Necessary and Sufficient Geometries for Gradient Methods

Daniel Levy
Stanford University
danilevy@stanford.edu

John C. Duchi
Stanford University
jduchi@stanford.edu

Abstract

We study the impact of the constraint set and gradient geometry on the convergence of online and stochastic methods for convex optimization, providing a characterization of the geometries for which stochastic gradient and adaptive gradient methods are (minimax) optimal. In particular, we show that when the constraint set is quadratically convex, diagonally pre-conditioned stochastic gradient methods are minimax optimal. We further provide a converse that shows that when the constraints are not quadratically convex—for example, any ℓ_p-ball for $p < 2$—the methods are far from optimal. Based on this, we can provide concrete recommendations for when one should use adaptive, mirror or stochastic gradient methods.

1 Introduction

We study stochastic and online convex optimization in the following setting: for a collection $\{F(\cdot, x), x \in \mathcal{X}\}$ of convex functions $F(\cdot, x) : \mathbf{R}^d \to \mathbf{R}$ and distribution P on $\mathcal{X}$, we wish to solve

$$\underset{\theta \in \Theta}{\text{minimize}} \ f_P(\theta) := \mathbf{E}_P\left[F(\theta, X)\right] = \int F(\theta, x) dP(x), \tag{1}$$

where $\Theta \subset \mathbf{R}^d$ is a closed convex set. The geometry of the underlying constraint set Θ and structure of subgradients $\partial F(\cdot, x)$ of course impact the performance of algorithms for problem (1). Thus, while stochastic subgradient methods are a *de facto* choice for their simplicity and scalability [22, 19, 5], their convergence guarantees depend on the ℓ_2-diameter of Θ and $\partial F(\cdot, x)$, so that for non-Euclidean geometries (e.g. when Θ is an ℓ_1-ball) one can obtain better convergence guarantees using mirror descent, dual averaging or the more recent adaptive gradient methods [18, 19, 4, 20, 13]. We revisit these ideas and precisely quantify optimal rates and gaps between the methods.

Our main contribution is to show that the geometry of the constraint set and gradients interact in a way completely analogous to Donoho et al.'s classical characterization of optimal estimation in Gaussian sequence models [10], where one observes a vector $\theta \in \Theta$ corrupted by Gaussian noise, $Y = \theta + \mathsf{N}(0, \sigma^2 I)$. For such problems, one can consider linear estimators—$\widehat{\theta} = AY$ for a $A \in \mathbf{R}^{d \times d}$—or potentially non-linear estimators—$\widehat{\theta} = \Phi(Y)$ where $\Phi : \mathbf{R}^d \to \Theta$. When Θ is quadratically convex, meaning the set $\Theta^2 := \{(\theta_j^2) \mid \theta \in \Theta\}$ is convex, Donoho et al. show there exists a minimax rate optimal linear estimator; conversely, there are non-quadratically convex Θ for which minimax rate optimal estimators $\widehat{\theta}$ must be nonlinear in Y.

To build our analogy, we turn to stochastic and online convex optimization. Consider Nesterov's dual averaging, where for a strongly convex $h : \Theta \to \mathbf{R}$, one iterates for $k = 1, 2, \ldots$ by receiving a (random) $X_k \in \mathcal{X}$, choosing $g_k \in \partial F(\theta_k, X_k)$, and for a stepsize $\alpha_k > 0$ updating

$$\theta_{k+1} := \underset{\theta \in \Theta}{\text{argmin}} \left\{ \sum_{i \le k} g_i^\top \theta + \frac{1}{\alpha_k} h(\theta) \right\}. \tag{2}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

When $\Theta = \mathbf{R}^d$ and h is Euclidean, that is, $h(\theta) = \frac{1}{2}\theta^\top A\theta$ for some $A \succ 0$, the updates are linear in the observed gradients g_i, as $\theta_k = -\alpha_k A^{-1}\sum_{i\le k} g_i$. Drawing a parallel between Φ in the Gaussian sequence model [10] and h in dual averaging (2), a natural conjecture is that a dichotomy similar to that for the Gaussian sequence model holds for stochastic and online convex optimization: if Θ is quadratically convex, there is a Euclidean h (yielding "linear" updates) that is minimax rate optimal, while there exist non-quadratically convex Θ for which Euclidean distance-generating h are arbitrarily suboptimal. We show that this analogy holds almost completely, with the caveat that we fully characterize minimax rates when the subgradients lie in a quadratically convex set or a weighted ℓ_r ball, $r \ge 1$. (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 compact, convex, quadratically convex, orthosymmetric constraint sets Θ, 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 $p \ge 2$, or hyperrectangles) a diagonal re-scaling suffices. This is important in machine learning problems of appropriate geometry, for example, in linear classification problems where the data (features) are sparse, so using a dense predictor θ is natural [13, 14]. Conversely, we show that if the constraint set Θ is a (scaled) ℓ_p ball, $1 \le p < 2$, then, considering unconstrained updates (2), the regret of the best method of linear type can be $\sqrt{d/\log d}$ times larger than the minimax rate. As part of this, we provide new information-theoretic lower bounds on optimization for general convex constraints Θ. In contrast to the frequent practice in literature of comparing regret upper bounds—prima facie illogical—we demonstrate the gap between linear and non-linear methods must hold.

Our conclusions relate to the growing literature in adaptive algorithms [3, 13, 21, 9]. Our results effectively prescribe that these adaptive algorithms are useful when the constraint set is quadratically convex as then there is a minimax optimal diagonal pre-conditioner. Even 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 $\sqrt{d}$ suboptimal on such problems. Conversely, our results strongly recommend against those methods for non-quadratically convex constraint sets. Our results thus clarify and explicate the work of Wilson et al. [28]: when the geometry of Θ and ∂F is appropriate for adaptive gradient methods or Euclidean algorithms, one should use them; when it is not—the constraints Θ are not quadratically convex—one should not.

Notation d always refers to dimension and n to sample size. For a norm $\gamma : \mathbf{R}^d \to \mathbf{R}_+$, $\mathbf{B}_\gamma(x_0, r) := \{x \mid \gamma(x - x_0) \le r\}$ 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 $\gamma^*(z) = \sup_{\gamma(x)\le 1} x^\top z$. For $\theta, \tau \in \mathbf{R}^d$, we abuse notation and define $\theta^2 := (\theta_j^2)_{j\le d}$, $|\theta| := (|\theta_j|)_{j\le d}$, $\frac{\theta}{\tau} := (\theta_j/\tau_j)_{j\le d}$ and $\theta \odot \tau := (\theta_j\tau_j)_{j\le d}$. The function $h : \mathbf{R}^d \to \mathbf{R}$ denotes a *distance generating function*, i.e. a function strongly convex w.r.t. a norm $\|\cdot\|$; $\mathrm{D}_h(x, y) = h(x) - h(y) - \nabla h(y)^\top(x - y)$ denotes the Bregman divergence, where h is strongly convex w.r.t. $\|\cdot\|$ if and only if $\mathrm{D}_h(x, y) \ge \frac{1}{2}\|x - y\|^2$. The subdifferential of $F(\cdot, x)$ at θ is $\partial_\theta F(\theta, x)$. $\mathsf{I}(X; Y)$ 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

We begin by defining the minimax framework in which we analyze procedures, review standard stochastic subgradient methods, and introduce the relevant geometric notions of convexity we require.

Minimax rate for convex stochastic optimization We measure the complexity of families of problems in two familiar ways: stochastic minimax complexity and regret [18, 1, 6]. Let $\Theta \subset \mathbf{R}^d$ be a closed convex set, $\mathcal{X}$ a sample space, and $\mathcal{F}$ a collection of functions $F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}$. For a collection $\mathcal{P}$ of distributions over $\mathcal{X}$, recall (1) that $f_P(\theta) := \int F(\theta, x)dP(x)$ is the expected loss of the point θ. Then the *minimax stochastic risk* is

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}, \mathcal{P}) := \inf_{\widehat{\theta}_n} \sup_{F\in\mathcal{F}} \sup_{P\in\mathcal{P}} \mathbf{E}\left[f_P(\widehat{\theta}_n(X_1^n)) - \inf_{\theta\in\Theta} f_P(\theta)\right],$$

where the expectation is taken over $X_1^n \stackrel{\text{iid}}{\sim} P$ and the infimum ranges over all measurable functions $\widehat{\theta}_n$ of $\mathcal{X}^n$. A related notion is the average *minimax regret*, which instead takes a supremum over samples $x_1^n \in \mathcal{X}^n$ and measures losses instantaneously. In this case, an algorithm consists of a sequence of decisions $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_n$, where $\widehat{\theta}_i$ is chosen conditional on samples x_1^{i-1}, so that

$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \mathcal{F}, \mathcal{X}) := \inf_{\widehat{\theta}_{1:n}} \sup_{F \in \mathcal{F}, x_1^n \in \mathcal{X}^n, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left[F\left(\widehat{\theta}_i \left(x_1^{i-1} \right), x_i \right) - F(\theta, x_i) \right].$$

In the regret case we may of course identify x_i with individual functions F, so this corresponds to the standard regret. In both of these definitions, we do not constrain the point estimates $\widehat{\theta}$ 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 [7], we always have $\mathfrak{M}_n^{\mathsf{S}} \le \mathfrak{M}_n^{\mathsf{R}}$; thus we typically provide lower bounds on $\mathfrak{M}_n^{\mathsf{S}}$ and upper bounds on $\mathfrak{M}_n^{\mathsf{R}}$.

We study functions whose continuity properties are specified by a norm γ over $\mathbf{R}^d$, defining

$$\mathcal{F}^{\gamma, r} := \left\{ F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R} \mid \text{for all } \theta \in \mathbf{R}^d, g \in \partial_\theta F(\theta, x), \gamma(g) \le r \right\}, \tag{3}$$

which is equivalent to the Lipschitz condition $|F(\theta, x) - F(\theta', x)| \le r\gamma^*(\theta - \theta')$, where γ^* is the dual norm to γ. For a given norm γ (γ as a mnemonic for gradient), we use the shorthands

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

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

Stochastic gradient methods, mirror descent, and regret Let us briefly review the canonical algorithms for solving the problem (1) and their associated convergence guarantees. For an algorithm outputing points $\theta_1, \ldots, \theta_n$, the *regret* on the sequence $F(\cdot, x_i)$ with respect to a point θ is

$$\mathsf{Regret}_n(\theta) := \sum_{i=1}^{n} [F(\theta_i, x_i) - F(\theta, x_i)].$$

Recalling the definition $\mathrm{D}_h(\theta, \theta_0) = h(\theta) - h(\theta_0) - \nabla h(\theta_0)^\top (\theta - \theta_0)$ of the Bregman divergence, the mirror descent algorithm [18, 4] iteratively sets

$$g_i \in \partial_\theta F(\theta_i, x_i) \text{ and updates } \theta_{i+1}^{\mathsf{MD}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ g_i^\top \theta + \frac{1}{\alpha} \mathrm{D}_h(\theta, \theta_i) \right\} \tag{4}$$

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 (4) and the iterates (2) of dual averaging satisfy (cf. [4, 6, 20])

$$\mathsf{Regret}_n(\theta) \le \frac{\mathrm{D}_h(\theta, \theta_0)}{\alpha} + \frac{\alpha}{2} \sum_{i \le n} \|g_i\|_*^2 \text{ for any } \theta \in \Theta. \tag{5}$$

One recovers the classical stochastic gradient method with the choice $h(\theta) = \frac{1}{2} \|\theta\|_2^2$, which is strongly convex with respect to the ℓ_2-norm, while the p-norm algorithms [15, 23], defined for $1 < p \le 2$, use $h(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2$, which is strongly convex with respect to the ℓ_p-norm $\|\cdot\|_p$.

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 Θ, which is equivalent to taking $\Theta = \mathbf{R}^d$ in the updates (4) or (2). The regret bound (5) still holds when considering unconstrained updates, whenever $\theta \in \Theta$, and the regret of the algorithm with respect to a constraint set Θ is simply $\sup_{\theta \in \Theta} \mathsf{Regret}_n(\theta)$. Even with unconstrained updates, the form (5) still captures small regret for all common constraint sets Θ [23]. To make clear, let $\Theta \subset \mathbf{R}^d$ be the ℓ_1-ball; taking $h(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2$ for $p = 1 + \frac{1}{\log(2d)}$, $q = \frac{p}{p-1} = 1 + \log(2d)$, and $\theta_0 = 0$ guarantees

$$\sup_{\|\theta\|_1 \le 1} \mathsf{Regret}_n(\theta) \le \frac{2}{\alpha} \sup_{\|\theta\|_1 \le 1} h(\theta) + \frac{\alpha}{2} \sum_{i \le n} \|g_i\|_q^2 \le \frac{2\log(2d)}{\alpha} + \frac{e^2 \alpha}{2} \sum_{i=1}^{n} \|g_i\|_\infty^2.$$

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

We frequently focus on distance generating functions of the form $h(\theta) = \frac{1}{2}\theta^\top A\theta$ 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**. It is important to note that, in this case, the mirror descent update is the stochastic gradient update with $A^{-1}g$, where g is a stochastic subgradient. We shall refer to all such methods as **methods of linear type**.

Quadratic convexity and orthosymmetry For a set Θ, we let $\Theta^2 := \{\theta^2, \theta \in \Theta\}$ denote its square. The set Θ is *quadratically convex* if Θ^2 is convex; typical examples of quadratically convex sets are weighted ℓ_p bodies for $p \geq 2$ or hyperrectangles. We let $\mathsf{QHull}(\Theta)$ be the quadratic convex hull of Θ, meaning the smallest convex and quadratically convex set containing Θ. The set $\Theta \subset \mathbf{R}^d$ is *orthosymmetric* if it is invariant to flipping the signs of any coordinate. Formally, if $\theta \in \Theta$ then $s \in \{\pm 1\}^d$ implies $(s_j\theta_j)_{j\leq d} \in \Theta$. We extend this notion to norms: we say that a norm γ is orthosymmetric if $\gamma(g) = \gamma(|g|)$ for all g. Similarly, we will say that a norm γ is quadratically convex if γ induces a quadratically convex unit ball.

3 Minimax optimality and quadratically convex constraint sets

We begin our contributions by considering quadratically convex constraint sets, providing lower bounds on the minimax risk and matching upper bounds on the minimax regret of convex optimization over such sets. We further show that these are attained by diagonally-scaled gradient methods. 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 Θ), 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 and preview our contributions, we begin our study with the familiar case of $\Theta = \mathbf{B}_p(0,1)$ and norm on the subgradients $\gamma = \|\cdot\|_r$ (mnemonically, γ for gradients), with $p \in [2,\infty]$ (so that Θ is quadratically convex) and $r \geq 1$. 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 $p \geq 2$

The results for the basic case that the constraints Θ are 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 B.1 and B.2, respectively.

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

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

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

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

In both cases, the stochastic gradient method achieves the regret upper bound via a straightforward optimization of the regret bounds (5) with $h(\theta) = \frac{1}{2}\|\theta\|_2^2$. That is, a method of linear type is optimal.

3.2 General quadratically convex constraints

We now turn to the more general case that Θ is an arbitrary convex, compact, quadratically convex and orthosymmetric set. We combine two techniques to develop the results. The first essentially builds out of the ideas of Donoho et al. [10] in Gaussian sequence estimation, which shows that the largest hyperrectangle in Θ 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 Θ—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 1 and 2).

We begin with the lower bound, which relies on rectangular structures in the primal Θ and dual gradient spaces. For the proposition, we use a specialization of the function families (3) to rectangular sets, where for $M \in \mathbf{R}_+^d$ we define

$$\mathcal{F}^M := \left\{ F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R} \mid \text{for all } \theta \in \mathbf{R}^d, g \in \partial_\theta f(\theta, x), \max_{j \le d} \frac{|g_j|}{M_j} \le 1 \right\}.$$

Proposition 3 (Duchi et al. [14], Proposition 1)**.** *Let $M \in \mathbf{R}_+^d$ and $\mathcal{F}^M$ be as above. Let $a \in \mathbf{R}_+^d$ and assume the hyperrectangular containment $\prod_{j=1}^d [-a_j, a_j] \subset \Theta$. Then*

$$\mathfrak{M}_n^{\mathsf{S}}(\Theta, \mathcal{F}^M) \ge \frac{1}{8\sqrt{n \log 3}} \sum_{j=1}^d M_j a_j.$$

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

As we discuss in Section 2, diagonally-scaled gradient methods (componentwise re-scaling of the subgradients) are equivalent to using $h_\Lambda(\theta) := \frac{1}{2}\theta^\top \Lambda \theta$ for $\Lambda = \operatorname{diag}(\lambda) \succeq 0$ in the mirror descent update (4). In this case, for any norm γ on the gradients, the minimax regret bound (5) becomes

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

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

$$\mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \le \frac{1}{2n} \inf_{\lambda \succeq 0} \sup_{\theta \in \Theta} \sup_{g \in \mathbf{B}_\gamma(0,1)} \left[\sum_{j \le d} \lambda_j \theta_j^2 + n \sum_{j \le d} \frac{1}{\lambda_j} g_j^2\right] \tag{6}$$

The regret bound (6) holds without assumptions on Θ or γ. However, in the case when Θ is quadratically convex, strong duality allows us to simplify this quantity:

Proposition 4. *Let $V, \Theta \subset \mathbf{R}^d$ be convex, quadratically convex and compact sets. Then*

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

Proof. The quadratic convexity of the sets Θ and V implies that a (weighted) squared 2-norm becomes a linear functional when lifted to the squared sets $\Theta^2 := \{\theta^2 \mid \theta \in \Theta\}$ and V^2. Indeed, defining $J : \mathbf{R}_+^{2d} \times \mathbf{R}_+^d \to \mathbf{R}$, $J(\tau, w, \lambda) := \lambda^\top \tau + (\frac{1}{\lambda})^\top w$, the function J is concave-convex: it is linear (a fortiori concave) in (τ, w) and convex in λ. Thus, using that the set $\{\lambda \in \mathbf{R}_+^d\}$ is convex and $\Theta^2 \times V^2$ is convex compact (because Θ and V are quadratically convex compact), Sion's minimax theorem [25] implies

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

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

Proposition 4 provides a powerful hammer for diagonally scaled Euclidean optimization algorithms, as we can choose an optimal scaling for any *fixed* pair θ, g, taking a worst case over such pairs:

Corollary 1. *Let Θ be a convex, quadratically convex, compact set. Then*

$$\mathfrak{M}_n^{\mathsf{R}}(\theta, \gamma) \le \frac{1}{\sqrt{n}} \sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1)), \theta \in \Theta} \theta^\top g,$$

and diagonally-scaled gradient methods achieves this regret.

Proof. We upper bound the minimax regret (6) 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 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 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 $\gamma(s \odot v) = \gamma(v)$ for all $s \in \{\pm 1\}^d$ 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. [10] 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 Θ be quadratically convex, orthosymmetric, and compact. Then*

$$\frac{1}{8\sqrt{\log 3}} \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \gamma^*(\theta) \le \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \le \mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \le \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \gamma^*(\theta).$$

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

We present the proof in Appendix C.1.

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 slightly less general. Nonetheless, when γ is a weighted ℓ_r-norm ball (for $r \in [1,2]$), diagonally-scaled gradient methods are minimax rate optimal, as Corollary 2 will show; when the norms γ are arbitrary we have a slightly more complex result.

Theorem 2. *Let Θ be an orthosymmetric, quadratically convex, convex and compact set and γ an arbitrary norm. Recall the definition $(\frac{\theta}{\gamma(e.)})_j = \theta_j / \gamma(e_j)$. Then for any $k \in \mathbf{N}$,*

$$\begin{aligned} \frac{1}{8\sqrt{n \log 3}} \left(1 - \frac{k}{n \log 3}\right) \sup_{\theta \in \Theta, \|\theta\|_0 \le k} \left\| \frac{\theta}{\gamma(e.)} \right\|_2 &\le \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \\ &\le \mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \le \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \sup_{g \in \mathsf{QHull}(\mathbf{B}_\gamma(0,1))} \theta^\top g. \end{aligned} \tag{7}$$

Corollary 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 C.2 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, meaning the results of the previous section do not apply, the previous theorem still guarantees that diagonally-scaled gradient methods are minimax rate optimal.

Corollary 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 $n \geq 2d$,*

$$\frac{1}{16} \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \left\| \frac{\theta}{\gamma(e.)} \right\|_2 \leq \mathfrak{M}_n^{\mathsf{S}}(\Theta, \gamma) \leq \mathfrak{M}_n^{\mathsf{R}}(\Theta, \gamma) \leq \frac{1}{\sqrt{n}} \sup_{\theta \in \Theta} \left\| \frac{\theta}{\gamma(e.)} \right\|_2 .$$

There exists $\lambda^* \in \mathbf{R}_+^d$ 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 C.3 for the (short) precise proof.

Theorem 1 and Corollary 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 constraint sets 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 [17]). Unsurprisingly, methods of linear type retain their optimality properties.

Corollary 3. *Let Θ_0 be a compact, orthosymmetric, convex and quadratically convex set. Let $U \in \mathcal{O}_n(\mathbf{R})$ be a rotation matrix and $\Theta := U\Theta_0 = \{U\theta \mid \theta \in \Theta_0\}$. Consider the collection*

$$\mathcal{F} := \{F : \mathbf{R}^d \times \mathcal{X} \to \mathbf{R}, \forall x \in \mathcal{X}, \forall \theta \in \mathbf{R}^d, \forall g \in \partial_\theta f(\theta, x), \gamma(U^T g) \leq 1\}.$$

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

We present the proof in Appendix C.4.

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

For $\Theta \subset \mathbf{R}^d$ 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 a collection of problem instances where the constraint set is orthosymmetric, compact, and convex but not quadratically convex. On such problems, the constraint set has substantial consequences; for some non-quadratically convex sets Θ, methods of linear type (e.g. the stochastic gradient method) can be minimax rate-optimal, while for other constraint sets, all methods of linear type must have regret at least a factor $\sqrt{d/\log d}$ worse than the minimax optimal rate, which (non-linear) mirror descent with appropriate distance generating function achieves.

To construct these problem instances, we turn to simple non-quadratically convex constraint sets: ℓ_p balls for $p \in [1, 2]$. We measure subgradient norms in the dual ℓ_{p^*} norm, $p^* = \frac{p-1}{p}$. 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. (5) with $h(\theta) = \frac{1}{2}\|\theta\|_2^2$. 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 $\sqrt{d/\log d}$.

4.1 Minimax rates for p-norm constraint sets, $p \in [1, 2]$

For $p \in [1, 2]$, we consider the constraint set $\Theta = \mathbf{B}_p(0, 1)$ 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* $p \in [1,2]$, $\Theta = \mathbf{B}_p(0,1)$ *and* $\gamma = \|\cdot\|_{p^*}$.

(i) If $1 \le p \le 1 + 1/\log(2d)$, *then*

$$1 \wedge \sqrt{\frac{\log(2d)}{n}} \lesssim \mathfrak{M}_n^{\mathsf{S}}(\Theta,\gamma) \le \mathfrak{M}_n^{\mathsf{R}}(\Theta,\gamma) \lesssim 1 \wedge \sqrt{\frac{\log(2d)}{n}}.$$

Mirror descent (4) *with* $h(\theta) := \frac{1}{2(a-1)}\|\theta\|_a^2$ *for* $a = 1 + \frac{1}{\log(2d)}$ *achieves the optimal rate.*

(ii) If $1 + 1/\log(2d) < p \le 2$, *then*

$$1 \wedge \sqrt{\frac{1}{n(p-1)}} \lesssim \mathfrak{M}_n^{\mathsf{S}}(\Theta,\gamma) \le \mathfrak{M}_n^{\mathsf{R}}(\Theta,\gamma) \lesssim 1 \wedge \sqrt{\frac{1}{n(p-1)}}.$$

Mirror descent with $h(\theta) := \frac{1}{2(p-1)}\|\theta\|_p^2$ *achieves the optimal rate.*

To prove the theorem, we upper bound the regret of mirror descent with norm-based distance generating functions (cf. [24, Corollary 2.18]), which follows immediately from the regret bound (5).

Proposition 5. *Let* Θ *be closed convex,* γ *a norm, and* $1 < a \le 2$, $a^* = \frac{a}{a-1}$. *Mirror descent with distance generating function* $h(\theta) := \frac{1}{2(a-1)}\|\theta\|_a^2$ *and stepsize* $\alpha = \frac{\sup_{\theta\in\Theta}\|\theta-\theta_0\|_a}{\sqrt{n}\sup_{g\in\mathbf{B}_\gamma(0,1)}\|g\|_{a^*}}$ *achieves regret*

$$\mathfrak{M}_n^{\mathsf{R}}(\Theta,\gamma) \le \frac{\sup_{\theta\in\Theta}\|\theta\|_a \sup_{g\in\mathbf{B}_\gamma(0,1)}\|g\|_{a^*}}{\sqrt{n(a-1)}}.$$

We present the full proof of Theorem 3 in Appendix D.1. We obtain the lower bound with the familiar reduction from estimation to testing and Assouad's method (see Appendix A.2).

4.2 Exhibiting hard problems for Euclidean gradient methods

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_{p^*}-norm ball ($p^* = \frac{p}{p-1}$). For problems and p, standard subgradient methods achieve worst-case regret $O(d^{1/2-1/p^*}/\sqrt{n})$. This 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 is at least this familiar upper bound of standard subgradient methods, precisely quantifying the gap between linear and non-linear methods for this problem class.

Theorem 4. *Let* $\mathsf{Regret}_{n,A}(\theta) = \sum_{i=1}^n g_i^\top(\theta_i - \theta)$ *denote the regret of the (Euclidean) online mirror descent method with distance generating function* $h_A(\theta) = \frac{1}{2}\theta^\top A\theta$ *for linear functions* $F_i(\theta) = g_i^\top\theta$. *For any* $A \succeq 0$ *and* $p \in [1,2]$ *with* $q = \frac{p}{p-1}$, *there exists a sequence of vectors* $g_i \in \mathbf{R}^d$, $\|g_i\|_q \le 1$, *and point* $\theta \in \mathbf{R}^d$ *with* $\|\theta\|_p \le 1$ *such that*

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

We provide the proof in Appendix D.2. 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 5 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 has regret at least $\Omega(1)\sqrt{d/n}$. On the same problem, non-linear mirror descent has regret at most $O(1)\sqrt{\log d/n}$, showing the advertised $\sqrt{d/\log d}$ gap. When $p \ge 2 - 1/\log d$ (so Θ is nearly quadratically convex), the gap reduces to at most a constant factor.

5 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, $\{g_l\}_{l\leq i}$. The algorithm then updates the iterate based on the distance generating function $h_i(\theta) := \frac{1}{2}\theta^\top \Lambda_i \theta$. 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 [13] achieves the minimax optimal regret while standard (non-adaptive) subgradient methods can be $\sqrt{d}$ suboptimal on the same problem.

We consider the following setting: $\Theta = \mathbf{B}_\infty(0,1)$ and $\gamma_\beta(g) = \|\beta \odot g\|_1$, for an arbitrary $\beta \in \mathbf{R}^d, \beta \succ 0$. Intuitively, β_j corresponds to the "scale" of the j-th dimension. On this problem, a straightforward optimization of the regret bound (5) guarantees that stochastic gradient methods achieve regret $\sqrt{dn}/\min_j \beta_j$. We exhibit a problem instance (in Appendix E) such that, for any stepsize α, online gradient descent attains this worst-case regret.

Theorem 5. *Let* $\mathsf{Regret}_{n,\alpha}(\theta) = \sum_{i\leq n} g_i^\top(\theta_i - \theta)$ *denote the regret of the online gradient descent method with stepsize* $\alpha \geq 0$ *for linear functions* $F_i(\theta) = g_i^\top \theta$*. For any choice of* $\alpha \geq 0$ *and* $\beta \succ 0$*, there exists a sequence of vectors* $\{g_i\}_{i\leq n} \subset \mathbf{R}^d$*,* $\gamma_\beta(g_i) \leq 1$ *and point* $\theta \in \Theta$ *such that*

$$\mathsf{Regret}_{n,\alpha}(\theta) \geq \frac{1}{2}\min\left\{\frac{dn}{2\|\beta\|_1}, \frac{\sqrt{2dn}}{\min_{j\leq d}\beta_j}\right\}.$$

In contrast, AdaGrad [13] achieves regret $\sqrt{n}\,\|1/\beta\|_2$, demonstrating suboptimality gap as large as $\sqrt{d}$ for some choices of β. Indeed, let $\mathsf{Regret}_{n,\mathsf{AdaGrad}}(\theta)$ be the regret of AdaGrad. Then

$$\mathsf{Regret}_{n,\mathsf{AdaGrad}}(\theta) \leq 2\sqrt{2}\sum_{j\leq d}\sqrt{\sum_{i\leq n} g_{i,j}^2}.$$

(see [13, Corollary 6]), and by Cauchy-Schwarz,

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

To concretely consider different scales across dimensions, we choose $\beta_j = j$. Theorem 5 guarantees that there exists a collection of linear functions such that stochastic gradient methods suffer regret $\Omega(1)\sqrt{dn}$. Given that $\|1/\beta\|_2 \leq \sqrt{\zeta(2)} \leq \pi/\sqrt{6}$, AdaGrad achieves regret $O(1)\sqrt{n}$—amounting to a suboptimality gap of order $\sqrt{d}$—exhibiting the need for adaptivity. This $\sqrt{d}$ gap is also the largest possible over subgradient methods, which may achieve regret $\sqrt{d\sum_{i\leq n}\|g_i\|_2^2} \leq \sqrt{d}\sum_{j\leq d}\sqrt{\sum_{i\leq n} g_{i,j}^2}$ for $\Theta = \mathbf{B}_\infty(0,1)$. Finally, we note in passing that AdaGrad is minimax optimal on this class of problems via a straightforward application of Theorem 1.

6 Discussion

In this paper, we provide concrete recommendations for when one should use adaptive, mirror or standard gradient methods depending on the geometry of the problem. While we emphasize 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? Finally, we close with a parting message about the value and costs of adaptive and related methods. One should turn to adaptive gradient methods (at most) in settings where methods of linear type are optimal. It is as our mothers told us when we were children: if you want steak, don't order chicken.

Acknowledgments This work was supported by NSF-CAREER Award 1553086, ONR-YIP N00014-19-1-2288, and the Stanford DAWN Project. We thank Aditya Grover, Annie Marsden and Hongseok Namkoong for valuable comments on the draft 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] P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20*, 2007.

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

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

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

[7] 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.

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

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

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

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

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

[13] 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.

[14] 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.

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

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

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

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

[19] 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.

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

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

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

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

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

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

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

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

[28] 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.

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