Annie Marsden and Hongseok Namkoong for valuable comments as well as Quentin Guignard for pointing us to the Banach-Mazur distance for Theorem 4.

References

[1] 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.

[2] P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences, Série I*, 296(23):1021–1024, 1983.

[3] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in numerical linear algebra. *SIAM Journal on Matrix Analysis and Applications*, 32(3):866–901, 2011.

[4] P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, 2007.

[5] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[6] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(1):1780–1815, 2013.

[7] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale learning. *SIAM Review*, 60(2):223–311, 2018.

[8] 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.

[9] T. T. Cai and Y. Wu. Statistical and computational limits for sparse matrix detection. *Annals of Statistics*, 48(3):1593–1614, 2020.

[10] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

[11] 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.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition.* Wiley, 2006.

[13] 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.

[14] A. Cutkosky and T. Sarlos. Matrix-free preconditioning in online learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[15] D. L. Donoho, R. C. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, 18(3):1416–1437, 1990.

[16] J. C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

[17] 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.

[18] 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.

[19] K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

[20] C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

[21] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[22] I. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. 2017. Available online at https://imjohnstone.su.domains/GE_08_09_17.pdf.

[23] A. Lewis. Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, 6: 164–177, 1996.

[24] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way (Third Edition)*. Academic Press, 2008.

[25] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[26] 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.

[27] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):261–283, 2009.

[28] F. Orabona and K. Crammer. New adaptive algorithms for online classification. In *Advances in Neural Information Processing Systems 23*, 2010.

[29] F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.

[30] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[31] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

[32] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[33] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

[34] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[35] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra.* SIAM, 1997.

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

[37] R. Vershynin. Lectures in geometric functional analysis. Unpublished manuscript, 2009. URL https://www.math.uci.edu/~rvershyn/papers/GFA-book.pdf.

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

[39] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems 30*, 2017.

[40] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

A The Assouad and Fano Methods for Minimax Lower Bounds

In this precursor to the appendix, we review the Le Cam, Fano and Assouad methods [2, 40, 1, 38] for proving lower bounds for stochastic optimization. Each reduces estimation to testing then uses information theoretic tools to bound the probability of error in various hypothesis tests.

A.1 Le Cam and Fano Methods

We start with a lemma that provides the standard reduction from estimation to testing that we extensively use in our proofs. Duchi [16, Ch. 5] essentially contains this result; we provide the proof for completeness.

Lemma A.1 (From estimation to testing)**.** *Let $\mathcal{P}$ be a collection of distributions over $\mathcal{S}$ and $L : X \times \mathcal{P} \to \mathbf{R}_+$ satisfy*

$$\inf_{x \in X} L(x, P) = 0 \quad \textit{for } P \in \mathcal{P}.$$

For distributions $P, Q \in \mathcal{P}$, define the separation

$$\mathsf{sep}_L(P, Q; X) := \sup \left\{ \delta \geq 0 \;\middle|\; \textit{for all } x \in X, \begin{array}{l} L(x, P) \leq \delta \textit{ implies } L(x, Q) \geq \delta \\ L(x, Q) \leq \delta \textit{ implies } L(x, P) \geq \delta \end{array} \right\}.$$

Let $\delta > 0$ and $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ be a family of distributions indexed by a finite set $\mathcal{V}$ satisfying the separation condition $\mathsf{sep}_L(P_v, P_{v'}; X) \geq \delta$ for $v \neq v' \in \mathcal{V}$. Then for $S_1^k \overset{\text{iid}}{\sim} P$,

$$\inf_{\widehat{x}} \sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{x}(S_1^k), P) \geq \delta \inf_{\psi} \mathbb{P}(\psi(S_1^k) \neq V),$$

where $\mathbb{P}$ is the joint distribution over the random index V chosen uniformly in $\mathcal{V}$ and $S_1^k \overset{\text{iid}}{\sim} P_v$ conditional on $V = v$.

Proof. Let $V \sim \mathsf{Uniform}(\mathcal{V})$ and $S_1^k \mid (V = v) \overset{\text{iid}}{\sim} P_v$. Then for any estimator $\widehat{x}$, we have

$$\sup_{P \in \mathcal{P}} \mathbf{E}_P L(\widehat{x}(S_1^k), P) \geq \frac{1}{|\mathcal{V}|} \sum_v \mathbf{E}_{P_v} L(\widehat{x}, P_v) \geq \delta \frac{1}{|\mathcal{V}|} \sum_v P_v(L(\widehat{x}, P_v) \geq \delta) = \delta \mathbb{P}(L(\widehat{x}(S_1^k), P_V) \geq \delta),$$

where $\mathbb{P}$ denotes the joint distribution of S_1^k and V. Define the test $\psi(s_1^k) := \mathrm{argmin}_{v\in\mathcal{V}} L(\widehat{x}(s_1^k), P_v)$. The separation assumption guarantees that if $\psi(x) \neq v$ then $L(x, P_v) \geq \delta$, so

$$\mathbb{P}(L(\widehat{x}(S_1^k), P_V) \geq \delta) \geq \mathbb{P}\left(\psi(S_1^k) \neq V\right).$$

Taking the infimum over all tests ψ yields the result. □

With this, the classical Le Cam and Fano methods are straightforward combinations of Lemma A.2 with (respectively) Le Cam's lemma [40, Lemma 1] and Fano's inequality [12, Theorem 2.10.1].

Proposition A.1 (Le Cam's method). *Let P_0 and P_1 be two distributions of $\mathcal{P}$ over $\mathcal{S}$. Let $\delta > 0$ be such that $\mathsf{sep}_L(P_0, P_1, X) \geq \delta$. Then*

$$\inf_{\widehat{x}} \sup_{P\in\mathcal{P}} \mathbf{E}_P L(\widehat{x}(S_1^k), P) \geq \frac{\delta}{2}(1 - \left\|P_0^k - P_1^k\right\|_{\mathsf{tv}}).$$

Proposition A.2 (Fano's method). *Let $\mathcal{V}$ be a finite index set and $\{P_v\}_{v\in\mathcal{V}}$ a collection of distributions contained by $\mathcal{P}$ such that $\min_{v\neq v'} \mathsf{sep}_L(P_v, P_{v'}, X) \geq \delta$, then*

$$\inf_{\widehat{x}} \sup_{P\in\mathcal{P}} \mathbf{E}_P L(\widehat{x}(S_1^k), P) \geq \delta\left(1 - \frac{\mathsf{I}(S_1^k; V) + \log 2}{\log|\mathcal{V}|}\right).$$

With these tools, minimax lower bounds on the stochastic risk $\mathfrak{M}_k^{\mathsf{S}}$ in Section 2 follow by (i) demonstrating an appropriate loss L and (ii) separation. The next lemma, essentially present in the paper [1] (cf. [16]), reduces optimization to testing by providing an appropriate separation function.

Lemma A.2 (From optimization to function estimation). *Let $\mathcal{S}$ be a sample space, $X \subset \mathbf{R}^n$, $\mathcal{F}$ be a collection a functions $\mathbf{R}^n \times \mathcal{S} \to \mathbf{R}$, and $\mathcal{P}$ be a collection of distributions over $\mathcal{S}$. Let $\mathcal{V}$ index $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$. For $F \in \mathcal{F}$, define $f_v(x) := \mathbf{E}_{P_v}[F(x, S)]$ and for each $v, v' \in \mathcal{V}$, set*

$$\mathsf{d}_{\mathrm{opt}}(v, v', X) := \inf_{x\in X}\left\{f_v(x) + f_{v'}(x) - \inf_{x\in X} f_v(x) - \inf_{x\in X} f_{v'}(x)\right\}.$$

If $\mathsf{d}_{\mathrm{opt}}(v, v', X) \geq \delta \geq 0$ for all $v \neq v' \in \mathcal{V}$, then

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}) \geq \mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}, \mathcal{P}) \geq \frac{\delta}{2} \inf_{\psi} \mathbb{P}(\psi(S_1^k) \neq V).$$

Proof. We construct an appropriate loss L and apply Lemma A.1. Define $L(x, P) := f_P(x) - \inf_{x\in X} f_P(x)$. By construction, $L(x, P) \geq 0$ and $\inf_{x\in X} L(x, P) = 0$ for all $x \in X$ and $P \in \mathcal{P}$. Let $v \neq v' \in \mathcal{V}$. Then if $L(x, P_v) = f_v(x) - \inf_{x\in X} f_v(x) \leq \frac{1}{2}\mathsf{d}_{\mathrm{opt}}(v, v', X)$, it is evidently the case that $f_{v'}(x) - \inf_{x\in X} f_{v'}(x) \geq \frac{1}{2}\mathsf{d}_{\mathrm{opt}}(v, v', X)$, so that $\mathsf{sep}_L(P_v, P_{v'}, X) \geq \frac{1}{2}\mathsf{d}_{\mathrm{opt}}(v, v', X)$. The distributions $\{P_v\}_{v\in\mathcal{V}}$ are $\delta/2$-separated, allowing application of Lemma A.1. □

Our general strategy for proving lower bounds on $\mathfrak{M}_k^{\mathsf{S}}$ is as follows:

- Choose a function $F \in \mathcal{F}$ and define $\mathcal{V}$ and $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$ such that $\mathsf{d}_{\mathrm{opt}}(v, v', X) \geq \delta > 0$.
- Lower bound the testing error $\inf_\psi \mathbb{P}(\psi(S_1^k) \neq V)$, and choose the largest separation δ to make this testing error a positive constant.

To showcase this proof technique, we prove that minimax stochastic risk for 1-dimensional optimization has lower bound $1/\sqrt{k}$; we use this to address technicalities in later proofs.

Lemma A.3. *Let $\mathcal{F}^{n=1} = \{f : \mathbf{R} \times \mathcal{S} \to \mathbf{R} \mid f(\cdot, s)$ is convex and 1-Lipschitz$\}$. Then*

$$\mathfrak{M}_k^{\mathsf{S}}([-1,1], \mathcal{F}^{n=1}) \geq \frac{1}{4\sqrt{6k}}.$$

Proof. Let $X = [-1,1]$ and $\mathcal{S} = \{\pm 1\}, \mathcal{V} = \{\pm 1\}$.

To see the separation condition, let $F(x,s) := |x - s|$. For $\delta \in [0, \frac{1}{2}]$, we define P_v s.t. if $S \sim P_v$ we have

$$S = \begin{cases} 1 \ \text{ with probability } \ \frac{1+v\delta}{2} \\ -1 \ \text{ with probability } \ \frac{1-v\delta}{2}. \end{cases}$$

We have $f_v(x) = \frac{1+\delta}{2}|x - v| + \frac{1-\delta}{2}|x + v|$ and $\inf_x f_v(x) = \frac{1-\delta}{2}$. To lower bound the separation, note that

$$f_1(x) + f_{-1}(x) - \inf_X f_1 - \inf_X f_{-1} = |x-1| + |x+1| - (1-\delta) \geq \delta.$$

This yields $\mathsf{d}_{\text{opt}}(1, -1, X) \geq \delta$.

We lower bound the testing error via Proposition A.1:

$$\inf_{\psi : \mathcal{S}^k \to \{\pm 1\}} \mathbb{P}(\psi(S_1^k) \neq V) = \frac{1}{2}\left(1 - \left\| P_1^k - P_{-1}^k \right\|_{\mathsf{tv}}\right) \geq \frac{1}{2}\left(1 - \sqrt{\frac{k}{2} D_{\text{kl}}\left(P_1 \| P_{-1}\right)}\right),$$

where the rightmost inequality is Pinsker's inequality. Noting that $D_{\text{kl}}\left(P_1 \| P_{-1}\right) = \delta \log \frac{1+\delta}{1-\delta} \leq 3\delta^2$ for $\delta \in [0, \frac{1}{2}]$ and setting $\delta = 1/\sqrt{6k}$ yields the result. □

A.2 The Assouad Method

Assouad's method reduces the problem of estimation (or optimization) to one of multiple binary hypothesis tests. In this case, we index a set of distributions $\mathcal{P} = \{P_v\}_{v \in \mathcal{V}}$ on a set $\mathcal{S}$ by the hypercube $\mathcal{V} = \{\pm 1\}^n$. For a function $F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R}$, we define $f_v(x) := \mathbf{E}_{P_v}[F(x, S)]$. Then for a vector $\delta \in \mathbf{R}_+^n$, following Duchi [16, Lemma 5.3.2], we say that the functions $\{f_v\}$ induce a δ-separation in Hamming metric if

$$f_v(x) - \inf_{x \in X} f_v(x) \geq \sum_{j=1}^n \delta_j 1(\text{sign}(x_j) \neq v_j)\,. \tag{16}$$

With this condition, we have the following generalized Assouad method [16, Lemma 5.3.2].

Lemma A.4 (Generalized Assouad's method)**.** *Let $S_1^k \overset{\text{iid}}{\sim} P_V$, where $V \sim \mathsf{Uniform}(\{\pm 1\}^n)$. Define the averages*

$$\mathbb{P}_{+j} := \frac{1}{2^{n-1}} \sum_{v : v_j = 1} P_v^k \quad \textit{and} \quad \mathbb{P}_{-j} := \frac{1}{2^{n-1}} \sum_{v : v_j = -1} P_v^k.$$

Assume that the collection $\{f_v\}$ for $f_v = \mathbf{E}_{P_v}[F(\cdot, S)]$ induces a δ-separation (16). Then letting $\mathcal{F} = \{F\}$, the single function F,

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}, \mathcal{P}) \geq \frac{1}{2} \sum_{j=1}^n \delta_j (1 - \|\mathbb{P}_{+j} - \mathbb{P}_{-j}\|_{\mathsf{tv}}).$$

B Duality results and minimax linear estimators

In this section, we prove many of the optimality results for the Gaussian sequence model in Section 2.1 that we use. The first result we require is a specialization of Sion's minimax theorem [34], which is originally due to Fan [19].

Lemma B.1. *Let X and Y be compact convex subsets of (possibly distinct) topological vector spaces and $L : X \times Y \to \mathbf{R}$ be convex in its first argument, concave in its second, and continuous. Then*

$$\inf_{x \in X} \sup_{y \in Y} L(x, y) = \sup_{y \in Y} \inf_{x \in X} L(x, y)$$

and the infimum and supremum are attained.

For a vector space X with norm $\|\cdot\|$ and function $f : X \to \mathbf{R}$, the Lipschitz norm on f is $\|f\|_{\text{Lip}} = \sup_{x \neq y \in X} |f(x) - f(y)| / \|x - y\|$. The Wasserstein distance between measures ν, μ on the space X is then

$$W(\mu, \nu) = \sup_{\|f\|_{\text{Lip}} \le 1} \int f(d\mu - d\nu).$$

The Wasserstein distance metrizes convergence in distribution $\mu_n \overset{d}{\to} \mu$ if the measures μ_n, μ have bounded first moment and turns the space of measures on X into a (topological) vector space.

B.1 Proof of Proposition 2.1

Recalling the form (4) of the risk $R(A, x) = x^T(A - I)^T(A - I)x + \sigma^2 \|A\|_{\text{Fr}}^2$, it is immediate that A is unique, as $\|A\|_{\text{Fr}}^2$ is strongly convex in A and $\sup_{x \in X} \|(A - I)x\|_2^2$ is convex in A. As X is compact, the risk $R^*(A, X)$ is finite for all A, and hence its minimum is attained.

We now consider the saddle point problem

$$\inf_A \sup_{x \in X} R(A, x) = \inf_A \sup_{x \in X} \left\{ x^T A^T A x - 2x^T A x + \sigma^2 \|A\|_{\text{Fr}}^2 \right\}.$$

By lifting to the space $\mathcal{P}(X)$ of probability measures defined on X, we have for each A that

$$R^*(A, X) = \sup_{\nu \in \mathcal{P}(X)} \int \text{tr}((A - I)^T(A - I)xx^T) d\nu(x) + \sigma^2 \|A\|_{\text{Fr}}^2 .$$

We shall apply Fan's minimax theorem (Lemma B.1) to construct the A minimizing this worst-case risk. In our case, we define

$$L(A, \nu) := \int \text{tr}((A - I)^T(A - I)xx^T) d\nu(x) + \sigma^2 \|A\|_{\text{Fr}}^2 ,$$

which is clearly convex/concave and continuous in A. To see the continuity of L in ν, we note that for any matrix B and any $x, y \in X$ we have

$$\langle xx^T - yy^T, B \rangle = (x - y)^T(B + B^T)(x + y)/2 \le 2 \left\| B \right\|_{\text{op}} \|x - y\|_2 \, \text{diam}(X),$$

so that $x \mapsto \langle xx^T, B \rangle$ is Lipschitz and

$$|L(A, \nu) - L(A, \mu)| \le 2 \left\| A - I \right\|_{\text{op}}^2 \text{diam}(X) \cdot W(\nu, \mu).$$

Thus L is continuous in ν for for the Wasserstein distance (or topology of convergence in distribution), and $\mathcal{P}(X)$ is compact for this topology by Prokhorov's theorem. In particular, if we use the shorthand $X_\nu = \int xx^T d\nu(x)$, there exist A, ν such that

$$\inf_B L(B,\nu) = L(A,\nu) = \mathrm{tr}((A-I)^T(A-I)X_\nu) + \sigma^2 \|A\|_{\mathrm{Fr}}^2 = \sup_{\mu\in\mathcal{P}(X)} L(A,\mu). \tag{17}$$

We now construct the A solving the saddle point problem (17), that is, given ν, we show the (unique) A minimizing $L(A,\nu)$. Taking derivatives of $L(A,\nu)$ and recalling the shorthand $X_\nu = \int xx^T d\nu(x)$, we see that A must satisfy

$$AX_\nu + X_\nu A^T - 2X_\nu + \sigma^2(A + A^T) = 0.$$

If X_ν has spectral decomposition $X_\nu = U\Lambda U^T$ we let $A = UDU^T$ for a diagonal matrix D to be determined, and it is evidently enough to solve

$$2D\Lambda - 2\Lambda + 2\sigma^2 D = 0, \quad \text{or} \quad D = (\Lambda + \sigma^2 I)^{-1}\Lambda.$$

In particular, the choice $A = (X_\nu + \sigma^2 I)^{-1/2} X_\nu (X_\nu + \sigma^2 I)^{-1/2}$ is optimal; it is also unique for the given ν as $A \mapsto L(A,\nu)$ is strongly convex in A.

Finally, we show that without loss of generality, we may take A to be diagonal. If S is a diagonal matrix of independent random signs, then $\mathbf{E}_\nu[(Sx)(Sx)^T] = \mathrm{diag}(X_\nu) = \mathbf{E}_\nu[\mathrm{diag}(x)^2]$. Let $\overline{\nu}$ be the measure on X induced by drawing $x \sim \nu$ and then multiplying x by the random signs Sx. Notably, we have $\mathrm{tr}(X_\nu) = \mathrm{tr}(X_{\overline{\nu}})$ and $\mathrm{tr}(DX_\nu) = \mathrm{tr}(DX_{\overline{\nu}})$ for any diagonal matrix D, as $\mathrm{diag}(X_\nu) = \mathrm{diag}(X_{\overline{\nu}})$. Suppose for the sake of contradiction that $L(\mathrm{diag}(A),\nu) > L(A,\nu)$. In this case, A must be non-diagonal, and so we have

$$\begin{aligned} L(\mathrm{diag}(A),\nu) &= \mathrm{tr}(\mathrm{diag}(A)^2 X_\nu) - 2\mathrm{tr}(\mathrm{diag}(A)X_\nu) + \mathrm{tr}(X_\nu) + \sigma^2 \|\mathrm{diag}(A)\|_{\mathrm{Fr}}^2 \\ &= \mathrm{tr}(\mathrm{diag}(A)^2 X_{\overline{\nu}}) - 2\mathrm{tr}(\mathrm{diag}(A)X_{\overline{\nu}}) + \mathrm{tr}(X_{\overline{\nu}}) + \sigma^2 \|\mathrm{diag}(A)\|_{\mathrm{Fr}}^2 \\ &\overset{(\star)}{<} \mathrm{tr}(A^2 X_{\overline{\nu}}) - 2\mathrm{tr}(\mathrm{diag}(A)X_{\overline{\nu}}) + \mathrm{tr}(X_{\overline{\nu}}) + \sigma^2 \|A\|_{\mathrm{Fr}}^2, \end{aligned}$$

where inequality $(\star)$ follows because $\|A\|_{\mathrm{Fr}} > \|\mathrm{diag}(A)\|_{\mathrm{Fr}}$ while $\mathrm{diag}(A)^2 \preceq \mathrm{diag}(A^2)$. Finally, noting that $\mathrm{tr}(\mathrm{diag}(A)X_{\overline{\nu}}) = \mathrm{tr}(AX_{\overline{\nu}})$, we see that $L(\mathrm{diag}(A),\nu) < L(A,\overline{\nu})$, and so we have demonstrated that $L(A,\nu) < L(A,\overline{\nu})$. But this contradicts the assumed maximality of ν, and so it must be the case that A is diagonal.

Now that we have A diagonal, the claimed equality is immediate, and we also notice that $L(D,\nu) = L(D,\overline{\nu})$ for any diagonal D.

B.2 Proof of Corollary 2.3

For any linear operator $A : \mathbf{R}^{\mathbf{N}} \to \mathbf{R}^{\mathbf{N}}$, we may write

$$(Ay)_j = a_j(y) = a_j(x+\xi) = a_j(x) + a_j(\xi)$$

for each j. Let $\Pi_n : \mathbf{R}^{\mathbf{N}} \to \mathbf{R}^n$ be the projection onto the first n coordinates of a vector and $Z_n : \mathbf{R}^{\mathbf{N}} \to \mathbf{R}^{\mathbf{N}}$ be the projection zeroing the first n elements. Then

$$\inf_A \sup_{x\in X} \mathbf{E}_x\left[\|Ay - x\|_2^2\right] \ge \inf_A \sup_{x\in \Pi_n X} \mathbf{E}_x[\|Ay - x\|_2^2] = \inf_A \sup_{x\in\Pi_n X} \sum_{j=1}^n \mathbf{E}_x[(a_j((x,\mathbf{0}) + \xi) - x_j)^2],$$

where $(x, \mathbf{0}) \in \Pi_n X \times \mathbf{R}^{\mathbf{N}}$. Then by linearity, for $x \in \mathbf{R}^n$ we write

$$a_j((x, \mathbf{0}) + \xi) = \varphi_j(x) + a_j(Z_n \xi) + a_j((I - Z_n)\xi)$$

where $\varphi_j : \mathbf{R}^n \to \mathbf{R}$ is the linear function $\varphi_j(x) = a_j((x, \mathbf{0}))$. But $Z_n \xi$ and $(I - Z_n)\xi = (\xi_1, \ldots, \xi_n, \mathbf{0})$ are independent, and because $\mathbf{E}[a_j(Z_n \xi)] = 0$ we obtain

$$\begin{aligned}\mathbf{E}_x[(a_j((x, \mathbf{0}) + \xi) - x_j)^2] &= \mathbf{E}[(\varphi_j(x + [\xi_j]_{j \le n}) - x_j + a_j(Z_n \xi))^2] \\ &= \mathbf{E}[(\varphi_j(x + [\xi_j]_{j \le n}) - x_j)^2] + \mathrm{Var}(a_j(Z_n \xi)).\end{aligned}$$

The optimal choice of a_j then necessarily satisfies $a_j(Z_n u) = 0$ for all u. Thus, by restricting to finite dimensions, we use Proposition 2.1 to see that for any n,

$$\begin{aligned}\inf_A \sup_{x \in X} \mathbf{E}_x \left[\|Ay - x\|_2^2\right] &\ge \inf_{d \in \mathbf{R}^n} \sup_{x \in X} \sum_{j=1}^n \left((d_j - 1)^2 x_j^2 + \sigma^2 d_j^2\right) \\ &= \sup_{v \in \mathsf{SqHull}(X)} \inf_d \sum_{j=1}^n \left((d_j - 1)^2 v_j + \sigma^2 d_j^2\right) = \sup_{v \in \mathsf{SqHull}(X)} \sum_{j=1}^n \frac{\sigma^2 v_j}{v_j + \sigma^2}\end{aligned}$$

as in the proof of Corollary 2.2.

By compactness, for each $\epsilon > 0$, we can choose $N < \infty$ such that $\sup_{x \in X} \sum_{j > N} x_j^2 < \epsilon$. We thus have upper bound

$$\inf_A R^*(A, X) \le \inf_A \sup_{x \in \Pi_N X} \mathbf{E}_x[\|Ay - x\|_2^2] + \epsilon.$$

Apply the same proof technique as that in Corollary 2.2 to obtain

$$\inf_A R^*(A, \Pi_N X) = \sup_{v \in \mathsf{SqHull}(\Pi_N X)} \sum_{j=1}^N \frac{\sigma^2 v_j}{v_j + \sigma^2} = \sup_{v \in \mathsf{SqHull}(X)} \sum_{j=1}^N \frac{\sigma^2 v_j}{v_j + \sigma^2}.$$

Now use that $\epsilon > 0$ was arbitrary, X is compact, and take n and N to infinity. □

B.3 Proof of Proposition 2.2

We apply Le Cam's two point method via the reduction from estimation to testing that Lemma A.1 implies. Consider for any fixed $b > 0$ the problem of estimating a single value $x \in [-b, b]$ with $y \sim \mathsf{N}(x, \sigma^2)$. Then for $v \in \{-1, 1\}$, letting P_v be the normal distribution $\mathsf{N}(\delta v, \sigma^2)$ for some $\delta \in [0, b]$ to be chosen, we have

$$\inf_{\widehat{x}} \sup_{x \in [-b, b]} \mathbf{E}_x[(\widehat{x} - x)^2] \ge \inf_{\widehat{x}} \frac{1}{2} \left\{\mathbf{E}_{P_1}[(\widehat{x} - x)^2] + \mathbf{E}_{P_{-1}}[(\widehat{x} - x)^2]\right\} \ge \frac{\delta^2}{2} \left(1 - \|P_{-1} - P_1\|_{\mathsf{tv}}\right).$$

Recalling the Hellinger distance $d_{\text{hel}}^2(P, Q) = 1 - \int \sqrt{dPdQ}$ between probabilities and the standard relationship $\|P_{-1} - P_1\|_{\mathsf{tv}} \le d_{\text{hel}}(P_1, P_{-1}) \sqrt{2 - d_{\text{hel}}^2(P_1, P_{-1})}$, we note that $d_{\text{hel}}^2(\mathsf{N}(\mu_0, \sigma^2), \mathsf{N}(\mu_1, \sigma^2)) = 1 - \exp(-\frac{1}{8\sigma^2}(\mu_0 - \mu_1)^2)$ to obtain

$$\inf_{\widehat{x}} \sup_{x \in [-b, b]} \mathbf{E}_x \left[(\widehat{x} - x)^2\right] \ge \sup_{0 \le \delta \le b} \frac{\delta^2}{2} \left(1 - \sqrt{1 - \exp(-\delta^2/\sigma^2)}\right) \ge \frac{\sigma^2 \wedge b^2}{10}$$

via the choice $\delta = \min\{\sigma, b\}$. We thus obtain for any hypercube $H = [-x_j, x_j]_{j \ge 1}$ that

$$R^*(X) \ge R^*(H) \ge \frac{1}{10} \sum_{j \ge 1} \sigma^2 \wedge x_j^2.$$

C Proofs for Section 3.1

C.1 Proof of Proposition 3.1

We use the general information-theoretic framework of reduction from estimation to testing presented in Section A.1 to prove the lower bound.

Separation Let us consider the sample space $\mathcal{S} = \{\pm e_j\}_{j \le n}$ and the function $F(x, s) := x^\top s$, so F belongs to $\mathcal{F}^{\gamma,1}$. Letting $\delta \in [0, 1/2]$ to be determined, for $v \in \{\pm 1\}^n$, we define P_v such that for $S \sim P_v$ we have

$$S = \begin{cases} v_j e_j & \text{with probability } \frac{1+\delta}{2n} \\ -v_j e_j & \text{with probability } \frac{1-\delta}{2n}. \end{cases}$$

We then have $f_v(x) = \frac{\delta}{n} x^\top v$. By duality,

$$f_v^* := \inf_X f_v = -\frac{\delta}{n} \sup_{x \in \mathbf{B}_p(0,1)} v^\top x = -\frac{\delta}{n} \|v\|_{p^*},$$

where p^* is such that $1/p + 1/p^* = 1$. For $v, v' \in \{\pm 1\}^n$, we thus have

$$\begin{aligned} \mathsf{d}_{\mathrm{opt}}(v, v', X) = \inf_{x \in X} f_v(x) + f_{v'}(x) - f_v^* - f_{v'}^* &= \inf_{x \in \mathbf{B}_p(0,1)} \frac{\delta}{n}(x^\top(v + v') + \|v\|_{p^*} + \|v'\|_{p^*}) \\ &= \frac{\delta}{n}(\|v\|_{p^*} + \|v'\|_{p^*} - \|v + v'\|_{p^*}) \\ &= 2\frac{\delta}{n}\left[n^{1/p^*} - (n - \mathrm{d}_{\mathrm{Ham}}(v, v'))^{1/p^*}\right], \end{aligned}$$

where $\mathrm{d}_{\mathrm{Ham}}(v, v')$ is the Hamming distance between v and v'. The Gilbert-Varshimov bound [38, Example 5.3] guarantees the existence of an $n/2$ ℓ_1-packing of $\{\pm 1\}^n$ of size at least $\exp(n/8)$. Let $\mathcal{V}$ be such a packing; we have for a numerical constant $c_0 > 0$ that

$$\mathsf{d}_{\mathrm{opt}}(v, v', X) \ge c_0 \delta n^{-1/p} \quad \text{for all } v \neq v' \in \mathcal{V}. \tag{18}$$

Applying Lemma A.2 yields

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \frac{c_0}{2} \delta n^{-1/p} \inf_\psi \mathbb{P}(\psi(S_1^k) \neq V).$$

Bounding the testing error We bound the testing error with Fano's inequality and upper bounding the mutual information $\mathsf{I}(S; V)$. Using the identity $\delta \log \frac{1+\delta}{1-\delta} \le 3\delta^2$, it holds

$$\mathsf{I}(S_1^k; V) \le n \max_{v, v'} D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) \le 3n\delta^2,$$

and, recalling that $\log|\mathcal{V}| \ge n/8$ yields

$$\inf_\psi \mathbb{P}(\psi(S_1^k) \neq V) \ge \left(1 - \frac{3k\delta^2 + \log 2}{n/8}\right).$$

In the case that $n \ge 32 \log 2$, choosing $\delta = \sqrt{\frac{n}{48k}}$ yields the desired lower-bound. In the case that $n < 32 \log 2$, with $\mathcal{F}^{n=1}$ as in Lemma A.3, that any 1-dimensional optimization problem may be embedded into a n-dimensional problem yields

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \mathfrak{M}_k^{\mathsf{S}}([-1, 1], \mathcal{F}^{n=1}) \gtrsim \frac{1}{\sqrt{k}}.$$

This gives the lower bound for all $n \in \mathbf{N}$.

To conclude the proof, we establish an upper bound on the minimax regret. We consider the regret guarantee of (9) for $h(x) = \frac{1}{2}\|x\|_2^2$. Since $p \geq 2$, it holds that for all $x \in \mathbf{R}^n$, we have $\|x\|_2 \leq n^{\frac{1}{2}-\frac{1}{p}}\|x\|_p$ and thus $\sup_{x,x' \in X} \mathrm{D}_h(x, x') \leq n^{\frac{1}{2}-\frac{1}{p}}$. On the other hand, since $r \in [1, 2]$, $\|g\|_2 \leq \|g\|_r \leq 1$. A straightforward optimization of the stepsize α yields the upper bound on $\mathfrak{M}_k^{\mathsf{R}}(X, \gamma)$. □

C.2 Proof of Proposition 3.2

The proof is similar to Proposition 3.1 so we forego some of the details.

Separation Let $\eta > 0$ to be determined, and consider the sample space and objectives $\mathcal{S} = \{\pm 1\}^n$ and $F(x, s) := \eta x^\top s$. For $v \in \{\pm 1\}^n$, we define P_v such that $S \sim P_v$ has coordinates satisfying

$$S_j = \begin{cases} v_j & \text{with probability } \frac{1+\delta}{2} \\ -v_j & \text{with probability } \frac{1-\delta}{2}. \end{cases}$$

This yields $f_v(x) = \eta\delta x^\top v$. Considering again the Gilbert-Varshimov packing $\mathcal{V} \subset \{\pm 1\}^n$, we lower bound the separation by noting that for all $v \neq v' \in \mathcal{V}$,

$$\mathsf{d}_{\mathrm{opt}}(v, v', X) = \inf_{x \in X} f_v(x) + f_{v'}(x) - f_v^* - f_{v'}^* \geq c_0 \eta \delta n^{1/p^*}.$$

Bounding the testing error Noting that

$$D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) = \sum_{j \leq n} \mathbf{1}_{v_j = v'_j} \delta \log \frac{1+\delta}{1-\delta} \leq 3n\delta^2,$$

and so $\mathsf{I}(S_1^k; V) \leq 3kn\delta^2$. For F to remain in $\mathcal{F}^{\gamma,1}$, we must have that for all $x \in \mathcal{S}, \eta\|x\|_r \leq 1$; noting that $\|x\|_r = n^{1/q}$, we choose $\eta = n^{-1/q}$. In the case that $n \geq 32 \log 2$, taking $\delta = 1/\sqrt{48k}$ yields the minimax lower-bound

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \gtrsim \frac{n^{\frac{1}{p^*}} n^{-\frac{1}{q}}}{\sqrt{k}} = \frac{n^{\frac{1}{2}-\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{q}}}{\sqrt{k}}.$$

In the case that $n < 32 \log 2$, we once again refer Lemma A.3, which concludes the proof for the lower bound on the minimax stochastic risk.

For the upper bound, we turn to (9), with $h(x) = \frac{1}{2}\|x\|_2^2$. It holds again that $\sup_{x,x' \in X} \mathrm{D}_h(x, x') \leq n^{1/2-1/p}$. Since $r \geq 2$, we have that $\sup_{\|g\|_r \leq 1} \|g\|_2 = n^{\frac{1}{2}-\frac{1}{r}}$ and choosing the stepsize α to optimize (9) yields the upper bound on the minimax regret. □

D Proofs for Section 3.2

D.1 Proof of Theorem 2

The upper bound is simply Corollary 3.1. For the lower bound, similar to our warm-up in Section 3.1, we consider "sparse" gradients, though instead of using Fano's method we use Assouad's method to more carefully relate the geometry of the norm γ and constraint set X.

Let $a \in \mathbf{R}_+^n$ be such that $\mathsf{Rec}(a) \subset X$. We consider the sample space $\mathcal{S} := \{\pm e_j\}_{j \le n}$ and functions

$$F(x, s) := \sum_{j \le n} \frac{1}{\gamma(e_j)} |s_j||x_j - a_j s_j|.$$

For any $s \in \mathcal{S}$, the subdifferential $\partial_x F(x, s)$ has at most one non-zero coordinate; the orthosymmetry of γ implies $F \in \mathcal{F}^{\gamma,1}$. Let $p \in \mathbf{R}_+^n$ (to be specified presently) be such that $\mathbf{1}^\top p = 1$ and for $1 \le j \le n$, let $\delta_j \in [0, 1/2]$. We define the distributions P_v on $\mathcal{S}$ by

$$S = \begin{cases} v_j e_j & \text{with probability } \frac{p_j(1+\delta_j)}{2} \\ -v_j e_j & \text{with probability } \frac{p_j(1-\delta_j)}{2}. \end{cases}$$

With this choice, we evidently have

$$f_v(x) = \mathbf{E}_{S \sim P_v} F(x, S) = \sum_{j \le n} \frac{p_j}{\gamma(e_j)} \left[\frac{1+\delta_j}{2} |x_j - a_j v_j| + \frac{1-\delta_j}{2} |x_j + a_j v_j| \right]$$

and immediately that $\inf_X f_v = \sum_{j \le n} \frac{p_j a_j}{\gamma(e_j)} (1 - \delta_j)$. As a consequence, we have the Hamming separation (recall Eq. (16))

$$f_v(x) - \inf_X f_v = \sum_{j \le n} \frac{p_j a_j \delta_j}{\gamma(e_j)} \mathbf{1}_{\mathrm{sign}(x_j) \ne v_j},$$

which allows us to apply Assouad's method via Lemma A.4.

Using the same notation as Lemma A.4, we have

$$\left\| \mathbb{P}_{+j}^k - \mathbb{P}_{-j}^k \right\|_{\mathsf{tv}}^2 \le \frac{1}{2} D_{\mathrm{kl}} \left(\mathbb{P}_{+j}^k \| \mathbb{P}_{-j}^k \right) \le \log 3 \cdot k p_j \delta_j^2.$$

Choosing $\delta_j = \min\{\frac{1}{2}, \frac{1}{2\sqrt{k p_j \log(3)}}\}$ yields the lower bound

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \frac{1}{8} \sum_{j \le n} \frac{a_j}{\gamma(e_j)} \min \left\{ p_j, \frac{\sqrt{p_j}}{\sqrt{k \log 3}} \right\},$$

and by taking $p_j = (\frac{a_j}{\gamma(e_j)})^2 / \left\| \mathrm{res}(a, \gamma) \right\|_2^2$, we obtain for any $a \in X$ that

$$\begin{aligned} \mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \mathfrak{M}_k^{\mathsf{S}}(\mathsf{Rec}(a), \gamma) &\ge \frac{1}{8} \sum_{j \le n} \frac{a_j}{\gamma(e_j)} \min \left\{ \frac{a_j^2}{\gamma(e_j)^2 \left\| \mathrm{res}(a, \gamma) \right\|_2^2}, \frac{1}{\sqrt{k \log 3}} \frac{a_j}{\gamma(e_j) \left\| \mathrm{res}(a, \gamma) \right\|_2} \right\} \\ &= \frac{1}{8 \left\| \mathrm{res}(a, \gamma) \right\|_2^2} \sum_{j=1}^n \frac{a_j^2}{\gamma(e_j)^2} \min \left\{ \frac{a_j}{\gamma(e_j)}, \frac{\left\| \mathrm{res}(a, \gamma) \right\|_2}{\sqrt{k \log 3}} \right\} \end{aligned}$$

For notational simplicity, define the set $T := \{\mathrm{res}(x, \gamma)) \mid x \in X\}$, which is evidently orthosymmetric and convex (it is a diagonal scaling of X). Then

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge \sup_{u \in T} \frac{1}{8 \left\| u \right\|_2^2} \sum_{j=1}^n u_j^2 \min \left\{ u_j, \frac{\left\| u \right\|_2}{\sqrt{k \log 3}} \right\}. \tag{19}$$

For any vector $u \in \mathbf{R}^n_+$ and $c < 1$, if we define $J = \{j \in [n] \mid u_j \ge \frac{c}{\sqrt{n}} \|u\|_2\}$, then

$$\|u\|_2^2 = \|u_J\|_2^2 + \|u_{J^c}\|_2^2 \le \|u_J\|_2^2 + \|u\|_2^2 \sum_{j \in J^c} \frac{c^2}{n} \le \|u_J\|_2^2 + c^2 \|u\|_2^2, \quad \text{i.e.} \quad \|u_J\|_2 \ge \sqrt{1-c^2}\,\|u\|_2 .$$

Now, fix $d \in \mathbf{N}$. If in the supremum (19) we consider any vector $u \in T, u \ge 0$ satisfying $\|u\|_0 \le d$, then setting the index set $J = \{j : u_j \ge \|u\|_2 / \sqrt{k \log 3}\} = \{j : u_j \ge \|u\|_2 / \sqrt{d(k/d) \log 3}\}$ we have

$$\mathfrak{M}^{\mathsf{S}}_k(X, \gamma) \ge \frac{1}{8\|u\|_2^2} \sum_{j=1}^n u_j^2 \min\left\{u_j, \frac{\|u\|_2}{\sqrt{k \log 3}}\right\} \ge \frac{1}{8\|u\|_2^2} \sum_{j \in J} u_j^2 \frac{\|u\|_2}{\sqrt{k \log 3}} \ge \frac{1}{8}\left(1 - \frac{d}{k \log 3}\right) \frac{\|u\|_2}{\sqrt{k \log 3}}.$$

Taking a supremum over u with $\|u\|_0 \le d$ gives the theorem.

D.2 Proof of Corollary 3.2

Given the proof of Theorem 2, the proof is nearly immediate. Let $p \in [1, 2]$, $\beta \in \mathbf{R}^n_{++}$, and $\gamma(v) = \|\beta \odot v\|_p$. For the lower bound, the final display of the proof of Theorem 2 above guarantees the lower bound $\mathfrak{M}^{\mathsf{S}}_k(X, \gamma) \ge \frac{1}{16} \|u\|_2 / \sqrt{k}$ for all $u \in \{\mathrm{res}(x, \gamma) \mid x \in X\}$ and $k \ge 2n$. We first observe that $\mathsf{QHull}(\mathbf{B}_\gamma(0, 1)) = \{v \in \mathbf{R}^n \mid \|\beta \odot v\|_2 \le 1\}$. Thus, the upper bound in Theorem 2 is

$$\mathfrak{M}^{\mathsf{R}}_k(X, \gamma) \le \frac{1}{\sqrt{k}} \sup_{x \in X} \sup_{g : \|\beta \odot g\|_2 \le 1} x^\top g.$$

Using

$$\sup_{g : \|\beta \odot g\|_2 \le 1} u^\top g = \sup_{z : \|z\|_2 \le 1} u^\top (z / \beta) = \|u / \beta\|_2 ,$$

and recalling $\beta_j = \gamma(e_j)$ concludes the proof. □

E Proofs for Section 4

E.1 Proof of Theorem 3

Let us tackle the first case stated in the theorem; we reduce the second case to the first one by scaling the dimension.

E.1.1 Case $1 \le p \le 1 + 1/\log(2n)$

We always have the lower bound $1/\sqrt{k}$ by Lemma A.3 by reducing to a lower-dimensional problem, so we assume without loss of generality that $n \ge 8$.

Separation Let us consider $\mathcal{V} = \{\pm e_j\}_{j \le n}$. For $v = \pm e_j \in \mathcal{V}$, we define P_v on $S \in \{\pm 1\}^n$ by choosing coordinates of S independently via

$$S_j = \begin{cases} 1 & \text{with probability } \frac{1 + \delta v_j}{2} \\ -1 & \text{with probability } \frac{1 - \delta v_j}{2}. \end{cases}$$

Immediately, we have $\mathbf{E}_{P_v} S = \delta v$. For $s \in \{\pm 1\}^n$, we define $F(x, s) := n^{-1/p^*} x^\top s$, so $F \in \mathcal{F}^{\gamma,1}$, $f_v(x) = \mathbf{E}_{P_v} F(x, S) = \delta n^{-1/p^*} x^\top v$, and a calculation gives that $f_v^* := \inf_X f_v = -\delta n^{-1/p^*}$. For $v \neq v' \in \mathcal{V}$, we have

$$\begin{aligned} \mathsf{d}_{\rm opt}(v, v', X) = \inf_{x \in X} f_v(x) + f_{v'}(x) - f_v^* - f_{v'}^* &= n^{-1/p^*} \delta \inf_{x \in X} \Big((v + v')^\top x + 2 \Big) \\ &= \delta n^{-1/p^*} (2 - \|v + v'\|_{p^*}) \\ &\geq (2 - \sqrt{2}) \delta n^{-1/p^*}. \end{aligned}$$

Lemma A.2 yields

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \geq \frac{2 - \sqrt{2}}{2} \delta n^{-1/p^*} \inf_{\psi : \mathcal{S}^k \to \mathcal{V}} \mathbb{P}(\psi(S_1^k) \neq V).$$

It now remains to bound the testing error.

Bounding the testing error Noting that $|\mathcal{V}| = \log(2n)$, we lower bound the testing error via Fano's inequality

$$\inf_{\psi : \mathcal{S}^k \to \mathcal{V}} \mathbb{P}(\psi(S_1^k) \neq V) \geq \left(1 - \frac{\mathsf{I}(S_1^k; V) + \log 2}{\log(2n)} \right).$$

For any $v \neq v' \in \mathcal{V}$, we have for $\delta \in [0, \frac{1}{2}]$ that

$$D_{\rm kl}\left(P_v \| P_{v'}\right) = \delta \log \frac{1 + \delta}{1 - \delta} \leq 3\delta^2.$$

We can thus bound the mutual information between S_1^n and V

$$\mathsf{I}(S_1^k; V) \leq k \max_{v \neq v'} D_{\rm kl}\left(P_v \| P_{v'}\right) \leq 3k\delta^2.$$

In the case that $k < 8$, the lower bound holds trivially via Lemma A.3. In the case that $k \geq 8$, assuming that choosing $\delta^2 = \frac{\log(2n)}{6k} \wedge \frac{1}{2}$ yields

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \geq \frac{2 - \sqrt{2}}{2} n^{-1/p^*} \min\left\{ \sqrt{\frac{\log(2n)}{6k}}, \frac{1}{2} \right\} \left(1 - \frac{1}{2} - \frac{1}{4} \right), \tag{20}$$

which is valid for all $p \in [1, 2]$. In the case that $1 \leq p \leq 1 + 1/\log(2n)$, we note that $n^{-1/p^*} = 1/n^{\frac{p-1}{p}} \geq 1/e$, which yields

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \geq c \cdot \sqrt{\frac{\log(2n)}{k}} \wedge 1$$

for a numerical constant $c_0 > 0$.

To conclude, we see that the upper bound is essentially Proposition 2.3. Choosing $a = 1 + 1/\log(2n)$, the quantity $\frac{\sup_{x \in X} \|x\|_a \sup_{g \in \mathbf{B}_\gamma(0,1)} \|g\|_{a^*}}{\sqrt{a-1}\sqrt{k}}$ upper bounds the minimax regret. As $a > p$, $\sup_{x \in X} \|x\|_a = 1$. We have $a^* = \log(2n) + 1$ and $p^* \geq a^*$, so that

$$\|g\|_{a^*} \leq n^{\frac{1}{a^*} - \frac{1}{p^*}} \|g\|_{p^*} \leq n^{\frac{1}{a^*}}$$

as $g \in \mathbf{B}_{p^*}(0, 1)$. Once we note that both $n^{1/a^*} = \exp(\frac{\log n}{\log(2n)+1}) \leq e$ and $1/\sqrt{2(a-1)} = \sqrt{\log(2n)/2}$, we conclude this case.

E.1.2 Case $1 + 1/\log(2n) < p \le 2$

Let $n_0 \le n$. We can embed a function $F_{n_0} : \mathbf{R}^{n_0} \times \mathcal{S} \to \mathbf{R}$ as a function $F : \mathbf{R}^n \times \mathcal{S} \to \mathbf{R}$ by letting π_{n_0} denote the projection onto the first n_0-components, and defining

$$F(x, s) = F_{n_0}(\pi_{n_0} x, s).$$

If the subgradients of F_{n_0} lie in $\mathbf{B}_{p^*}(0, 1)$, so do those of F. Similarly, if $x_0 \in \{\tau \in \mathbf{R}^{n_0}, \|\tau\|_p \le 1\}$ then $x = (x_0, \mathbf{0}_{n_0+1:n}) \in \mathbf{B}_p(0, 1)$. As such, any lower bound for the n_0-dimensional problem implies an identical one for all $n \ge n_0$-dimensional problems. For $1 + 1/\log(2n) < p \le 2$, let us define $n_0 = \lceil \exp(\frac{1}{p-1})/2 \rceil$, so $n_0 \le n$ as desired. In the case that $p > 1 + 1/\log 16$, Lemma A.3 yields the desired lower bound. In the case that $p \le 1 + 1/\log 16$, we have that $n_0 \ge 8$, and the lower bound (20) holds so that for a numerical constant $c > 0$,

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge c n_0^{-1/p^*} \cdot \sqrt{\frac{\log(2n_0)}{n}} \wedge 1.$$

Once we use that $n_0^{-1/p^*} \ge (1/2)^{\frac{1}{p}-1} \exp(-1/p) \ge \sqrt{2/e}$, substituting $n_0 = \lceil \exp(\frac{1}{p-1})/2 \rceil$ above gives the final lower bound

$$\mathfrak{M}_k^{\mathsf{S}}(X, \gamma) \ge c \cdot \frac{1}{\sqrt{2(p-1)n}} \wedge 1.$$

Proposition 2.3 yields the upper bound and concludes this proof. □

E.2 Proof of Proposition 4.1

To see the lower bound on the stochastic minimax risk, we apply Assouad's method (see Section A.2). Fix any $x \in X$, and let the sampled vectors $s \in \{-1, 1\}^n$ generate functions via

$$F(x; s) = \sum_{j=1}^{n} \left| x_j - s_j x_j^\star \right|,$$

which evidently satisfy $\|\partial F(x; s)\|_\infty \le 1$. For $v \in \{\pm 1\}^n$, let the sampling distribution be $P_v(S = s) = \prod_{j=1}^n \frac{1+\delta v_j s_j}{2}$, that is, independent Bernoulli coordinates with biases $\frac{1+\delta v_j}{2}$. Then

$$f_v(x) := \mathbf{E}_{P_v}[F(x; S)] = \sum_{j=1}^{n} \frac{1+\delta}{2} |x_j - v_j x_j^\star| + \frac{1-\delta}{2} |x_j + v_j x_j^\star|,$$

which induces the Hamming separation (16)

$$f_v(x) - \inf_{x \in X} f_v(x) \ge \delta \sum_{j=1}^{n} |x_j^\star| 1(\operatorname{sign}(x_j) \ne v_j),$$

as $1 - (1 - \delta) = \delta$. Thus Lemma A.4 implies

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}_\infty) \ge \frac{\delta}{2} \sum_{j=1}^{n} |x_j^\star| \left(1 - \|\mathbb{P}_j - \mathbb{P}_{-j}\|_{\mathsf{tv}}\right),$$

where $\mathbb{P}_j = \frac{1}{2^{n-1}} \sum_{v_j=1} P_v^k$ and $\mathbb{P}_{-j} = \frac{1}{2^{n-1}} \sum_{v_j=-1} P_v^k$. By Jensen's inequality followed by Pinsker's inequality, we have

$$\|\mathbb{P}_j - \mathbb{P}_{-j}\|_{\mathsf{tv}}^2 \le \frac{k}{2} \max_{v, v' : v_j \ne v_j'} D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) = \frac{k}{2} D_{\mathrm{kl}}\left(\mathsf{Ber}\left(\frac{1+\delta}{2}\right) \| \mathsf{Ber}\left(\frac{1-\delta}{2}\right)\right),$$

where $\mathsf{Ber}(p)$ denotes a Bernoulli p distribution. Of course, this final quantity has upper bound

$$D_{\mathrm{kl}}\left(P_v \| P_{v'}\right) = \frac{1+\delta}{2} \log \frac{1+\delta}{1-\delta} + \frac{1-\delta}{2} \log \frac{1-\delta}{1+\delta} \le 2 \log 3 \cdot \delta^2$$

for $\delta \in [0, \frac{1}{2}]$. That is, for any $x^\star \in X$, we have

$$\mathfrak{M}_k^{\mathsf{S}}(X, \mathcal{F}_\infty) \ge \frac{\delta}{2} \left\|x^\star\right\|_1 \left(1 - \sqrt{k \log 3 \cdot \delta^2}\right)$$

for $\delta \le \frac{1}{2}$. Take $\delta^2 = \frac{1}{4 \log 3}$.

To demonstrate the upper bound, we take $X \subset \mathbf{R}^{\mathbf{N}}$, as the special case of finite dimensions follows immediately. For notational simplicity, for a positive sequence $(a_j)_{j \in \mathbf{N}}$ to be chosen we let $A : \mathbf{R}^{\mathbf{N}} \to \mathbf{R}^{\mathbf{N}}$ be the diagonal linear operator $Ax = (a_j x_j)_{j \in \mathbf{N}}$. Now, for $p \in (1, 2]$ to be chosen as well, define the distance generating function

$$h(x) = \frac{1}{2(p-1)} \left\|Ax\right\|_p^2.$$

By an analogous argument to the finite-dimensional case, this is strongly convex with respect to the norm $\|x\| = \|Ax\|_p$ and has dual norm $\|g\|_* = \left\|A^{-1} g\right\|_q$ for $q = \frac{p}{p-1}$. In this case, for any sequence of subgradients $g_i \in [-1, 1]^{\mathbf{N}}$, we thus obtain regret bound

$$\mathsf{Regret}_k(x) \le \frac{1}{2(p-1)\alpha} \left\|Ax\right\|_p^2 + \frac{\alpha}{2} \sum_{i=1}^k \left\|A^{-1} g_i\right\|_q^2.$$

Notably, for $g \in [-1, 1]^{\mathbf{N}}$ we have

$$\left\|A^{-1} g\right\|_q^q \le \sum_{j=1}^\infty a_j^{-q}.$$

We use the assumption in the proposition that X has finite effective dimension (12) N, so that for $\beta = \frac{1}{\log N} \le 1$, $\sup_{x \in X} \sum_{j=1}^\infty j^\beta |x_j| \le e \sup_{x \in X} \|x\|_1$. Letting $a_j = j^\beta$, we take $q = \frac{2}{\beta}$, whence $\|A^{-1} g\|_q^q = \sum_{j=1}^\infty j^{-2} = \frac{\pi^2}{6}$. As this also gives conjugate $p = \frac{q}{q-1} = \frac{2}{2-\beta}$ with $p - 1 = \frac{\beta}{2-\beta}$, we obtain the regret bound

$$\mathsf{Regret}_k(x) \le \frac{1}{\alpha} \frac{2-\beta}{2\beta} \left\|Ax\right\|_p^2 + O(1) \cdot \alpha k \le O(1) \left[\frac{1}{\alpha\beta} \sup_{x \in X} \|x\|_1^2 + \alpha k\right].$$

Take $\alpha = \sup_{x \in X} \|x\|_1 / \sqrt{k\beta}$ to obtain the convergence upper bound. □

E.3 Proof of Observation 4.1

The first claim of the observation is trivial. For the second and third, we use a bit of additional notation, saying that X is β-self-similar if

$$\sup_{x \in X} \sum_{j=1}^\infty j^\beta |x_j| \le e \sup_{x \in X} \|x\|_1. \tag{21}$$

We show that if $\lim_{\beta \downarrow 0} \sup_{x \in X} \sum_{j=1}^\infty j^\beta |x_j| < \infty$, then there is some $\beta > 0$ for which X is β-self-similar (21). The key is the following claim:

Lemma E.1. *Let X satisfy $\lim_{\beta\downarrow 0}\sup_{x\in X}\sum_{j=1}^{\infty} j^{\beta}|x_j| < \infty$. Then there exists $\beta_0 > 0$ such that $\beta \mapsto s(\beta) := \sup_{x\in X}\sum_{j=1}^{\infty} j^{\beta}|x_j|$ is continuous on $[0,\beta_0)$, and $\lim_{\beta\downarrow 0} s(\beta) = \sup_{x\in X}\|x\|_1$.*

Proof. Let $\beta_0 = \sup\{\beta > 0 \mid s(\beta) < \infty\}$. Take $\beta \in [0,\beta_0)$, and let $\beta' \to \beta$ in $[0,\beta_0)$. By the assumption that $\sup_{x\in X}\sum_{j=1}^{\infty} j^{\beta}|x_j| < \infty$ for all $\beta < \beta_0$, for any $\epsilon > 0$ and all suitably small $\gamma > 0$, we may take N such that $\sup_X \sum_{j>N} j^{\beta+\gamma}|x_j| < \epsilon$. Then for β' close enough to β, we obtain

$$|s(\beta) - s(\beta')| \le \sup_{x\in X}\sum_{j\le N} |j^{\beta} - j^{\beta'}||x_j| + \sup_{x\in X}\sum_{j>N}(j^{\beta} + j^{\beta'})|x_j| \le \sup_{x\in X}\sum_{j\le N}|j^{\beta} - j^{\beta'}||x_j| + 2\epsilon.$$

As $\beta \mapsto j^{\beta}$ is continuous in β, when β' is close enough to β we obtain $N|j^{\beta} - j^{\beta'}| \le \epsilon$ for all $j \le N$. So $|s(\beta) - s(\beta')| \le \epsilon \sup_{x\in X}\|x\|_1 + 2\epsilon$. As $\epsilon > 0$ was arbitrary this completes the proof. □

The preceding lemma demonstrates that $\lim_{\beta\downarrow 0}\sup_{x\in X}\sum_{j=1}^{\infty} j^{\beta}|x_j| = \sup_{x\in X}\|x\|_1$, so there is some $\beta > 0$ for which $\sup_{x\in X}\sum_{j=1}^{\infty} j^{\beta}|x_j| \le e\sup_{x\in X}\|x\|_1$. Thus $\operatorname{effdim}(X) < \infty$.

Finally, for the final claim of the observation, let $\mathbf{1}_n$ denote the linear functional $\langle \mathbf{1}_n, x\rangle = \sum_{j=1}^{n} x_j$, and let N_γ be the smallest N such that $\sup_{x\in X}\langle \mathbf{1}_N, x\rangle \ge \sup_{x\in X} n^{2\gamma}\sum_{j>n}|x_j|$ for all $n \ge N$. Define

$$N := \max\left\{N_\gamma, \exp\left(1 + \frac{3}{2\gamma}\log\frac{3}{2\gamma}\right), \exp\left(\frac{1}{2\log(e-\sqrt{e})}\right)\right\}.$$

We claim that X is β-self-similar (21) for $\beta = \frac{1}{2\log N}$, so that $\operatorname{effdim}(X) \le \exp(1/\beta) = N^2$ as desired. To see the claim of self-similarity, note that the choice $\beta = \frac{1}{2\log N}$ guarantees $\beta - 2\gamma \le -\frac{3\gamma}{2}$, and

$$\sum_{j=1}^{\infty} j^{\beta}|x_j| \le N^{\frac{1}{2\log N}}\sum_{j=1}^{N}|x_j| + \sum_{j=N+1}^{\infty} j^{\beta}|x_j| \le \sqrt{e}\sup_{x\in X}\langle \mathbf{1}_N, x\rangle + \sum_{j=N+1}^{\infty} j^{\beta}|x_j|.$$

Define the index blocks $\mathcal{B}_k := \{i \in \mathbf{N} \mid e^k \le i < e^{k+1}\}$, so that

$$\sum_{j=N+1}^{\infty} j^{\beta}|x_j| \le \sum_{k=\lfloor \log N\rfloor}^{\infty}\sum_{j\in\mathcal{B}_k} j^{\beta}|x_j| \le \sum_{k=\lfloor\log N\rfloor}^{\infty} e^{\beta(k+1)}\frac{1}{e^{2k\gamma}}\sup_{x\in X}\langle \mathbf{1}_N, x\rangle$$

because $\sum_{j\in\mathcal{B}_k}|x_j| \le e^{-2k\gamma}\sup_{x\in X}\langle \mathbf{1}_N, x\rangle$ by assumption. Computing the infinite sums, we have that because $\frac{1}{2\log N} \le \frac{\gamma}{2}$,

$$\begin{aligned}\sum_{k=\lfloor\log N\rfloor}^{\infty} e^{\beta(k+1)}\frac{1}{e^{2k\gamma}} &= e^{\beta}\sum_{k=\lfloor\log N\rfloor}^{\infty} e^{k(\beta-2\gamma)} \le e^{\beta}e^{-3\gamma\lfloor\log N\rfloor/2}\sum_{k=0}^{\infty} e^{-3k\gamma/2} \\ &= \frac{e^{\beta-3\gamma\lfloor\log N\rfloor/2}}{1-e^{-3\gamma/2}} \le \left(\frac{N}{e}\right)^{-\frac{3}{2}\gamma}\frac{2e^{\beta}}{3\gamma},\end{aligned}$$

where we used that $e^x \ge 1+x$, or $1-e^x \le -x$. Finally, noting that $(N/e)^{-\kappa} \le \kappa$ iff $\log N \ge 1 + \frac{1}{\kappa}\log\frac{1}{\kappa}$, we substitute $\kappa = \frac{3\gamma}{2}$ to obtain $\sum_{j=N+1}^{\infty} j^{\beta}|x_j| \le e^{\beta}\sup_{x\in X}\langle \mathbf{1}_N, x\rangle$. We have shown that

$$\sum_{j=1}^{\infty} j^{\beta}|x_j| \le (\sqrt{e} + e^{\beta})\sup_{x\in X}\langle \mathbf{1}_N, x\rangle$$

for $\beta = \frac{1}{2\log N}$. So long as $\beta \le \log(e - \sqrt{e})$ we have $\sqrt{e} + e^{\beta} \le e$.

□

F Proofs of regret lower bounds

F.1 Proof of Theorem 4

Let $A \succ 0$ be a positive semi-definite matrix for the distance generating function $h_A(x) = \frac{1}{2}x^\top Ax$ defined above, and let $q = \frac{p}{p-1}$ be the conjugate to p. We choose linear functions $F_i(x) := g_i^\top x$ where $g_i \in \mathbf{B}_q(0,1)$. In this case, letting $\{x_i\}_{i \le k}$ be the points mirror descent plays, the regret with respect to $x \in \mathbf{R}^n$ is

$$\mathsf{Regret}_{k,A}(x) = \sum_{i \le k} F_i(x_i) - F_i(x) = \sum_{i \le k} g_i^\top (x_i - x),$$

so that

$$\mathsf{Regret}^*_{k,A} := \sup_{\|x\|_p \le 1} \mathsf{Regret}_{k,A}(x) = \left\|\sum_{i \le k} g_i\right\|_q + \frac{1}{2}\sum_{i \le k} \|g_i\|_{A^{-1}}^2 - \frac{1}{2}\left\|\sum_{i \le k} g_i\right\|_{A^{-1}}^2.$$

Now, we choose linear functions f_i so that the regret is large. To do so, choose vectors

$$u \in \operatorname*{argmax}_{\|s\|_q \le 1} s^\top A^{-1} s \quad \text{and} \quad v \in \operatorname*{argmin}_{\|s\|_q = 1} s^\top A^{-1} s. \tag{22}$$

Then set the (gradient) vectors $g_i \in \mathbf{R}^n$ so that for a $\delta \in [0,1]$ to be chosen,

(a) $g_i = u$ for $k/4$ of the indices $i \in [k]$

(b) $g_i = -u$ for $k/4$ of the indices $i \in [k]$

(c) $g_i = v$ for $\frac{k}{4}(1+\delta)$ of the indices $i \in [k]$

(d) $g_i = -v$ for $\frac{k}{4}(1-\delta)$ of the indices $i \in [k]$.

With these choices, we obtain the regret lower bound

$$\begin{aligned}\mathsf{Regret}^*_{k,A} &\ge \sup_{\delta \le 1}\left[\frac{k}{2}\delta\|v\|_q + \frac{k}{4}u^\top A^{-1}u - \frac{\delta^2 k^2}{8}v^\top A^{-1}v\right] \\ &\ge \frac{k}{4}\cdot\left[u^\top A^{-1}u + \min\left\{1, \frac{2\|v\|_q}{kv^\top A^{-1}v}\right\}\|v\|_q\right].\end{aligned} \tag{23}$$

We now consider two cases. In the first, A is large enough that $\|v\|_q \ge \frac{1}{2}kv^\top A^{-1}v$. Then the regret bound (23) becomes

$$\mathsf{Regret}^*_{k,A} \ge \frac{k}{4}\left[u^\top A^{-1}u + \|v\|_q\right] \ge \frac{k}{4},$$

as $\|v\|_q = 1$ by the construction (22). This gives the first result of the theorem. For the second claim, which holds in the case that $\|v\|_q < \frac{1}{2}kv^\top A^{-1}v$, we consider the operator norms of general invertible linear operators. For a mapping $T : \mathbf{R}^n \to \mathbf{R}^n$, define the ℓ_p to ℓ_q operator norm

$$\|T\|_{\ell_p \to \ell_q} := \sup_{x \ne 0}\frac{\|T(x)\|_q}{\|x\|_p}.$$

Then the construction (22) evidently yields

$$u^\top A^{-1} u = \|A^{-1/2}\|_{\ell_q \to \ell_2}^2 \quad \text{and} \quad \frac{\|v\|_q^2}{v^\top A^{-1} v} = \sup_{x \neq 0} \frac{\|A^{1/2} x\|_q^2}{\|x\|_2^2} = \|A^{1/2}\|_{\ell_2 \to \ell_q}^2.$$

Revisiting the regret (23), we obtain

$$\mathsf{Regret}_{k,A}^* \ge \frac{k}{4} \cdot \left[\left\| A^{-1/2} \right\|_{\ell_q \to \ell_2}^2 + \frac{2}{k} \left\| A^{1/2} \right\|_{\ell_2 \to \ell_q}^2 \right] \ge \sqrt{\frac{k}{2}} \|A^{-1/2}\|_{\ell_q \to \ell_2} \|A^{1/2}\|_{\ell_2 \to \ell_q},$$

where we have used that $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ for all a, b. But for any invertible linear operator, standard results on the Banach-Mazur distance [37, Corollary 2.3.2] imply that

$$\inf_{A \succ 0} \|A\|_{\ell_2 \to \ell_q} \left\| A^{-1} \right\|_{\ell_q \to \ell_2} \ge n^{1/2 - 1/q}.$$

This gives the lower bound.

For the claimed upper bound, note for $h(x) = \frac{1}{2\alpha} \|x\|_2^2$ (i.e. $A = \frac{1}{\alpha} I_n$) and initial point $x_0 = \mathbf{0}$, we have $\mathsf{Regret}_k(x) \le \frac{1}{2\alpha} \|x\|_2^2 + \frac{\alpha}{2} \sum_{i=1}^k \|g_i\|_2^2$. As $\|x\|_2 \le 1$ whenever $\|x\|_p \le 1$ and and $\|g\|_2 \le n^{1/2-1/q}$ whenever $\|g\|_q \le 1$, we have $\mathsf{Regret}_k(x) \le \frac{1}{2\alpha} + \frac{\alpha}{2} k n^{1-2/q}$. Choose $\alpha = (kn^{1-2/q})^{-1/2}$ to minimize this bound and achieve $\sup_{\|x\|_p \le 1} \mathsf{Regret}_k(x) \le \sqrt{k} n^{1/2-1/q}$. □

F.2 Proof of Theorem 5

As in the proof of Theorem 4, let $A \succ 0$ be a positive definite matrix, so that the regret of (Euclidean) mirror descent with distance generating function $h_A(x) = \frac{1}{2} x^\top A x$ is

$$\mathsf{Regret}_{k,A}(x) = \left\langle \sum_{i \le k} g_i, x \right\rangle + \frac{1}{2} \sum_{i \le k} \|g_i\|_{A^{-1}}^2 - \frac{1}{2} \left\| \sum_{i \le k} g_i \right\|_{A^{-1}}^2.$$

For vectors u, v with $\|u\|_\infty \le 1, \|v\|_\infty \le 1$ to be chosen and $p \in (0,1)$ to be chosen as well, we set

(a) $g_i = u$ for $pk/2$ of the indices $i \in [k]$

(b) $g_i = -u$ for $pk/2$ of the indices $i \in [k]$

(c) $g_i = v$ for $(1-p)k$ of the indices $i \in [k]$.

This then yields regret lower bound

$$\mathsf{Regret}_{k,A}(x) \ge (1-p) k v^\top x - \frac{(1-p)^2 k^2}{2} v^\top A^{-1} v + \frac{pk}{2} u^\top A^{-1} u. \tag{24}$$

We argue we may assume w.l.o.g. that $\limsup_k \|A(k)\|_{\text{op}} / \sqrt{k} < \infty$. Suppose to the contrary that along a subsequence, which for simplicity we take to be the entire sequence, that $\|A(k)\|_{\text{op}} \gg \sqrt{k}$. Let $C < \infty$ be arbitrary, take k large enough that $\|A(k)\|_{\text{op}} \ge C\sqrt{k}$, and assume w.l.o.g. that $C \le \sqrt{k}$ (we can always take k larger). Let w be the unit eigenvector of $A = A(k)$ achieving $w^\top A w = \|A\|_{\text{op}}$, and set $v = \delta w$ for a $0 \le \delta \le 1$ to be chosen, so that $\|v\|_\infty \le \delta \le 1$. Then at such indices k the lower bound (24) implies

$$\mathsf{Regret}_{k,A(k)}(x) \ge (1-p) k \delta w^\top x - \frac{(1-p)^2 k^2 \delta^2}{2 \|A(k)\|_{\text{op}}} + \frac{kp}{2} u^\top A^{-1} u \ge \frac{k(1-p)\delta}{2} \left[w^\top x - \frac{(1-p)\sqrt{k}\delta}{2C} \right].$$

Taking a supremum over $x \in X$, which is a convex body (so that it has interior), we have $\sup_{x \in X} w^\top x = c(X) > 0$, whence

$$\sup_{x \in X} \mathsf{Regret}_{k,A(k)}(x) \ge \sup_{0 \le \delta \le 1} k(1-p)\delta \left[c(X) - \frac{(1-p)\sqrt{k}\delta}{2C} \right] \ge \frac{C(1-p) \cdot \min\{c(X), 1\}}{2} \sqrt{k},$$

where the second inequality sets $\delta = C \cdot \min\{c(X), 1\}/\sqrt{k}$. As $c(X) > 0$ and $C < \infty$ was otherwise arbitrary, whenever $p < 1$ the preceding lower bound is stronger than that the theorem claims.

Returning to the main thread, we may therefore assume that $\sup_k \left\| A(k) \right\|_{\rm op} / \sqrt{k} \le C$ for some finite C. Returning to the regret bound (24), we optimize over v and u to obtain

$$\mathsf{Regret}_{k,A}(x) \ge \sup_{\|v\|_\infty \le 1, \|u\|_\infty \le 1} \left[k(1-p) v^\top x - \frac{k^2(1-p)^2}{2} v^\top A^{-1} v + \frac{kp}{2} u^\top A^{-1} u \right].$$

Considering the supremum over v, we have for any $A \succ 0$ that

$$\operatorname*{argmax}_v \left\{ k(1-p) v^\top x - \frac{k^2(1-p)^2}{2} v^\top A^{-1} v \right\} = \frac{1}{k(1-p)} Ax.$$

Because X is bounded and $\left\| A \right\|_{\rm op} \le C/\sqrt{k}$, for $p < 1$ and suitably large k the v achieving this supremum evidently satisfies $\|v\|_\infty = \|Ax\|_\infty / (k(1-p)) \le 1$, and so for any $p \in (0,1)$ and large enough k we obtain

$$\mathsf{Regret}_{k,A}(x) \ge \sup_{\|u\|_\infty \le 1} \frac{1}{2} \left[kpu^\top A^{-1} u + x^\top A x \right]. \tag{25}$$

We use a duality argument to lower bound the quantity (25). Let $\mathcal{P}(X)$ denote the collection of probability measures on X, and let u be a random vector, uniform on $\{-1, 1\}^n$. Then

$$\begin{aligned} \inf_{A \succeq 0} \sup_{x \in X} \sup_{\|u\|_\infty \le 1} \left\{ kpu^\top A^{-1} u + x^\top A x \right\} &\ge \inf_{A \succeq 0} \sup_{\nu \in \mathcal{P}(X)} \left[kp \langle A^{-1}, \mathbf{E}[uu^\top] \rangle + \langle A, \mathbf{E}_\nu[xx^\top] \rangle \right] \\ &\ge \sup_{\nu \in \mathcal{P}(X)} \inf_{A \succeq 0} \left\{ kp \cdot \mathrm{tr}(A^{-1}) + \langle A, \mathbf{E}_\nu[xx^\top] \rangle \right\}. \end{aligned}$$

Taking derivatives with respect to A, we see that the inner infimum is achieved whenever

$$-kpA^{-2} + \mathbf{E}_\nu[xx^\top] = 0, \quad \text{i.e.} \quad A = \sqrt{kp} \mathbf{E}_\nu[xx^\top]^{-1/2}.$$

So long as $\mathbf{E}_\nu[xx^\top] \succeq C^{-1} I_n$, which we may choose, this satisfies the constraint that $\left\| A \right\|_{\rm op} \le C\sqrt{k}$. Substituting into the regret lower bound (25), we have for any $C < \infty$ and $p \in (0,1)$ that for all large enough k

$$\sup_{x \in X} \mathsf{Regret}_{k,A}(x) \ge \sqrt{kp} \sup_{\nu \in \mathcal{P}(X)} \left\{ \mathrm{tr}\left(\mathbf{E}_\nu[xx^\top]^{1/2} \right) \mid \mathbf{E}_\nu[xx^\top] \succeq C^{-1} I_n \right\}.$$

Finally, we use the following lemma relating quadratic hulls and measures.

Lemma F.1. *Let $\mathcal{P}(X)$ denote the collection of probability measures on an orthosymmetric convex set $X \subset \mathbf{R}^n$. Then*

$$\sup_{\nu \in \mathcal{P}(X)} \mathrm{tr}\left(\mathbf{E}_\nu[xx^\top]^{1/2} \right) = \sup_{\nu \in \mathcal{P}(X)} \sum_{j=1}^n \sqrt{\mathbf{E}_\nu[x_j^2]} = \sup_{q \in \mathsf{QHull}(X)} \langle \mathbf{1}, q \rangle.$$

Moreover, the suprema can be taken over symmetric measures.

Proof. We prove the first equality first. The function $\sqrt{\cdot}$ is concave, and so the function $A \mapsto \operatorname{tr}(A^{1/2})$, as a permutation-symmetric function only of the eigenvalues of $A \succeq 0$, is concave on $A \succeq 0$ as well [23], with derivative $\nabla \operatorname{tr}(A^{1/2}) = \frac{1}{2}A^{-1/2}$. Letting $D = \operatorname{diag}(A)$ be the diagonal of A, we thus have by the standard first-order concavity inequality

$$\operatorname{tr}(A^{1/2}) \le \operatorname{tr}(D^{1/2}) + \langle \nabla \operatorname{tr}(D^{1/2}), D - A\rangle = \operatorname{tr}(D^{1/2}) + \frac{1}{2}\langle D^{-1/2}, D - A\rangle = \operatorname{tr}(D^{1/2}),$$

so that for any positive semidefinite matrix A we have $\operatorname{tr}(A^{1/2}) \le \operatorname{tr}(\operatorname{diag}(A)^{1/2})$. Taking $A = \mathbf{E}_\nu[xx^\top]^{1/2}$ then gives the first equality once we recognize that if ν is symmetric, so that $\mathbf{E}_\nu[xx^\top]$ is diagonal, then $\operatorname{tr}(\mathbf{E}_\nu[xx^\top]^{1/2}) = \sum_{j=1}^n \sqrt{\mathbf{E}_\nu[x_j^2]}$.

For the second equality, recall that for any vector $q \in \mathsf{QHull}(X)$ with $q \succeq 0$, we may write $q = \sqrt{z}$ (applied elementwise), where z is a convex combination of vectors of the form $[x_j^2]_{j=1}^n$, $x \in X$. Letting $x^i, i = 1, \ldots, m$ be these vectors, with $z = \sum_{i=1}^m \lambda_i (x^i)^2$ for some $\lambda \ge 0$ and $\mathbf{1}^\top \lambda = 1$, we let ν be the distribution on X assigning probabilities $\frac{\lambda_i}{2}$ to x^i and $\frac{\lambda_i}{2}$ to $-x^i$, which is evidently symmetric and satisfies $\sum_{j=1}^n \sqrt{\mathbf{E}_\nu[x_j^2]} = \mathbf{1}^\top q$. □

By a slight perturbation, we therefore obtain that for any $\epsilon > 0$, we can choose C large enough that for all large k, we have

$$\begin{aligned}\sup_{x \in X} \mathsf{Regret}_{k,A(k)}(x) &\ge \sqrt{kp} \sup_{\nu \in \mathcal{P}(X)} \left\{ \operatorname{tr}\left(\mathbf{E}_\nu[xx^\top]^{1/2}\right) \mid \mathbf{E}_\nu[xx^\top] \succeq C^{-1} I_n \right\} \\ &\ge (1-\epsilon)\sqrt{kp} \sup_{q \in \mathsf{QHull}(X)} \langle \mathbf{1}, q\rangle.\end{aligned}$$

As $\epsilon > 0$ and $p < 1$ were arbitrary, this completes the proof of the lower bound.

For the upper bound on the regret, note that with the updates $x_{k+1} = x_k - D^{-1} g_k$, we have

$$\mathsf{Regret}_k(x) \le \frac{1}{2}(x_0 - x)^\top D (x_0 - x) + \frac{1}{2}\sum_{i=1}^k g_i^\top D^{-1} g_i.$$

Take $x_0 = \mathbf{0}$ and $g_i = \mathbf{1}$ to obtain the upper bound

$$\sup_{x \in X} \mathsf{Regret}_k(x) \le \sup_{x \in X} \frac{1}{2}\sum_{j=1}^n d_j x_j^2 + \frac{k}{2}\sum_{j=1}^n \frac{1}{d_j}$$

for $D = \operatorname{diag}(d)$. Applying Sion's minimax theorem, we have

$$\inf_{D \succeq 0} \sup_{x \in X} \mathsf{Regret}_k(x) = \sup_{\nu \in \mathcal{P}(X)} \inf_{d \succeq 0} \frac{1}{2}\sum_{j=1}^n d_j \mathbf{E}_\nu[x_j^2] + \frac{k}{2}\sum_{j=1}^n \frac{1}{d_j} = \sqrt{k} \cdot \sup_{\nu \in \mathcal{P}(X)} \sum_{j=1}^n \sqrt{\mathbf{E}_\nu[x_j^2]},$$

which is the claimed result. □

F.3 Proof of Theorem 6

The proof follows similar lines as the one we show in Appendix F.1 but choosing different $u, v \in \mathbf{R}^n$. Let $\alpha \ge 0$ be a stepsize. We consider linear functions $F_i(x) := g_i^\top x$ with $\|\beta \odot g_i\|_1 \le 1$. Let $\{x_i\}_{i \le k}$ be the iterates of online gradient descent. The regret with respect to $x \in \mathbf{R}^n$ is

$$\mathsf{Regret}_{k,\alpha}(x) = \sum_{i \le k} g_i^\top (x_i - x).$$

This yields

$$\mathsf{Regret}^*_{k,\alpha} = \sup_{\|x\|_\infty \le 1} \mathsf{Regret}_{k,\alpha}(x) = \left\|\sum_{i\le k} g_i\right\|_1 + \frac{\alpha}{2}\sum_{i\le k}\|g_i\|_2^2 - \frac{\alpha}{2}\left\|\sum_{i\le k} g_i\right\|_2^2.$$

Let $d = \arg\min_{j\le n}\beta_j$, we choose

$$u = e_d/\beta_d \;\text{ and }\; v = \frac{\mathbf{1}}{\|\beta\|_1}.$$

For $\delta \in [0,1]$, we now choose the vectors $g_i \in \mathbf{R}^n$ as follows:

(a) $g_i = u$ for $k/4$ of the indices $i \in [k]$.

(b) $g_i = -u$ for $k/4$ of the indices $i \in [k]$.

(c) $g_i = v$ for $\frac{k}{4}(1+\delta)$ of the indices $i \in [k]$.

(d) $g_i = -v$ for $\frac{k}{4}(1-\delta)$ of the indices $i \in [k]$.

For this construction, we lower bound the regret

$$\begin{aligned}\mathsf{Regret}^*_{k,\alpha} &\ge \sup_{0\le\delta\le 1}\left\{\frac{k\delta}{2}\|v\|_1 + \frac{k\alpha}{4}\|u\|_2^2 - \frac{\alpha\delta^2k^2}{8}\|v\|_2^2\right\} \\ &\ge \frac{k\alpha}{4}\|u\|_2^2 + \frac{k\|v\|_1}{4}\min\left\{1, \frac{2\|v\|_1}{k\alpha\|v\|_2^2}\right\}.\end{aligned} \tag{26}$$

If the stepsize is too small (i.e. $\alpha \le \frac{2}{k}\frac{\|v\|_1}{\|v\|_2^2}$) then (26) becomes

$$\mathsf{Regret}^*_{k,\alpha} \ge \frac{kn}{4\|\beta\|_1}.$$

In the other case that $\alpha > \frac{2}{k}\frac{\|v\|_1}{\|v\|_2^2}$, (26) yields

$$\mathsf{Regret}^*_{k,\alpha} \ge \frac{k}{4\alpha}\|u\|_2^2 + \frac{\|v\|_1^2}{\|v\|_2^2}\frac{\alpha}{2} \ge \frac{\sqrt{2}}{2}\frac{\sqrt{kn}}{\min_{j\le n}\beta_j},$$

which is the desired result. □

G Proofs related to ℓ_1-diameters and n-widths

Here, we collect the lemmas necessary to prove Corollary 5.2. Recall that $\mathbf{1}_n \in \mathbf{R}^{\mathbf{N}}$ denotes the vector with 1 in the first n positions and 0 elsewhere.

Lemma G.1. *For the set $X = \mathrm{Conv}(C_0 \cup C_1)$ that equation (15) defines, we have*

$$\sup_{q\in\mathsf{QHull}(X)}\langle \mathbf{1}_n, q\rangle = \sqrt{\sum_{j=1}^n a_j^{-2}} \quad \textit{and} \quad \sup_{x\in X}\langle \mathbf{1}_n, x\rangle = \max_{m\le n}\frac{\sum_{j=1}^m b_j}{\sqrt{\sum_{j=1}^m a_j^2 b_j^2}} = \max_{m\le n}\frac{\langle \mathbf{1}_m, b\rangle}{Z(m)}.$$

Proof. Let $Q = \mathsf{QHull}(X) = \{q \mid \sum_{j=1}^{\infty} a_j^2 q_j^2 \le 1\}$. By Cauchy-Schwarz, the suprema of $\langle \mathbf{1}_n, q\rangle$ over $q \in Q$ satisfy $q_j = \frac{\lambda}{a_j^2}$ where $\lambda > 0$ normalizes q so that $\sum_{j=1}^{n} a_j^2 q_j^2 = 1$, that is, $\lambda = (\sum_{j=1}^{n} a_j^{-2})^{-1/2}$. For the second equality, note that $\langle \mathbf{1}_n, x\rangle$ is linear in x, and so the supremum is achieved at one of the vertices of C_0 or C_1. Thus

$$\sup_{x\in X}\langle \mathbf{1}_n, x\rangle = \max_{x\in C_0}\langle \mathbf{1}_n, x\rangle \vee \max_{x\in C_1}\langle \mathbf{1}_n, x\rangle = \max_{j\le n}\frac{1}{a_j} \vee \max_{m\le n}\frac{1}{Z(m)}\langle \mathbf{1}_m, b\rangle.$$

Substitute $1 = \max_{j\le n}\frac{1}{a_j}$, as $a_1 = 1$ and a_j are nondecreasing, then recognize that $b_1/\sqrt{b_1^2} = 1$ to obtain the lemma. □

We now give rough bounds on the widths of the set X and its hull.

Lemma G.2. *For the set $X = \mathrm{Conv}(C_0 \cup C_1)$, we have*

$$w^2(n) = \sup_{m\ge n}\frac{m-n}{\sum_{j=1}^{m} a_j^2}$$

and

$$w^2(n) \ge w_{\mathrm{nl}}^2(n) \ge \sup_{m\ge n}\frac{1}{Z(m)^2}\sum_{j=n+1}^{m} b_j^2 = \sup_{m\ge n}\frac{\sum_{j=n+1}^{m} b_j^2}{\sum_{j=1}^{m} a_j^2 b_j^2}.$$

Proof. For the linear width, we recognize that $Q = \{q \mid \sum_{j=1}^{\infty} a_j^2 q_j^2 \le 1\}$ is elliptical, so using the characterization (14) of $w^2(n)$ gives the first claim of the lemma. For the second we can take as a lower bound the nonlinear width of the set C_1, so that

$$w_{\mathrm{nl}}^2(n) \ge \sup_{x\in C_1}\left\{\sum_{j>n} x_{(j)}^2\right\} = \sup_{m\ge n}\left\{\frac{1}{Z(m)^2}\sum_{j=n+1}^{m} b_j^2\right\}$$

as desired. □

Finally, we take the scalars a_j, b_j as in the statement of Corollary 5.2. Set

$$a_j = j^{\alpha/2} \quad\text{and}\quad b_j^2 = 2^j,$$

where $0 < \alpha < 1$. Then direct calculations yield the asymptotics that for $m \ge n$,

$$\begin{aligned}
\sum_{j=1}^{m} a_j^2 = \sum_{j=1}^{m} j^\alpha &\in \left[\int_0^m t^\alpha dt, \int_1^{m+1} t^\alpha dt\right] \asymp m^{\alpha+1},\\
2^m m^\alpha = a_m^2 b_m^2 \le \sum_{j=1}^{m} a_j^2 b_j^2 &= 2^m m^\alpha \sum_{j=1}^{m}\left(\frac{j}{m}\right)^\alpha 2^{j-m} \le 2^m m^\alpha \sum_{j=0}^{m-1} 2^{-j} \le 2^{m+1} m^\alpha \qquad (27)\\
2^m = b_m^2 \le \sum_{j=n+1}^{m} b_j^2 &\le \sum_{j=1}^{m} 2^j \le 2^{m+1}.
\end{aligned}$$

The first equation in (27) implies that

$$w^2(n) = \sup_{m\ge n}\frac{m-n}{\sum_{j=1}^{m} a_j^2} \asymp \sup_{m\ge n}\frac{m-n}{m^{1+\alpha}} \asymp n^{-\alpha},$$

while the last two equations lower bound the nonlinear width (via Lemma G.2) by

$$w_{\mathrm{nl}}^2(n) \geq \sup_{m \geq n} \frac{\sum_{j=n+1}^m b_j^2}{\sum_{j=1}^m a_j^2 b_j^2} \geq \sup_{m \geq n} \frac{2^m}{2^m m^\alpha} = n^{-\alpha},$$

and so we have $w^2(n) \lesssim w_{\mathrm{nl}}^2(n) \leq w^2(n)$ for all n.

To prove Corollary 5.2, it remains to compute ℓ_1 diameter ratio. Applying Lemma G.1, for $\alpha < 1$ we have

$$\sup_{q \in \mathsf{QHull}(X)} \langle \mathbf{1}_n, q \rangle = \sqrt{\sum_{j=1}^n a_j^{-2}} = \sqrt{\sum_{j=1}^n j^{-\alpha}} \asymp n^{\frac{1-\alpha}{2}}.$$

On the other hand, because $\sum_{j=1}^m b_j \asymp 2^{m/2}$, the bounds (27) and Lemma G.1 give

$$\sup_{x \in X} \langle \mathbf{1}_n, x \rangle = \max_{m \leq n} \frac{\sum_{j=1}^m b_j}{\sqrt{\sum_{j=1}^m a_j^2 b_j^2}} \lesssim \max_{m \leq n} \frac{2^{m/2}}{m^{\alpha/2} 2^{m/2}} = 1.$$