Finite Sample Convergence Rates of Zero-Order Stochastic Optimization Methods

John C. Duchi[1] **Michael I. Jordan**[1,2] **Martin J. Wainwright**[1,2] **Andre Wibisono**[1]
[1]Department of Electrical Engineering and Computer Science and [2]Department of Statistics
University of California, Berkeley
Berkeley, CA USA 94720
{jduchi,jordan,wainwrig,wibisono}@eecs.berkeley.edu

Abstract

We consider derivative-free algorithms for stochastic optimization problems that use only noisy function values rather than gradients, analyzing their finite-sample convergence rates. We show that if pairs of function values are available, algorithms that use gradient estimates based on random perturbations suffer a factor of at most $\sqrt{d}$ in convergence rate over traditional stochastic gradient methods, where d is the problem dimension. We complement our algorithmic development with information-theoretic lower bounds on the minimax convergence rate of such problems, which show that our bounds are sharp with respect to all problem-dependent quantities: they cannot be improved by more than constant factors.

1 Introduction

Derivative-free optimization schemes have a long history in optimization (see, for example, the book by Spall [21]), and they have the clearly desirable property of never requiring explicit gradient calculations. Classical techniques in stochastic and non-stochastic optimization, including Kiefer-Wolfowitz-type procedures [e.g. 17], use function difference information to approximate gradients of the function to be minimized rather than calculating gradients. Researchers in machine learning and statistics have studied online convex optimization problems in the bandit setting, where a player and adversary compete, with the player choosing points θ in some domain Θ and an adversary choosing a point x, forcing the player to suffer a loss $F(\theta; x)$, where $F(\cdot; x) : \Theta \to \mathbb{R}$ is a convex function [13, 5, 1]. The goal is to choose optimal θ based only on observations of function values $F(\theta; x)$. Applications including online auctions and advertisement selection in search engine results. Additionally, the field of simulation-based optimization provides many examples of problems in which optimization is performed based only on function values [21, 10], and problems in which the objective is defined variationally (as the maximum of a family of functions), such as certain graphical model and structured-prediction problems, are also natural because explicit differentiation may be difficult [23].

Despite the long history and recent renewed interest in such procedures, an understanding of their finite-sample convergence rates remains elusive. In this paper, we study algorithms for solving stochastic convex optimization problems of the form

$$\min_{\theta \in \Theta} f(\theta) := \mathbb{E}_P[F(\theta; X)] = \int_{\mathcal{X}} F(\theta; x) dP(x), \tag{1}$$

where $\Theta \subseteq \mathbb{R}^d$ is a compact convex set, P is a distribution over the space $\mathcal{X}$, and for P-almost every $x \in \mathcal{X}$, the function $F(\cdot; x)$ is closed convex. Our focus is on the convergence rates of algorithms that observe only stochastic realizations of the function values $f(\theta)$.

Work on this problem includes Nemirovski and Yudin [18, Chapter 9.3], who develop a randomized sampling strategy that estimates $\nabla F(\theta; x)$ using samples from the surface of the ℓ_2-sphere, and

Flaxman et al. [13], who build on this approach, applying it to bandit convex optimization problems. The convergence rates in these works are (retrospectively) sub-optimal [20, 2]: Agarwal et al. [2] provide algorithms that achieve convergence rates (ignoring logarithmic factors) of $\mathcal{O}(\mathrm{poly}(d)/\sqrt{k})$, where $\mathrm{poly}(d)$ is a polynomial in the dimension d, for stochastic algorithms receiving only single function values, but (as the authors themselves note) the algorithms are quite complicated.

Some of the difficulties inherent in optimization using only a single function evaluation can be alleviated when the function $F(\cdot; x)$ can be evaluated at *two* points, as noted independently by Agarwal et al. [1] and Nesterov [20]. The insight is that for small u, the quantity $(F(\theta + uZ; x) - F(\theta; x))/u$ approximates a directional derivative of $F(\theta; x)$ and can thus be used in first-order optimization schemes. Such two-sample-based gradient estimators allow simpler analyses, with sharper convergence rates [1, 20], than algorithms that have access to only a single function evaluation in each iteration. In the current paper, we take this line of work further, finding the *optimal* rate of convergence for procedures that are only able to obtain function evaluations, $F(\cdot; X)$, for samples X. Moreover, adopting the two-point perspective, we present simple randomization-based algorithms that achieve these optimal rates.

More formally, we study algorithms that receive paired observations $Y(\theta, \tau) \in \mathbb{R}^2$, where θ and τ are points the algorithm selects, and the tth sample is

$$Y^t(\theta^t, \tau^t) := \begin{bmatrix} F(\theta^t; X^t) \\ F(\tau^t; X^t) \end{bmatrix} \tag{2}$$

where X^t is a sample drawn from the distribution P. After k iterations, the algorithm returns a vector $\widehat{\theta}(k) \in \Theta$. In this setting, we analyze stochastic gradient and mirror-descent procedures [27, 18, 6, 19] that construct gradient estimators using the two-point observations Y^t. By a careful analysis of the dimension dependence of certain random perturbation schemes, we show that the convergence rate attained by our stochastic gradient methods is roughly a factor of $\sqrt{d}$ worse than that attained by stochastic methods that observe the full gradient $\nabla F(\theta; X)$. Under appropriate conditions, our convergence rates are a factor of $\sqrt{d}$ better than those attained by Agarwal et al. [1] and Nesterov [20]. In addition, though we present our results in the framework of stochastic optimization, our analysis applies to (two-point) bandit online convex optimization problems [13, 5, 1], and we consequently obtain the sharpest rates for such problems. Finally, we show that the convergence rates we provide are tight—meaning sharp to within constant factors—by using information-theoretic techniques for constructing lower bounds on statistical estimators.

2 Algorithms

Stochastic mirror descent methods are a class of stochastic gradient methods for solving the problem $\min_{\theta \in \Theta} f(\theta)$. They are based on a proximal function ψ, which is a differentiable convex function defined over Θ that is assumed (w.l.o.g. by scaling) to be 1-strongly convex with respect to the norm $\|\cdot\|$ over Θ. The proximal function defines a Bregman divergence $D_\psi : \Theta \times \Theta \to \mathbb{R}_+$ via

$$D_\psi(\theta, \tau) := \psi(\theta) - \psi(\tau) - \langle \nabla\psi(\tau), \theta - \tau \rangle \geq \frac{1}{2} \|\theta - \tau\|^2 , \tag{3}$$

where the inequality follows from the strong convexity of ψ over Θ. The mirror descent (MD) method proceeds in a sequence of iterations that we index by t, updating the parameter vector $\theta^t \in \Theta$ using stochastic gradient information to form θ^{t+1}. At iteration t the MD method receives a (subgradient) vector $g^t \in \mathbb{R}^d$, which it uses to update θ^t via

$$\theta^{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \langle g^t, \theta \rangle + \frac{1}{\alpha(t)} D_\psi(\theta, \theta^t) \right\} , \tag{4}$$

where $\{\alpha(t)\}$ is a non-increasing sequence of positive stepsizes.

We make two standard assumptions throughout the paper. Let θ^* denote a minimizer of the problem (1). The first assumption [18, 6, 19] describes the properties of ψ and the domain.

Assumption A. *The proximal function ψ is strongly convex with respect to the norm $\|\cdot\|$. The domain Θ is compact, and there exists $R < \infty$ such that $D_\psi(\theta^*, \theta) \leq \frac{1}{2} R^2$ for $\theta \in \Theta$.*

Our second assumption is standard for almost all first-order stochastic gradient methods [19, 24, 20], and it holds whenever the functions $F(\cdot; x)$ are G-Lipschitz with respect to the norm $\|\cdot\|$. We use $\|\cdot\|_*$ to denote the dual norm to $\|\cdot\|$, and let $\mathsf{g} : \Theta \times \mathcal{X} \to \mathbb{R}^d$ denote a measurable subgradient selection for the functions F; that is, $\mathsf{g}(\theta; x) \in \partial F(\theta; x)$ with $\mathbb{E}[\mathsf{g}(\theta; X)] \in \partial f(\theta)$.

Assumption B. *There is a constant $G < \infty$ such that the (sub)gradient selection g satisfies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_*^2] \le G^2$ for $\theta \in \Theta$.*

When Assumptions A and B hold, the convergence rate of stochastic mirror descent methods is well understood [6, 19, Section 2.3]. Indeed, let the variables $X^t \in \mathcal{X}$ be sampled i.i.d. according to P, set $g^t = \mathsf{g}(\theta^t; X^t)$, and let θ^t be generated by the mirror descent iteration (4) with stepsize $\alpha(t) = \alpha/\sqrt{t}$. Then one obtains

$$\mathbb{E}[f(\widehat{\theta}(k))] - f(\theta^*) \le \frac{1}{2\alpha\sqrt{k}} R^2 + \frac{\alpha}{\sqrt{k}} G^2. \tag{5}$$

For the remainder of this section, we explore the use of function difference information to obtain subgradient estimates that can be used in mirror descent methods to achieve statements similar to the convergence guarantee (5).

2.1 Two-point gradient estimates and general convergence rates

In this section, we show—under a reasonable additional assumption—how to use two samples of the random function values $F(\theta; X)$ to construct nearly unbiased estimators of the gradient $\nabla f(\theta)$ of the expected function f. Our analytic techniques are somewhat different than methods employed in past work [1, 20]; as a consequence, we are able to achieve optimal dimension dependence.

Our method is based on an estimator of $\nabla f(\theta)$. Our algorithm uses a non-increasing sequence of positive smoothing parameters $\{u_t\}$ and a distribution μ on $\mathbb{R}^d$ (which we specify) satisfying $\mathbb{E}_\mu[ZZ^\top] = I$. Upon receiving the point $X^t \in \mathcal{X}$, we sample an independent vector Z^t and set

$$g^t = \frac{F(\theta^t + u_t Z^t; X^t) - F(\theta^t; X^t)}{u_t} Z^t. \tag{6}$$

We then apply the mirror descent update (4) to the quantity g^t.

The intuition for the estimator (6) of $\nabla f(\theta)$ follows from an understanding of the directional derivatives of the random function realizations $F(\theta; X)$. The directional derivative $f'(\theta, z)$ of the function f at the point θ in the direction z is $f'(\theta, z) := \lim_{u \downarrow 0} \frac{f(\theta+uz)-f(\theta)}{u}$. The limit always exists when f is convex [15, Chapter VI], and if f is differentiable at θ, then $f'(\theta, z) = \langle \nabla f(\theta), z \rangle$. In addition, we have the following key insight (see also Nesterov [20, Eq. (32)]): whenever $\nabla f(\theta)$ exists,

$$\mathbb{E}[f'(\theta, Z)Z] = \mathbb{E}[\langle \nabla f(\theta), Z \rangle Z] = \mathbb{E}[ZZ^\top \nabla f(\theta)] = \nabla f(\theta)$$

if the random vector $Z \in \mathbb{R}^d$ has $\mathbb{E}[ZZ^\top] = I$. Intuitively, for u_t small enough in the construction (6), the vector g^t should be a nearly unbiased estimator of the gradient $\nabla f(\theta)$.

To formalize our intuition, we make the following assumption.

Assumption C. *There is a function $L : \mathcal{X} \to \mathbb{R}_+$ such that for (P-almost every) $x \in \mathcal{X}$, the function $F(\cdot; x)$ has $L(x)$-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$, and the quantity $L(P)^2 := \mathbb{E}[L(X)^2] < \infty$.*

With Assumption C, we can show that g^t is (nearly) an unbiased estimator of $\nabla f(\theta^t)$. Furthermore, for appropriate random vectors Z, we can also show that g^t has small norm, which yields better convergence rates for mirror descent-type methods. (See the proof of Theorem 1.) In order to study the convergence of mirror descent methods using the estimator (6), we make the following additional assumption on the distribution μ.

Assumption D. *Let Z be sampled according to the distribution μ, where $\mathbb{E}[ZZ^\top] = I$. The quantity $M(\mu)^2 := \mathbb{E}[\|Z\|^4 \|Z\|_*^2] < \infty$, and there is a constant $s(d)$ such that for any vector $g \in \mathbb{R}^d$, $\mathbb{E}[\|\langle g, Z \rangle Z\|_*^2] \le s(d) \|g\|_*^2$.*

As the next theorem shows, Assumption D is somewhat innocuous, the constant $M(\mu)$ not even appearing in the final bound. The dimension (and norm) dependent term $s(d)$, however, is important for our results. In Section 2.2 we give explicit constructions of random variables that satisfy Assumption D. For now, we present the following result.

Theorem 1. *Let $\{u_t\} \subset \mathbb{R}_+$ be a non-increasing sequence of positive numbers, and let θ^t be generated according to the mirror descent update (4) using the gradient estimator (6). Under Assumptions A, B, C, and D, if we set the step and perturbation sizes*

$$\alpha(t) = \alpha \frac{R}{2G\sqrt{s(d)}\sqrt{t}} \quad \text{and} \quad u_t = u \frac{G\sqrt{s(d)}}{L(P)M(\mu)} \cdot \frac{1}{t},$$

then

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq 2\frac{RG\sqrt{s(d)}}{\sqrt{k}} \max\left\{\alpha, \alpha^{-1}\right\} + \alpha u^2 \frac{RG\sqrt{s(d)}}{k} + u\frac{RG\sqrt{s(d)}\log k}{k},$$

where $\widehat{\theta}(k) = \frac{1}{k}\sum_{t=1}^k \theta^t$, and the expectation is taken with respect to the samples X and Z.

The proof of Theorem 1 requires some technical care—we never truly receive unbiased gradients—and it builds on convergence proofs developed in the analysis of online and stochastic convex optimization [27, 19, 1, 12, 20] to achieve bounds of the form (5). Though we defer proof to Appendix A.1, at a very high level, the argument is as follows. By using Assumption C, we see that for small enough u_t, the gradient estimator g^t from (6) is close (in expectation with respect to X^t) to $f'(\theta^t, Z^t)Z^t$, which is an unbiased estimate of $\nabla f(\theta^t)$. Assumption C allows us to bound the moments of the gradient estimator g^t. By carefully showing that taking care to make sure that the errors in g^t as an estimator of $\nabla f(\theta^t)$ scale with u_t, we given an analysis similar to that used to derive the bound (5) to obtain Theorem 1.

Before continuing, we make a few remarks. First, the method is reasonably robust to the selection of the step-size multiplier α (as noted by Nemirovski et al. [19] for gradient-based MD methods). So long as $\alpha(t) \propto 1/\sqrt{t}$, mis-specifying the multiplier α results in a scaling at worst linear in $\max\{\alpha, \alpha^{-1}\}$. Perhaps more interestingly, our setting of u_t was chosen mostly for convenience and elegance of the final bound. In a sense, we can simply take u to be extremely close to zero (in practice, we must avoid numerical precision issues, and the stochasticity in the method makes such choices somewhat unnecessary). In addition, the convergence rate of the method is independent of the Lipschitz continuity constant $L(P)$ of the instantaneous gradients $\nabla F(\cdot; X)$; the penalty for nearly non-smooth objective functions comes into the bound only as a second-order term. This suggests similar results should hold for non-differentiable functions; we have been able to show that in some cases this is true, but a fully general result has proved elusive thus far. We are currently investigating strategies for the non-differentiable case.

Using similar arguments based on Azuma-Hoeffding-type inequalities, it is possible to give high-probability convergence guarantees [cf. 9, 19] under additional tail conditions on g, for example, that $\mathbb{E}[\exp(\|\mathsf{g}(\theta; X)\|_*^2 / G^2)] \leq \exp(1)$. Additionally, though we have presented our results as convergence guarantees for stochastic optimization problems, an inspection of our analysis in Appendix A.1 shows that we obtain (expected) regret bounds for bandit online convex optimization problems [e.g. 13, 5, 1].

2.2 Examples and corollaries

In this section, we provide examples of random sampling strategies that give direct convergence rate estimates for the mirror descent algorithm with subgradient samples (6). For each corollary, we specify the norm $\|\cdot\|$, proximal function ψ, and distribution μ, verify that Assumptions A, B, C, and D hold, and then apply Theorem 1 to obtain a convergence rate.

We begin with a corollary that describes the convergence rate of our algorithm when the expected function f is Lipschitz continuous with respect to the Euclidean norm $\|\cdot\|_2$.

Corollary 1. *Given the proximal function $\psi(\theta) := \frac{1}{2}\|\theta\|_2^2$, suppose that $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_2^2] \leq G^2$ and that μ is uniform on the surface of the ℓ_2-ball of radius $\sqrt{d}$. With the step size choices in Theorem 1,*

we have

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq 2\frac{RG\sqrt{d}}{\sqrt{k}}\max\{\alpha, \alpha^{-1}\} + \alpha u^2\frac{RG\sqrt{d}}{k} + u\frac{RG\sqrt{d}\log k}{k}.$$

Proof Note that $\|Z\|_2 = \sqrt{d}$, which implies $M(\mu)^2 = \mathbb{E}[\|Z\|_2^6] = d^3$. Furthermore, it is easy to see that $\mathbb{E}[ZZ^\top] = I$. Thus, for $g \in \mathbb{R}^d$ we have

$$\mathbb{E}[\|\langle g, Z\rangle Z\|_2^2] = d\mathbb{E}[\langle g, Z\rangle^2] = d\mathbb{E}[g^\top ZZ^\top g] = d\|g\|_2^2,$$

which gives us $s(d) = d$. □

The rate provided by Corollary 1 is the fastest derived to date for zero-order stochastic optimization using two function evaluations. Both Agarwal et al. [1] and Nesterov [20] achieve rates of convergence of order $RGd/\sqrt{k}$. Admittedly, neither requires that the random functions $F(\cdot; X)$ be continuously differentiable. Nonetheless, Assumption C does not require a uniform bound on the Lipschitz constant $L(X)$ of the gradients $\nabla F(\cdot; X)$; moreover, the convergence rate of the method is essentially independent of $L(P)$.

In high-dimensional scenarios, appropriate choices for the proximal function ψ yield better scaling on the norm of the gradients [18, 14, 19, 12]. In online learning and stochastic optimization settings where one observes gradients $\mathsf{g}(\theta; X)$, if the domain Θ is the simplex, then exponentiated gradient algorithms [16, 6] using the proximal function $\psi(\theta) = \sum_j \theta_j \log\theta_j$ obtain rates of convergence dependent on the ℓ_∞-norm of the gradients $\|\mathsf{g}(\theta; X)\|_\infty$. This scaling is more palatable than dependence on Euclidean norms applied to the gradient vectors, which may be a factor of $\sqrt{d}$ larger. Similar results apply [7, 6] when using proximal functions based on ℓ_p-norms. Indeed, making the choice $p = 1 + 1/\log d$ and $\psi(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$, we obtain the following corollary.

Corollary 2. *Assume that $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_\infty^2] \leq G^2$ and that $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq R\}$. Set μ to be uniform on the surface of the ℓ_2-ball of radius $\sqrt{d}$. Use the step sizes specified in Theorem 1. There are universal constants $C_1 < 20e$ and $C_2 < 10e$ such that*

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq C_1\frac{RG\sqrt{d}\log d}{\sqrt{k}}\max\left\{\alpha, \alpha^{-1}\right\} + C_2\frac{RG\sqrt{d}\log d}{k}\left(\alpha u^2 + u\log k\right).$$

Proof The proof of this corollary is somewhat involved. The main argument involves showing that the constants $M(\mu)$ and $s(d)$ may be taken as $M(\mu) \leq d^6$ and $s(d) \leq 24d\log d$.

First, we recall [18, 7, Appendix 1] that our choice of ψ is strongly convex with respect to the norm $\|\cdot\|_p$. In addition, if we define $q = 1 + \log d$, then we have $1/p + 1/q = 1$, and $\|v\|_q \leq e\|v\|_\infty$ for any $v \in \mathbb{R}^d$ and any d. As a consequence, we see that we may take the norm $\|\cdot\| = \|\cdot\|_1$ and the dual norm $\|\cdot\|_* = \|\cdot\|_\infty$, and $\mathbb{E}[\|\langle g, Z\rangle Z\|_q^2] \leq e^2\mathbb{E}[\|\langle g, Z\rangle Z\|_\infty^2]$. To apply Theorem 1 with appropriate values from Assumption D, we now bound $\mathbb{E}[\|\langle g, Z\rangle Z\|_\infty^2]$; see Appendix A.3 for a proof.

Lemma 3. *Let Z be distributed uniformly on the ℓ_2-sphere of radius $\sqrt{d}$. For any $g \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\langle g, Z\rangle Z\|_\infty^2] \leq C \cdot d\log d\,\|g\|_\infty^2,$$

where $C \leq 24$ is a universal constant.

As a consequence of Lemma 3, the constant $s(d)$ of Assumption D satisfies $s(d) \leq Cd\log d$. Finally, we have the essentially trivial bound $M(\mu)^2 = \mathbb{E}[\|Z\|_1^4\|Z\|_\infty^2] \leq d^6$ (we only need the quantity $M(\mu)$ to be finite to apply Theorem 1). Recalling that the set $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq R\}$, our choice of ψ yields [e.g., 14, Lemma 3]

$$(p-1)D_\psi(\theta, \tau) \leq \frac{1}{2}\|\theta\|_p^2 + \frac{1}{2}\|\tau\|_p^2 + \|\theta\|_p\|\tau\|_p.$$

We thus find that $D_\psi(\theta, \tau) \leq 2R^2\log d$ for any $\theta, \tau \in \Theta$, and using the step and perturbation size choices of Theorem 1 gives the result. □

Corollary 2 attains a convergence rate that scales with dimension as $\sqrt{d}\log d$. This dependence on dimension is much worse than that of (stochastic) mirror descent using full gradient information [18, 19]. The additional dependence on d suggests that while $\mathcal{O}(1/\epsilon^2)$ iterations are required to achieve ϵ-optimization accuracy for mirror descent methods (ignoring logarithmic factors), the two-point method requires $\mathcal{O}(d/\epsilon^2)$ iterations to obtain the same accuracy. A similar statement holds for the results of Corollary 1. In the next section, we show that this dependence is sharp: except for logarithmic factors, no algorithm can attain better convergence rates, including the problem-dependent constants R and G.

3 Lower bounds on zero-order optimization

We turn to providing lower bounds on the rate of convergence for any method that receives random function values. For our lower bounds, we fix a norm $\|\cdot\|$ on $\mathbb{R}^d$ and as usual let $\|\cdot\|_*$ denote its dual norm. We assume that $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \le R\}$ is the norm ball of radius R. We study all optimization methods that receive function values of random convex functions, building on the analysis of stochastic gradient methods by Agarwal et al. [3].

More formally, let $\mathbb{A}_k$ denote the collection of all methods that observe a sequence of data points $(Y^1, \ldots, Y^k) \subset \mathbb{R}^2$ with $Y^t = [F(\theta^t, X^t)\; F(\tau^t, X^t)]$ and return an estimate $\widehat{\theta}(k) \in \Theta$. The classes of functions over which we prove our lower bounds consist of those satisfying Assumption B, that is, for a given Lipschitz constant $G > 0$, optimization problems over the set $\mathcal{F}_G$. The set $\mathcal{F}_G$ consists of pairs (F, P) as described in the objective (1), and for $(F, P) \in \mathcal{F}_G$ we assume there is a measurable subgradient selection $\mathsf{g}(\theta; X) \in \partial F(\theta; X)$ satisfying $\mathbb{E}_P[\|\mathsf{g}(\theta; X)\|_*^2] \le G^2$ for $\theta \in \Theta$.

Given an algorithm $\mathcal{A} \in \mathbb{A}_k$ and a pair $(F, P) \in \mathcal{F}_G$, we define the optimality gap

$$\epsilon_k(\mathcal{A}, F, P, \Theta) := f(\widehat{\theta}(k)) - \inf_{\theta \in \Theta} f(\theta) = \mathbb{E}_P\left[F(\widehat{\theta}(k); X)\right] - \inf_{\theta \in \Theta} \mathbb{E}_P\left[F(\theta; X)\right], \tag{7}$$

where $\widehat{\theta}(k)$ is the output of $\mathcal{A}$ on the sequence of observed function values. The quantity (7) is a random variable, since the Y^t are random and $\mathcal{A}$ may use additional randomness. We we are thus interested in its expected value, and we define the minimax error

$$\epsilon_k^*(\mathcal{F}_G, \Theta) := \inf_{\mathcal{A} \in \mathbb{A}_k} \sup_{(F,P) \in \mathcal{F}_G} \mathbb{E}[\epsilon_k(\mathcal{A}, F, P, \Theta)], \tag{8}$$

where the expectation is over the observations $(Y^1, \ldots, Y^k)$ and randomness in $\mathcal{A}$.

3.1 Lower bounds and optimality

In this section, we give lower bounds on the minimax rate of optimization for a few specific settings. We present our main results, then recall Corollaries 1 and 2, which imply we have attained the minimax rates of convergence for zero-order (stochastic) optimization schemes. The following sections contain proof sketches; we defer technical arguments to appendices.

We begin by providing minimax lower bounds when the expected function $f(\theta) = \mathbb{E}[F(\theta; X)]$ is Lipschitz continuous with respect to the ℓ_2-norm. We have the following proposition.

Proposition 1. *Let $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \le R\right\}$ and $\mathcal{F}_G$ consist of pairs (F, P) for which the subgradient mapping g satisfies $\mathbb{E}_P[\|\mathsf{g}(\theta; X)\|_2^2] \le G^2$ for $\theta \in \Theta$. There exists a universal constant $c > 0$ such that for $k \ge d$,*

$$\epsilon_k^*(\mathcal{F}_G, \Theta) \ge c\frac{GR\sqrt{d}}{\sqrt{k}}.$$

Combining the lower bounds provided by Proposition 1 with our algorithmic scheme in Section 2 shows that our analysis is essentially sharp, since Corollary 1 provides an upper bound for the minimax optimality of $RG\sqrt{d}/\sqrt{k}$. The stochastic gradient descent algorithm (4) coupled with the sampling strategy (6) is thus optimal for stochastic problems with two-point feedback.

Now we investigate the minimax rates at which it is possible to solve stochastic convex optimization problems whose objectives are Lipschitz continuous with respect to the ℓ_1-norm. As noted earlier, such scenarios are suitable for high-dimensional problems [e.g. 19].

Proposition 2. *Let* $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le R\}$ *and* $\mathcal{F}_G$ *consist of pairs* (F, P) *for which the subgradient mapping* g *satisfies* $\mathbb{E}_P[\|\mathsf{g}(\theta; X)\|_\infty^2] \le G^2$ *for* $\theta \in \Theta$. *There exists a universal constant* $c > 0$ *such that for* $k \ge d$,

$$\epsilon_k^*(\mathcal{F}_G, \Theta) \ge c \frac{GR\sqrt{d}}{\sqrt{k}}.$$

We may again consider the optimality of our mirror descent algorithms, recalling Corollary 2. In this case, the MD algorithm (4) with the choice $\psi(\theta) = \frac{1}{2(p-1)} \|\theta\|_p^2$, where $p = 1 + 1/\log d$, implies that there exist universal constants c and C such that

$$c \frac{GR\sqrt{d}}{\sqrt{k}} \le \epsilon_k^*(\mathcal{F}_G, \Theta) \le C \frac{GR\sqrt{d} \log d}{\sqrt{k}}$$

for the problem class described in Proposition 2. Here the upper bound is again attained by our two-point mirror-descent procedure. Thus, to within logarithmic factors, our mirror-descent based algorithm is optimal for these zero-order optimization problems.

When full gradient information is available, that is, one has access to the subgradient selection $\mathsf{g}(\theta; X)$, the $\sqrt{d}$ factors appearing in the lower bounds in Proposition 1 and 2 are not present [3]. The $\sqrt{d}$ factors similarly disappear from the convergence rates in Corollaries 1 and 2 when one uses $g^t = \mathsf{g}(\theta; X)$ in the mirror descent updates (4); said differently, the constant $s(d) = 1$ in Theorem 1 [19, 6]. As noted in Section 2, our lower bounds consequently show that in addition to dependence on the radius R and second moment G^2 in the case when gradients are available [3], all algorithms must suffer an additional $\mathcal{O}(\sqrt{d})$ penalty in convergence rate. This suggests that for high-dimensional problems, it is preferable to use full gradient information if possible, even when the cost of obtaining the gradients is somewhat high.

3.2 Proofs of lower bounds

We sketch proofs for our lower bounds on the minimax error (8), which are based on the framework introduced by Agarwal et al. [3]. The strategy is to reduce the optimization problem to a testing problem: we choose a finite set of (well-separated) functions, show that optimizing well implies that one can identify the function being optimized, and then, as in statistical minimax theory [26, 25], apply information-theoretic lower bounds on the probability of error in hypothesis testing problems.

We begin with a finite set $\mathcal{V} \subseteq \mathbb{R}^d$, to be chosen depending on the characteristics of the function class $\mathcal{F}_G$, and a collection of functions and distributions $\mathcal{G} = \{(F_v, P_v) : v \in \mathcal{V}\} \subseteq \mathcal{F}_G$ indexed by $\mathcal{V}$. Define $f_v(\theta) = \mathbb{E}_{P_v}[F_v(\theta; X)]$, and let $\theta_v^* \in \operatorname{argmin}_{\theta \in \Theta} f_v(\theta)$. We also let $\delta > 0$ be an accuracy parameter upon which P_v and the following quantities implicitly depend. Following Agarwal et al. [3], we define the separation between two functions as

$$\rho(f_v, f_w) := \inf_{\theta \in \Theta} \left[f_v(\theta) + f_w(\theta)\right] - f_v(\theta_v^*) - f_w(\theta_w^*),$$

and the minimal separation of the set $\mathcal{V}$ (this may depend on the accuracy parameter δ) as

$$\rho^*(\mathcal{V}) := \min\{\rho(f_v, f_w) : v, w \in \mathcal{V}, v \ne w\}.$$

For any algorithm $\mathcal{A} \in \mathbb{A}_k$, there exists a hypothesis test $\widehat{v} : (Y^1, \ldots, Y^k) \to \mathcal{V}$ such that for V sampled uniformly from $\mathcal{V}$ (see [3, Lemma 2]),

$$\mathbb{P}(\widehat{v}(Y^1, \ldots, Y^k) \ne V) \le \frac{2}{\rho^*(\mathcal{V})} \mathbb{E}[\epsilon_k(\mathcal{A}, F_V, P_V, \Theta)] \le \frac{2}{\rho^*(\mathcal{V})} \max_{v \in \mathcal{V}} \mathbb{E}[\epsilon_k(\mathcal{A}, F_v, P_v, \Theta)], \quad (9)$$

where the expectation is taken over the observations $(Y^1, \ldots, Y^k)$. By Fano's inequality [11],

$$\mathbb{P}(\widehat{v} \ne V) \ge 1 - \frac{I(Y^1, \ldots, Y^k; V) + \log 2}{\log |\mathcal{V}|}. \quad (10)$$

We thus must upper bound the mutual information $I(Y^1, \ldots, Y^k; V)$, which leads us to the following. (See Appendix B.3 for the somewhat technical proof of the lemma.)

Lemma 4. *Let $X \mid V = v$ be distributed as $N(\delta v, \sigma^2 I)$, and let $F(\theta; x) = \langle \theta, x \rangle$. Let V be a uniform random variable on $\mathcal{V} \subset \mathbb{R}^d$, and assume that $\mathrm{Cov}(V) \preceq \lambda I$ for some $\lambda \geq 0$. Then*

$$I(Y^1, Y^2, \ldots, Y^k; V) \leq \frac{\lambda k \delta^2}{\sigma^2}.$$

Using Lemma 4, we can obtain a lower bound on the minimax optimization error whenever the instantaneous objective functions are of the form $F(\theta; x) = \langle \theta, x \rangle$. Combining inequalities (9), (10), and Lemma 4, we find that if we choose the accuracy parameter

$$\delta = \frac{\sigma}{\sqrt{k\lambda}} \left(\frac{\log |\mathcal{V}|}{2} - \log 2 \right)^{1/2}, \tag{11}$$

(we assume that $|\mathcal{V}| > 4$) we find that there exist a pair $(F, P) \in \mathcal{F}_G$ such that

$$\mathbb{E}[\epsilon_k(\mathcal{A}, F, P, \Theta)] \geq \rho^*(\mathcal{V})/4. \tag{12}$$

The inequality (12) can give concrete lower bounds on the minimax optimization error. In our lower bounds, we use $F_v(\theta; x) = \langle \theta, x \rangle$ and set P_v to be the $N(\delta v, \sigma^2 I)$ distribution, which allows us to apply Lemma 4. Proving Propositions 1 and 2 thus requires three steps:

1. Choose the set $\mathcal{V}$ with the property that $\mathrm{Cov}(V) \preceq \lambda I$ when $V \sim \mathrm{Uniform}(\mathcal{V})$.
2. Choose the variance parameter σ^2 such that for each $v \in \mathcal{V}$, the pair $(F_v, P_v) \in \mathcal{F}_G$.
3. Calculate the separation value $\rho^*(\mathcal{V})$ as a function of the accuracy parameter δ.

Enforcing $(F_v, P_v) \in \mathcal{F}_G$ amounts to choosing σ^2 so that $\mathbb{E}[\|X\|_*^2] \leq G^2$ for $X \sim N(\delta v, \sigma^2 I)$. By construction $f_v(\theta) = \delta \langle \theta, v \rangle$, which allows us to give lower bounds on the minimal separation $\rho^*(\mathcal{V})$ for the choices of the norm constraint $\|\theta\| \leq R$ in Propositions 1 and 2. We defer formal proofs to Appendices B.1 and B.2, providing sketches here.

For Proposition 1, an argument using the probabilistic method implies that there are universal constants $c_1, c_2 > 0$ for which there is a $\frac{1}{2}$ packing $\mathcal{V}$ of the ℓ_2-sphere of radius 1 with size at least $|\mathcal{V}| \geq \exp(c_1 d)$ and such that $(1/|\mathcal{V}|) \sum_{v \in \mathcal{V}} v v^\top \preceq c_2 I_{d \times d}/d$. By the linearity of f_v, we find $\rho(f_v, f_w) \geq \delta R/16$, and setting $\sigma^2 = G^2/(2d)$ and δ as in the choice (11) implies that $\mathbb{E}[\|X\|_2^2] \leq G^2$. Substituting δ and $\rho^*(\mathcal{V})$ into the bound (12) proves Proposition 1.

For Proposition 2, we use the packing set $\mathcal{V} = \{\pm e_i : i = 1, \ldots, d\}$. Standard bounds [8] on the normal distribution imply that for $Z \sim N(0, I)$, we have $\mathbb{E}[\|Z\|_\infty^2] = \mathcal{O}(\log d)$. Thus we find that for $\sigma^2 = \mathcal{O}(G^2/\log(d))$ and suitably small δ, we have $\mathbb{E}[\|X\|_\infty^2] = \mathcal{O}(G^2)$; linearity yields $\rho(f_v, f_w) \geq \delta R$ for $v \neq w \in \mathcal{V}$. Setting δ as in the expression (11) yields Proposition 2.

4 Discussion

We have analyzed algorithms for stochastic optimization problems that use only random function values—as opposed to gradient computations—to minimize an objective function. As our development of minimax lower bounds shows, the algorithms we present, which build on those proposed by Agarwal et al. [1] and Nesterov [20], are optimal: their convergence rates cannot be improved (in a minimax sense) by more than numerical constant factors. As a consequence of our results, we have attained sharp rates for bandit online convex optimization problems with multi-point feedback. We have also shown that there is a necessary sharp transition in convergence rates between stochastic gradient algorithms and algorithms that compute only function values. This result highlights the advantages of using gradient information when it is available, but we recall that there are many applications in which gradients are not available.

Finally, one question that this work leaves open, and which we are actively attempting to address, is whether our convergence rates extend to non-smooth optimization problems. We conjecture that they do, though it will be interesting to understand the differences between smooth and non-smooth problems when only zeroth-order feedback is available.

Acknowledgments This material supported in part by ONR MURI grant N00014-11-1-0688 and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant no. W911NF-11-1-0391. JCD was also supported by an NDSEG fellowship and a Facebook PhD fellowship.

References

[1] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[2] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, To appear, 2011. URL http://arxiv.org/abs/1107.1744.

[3] 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, May 2012.

[4] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. MSRI Publications, 1997.

[5] P. L. Bartlett, V. Dani, T. P. Hayes, S. M. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, 2008.

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

[7] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12:79–108, 2001.

[8] V. Buldygin and Y. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[9] N. Cesa-Bianchi, A. Conconi, and C.Gentile. On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366, 2002.

[10] A. Conn, K. Scheinberg, and L. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MPS-SIAM Series on Optimization*. SIAM, 2009.

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

[12] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[13] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[14] C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3), 2002.

[15] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, 1996.

[16] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, Jan. 1997.

[17] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, Second edition, 2003.

[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. Random gradient-free minimization of convex functions. URL http://www.ecore.be/DPs/dp_1297333890.pdf, 2011.

[21] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, 2003.

[22] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[23] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[24] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

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

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

[27] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Proofs of Convergence

A.1 Proof of Theorem 1

Before giving the proof of Theorem 1, we state two lemmas that we will need. The first is essentially standard [e.g. 19, Section 2.3], and we provide a proof of the lemma in the long version of this paper.

Lemma 5. *Let $\{g^t\}_{t=1}^k \subset \mathbb{R}^d$ be a sequence of vectors, and let θ^t be generated by the mirror descent iteration (4). If Assumption A holds, for any $\theta^* \in \Theta$ we have*

$$\sum_{t=1}^k \left\langle g^t, \theta^t - \theta^* \right\rangle \le \frac{1}{2\alpha(k)} R^2 + \sum_{t=1}^k \frac{\alpha(t)}{2} \left\|g^t\right\|_*^2.$$

We provide the proof of this next lemma, which is required to control the norms of the observed gradient vectors, in Appendix A.2.

Lemma 6. *Let the vector g^t be defined as in the construction (6) and let $\mathcal{F}_{t-1}$ denote the σ-field of $X^1, \ldots, X^{t-1}$ and $Z^1, \ldots, Z^{t-1}$. Let Assumption C hold. Then for some vector v with $\|v\|_* \le (1/2)\mathbb{E}[\|Z\|^2 \|Z\|_*]$,*

$$\mathbb{E}[g^t \mid \mathcal{F}_{t-1}] = \nabla f(\theta^t) + u_t L(P) v$$

and

$$\mathbb{E}[\left\|g^t\right\|_*^2 \mid \mathcal{F}_{t-1}] \le 2\mathbb{E}\left[\left\|\left\langle \mathsf{g}(\theta^t; X), Z \right\rangle Z\right\|_*^2 \mid \mathcal{F}_{t-1}\right] + \frac{u_t^2 L(P)^2}{2} \mathbb{E}\left[\|Z\|^4 \|Z\|_*^2\right].$$

Proof of Theorem 1 The proof of the theorem follows from Lemma 5. Indeed, defining the error vector $e^t := \nabla f(\theta^t) - g^t$, we have

$$\sum_{t=1}^k \left(f(\theta^t) - f(\theta^*)\right) \le \sum_{t=1}^k \left\langle \nabla f(\theta^t), \theta^t - \theta^* \right\rangle = \sum_{t=1}^k \left\langle g^t, \theta^t - \theta^* \right\rangle + \sum_{t=1}^k \left\langle e^t, \theta^t - \theta^* \right\rangle.$$

Applying Lemma 5, we find that

$$\sum_{t=1}^k \left(f(\theta^t) - f(\theta^*)\right) \le \frac{1}{2\alpha(k)} R^2 + \sum_{t=1}^k \frac{\alpha(t)}{2} \left\|g^t\right\|_*^2 + \sum_{t=1}^k \left\langle e^t, \theta^t - \theta^* \right\rangle. \tag{13}$$

Now we use Lemma 6, which implies that

$$\mathbb{E}[e^t \mid \mathcal{F}_{t-1}] = u_t L(P) v(\theta^t, u_t),$$

where $\|v\|_* \le \frac{1}{2}\mathbb{E}[\|Z\|^2 \|Z\|_*] \le \frac{1}{2}M(\mu)$. Using the compactness assumption and fact that $\theta^t \in \mathcal{F}_{t-1}$, we thus have

$$\sum_{t=1}^k \mathbb{E}[\left\langle e^t, \theta^t - \theta^* \right\rangle] \le \frac{1}{2} M(\mu) R L(P) \sum_{t=1}^k u_t.$$

In addition, Lemma 6 implies that

$$\begin{aligned}
\mathbb{E}[\left\|g^t\right\|_*^2] = \mathbb{E}[\mathbb{E}[\left\|g^t\right\|_*^2 \mid \mathcal{F}_{t-1}]] &\le 2\mathbb{E}\left[\mathbb{E}[\left\|\left\langle \mathsf{g}(\theta^t; X), Z \right\rangle Z\right\|_*^2 \mid \mathcal{F}_{t-1}]\right] + \frac{1}{2} u_t^2 L(P)^2 M(\mu)^2 \\
&\le 2s(d)\mathbb{E}\left[\mathbb{E}[\left\|\mathsf{g}(\theta^t; X)\right\|_*^2 \mid \mathcal{F}_{t-1}]\right] + \frac{1}{2} u_t^2 L(P)^2 M(\mu)^2.
\end{aligned}$$

Since $\mathbb{E}[\|\mathsf{g}(\theta^t; X)\|_*^2 \mid \mathcal{F}_{t-1}] \le G^2$ by assumption, we may apply our initial bound (13) to see

$$\begin{aligned}
&\sum_{t=1}^k \mathbb{E}\left[f(\theta^t) - f(\theta^*)\right] \\
&\le \frac{1}{2\alpha(k)} R^2 + s(d) G^2 \sum_{t=1}^k \alpha(t) + \frac{1}{4} L(P)^2 M(\mu)^2 \sum_{t=1}^k u_t^2 \alpha(t) + \frac{M(\mu) R L(P)}{2} \sum_{t=1}^k u_t.
\end{aligned} \tag{14}$$

Now we use our choices of the sample size $\alpha(t)$ and u_t to complete the proof. For the former, we have $\alpha(t) = \alpha R/2G\sqrt{s(d)}\sqrt{t}$. Since

$$\sum_{t=1}^{k} i^{-\frac{1}{2}} \leq \int_0^k t^{-\frac{1}{2}} dt = 2\sqrt{k},$$

we have

$$\frac{1}{2\alpha(k)} R^2 + s(d)G^2 \sum_{t=1}^{k} \alpha(t) \leq \frac{RG\sqrt{s(d)}}{\alpha}\sqrt{k} + \alpha RG\sqrt{s(d)}\sqrt{k} \leq 2RG\sqrt{s(d)}\sqrt{k}\max\{\alpha, \alpha^{-1}\}.$$

For the second summation in the quantity (14), we have the bound

$$\alpha u^2 \left(\frac{G^2 s(d)}{L(P)^2 M(\mu)^2}\right) \frac{RL(P)^2 M(\mu)^2}{4G\sqrt{s(d)}} \sum_{t=1}^{k} \frac{1}{t^{3/2}} \leq \alpha u^2 RG\sqrt{s(d)}$$

since $\sum_{t=1}^{k} i^{-\gamma} \leq 1 + \int_1^k t^{-\gamma} dt$. The final term in the inequality (14) is similarly bounded by

$$u\left(\frac{G\sqrt{s(d)}}{L(P)M(\mu)}\right)\frac{RL(P)M(\mu)}{2}(\log k + 1) = u\frac{RG\sqrt{s(d)}}{2}(\log k + 1) \leq uRG\sqrt{s(d)}\log k$$

since $k \geq 3$.

By combining the preceding inequalities and using Jensen's inequality to note that

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq \frac{1}{k}\sum_{t=1}^{k} \mathbb{E}\left[f(\theta^t) - f(\theta^*)\right],$$

we obtain the statement of the theorem. □

A.2 Proof of Lemma 6

Let h be an arbitrary convex function with L_h-Lipschitz continuous gradient with respect to the norm $\|\cdot\|$. Then for any $u > 0$

$$\begin{aligned} h'(\theta, z) = \frac{\langle \nabla h(\theta), uz\rangle}{u} &\overset{(i)}{\leq} \frac{h(\theta + uz) - h(\theta)}{u} \\ &\overset{(ii)}{\leq} \frac{\langle \nabla h(\theta), uz\rangle + (L_h/2)\|uz\|^2}{u} = h'(\theta, z) + \frac{L_h u}{2}\|z\|^2, \end{aligned}$$

where inequality (i) follows from the first-order tangent bound for a convex function, and inequality (ii) uses the L_h-Lipschitz continuity of the gradient. Thus we see that for any point $\theta \in \operatorname{relint}\operatorname{dom} h$ and for any $z \in \mathbb{R}^d$, we have

$$\frac{h(\theta + uz) - h(\theta)}{u} z = h'(\theta, z)z + \frac{L_h u}{2}\|z\|^2 \gamma(u, \theta, z)z, \tag{15}$$

where γ is some function with range contained in $[0, 1]$.

Now, assume that the function f has L-Lipschitz derivative with respect to the norm $\|\cdot\|$. Then for any Z with $\mathbb{E}[ZZ^\top] = I$, we find as a consequence of the equality (15) that

$$\mathbb{E}\left[\frac{h(\theta + uZ) - h(\theta)}{u} Z\right] = \mathbb{E}\left[h'(\theta, Z)Z + \frac{L_h u}{2}\|Z\|^2 \gamma(u, \theta, Z)Z\right] = \nabla h(\theta) + uL_h v(\theta, u), \tag{16}$$

where $v(\theta, u) \in \mathbb{R}^d$ is an error vector with $\|v(\theta, u)\|_* \leq (1/2)\mathbb{E}[\|Z\|^2 \|Z\|_*]$. Thus, for $u > 0$ small enough, we see that $[h(\theta + uZ) - h(\theta)]/u$ is an approximately unbiased estimator of $\nabla h(\theta)$.

With the general identities (15) and (16), we now turn to proving the statements of the lemma. The first statement follows because (with probability 1 over the samples X)

$$\mathbb{E}[g^t \mid X(t), \mathcal{F}_{t-1}] = \nabla F(\theta^t; X(t)) + u_t L(X(t)) v_t$$

for some vector v_t with $2 \left\|v_t\right\|_* \le \mathbb{E}[\left\|Z\right\|^2 \left\|Z\right\|_*]$, by the expression (16). Noting that

$$\mathbb{E}[L(X(t)) \left\|v_t\right\|_* \mid \mathcal{F}_{t-1}] \le \sqrt{\mathbb{E}[L(X)^2]} \sqrt{\mathbb{E}[\left\|v_t\right\|_*^2 \mid \mathcal{F}_{t-1}]} \le L(P) \mathbb{E}[\left\|Z\right\|^2 \left\|Z\right\|_*]$$

(by independence) and that $\mathbb{E}[\nabla F(\theta; X)] = \nabla f(\theta)$ completes the first argument.

For the second statement of the lemma, we use the equality (15) applied to $F(\cdot; X)$. Ignoring the indexing by t, we have in this case that

$$g = \langle \mathrm{g}(\theta, X), Z \rangle \, Z + \frac{L(X) u}{2} \left\|Z\right\|^2 \gamma(u, \theta, Z, X) Z$$

for a function $\gamma \in [0, 1]$. Applying the inequality $(a + b) \le 2a^2 + 2b^2$ to the upper bound

$$\mathbb{E}[\left\|g\right\|_*^2] \le \mathbb{E}\left[\left(\left\|\langle \mathrm{g}(\theta, X), Z \rangle \, Z\right\|_* + \left\|\frac{1}{2} L(X) u \left\|Z\right\|^2 \gamma(u, \theta, Z, X) Z\right\|_*\right)^2\right]$$

yields the result. □

A.3 Proof of Lemma 3

By using Levy's theorem on concentration of Haar measure on the unit sphere [4, Lemma 2.2], we have that for U uniform on the ℓ_2-sphere, for any vector v with $\left\|v\right\|_2 = 1$,

$$\mathbb{P}(\langle U, v \rangle \ge \epsilon) \le \exp\left(-\frac{d\epsilon^2}{2}\right). \tag{17}$$

Now, set $v = g / \left\|g\right\|_2$, which gives

$$\mathbb{E}[\left\|\langle g, Z \rangle \, Z\right\|_\infty^2] = d^2 \left\|g\right\|_2^2 \mathbb{E}[\langle v, U \rangle^2 \left\|U\right\|_\infty^2] \le d^2 \left\|g\right\|_2^2 \sqrt{\mathbb{E}[\langle v, U \rangle^4]} \sqrt{\mathbb{E}[\left\|U\right\|_\infty^4]}. \tag{18}$$

We bound the final two expectations in the expression (18) in turn.

Recall the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \ge t) dt$ valid for any non-negative random variable with finite mean. Since we have $\mathbb{P}(\langle U, v \rangle^4 \ge \epsilon) \le 2 \exp(-d\epsilon^{\frac{1}{2}}/2)$, we find that

$$\mathbb{E}[\langle v, U \rangle^4] = \int_0^1 \mathbb{P}(\langle v, U \rangle^4 \ge t) dt \le 2 \int_0^1 \exp\left(-\frac{d\sqrt{t}}{2}\right) dt = \frac{16}{d^2} \int_0^{d/2} u e^{-u} du$$

by making the substitution $u = d\sqrt{t}/2$. Since $\int u e^{-u} du = -e^{-u} - u e^{-u}$, we find that

$$\mathbb{E}[\langle U, v \rangle^4] \le \frac{16}{d^2} \left[-e^{-t} - t e^{-t}\right]\Big|_{t=0}^{d/2} = \frac{16}{d^2} \left[1 - e^{-d/2} - (d/2) e^{-d/2}\right] \le \frac{16}{d^2}.$$

The second expectation is somewhat more challenging to bound, though the technique is the same. Using a union bound, we have $\mathbb{P}(\left\|U\right\|_\infty^4 \ge \epsilon) \le 2d \exp(-d\epsilon^{\frac{1}{2}}/2)$. Noting that $\epsilon = 4 \log^2(2d)/d^2$ sets the upper bound to be equal to 1, we have

$$\mathbb{E}[\left\|U\right\|_\infty^4] = \int_0^1 \mathbb{P}(\left\|U\right\|_\infty \ge t) dt \le \int_0^{\frac{4 \log^2(2d)}{d^2}} 1 dt + 2d \int_{\frac{4 \log^2(2d)}{d^2}}^1 \exp\left(-\frac{d\sqrt{t}}{2}\right) dt.$$

Making the same change of variables as earlier, the second term above is equal to

$$\frac{16}{d} \int_{\log(2d)}^{d/2} t \exp(-t) dt = \frac{16}{d} \left[-e^{-t} - t e^{-t}\right]\Big|_{t=\log(2d)}^{d/2} \le \frac{16}{d} \left[\frac{1}{2d} + \frac{\log(2d)}{2d}\right].$$

In particular, we see that

$$\mathbb{E}[\left\|U\right\|_\infty^4] \le \frac{4 \log^2(2d)}{d^2} + \frac{8}{d^2} + \frac{8 \log(2d)}{d^2},$$

which is bounded by $36 \log^2 d / d^2$ for $d \ge 3$.

Recalling the inequality (18), we find that for $d \ge 3$

$$\mathbb{E}[\left\|\langle g, Z \rangle \, Z\right\|_\infty^2] \le d^2 \left\|g\right\|_2^2 \frac{4}{d} \cdot \frac{6 \log d}{d} = 24 \log d \left\|g\right\|_2^2 \le 24 d \log d \left\|g\right\|_\infty^2,$$

since $\left\|g\right\|_2 \le \sqrt{d} \left\|g\right\|_\infty$. □

B Lower bound proofs

B.1 Proof of Proposition 1

The proof of the proposition requires the choice of packing set $\mathcal{V}$, which necessitates some care. We can use the ℓ_2-ball in $\mathbb{R}^d$, however, as described by the next lemma (see Appendix C for a proof).

Lemma 7. *Let $d \geq 2$. Let $S^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^d$. There is $1/2$-packing $\mathcal{V}$ of S^{d-1} (that is, for any pair $v, w \in \mathcal{V}$ with $w \neq v$, $\|w - v\|_2 \geq \frac{1}{2}$) such that $|\mathcal{V}| \geq e^{49d/256}$ and*

$$\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} vv^\top \preceq \frac{5}{d} I_{d \times d}.$$

With Lemma 7 in hand, Proposition 1's proof follows the outline we provide in Section 3.2.

Proof of Proposition 1 According to Lemma 7, there exists a $\frac{1}{2}$-packing $\mathcal{V}$ of S^{d-1} of size $\log |\mathcal{V}| \geq 3d/16$ and a constant $c \leq 12$ such that for V sampled uniformly from $\mathcal{V}$, $\mathrm{Cov}(V) \preceq cI/d$. Computing the separation $\rho(f_v, f_w)$, we have

$$\begin{aligned}
\rho(f_v, f_w) &= \delta \inf_{\theta \in \Theta} \{\langle \theta, v + w \rangle\} + \delta R \|v\|_2 + \delta R \|w\|_2 \\
&= 2\delta R - \delta R \|v + w\|_2 = \delta R \left(2 - \sqrt{4 - \|v - w\|_2^2}\right)
\end{aligned}$$

since $\|v\|_2 = \|w\|_2 = 1$. Since $\|v - w\|_2 \geq 1/2$, we find that $\sqrt{4 - \|v - w\|_2^2} \leq 31/16$, so

$$\rho^*(\mathcal{V}) \geq \frac{\delta R}{16}. \tag{19}$$

Now we turn to finding a setting for the variance σ^2 of the samples X so that $\mathbb{E}[\|X\|_2^2] \leq G^2$. Since $X \sim N(\delta v, \sigma^2 I)$ and $\|v\|_2^2 = 1$, we have $\mathbb{E}[\|X\|_2^2] = \delta^2 + d\sigma^2$. If $\delta^2 \leq d\sigma^2$, then we may take $\sigma^2 = G^2/2d$, and the setting (11) yields for $d \geq 15$ that

$$\delta = \frac{G/\sqrt{2d}}{\sqrt{kc/d}} \left(\frac{3d}{32} - \log 2\right)^{\frac{1}{2}} = \frac{G}{\sqrt{2kc}} \left(\frac{3d}{32} - \log 2\right)^{\frac{1}{2}} \geq \frac{G}{\sqrt{2kc}} \cdot \frac{\sqrt{3d}}{8}.$$

Now we apply our bound (12), which shows that $\rho^*(\mathcal{V})$ lower bounds the minimax optimization error, note that for $k \geq d$, we have $\delta^2 \leq d\sigma^2$, so that the bound (19) applies with this choice of δ.

When $d \leq 15$, we recall the analysis of Agarwal et al. [3], which gives a lower bound of $\Omega(GR/\sqrt{k})$, which is within a constant factor of $GR\sqrt{d}/\sqrt{k}$. □

B.2 Proof of Proposition 2

The proof of Proposition 2 requires an auxiliary result, which we state here, to guarantee that our choice of distribution for X satisfies $\mathbb{E}[\|\mathsf{g}(\theta; X)\|_\infty^2] \leq G^2$.

Lemma 8. *Let $X \sim N(\delta v, \sigma^2 I)$ where $\|v\|_\infty \leq 1$. For $d \geq 3$,*

$$\mathbb{E}[\|X\|_\infty^2] \leq 8\sigma^2(1 + \log d) + 2\delta^2.$$

Proof Let $Z = X - \delta v$, so $Z \sim N(0, \sigma^2 I)$. Letting $(X_1, \ldots, X_d)$ and $(Z_1, \ldots, Z_d)$ denote the components of X and Z, respectively, we see that we have $X_i^2 \leq 2Z_i^2 + 2\delta^2 v_i^2$, so

$$\|X\|_\infty^2 \leq 2 \max\{Z_1^2, \ldots, Z_d^2\} + 2\delta^2 \max\{v_1^2, \ldots, v_d^2\} \leq 2 \|Z\|_\infty^2 + 2\delta^2. \tag{20}$$

Each Z_i is a random variable with $N(0, \sigma^2)$ distribution, and standard results on sub-Gaussian random variables [8, Chapter 2] imply that $\mathbb{E}[\|Z\|_\infty^2] \leq 4\sigma^2(1 + \log d)$. Applying the inequality (20)

implies the result of the lemma. $\square$

With Lemma 8 in hand, we can now prove the proposition.

Proof of Proposition 2 We take the packing set $\mathcal{V} = \{\pm e_i : i = 1, \ldots, d\}$. Now let $v \in \mathcal{V}$, and suppose $X \sim N(\delta v, \sigma^2 I)$. We must choose σ^2 so that $\mathbb{E}[\|X\|_\infty^2] \le G^2$. Recalling our choice (11) of δ following Lemma 4, we apply Lemma 8 and substitute δ to obtain

$$\mathbb{E}[\|X\|_\infty^2] \le 8\sigma^2(1 + \log d) + 2\delta^2 = 8\sigma^2(1 + \log d) + \frac{\sigma^2}{k\lambda}\left(\frac{\log|\mathcal{V}|}{2} - \log 2\right).$$

Now, if we sample $V \sim \text{Uniform}(\mathcal{V})$, then $\mathbb{E}[V] = 0$ and $\text{Cov}(V) = \mathbb{E}[VV^\top] = I/d$, which gives us $\lambda = 1/d$ in Lemma 4. Substituting $\lambda = 1/d$ and $|\mathcal{V}| = 2d$, then using our assumption $k \ge d$, we find that

$$\mathbb{E}[\|X\|_\infty^2] \le 8\sigma^2(1 + \log d) + \frac{\sigma^2 d \log(2d)}{2k} \le 8\sigma^2(1 + \log d) + \frac{\sigma^2 \log(2d)}{2}.$$

Therefore, if we set $\sigma^2 = G^2/(\frac{1}{2}\log(2d) + 8\log(d) + 8)$, we find that $\mathbb{E}[\|X\|_\infty^2] \le G^2$, as desired.

Lastly, we compute the separation $\rho^*(\mathcal{V})$. For each $v \in \mathcal{V}$ we have

$$\inf_{\theta\in\Theta} f_v(\theta) = \inf_{\theta\in\Theta} \delta\langle\theta, v\rangle = -\delta R\|v\|_\infty = -\delta R,$$

and the unique minimizer is $\theta_v^* = -Rv$. Now for each $v, w \in \mathcal{V}$, $v \ne w$, we have

$$\begin{aligned}\rho(f_v, f_w) &= \inf_{\theta\in\Theta}\{f_v(\theta) + f_w(\theta)\} - f_v(\theta_v^*) - f_w(\theta_w^*) = \inf_{\theta\in\Theta} \delta\langle v + w, \theta\rangle + 2\delta R \\ &= -\delta R\|v + w\|_\infty + 2\delta R \ge \delta R,\end{aligned}$$

whence the minimum separation is $\rho^*(\mathcal{V}) \ge \delta R$. Using the lower bound (12) on the minimax error, then substituting our chosen variable values, we obtain the lower bound

$$\begin{aligned}\sup_{(F,P)\in\mathcal{F}_G} \mathbb{E}[\epsilon_k(\mathcal{A}, F, P, \Theta)] &\ge \frac{\delta R}{4} = \left(\frac{G}{\sqrt{\log(2d)/2 + 8\log(d) + 8}}\right)\frac{R\sqrt{d}}{4\sqrt{k}}\left(\frac{\log(2d)}{2} - \log(2)\right)^{1/2} \\ &\ge \frac{GR\sqrt{d}}{12\sqrt{\log(2d)/2 + 8\log d + 8}}\sqrt{\log(2d)} > \frac{GR}{40}\sqrt{\frac{d}{k}},\end{aligned}$$

where the final two inequalities hold for $d \ge 3$.

When $d \le 3$, we may (as in Proposition 1 earlier) apply the lower bounds provided by Agarwal et al. [3]. $\square$

B.3 Proof of Lemma 4

Since the pair (θ^t, τ^t) is measurable with respect to $\mathcal{F}_{t-1} = \sigma(Y^1, \ldots, Y^{t-1})$ and Y^t is independent of the first $t-1$ observations $(Y^1, \ldots, Y^{t-1})$ given the pair (θ^t, τ^t), we see by the chain rule for mutual information [11] that

$$\begin{aligned}I(Y^1, \ldots, Y^k; V) = \sum_{t=1}^k I(Y^t; V \mid Y^1, \ldots, Y^{t-1}) &= \sum_{t=1}^k I(Y^t; V \mid \theta^t, \tau^t, Y^1, \ldots, Y^{t-1}) \\ &= \sum_{t=1}^k I(Y^t; V \mid \theta^t, \tau^t). \qquad (21)\end{aligned}$$

Thus, if we can bound $I(Y; V \mid T)$ by $\lambda\delta^2/\sigma^2$ for any pair $T = [\theta\ \tau] \in \mathbb{R}^{d\times 2}$ (formally, any distribution over the pair, though taking a supremum shows that we may focus on an arbitrary pair

T), the equality (21) will yield the desired result. To that end, we spend the remainder of the proof studying the differential entropies in the representation

$$I(Y; V \mid T) = h(Y \mid T) - h(Y \mid T, V).$$

The remainder of our proof will be based on the fact that the normal distribution maximizes the differential entropy across all distributions with the same covariance [11].

By construction we have $Y = T^\top X$ and $X \mid V \sim N(\delta V, \sigma^2 I)$, so conditioned on V, the vector $Y \in \mathbb{R}^2$ has normal distribution with mean $\mathbb{E}[Y \mid V, T] = \delta T^\top V$ and covariance

$$\begin{aligned}\operatorname{Cov}(Y \mid V, T) &= \mathbb{E}[YY^\top \mid V, T] - \mathbb{E}[Y \mid V, T]\,\mathbb{E}[Y \mid V, T]^\top \\ &= T^\top \left(\sigma^2 I + \delta^2 V V^\top\right) T - \delta^2 T^\top V V^\top T = \sigma^2 T^\top T = \sigma^2 \begin{bmatrix} \|\theta\|_2^2 & \langle \theta, \tau \rangle \\ \langle \theta, \tau \rangle & \|\tau\|_2^2 \end{bmatrix}.\end{aligned}$$

Let $\Sigma = \sigma^2 T^\top T$ be shorthand for this covariance. Now using the law of total covariance, we can compute the covariance of Y conditional on T:

$$\begin{aligned}\operatorname{Cov}(Y \mid T) &= \mathbb{E}[\operatorname{Cov}(Y \mid V, T)] + \operatorname{Cov}(\mathbb{E}[Y \mid V, T]) \\ &= \Sigma + \operatorname{Cov}(\delta T^\top V) = \Sigma + \delta^2 T^\top \operatorname{Cov}(V) T.\end{aligned}$$

Since $\operatorname{Cov}(V) \preceq \lambda I$ for V uniform on $\mathcal{V}$ by assumption, we find that

$$\operatorname{Cov}(Y \mid T) \preceq \Sigma + \delta^2 \lambda T^\top T = \Sigma \left(1 + \frac{\delta^2 \lambda}{\sigma^2}\right). \tag{22}$$

Using the fact that for the normal distribution $Z \sim N(\mu, \Gamma)$ in $\mathbb{R}^d$ we have $h(Z) = (d/2)\log(2\pi e) + (1/2)\log\det(\Gamma)$, we find the differential entropy bound

$$h(Y \mid T) \le \log(2\pi e) + \frac{1}{2}\log\det\left(\Sigma\left(1 + \frac{\delta^2\lambda}{\sigma^2}\right)\right) = \log(2\pi e) + \frac{1}{2}\log\det(\Sigma) + \log\left(1 + \frac{\delta^2\lambda}{\sigma^2}\right)$$

as a consequence of the inequality (22). But of course, we have via standard entropy calculations for the normal distribution that $h(Y \mid V, T) = \log(2\pi e) + \frac{1}{2}\log\det(\Sigma)$. Therefore

$$I(Y; V \mid T) = h(Y \mid T) - h(Y \mid V, T) \le \log\left(1 + \frac{\delta^2\lambda}{\sigma^2}\right).$$

Noting that $\log(1 + a) \le a$ completes the proof. □

C Proof of Lemma 7

We prove the result using the probabilistic method, showing that there is positive probability that a set of N (to be specified) vectors sampled uniformly at random from the ℓ_2-sphere S^{d-1} satisfy the conclusions of the lemma. As we noted in (17), if U is sampled uniformly from S^{d-1}, then for any $\epsilon \ge 0$ and $v \in S^{d-1}$ we have

$$\mathbb{P}(\langle U, v \rangle \ge \epsilon) \le \exp\left(-\frac{d\epsilon^2}{2}\right).$$

Now, consider a set of N points $\{U_1, \ldots, U_N\}$ sampled independently and uniformly at random from S^{d-1}. Then a pair U_i, U_j satisfying $\|U_i - U_j\|_2 \le \epsilon$ is equivalent to $\langle U_i, U_j \rangle \ge 1 - \epsilon^2/2$, since $\|U_i\|_2 = \|U_j\|_2 = 1$. Using a union bound, we thus see that

$$\mathbb{P}\left(\exists\, i \ne j \text{ s.t. } \|U_i - U_j\|_2 \le \epsilon\right) \le \sum_{i<j}^{N} \mathbb{P}\left(\langle U_i, U_j \rangle \ge 1 - \frac{\epsilon^2}{2}\right).$$

Since U_i and U_j are independent, we can condition on the value of $U_j = v \in S^{d-1}$ to obtain

$$\mathbb{P}\left(\langle U_i, U_j \rangle \ge 1 - \frac{\epsilon^2}{2}\right) \le \exp\left(-\frac{d(1 - \epsilon^2/2)^2}{2}\right),$$

which yields

$$\mathbb{P}\left(\exists\, i \neq j \text{ s.t. } \|U_i - U_j\|_2 \leq \epsilon\right) \leq \binom{N}{2} \exp\left(-\frac{d(1-\epsilon^2/2)^2}{2}\right).$$

If we choose $\epsilon = 1/2$ and $N = e^{49d/256}$, we obtain

$$\mathbb{P}\left(\exists\, i \neq j \text{ s.t. } \|U_i - U_j\|_2 \leq \frac{1}{2}\right) < \frac{N^2}{2} \exp\left(-\frac{49d}{128}\right) = \frac{1}{2}. \tag{23}$$

Now we show that since N is suitably large, the probability that the set $\{U_1, \ldots, U_N\}$ satisfies $\frac{1}{N}\sum_{i=1}^N U_i U_i^\top \approx (1/d)I$ is high. By the Levy bound (17), each vector $\sqrt{d}U_i$ is sub-Gaussian with parameter 1. Consequently, by standard non-asymptotic bounds on random matrices [22], we have

$$\mathbb{P}\left[\left|\left|\left|\frac{1}{N}\sum_{i=1}^N U_i U_i^T - \frac{1}{d}I\right|\right|\right|_{\text{op}} \geq \frac{\delta}{d}\right] \leq 2\exp\left(-\frac{N\delta^2}{16}\right) \qquad \text{for all } \delta \in (0,1).$$

Substituting the value $N = e^{49d/256}$ and setting $\delta = 4\sqrt{\log 4}\, e^{-49d/512}$, we obtain the inequatliy

$$2\exp\left(-\frac{N\delta^2}{16}\right) \leq 2\exp\left(-\log(4) e^{49d/256} e^{-49d/256}\right) = \frac{1}{2}.$$

Since $4\sqrt{\log 4}\, e^{-49d/512} < 4$ for $d \geq 2$, we have in particular that with probability strictly greater than $1/2$

$$\frac{1}{N}\sum_{i=1}^N U_i U_i^\top \preceq \frac{5}{d} I.$$

By a union bound argument, we see that for U_i sampled uniformly at random from S^{d-1} and for $N = e^{49d/256}$, we have

$$\mathbb{P}\left(\|U_i - U_j\|_2 \geq \frac{1}{2} \text{ for all } i \neq j \text{ and } \frac{1}{N}\sum_{i=1}^N U_i U_i^\top \preceq \frac{5}{d} I\right) > 0.$$

Hence a packing as claimed in the statement of the lemma must exist. □