Information-theoretic lower bounds on the oracle complexity of convex optimization

Alekh Agarwal[1]
alekh@cs.berkeley.edu

Peter L. Bartlett[1,2]
peter@berkeley.edu

Pradeep Ravikumar[3]
pradeepr@cs.utexas.edu

Martin J. Wainwright[1,2]
wainwrig@stat.berkeley.edu

Dept. of Electrical Engineering and Computer Sciences[1]
Dept. of Statistics[2]
UC Berkeley, Berkeley, CA

Dept. of Computer Sciences[2]
UT Austin, Austin, TX

September 2, 2010

Abstract

Relative to the large literature on upper bounds on complexity of convex optimization, lesser attention has been paid to the fundamental hardness of these problems. Given the extensive use of convex optimization in machine learning and statistics, gaining an understanding of these complexity-theoretic issues is important. In this paper, we study the complexity of stochastic convex optimization in an oracle model of computation. We improve upon known results and obtain tight minimax complexity estimates for various function classes.

1 Introduction

Convex optimization forms the backbone of many algorithms for statistical learning and estimation. Given that many statistical estimation problems are large-scale in nature—with the problem dimension and/or sample size being large—it is essential to use bounded computational resources as efficiently as possible. Understanding the computational complexity of convex optimization is thus a key issue for large-scale learning. A large body of literature is devoted to obtaining rates of convergence of specific procedures for various classes of convex optimization problems. A typical outcome of such analysis is an upper bound on the error—for instance, gap to the optimal cost—as a function of the number of iterations. Such analyses have been performed for many standard optimization algorithms, among them gradient descent, mirror descent, interior point programming, and stochastic gradient descent, to name a few. We refer the reader to various standard texts on optimization (e.g., [1, 2, 3]) for further details on such results.

On the other hand, there has been relatively little study of the inherent complexity of convex optimization problems. To the best of our knowledge, the first formal study in this area was undertaken in the seminal work of Nemirovski and Yudin [4], hereafter referred to as NY. One obstacle to a classical complexity-theoretic analysis, as these authors observed, is that of casting convex optimization problems in a Turing Machine model. They avoided this problem by instead considering a natural oracle model of complexity, in which at every round the optimization procedure queries an oracle for certain information on the function being optimized. Working within this framework, the authors obtained a series of lower bounds on the computational complexity of convex optimization problems. In addition to the original text NY [4], we refer the reader to the book by Nesterov [3], and the lecture notes by Nemirovski [5] for background.

In this paper, we consider the computational complexity of stochastic convex optimization within the oracle model. Our results lead to a characterization of the inherent difficulty of learning and estimation problems when computational resources are constrained. In particular, we improve upon the work of NY [4] in two ways. First, our lower bounds have an improved dependence on the dimension of the space. In the context of statistical estimation, these bounds show how the difficulty of the estimation problem increases with the number of parameters. Second, our techniques naturally extend to give sharper results for optimization over simpler function classes. For instance, they show that the optimal oracle complexity of statistical estimation with quadratic loss is significantly smaller than the corresponding complexity with absolute loss. Third, we show that for a fixed function class, if the set of optimizers is assumed to have special structure such as sparsity, then the fundamental complexity of optimization can be significantly smaller. All of our proofs exploit a new notion of the discrepancy between two functions that appears to be natural for optimization problems. They involve a reduction from a statistical parameter estimation problem to the stochastic optimization problem, and an application of information-theoretic lower bounds for the estimation problem. We note that the first two results of this paper appeared in the extended abstract [6].

The remainder of this paper is organized as follows. We begin in Section 2 with background on oracle complexity, and a precise formulation of the problems addressed in this paper. Section 3 is devoted to the statement of our main results, and discussion of their consequences. In Section 4, we provide the proofs of our main results, which all exploit a common framework of four steps. More technical aspects of these proofs are deferred to the appendices.

Notation: For the convenience of the reader, we collect here some notation used throughout the paper. For $p \in [1, \infty]$, we use $\|x\|_p$ to denote the ℓ_p-norm of a vector $x \in \mathbb{R}^p$, and we let q denote the conjugate exponent, satisfying $\frac{1}{p} + \frac{1}{q} = 1$. For two distributions $\mathbb{P}$ and $\mathbb{Q}$, we use $D(\mathbb{P} \| \mathbb{Q})$ to denote the Kullback-Leibler divergence between the distributions. The notation $\mathbb{I}(A)$ refers to the 0-1 valued indicator random variable of the set A. For two vectors $\alpha, \beta \in \{-1, +1\}^d$, we define the Hamming distance $\Delta_H(\alpha, \beta) := \sum_{i=1}^d \mathbb{I}[\alpha_i \neq \beta_i]$.

2 Background and problem formulation

We begin by introducing background on the oracle model of convex optimization, and precisely defining the problem to be studied.

2.1 Convex optimization in the oracle model

Convex optimization is the task of minimizing a convex function f over a convex set $\mathbb{S} \subseteq \mathbb{R}^d$. Assuming that the minimum is achieved, it corresponds to computing an element x_f^* that achieves the minimum—that is, $x_f^* \in \arg\min_{x \in \mathbb{S}} f(x)$. An *optimization method* is any procedure that solves this task, typically by repeatedly selecting values from $\mathbb{S}$. Our primary focus in this paper is the following question: given any class of convex functions $\mathcal{F}$, what is the minimum computational labor any such optimization method would expend for any function in $\mathcal{F}$?

In order to address this question, we follow the approach of Nemirovski and Yudin [4], and measure computational labor based on the oracle model of optimization. The main components

of this model are an *oracle* and an *information set*. An *oracle* is a (possibly random) function $\phi : \mathbb{S} \mapsto \mathcal{I}$ that answers any query $x \in \mathbb{S}$ by returning an element $\phi(x)$ in an information set $\mathcal{I}$. The information set varies depending on the oracle; for instance, for an exact oracle of m^{th} order, the answer to a query x_t consists of $f(x_t)$ and the first m derivatives of f at x_t. For the case of stochastic oracles studied in this paper, these values are corrupted with zero-mean noise with bounded variance. We then measure the computational labor of any optimization method as the number of queries it poses to the oracle.

In particular, given a positive integer T corresponding to the number of iterations, an optimization method $\mathcal{M}$ designed to approximately minimize the convex function f over the convex set $\mathbb{S}$ proceeds as follows. At any given iteration $t = 1, \ldots, T$, the method $\mathcal{M}$ queries at $x_t \in \mathbb{S}$, and the oracle reveals the information $\phi(x_t, f)$. The method then uses this information to decide at which point x_{t+1} the next query should be made. For a given oracle function ϕ, let $\mathbb{M}_T$ denote the class of all optimization methods $\mathcal{M}$ that make T queries according to the procedure outlined above. For any method $\mathcal{M} \in \mathbb{M}_T$, we define its error on function f after T steps as

$$\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi) := f(x_T) - \min_{x \in \mathbb{S}} f(x) \;=\; f(x_T) - f(x_f^*), \tag{1}$$

where x_T is the method's query at time T. Note that by definition of x_f^* as a minimizing argument, this error is a non-negative quantity.

When the oracle is stochastic, the method's query x_T at time T is itself random, since it depends on the random answers provided by the oracle. In this case, the optimization error $\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)$ is also a random variable. Accordingly, for the case of stochastic oracles, we measure the accuracy in terms of the expected value $\mathbb{E}_\phi[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]$, where the expectation is taken over the oracle randomness. Given a class of functions $\mathcal{F}$ defined over a convex set $\mathbb{S}$ and a class $\mathbb{M}_T$ of all optimization methods based on T oracle queries, we define the minimax error

$$\epsilon_T^*(\mathcal{F}, \mathbb{S}; \phi) := \inf_{\mathcal{M} \in \mathbb{M}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]. \tag{2}$$

In the sequel, we provide results for particular classes of oracles. So as to ease the notation, when the oracle ϕ is clear from the context, we simply write $\epsilon_T^*(\mathcal{F}, \mathbb{S})$.

2.2 Stochastic first-order oracles

In this paper, we study stochastic oracles for which the information set $\mathcal{I} \subset \mathbb{R} \times \mathbb{R}^d$ consists of pairs of noisy function and subgradient evaluations. More precisely, we have:

Definition 1. *For a given set $\mathbb{S}$ and function class $\mathcal{F}$, the class of first-order stochastic oracles consists of mappings $\phi : S \times \mathcal{F} \to \mathcal{I}$ of the form $\phi(x, f) = (\widehat{f}(x), \widehat{z}(x))$ such that*

$$\mathbb{E}[\widehat{f}(x)] = f(x), \qquad \mathbb{E}[\widehat{z}(x)] \in \partial f(x), \qquad \text{and} \qquad \mathbb{E}\big[\, \|\widehat{z}(x)\|_p^2 \,\big] \leq \sigma^2. \tag{3}$$

We use $\mathbb{O}_{p,\sigma}$ to denote the class of all stochastic first-order oracles with parameters (p, σ). Note that the first two conditions imply that $\widehat{f}(x)$ is an unbiased estimate of the function value $f(x)$, and that $\widehat{z}(x)$ is an unbiased estimate of a subgradient $z \in \partial f(x)$. When f is actually differentiable, then $\widehat{z}(x)$ is an unbiased estimate of the gradient $\nabla f(x)$. The third condition in equation (3) controls the "noisiness" of the subgradient estimates in terms of the ℓ_p-norm.

Stochastic gradient methods are a widely used class of algorithms that can be understood as operating based on information provided by a stochastic first-order oracle. As a particular example, consider a function of the separable form $f(x) = \frac{1}{n}\sum_{i=1}^n h_i(x)$, where each h_i is differentiable. Problems of this form arise very frequently in statistical problems, where each term i corresponds to a different sample and the overall cost function is some type of statistical loss (e.g., maximum likelihood, support vector machines, boosting etc.) The natural stochastic gradient method for this problem is to choose an index $i \in \{1, 2, \ldots, n\}$ uniformly at random, and then to return the pair $(h_i(x), \nabla h_i(x))$. Taking averages over the randomly chosen index i yields $\frac{1}{n}\sum_{i=1}^n h_i(x) = f(x)$, so that $h_i(x)$ is an unbiased estimate of $f(x)$, with an analogous unbiased property holding for the gradient of $h_i(x)$.

2.3 Function classes of interest

We now turn to the classes $\mathcal{F}$ of convex functions for which we study oracle complexity. In all cases, we consider real-valued convex functions defined over some convex set $\mathbb{S}$. We assume without loss of generality that $\mathbb{S}$ contains an open set around 0, and many of our lower bounds involve the maximum radius $r = r(\mathbb{S}) > 0$ such that

$$\mathbb{S} \supseteq \mathbb{B}_\infty(r) := \big\{x \in \mathbb{R}^d \mid \|x\|_\infty \leq r\big\}. \tag{4}$$

Our first class consists of bounded Lipschitz functions:

Definition 2. *For a given convex set $\mathbb{S} \subseteq \mathbb{R}^d$ and parameter $p \in [1, \infty]$, the class $\mathcal{F}_{\mathrm{cv}}(\mathbb{S}, L, p)$ consists of all convex functions $f : \mathbb{S} \to \mathbb{R}$ such that*

$$\big|f(x) - f(y)\big| \leq L\,\|x - y\|_q \qquad \text{for all } x, y \in \mathbb{S} \tag{5}$$

where $\frac{1}{q} = 1 - \frac{1}{p}$.

Note that we have defined the Lipschitz condition (5) in terms of the conjugate exponent $q \in [1, \infty]$, defined by the relation $\frac{1}{q} = 1 - \frac{1}{p}$. Our motivation in doing so is to maintain consistency with our definition of the stochastic first-order oracle, in which we assumed that $\mathbb{E}\big[\,\|\widehat{z}(x)\|_p^2\,\big] \leq \sigma^2$. We note that the Lipschitz condition (5) is equivalent to:

$$\|z\|_p \leq L \quad \forall z \in \partial f(x), \quad \text{and for all } x \in \mathbb{S}.$$

If we consider the case of a differentiable function f, the unbiasedness condition in Definition 1 implies that

$$\|\nabla f(x)\|_p = \|\mathbb{E}[\widehat{z}(x)]\|_p \overset{(a)}{\leq} \mathbb{E}\|\widehat{z}(x)\|_p \overset{(b)}{\leq} \sqrt{\mathbb{E}\|\widehat{z}(x)\|_p^2} \leq \sigma,$$

where inequality (a) follows the convexity of the ℓ_p-norm and Jensen's inequality, and inequality (b) is a result of Jensen's inequality applied to the concave function $\sqrt{x}$. This bound implies that f must be Lipschitz with constant at most σ with respect to the dual ℓ_q-norm. Therefore, we necessarily must have $L \leq \sigma$, in order for the function class from Definition 2 to be consistent with the stochastic first-order oracle.

The second function class we consider is that of strongly convex functions.

Definition 3. *For a given convex set $\mathbb{S} \subseteq \mathbb{R}^d$ and parameter $p \in [1, \infty]$, the class $\mathcal{F}_{\mathrm{scv}}(\mathbb{S}, p; L, \gamma)$ consists of all convex functions $f : \mathbb{S} \to \mathbb{R}$ such that the Lipschitz condition (5) holds, and such that f satisfies the ℓ_2-strong convexity condition*

$$f(x) \geq f(y) + \langle z, x - y \rangle + \frac{\gamma^2}{2}\|x - y\|^2 \qquad \text{for all } x, y \in \mathbb{S} \text{ and } z \in \partial f(y). \tag{6}$$

In this paper, we restrict our attention to the case of strong convexity with respect to the ℓ_2-norm. (Similar results on the oracle complexity for strong convexity with respect to different norms can be obtained by straightforward modifications of the arguments given here). For future reference, it should be noted that the Lipschitz constant L and strong convexity constant γ interact with one another. In particular, whenever $\mathbb{S} \subset \mathbb{R}^d$ contains the ℓ_∞-ball of radius r, the Lipschitz L and strong convexity γ constants must satisfy the inequality

$$\frac{L}{\gamma^2} \;\geq\; \frac{1}{4r}\, d^{1/p}. \tag{7}$$

In order to establish this inequality, we note that strong convexity implies that

$$\frac{\gamma^2}{2} \leq \frac{f(x) - f(y) - \langle z, x - y \rangle}{\|x - y\|_2^2} \;\leq \frac{2L\|x - y\|_q}{\|x - y\|_2^2}$$

We now choose the pair $x, y \in \mathbb{S}$ such that $\|x - y\|_\infty = r$ and $\|x - y\|_2 = r\sqrt{d}$. Such a choice is possible whenever $\mathbb{S}$ contains the ℓ_∞ ball of radius r. Since we have $\|x - y\|_q \leq d^{1/q}\|x - y\|_\infty$, this choice yields $\frac{\gamma^2}{2} \leq \frac{2Ld^{\frac{1}{q}-1}}{r}$, which establishes the claim (7).

As a third example, we study the oracle complexity of optimization over the class of convex functions that have sparse minimizers. This class of functions is well-motivated, since a large body of statistical work has studied the estimation of vectors, matrices and functions under various types of sparsity constraints. A common theme in this line of work is that the ambient dimension d enters only logarithmically, and so has a mild effect. Consequently, it is natural to investigate whether the complexity of optimization methods also enjoys such a mild dependence on ambient dimension under sparsity assumptions.

For a vector $x \in \mathbb{R}^d$, we use $\|x\|_0$ to denote the number of non-zero elements in x. Recalling the set $\mathcal{F}_{\mathrm{cv}}(\mathbb{S}, L, p)$ from Definition 2, we now define a class of Lipschitz functions with sparse minimizers.

Definition 4. *For a convex set $\mathbb{S} \subset \mathbb{R}^d$ and positive integer $k \leq \lfloor d/2 \rfloor$, we define*

$$\mathcal{F}_{\mathrm{sp}}(k; \mathbb{S}, L) := \big\{ f \in \mathcal{F}_{\mathrm{cv}}(\mathbb{S}, L, \infty) \mid \exists \quad x^* \in \arg\min_{x \in \mathbb{S}} f(x) \quad \text{satisfying } \|x^*\|_0 \leq k. \big\} \tag{8}$$

We frequently use the shorthand notation $\mathcal{F}_{\mathrm{sp}}(k)$ when the set $\mathbb{S}$ and parameters L are clear from context. In words, the set $\mathcal{F}_{\mathrm{sp}}(k)$ consists of all convex functions that are L-Lipschitz in the ℓ_∞-norm, and have at least one k-sparse optimizer.

3 Main results and their consequences

With the setup of stochastic convex optimization in place, we are now in a position to state the main results of this paper, and to discuss some of their consequences. As previously mentioned, a subset of our results assume that the set $\mathbb{S}$ contains an ℓ_∞ ball of radius $r = r(\mathbb{S})$. Our bounds scale with r, thereby reflecting the natural dependence on the size of the set $\mathbb{S}$. Also, we set the oracle variance bound σ to be the same as the Lipschitz constant L in our results.

3.1 Oracle complexity for convex Lipschitz functions

We begin by analyzing the minimax oracle complexity of optimization for the class of bounded and convex Lipschitz functions $\mathcal{F}_{\mathrm{cv}}$ from Definition 2.

Theorem 1. *Let $\mathbb{S} \subset \mathbb{R}^d$ be a convex set such that $\mathbb{S} \supseteq \mathbb{B}_\infty(r)$ for some $r > 0$. Then the minimax oracle complexity over the class $\mathcal{F}_{\mathrm{cv}}(\mathbb{S}, L, p)$ satisfies the following lower bounds:*

(a) For $1 \leq p \leq 2$,

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{S}; \phi) = \Omega\Big(L\, r\, \sqrt{\frac{d}{T}}\Big). \tag{9}$$

(b) For $p > 2$,

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{S}; \phi) = \Omega\Big(L\, r\, \frac{d^{1-\frac{1}{p}}}{\sqrt{T}}\Big). \tag{10}$$

Remarks: Nemirovski and Yudin [4] proved the lower bound $\Omega\big(\frac{1}{\sqrt{T}}\big)$ for the class $\mathcal{F}_{\mathrm{cv}}$, where the set $\mathbb{S}$ was allowed to vary arbitrarily. Their proof technique and resulting bound did not provide any dimension dependence, as opposed to the bounds provided here. Obtaining this correct dependence highlights the role of the geometry of the set $\mathbb{S}$ in determining the oracle complexity.

In general, our lower bounds cannot be improved, and hence specify the optimal minimax oracle complexity. We consider here some examples to illustrate their sharpness.

(a) Suppose that we make the choice $\mathbb{S} = \mathbb{B}_\infty(1)$. For this choice, we have $r(\mathbb{S}) = 1$, so that we conclude that for all $p \in [1, 2]$,

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{B}_\infty(1); \phi) = \Omega\Big(L\, \sqrt{\frac{d}{T}}\Big).$$

Up to constant factors, this lower bound is sharp for all $p \in [1, 2]$. Indeed, for any convex set $\mathbb{S}$, stochastic gradient descent achieves a matching upper bound (see Section 5.2.4, p. 196 of NY [4], as well as Appendix B in this paper for further discussion).

(b) As another example, suppose that $\mathbb{S} = \mathbb{B}_2(1)$. Observe that this ℓ_2-norm unit ball satisfies the relation $\mathbb{B}_2(1) \supset \frac{1}{\sqrt{d}}\mathbb{B}_\infty(1)$, so that we have $r(\mathbb{B}_2(1)) = 1/\sqrt{d}$. Consequently, for this choice, the lower bound (9) takes the form

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon_T^*(\mathcal{F}_{\mathrm{cv}}, \mathbb{B}_2(1); \phi) = \Omega\Big(L\, \frac{1}{\sqrt{T}}\Big),$$

which is a dimension-independent lower bound. This lower bound for $\mathbb{B}_2(1)$ is indeed tight for $p \in [1, 2]$, and as before, this rate is achieved by stochastic gradient descent [4].

(c) Turning to the case of $p > 2$, the lower bound (10) is matched up to constants by the method of mirror descent using the dual norm $\|\cdot\|_q$; for further discussion, we again refer the reader to Section 5.2.1, p. 190 of NY [4], as well as to Appendix B of this paper. Also, even though this lower bound requires the oracle to have only bounded variance, our proof actually uses a stochastic oracle based on Bernoulli random variables, for which all moments exist. Consequently, at least in general, our results show that there is no hope of achieving faster rates by restricting to oracles with bounds on higher-order moments. This is an interesting contrast to the case of having *less* than two moments, in which the rates are slower. For instance, as shown in Section 5.3.1 of NY [4], suppose that the gradient estimates in a stochastic oracle satisfy the moment bound $\mathbb{E}\|\widehat{z}(x)\|_p^b \leq \sigma^2$ for some $b \in [1, 2)$. In this setting, the oracle complexity is lower bounded by $\Omega\big(T^{-(b-1)/b}\big)$. Since $T^{\frac{b-1}{b}} \ll T^{\frac{1}{2}}$ for all $b \in [1, 2)$, there is a significant penalty in convergence rates for having less than two bounded moments.

3.2 Oracle complexity for strongly convex Lipschitz functions

We now turn to the statement of lower bounds over the class of Lipschitz and strongly convex functions $\mathcal{F}_{\text{scv}}$ from Definition 3. In all these statements, we assume that $\gamma^2 \leq \frac{4Ld^{-1/p}}{r}$, as is required for the definition of $\mathcal{F}_{\text{scv}}$ to be sensible.

Theorem 2. *Let $\mathbb{S} \subset \mathbb{R}^d$ be a convex set with $\mathbb{B}_\infty(r) \subseteq \mathbb{S}$. Then the minimax oracle complexity over the class $\mathcal{F}_{\text{scv}}(\mathbb{S}, p; L, \gamma)$ satisfies the following lower bounds:*

(a) For $p = 1$, we have

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi) = \Omega\left(\frac{L^2}{\gamma^2 T}\right). \tag{11}$$

(b) For $p > 2$, we have

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\text{scv}}, \phi) = \Omega\left(\frac{L^2 d^{1-2/p}}{\gamma^2 T}\right). \tag{12}$$

As with Theorem 1, these lower bounds are sharp. In particular, for $S = \mathbb{B}_\infty(1)$, stochastic gradient descent achieves the rate (11) up to logarithmic factors [7], and a closely related algorithm proposed in very recent work [8] matches the lower bound exactly. Note how the geometry of the set does not come into play, but rather the rate depends only on the Lipschitz parameter L and strong convexity parameter γ^2. For strongly convex functions, these two quantities fully capture the hardness of the problem.

3.3 Oracle complexity for convex Lipschitz functions with sparse optima

Finally, we turn to the oracle complexity of optimization over the class $\mathcal{F}_{\text{sp}}$ from Definition 4.

Theorem 3. *Let $\mathcal{F}_{\mathrm{sp}}$ be the class of all convex functions that are L-Lipschitz with respect to the $\|\cdot\|_\infty$ norm that have a k-sparse optimizer. Then for $k \leq \lfloor \frac{d}{2} \rfloor$, the oracle complexity satisfies the lower bound*

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\mathrm{sp}}, \phi) = \Omega\Big(L\sqrt{\frac{k^2 \log \frac{d}{k}}{T}}\Big). \tag{13}$$

Remark: If $k = \mathcal{O}(d^{1-\delta})$ for some $\delta \in (0,1)$ (so that $\log \frac{d}{k} = \Theta(\log d)$), then this bound is sharp up to constant factors. In particular, suppose that we use mirror descent based on the $\|\cdot\|_{1+\varepsilon}$ norm with $\varepsilon = 2\log d/(2\log d - 1)$. As we discuss in more detail in Appendix B, it can be shown that this technique will achieve a solution accurate to $\mathcal{O}(\sqrt{\frac{k^2 \log d}{T}})$ within T iterations, which matches our lower bound (13) up to constant factors whenever $k = \mathcal{O}(d^{1-\delta})$. To the best of our knowledge, Theorem 3 provides the first tight lower bound on the oracle complexity of sparse optimization.

4 Proofs of results

We now turn to the proofs of our main results. We begin in Section 4.1 by outlining the framework of basic results on which our proofs are based. Sections 4.2 through 4.4 are devoted to the proofs of Theorems 1 through 3 respectively.

4.1 Basic results

We begin by establishing a basic set of results that are exploited in the proofs of the main results. At a high-level, our main idea is to show that the problem of convex optimization is at least as hard as estimating the parameters of Bernoulli variables—that is, the biases of d independent coins. In order to perform this embedding, for a given error tolerance ϵ, we start with an appropriately chosen subset of the vertices of a d-dimensional hypercube, each of which corresponds to some values of the d Bernoulli parameters. For a given function class, we then construct a "difficult" subclass of functions that are indexed by these vertices of the hypercube. We then show that being able to optimize any function in this subclass to ϵ-accuracy requires identifying the hypercube vertex. This is a multiway hypothesis test based on the observations provided by T queries to the stochastic oracle, and we apply the Fano inequality [9] to lower bound the probability of error. In the remainder of this section, we provide more detail on each of steps involved in this embedding.

4.1.1 Constructing a difficult subclass of functions

Our first step is to construct a subclass of functions $\mathcal{G} \subseteq \mathcal{F}$ that we use to derive lower bounds. Any such subclass is parameterized by a subset $\mathcal{V} \subseteq \{-1,+1\}^d$ of the hypercube, chosen as follows. Recalling that Δ_H denotes the Hamming metric, we let $\mathcal{V} = \{\alpha^1, \ldots, \alpha^M\}$ be a subset of the vertices of the hypercube such that

$$\Delta_H(\alpha^j, \alpha^k) \geq \frac{d}{4} \quad \text{for all } j \neq k, \tag{14}$$

meaning that $\mathcal{V}$ is a $\frac{d}{4}$-packing in the Hamming norm. It is a classical fact (e.g., [13]) that one can construct such a set with cardinality $|\mathcal{V}| \geq (2/\sqrt{e})^{d/2}$.

Now let $\mathcal{G}_{\text{base}} = \{f_i^+, f_i^-, \, i = 1, \ldots, d\}$ denote some base set of $2d$ functions defined on the convex set $\mathbb{S}$, to be chosen appropriately depending on the problem at hand. For a given tolerance $\delta \in (0, \frac{1}{4}]$, we define, for each vertex $\alpha \in \mathcal{V}$, the function

$$g_\alpha(x) := \frac{c}{d} \sum_{i=1}^{d} \big\{(1/2 + \alpha_i \delta) f_i^+(x) + (1/2 - \alpha_i \delta) \, f_i^-(x)\big\}. \tag{15}$$

Depending on the result to be proven, our choice of the base functions $\{f_i^+, f_i^-\}$ and the pre-factor c will ensure that each g_α satisfies the appropriate Lipschitz and/or strong convexity properties over $\mathbb{S}$. Moreover, we will ensure that that all minimizers x_α of each g_α are contained within the set $\mathbb{S}$.

Based on these functions and the packing set $\mathcal{V}$, we define the function class

$$\mathcal{G}(\delta) := \big\{g_\alpha, \; \alpha \in \mathcal{V}\big\}. \tag{16}$$

Note that $\mathcal{G}(\delta)$ contains a total of $|\mathcal{V}|$ functions by construction, and as mentioned previously, our choices of the base functions etc. will ensure that $\mathcal{G}(\delta) \subseteq \mathcal{F}$. We demonstrate specific choices of the class $\mathcal{G}(\delta)$ in the proofs of Theorems 1 and 2 to follow.

4.1.2 Optimizing well is equivalent to function identification

We now claim that if a method can optimize over the subclass $\mathcal{G}(\delta)$ up to a certain tolerance, then it must be capable of identifying which function $g_\alpha \in \mathcal{G}(\delta)$ was chosen. We first require a measure for the *closeness* of functions in terms of their behavior near each others' minima. Recall that we use $x_f^* \in \mathbb{R}^d$ to denote a minimizing point of the function f. Given a convex set $S \subseteq \mathbb{R}^d$ and two functions f, g, we define

$$\rho(f, g) := \inf_{x \in S} \big[f(x) + g(x) - f(x_f^*) - g(x_g^*)\big]. \tag{17}$$

The discrepancy measure is non-negative, symmetric in its arguments, and satisfies $\rho(f, g) = 0$ if and only if $x_f^* = x_g^*$, so that we may refer to it as a semi-metric. (It does not satisfy the triangle inequality required to be a metric.)

Given the subclass $\mathcal{G}(\delta)$, we quantify how densely it is packed with respect to the semimetric ρ using the quantity

$$\psi(\mathcal{G}(\delta)) := \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta). \tag{18}$$

We denote this quantity by $\psi(\delta)$ when the class $\mathcal{G}$ is clear from the context. We now state a simple result that demonstrates the utility of maintaining a separation under ρ among functions in $\mathcal{G}(\delta)$.

Lemma 1. *For any $\widetilde{x} \in \mathbb{S}$, there can be at most one function $g_\alpha \in \mathcal{G}(\delta)$ such that*

$$g_\alpha(\widetilde{x}) - \inf_{x \in \mathbb{S}} g_\alpha(x) \leq \frac{\psi(\delta)}{3}. \tag{19}$$

Thus, if we have an element $\widetilde{x} \in \mathbb{S}$ that approximately minimizes one function in the set $\mathcal{G}(\delta)$ up to tolerance $\psi(\delta)$, then it cannot approximately minimize any other function in the set.

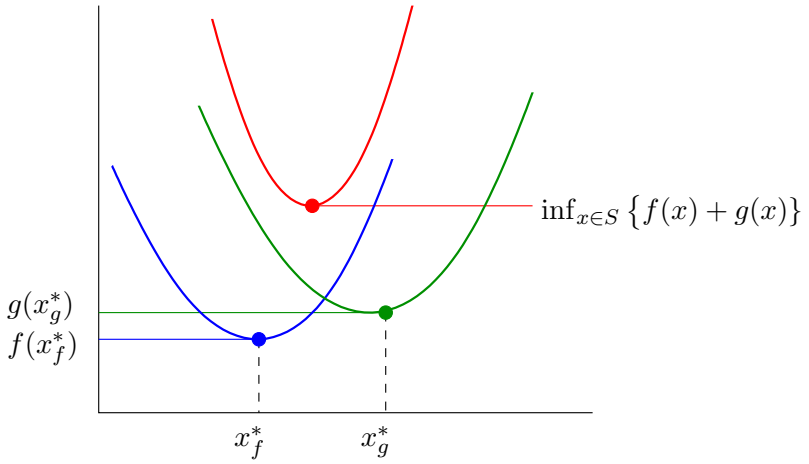

Figure 1. Illustration of the discrepancy function $\rho(f,g)$. The functions f and g achieve their minimum values $f(x_f^*)$ and $g(x_g^*)$ at the points x_f^* and x_g^* respectively.

Proof. For a given $\widetilde{x} \in \mathbb{S}$, suppose that there exists an $\alpha \in \mathcal{V}$ such that $g_\alpha(\widetilde{x}) - g_\alpha(x_\alpha^*) \leq \frac{\psi(\delta)}{3}$. From the definition of $\psi(\delta)$ in (18), for any $\beta \in \mathcal{V},\ \beta \neq \alpha$, we have

$$\psi(\delta) \quad \leq \quad g_\alpha(\widetilde{x}) - \inf_{x\in\mathbb{S}} g_\alpha(x) + g_\beta(\widetilde{x}) - \inf_{x\in\mathbb{S}} g_\beta(x) \quad \leq \quad \frac{\psi(\delta)}{3} + g_\beta(\widetilde{x}) - \inf_{x\in\mathbb{S}} g_\beta(x).$$

Re-arranging yields the inequality $g_\beta(\widetilde{x}) - g_\beta(x_\beta^*) \geq \frac{2}{3}\,\psi(\delta)$, from which the claim (19) follows. □

Suppose that for some fixed but unknown function $g_{\alpha^*} \in \mathcal{G}(\delta)$, some method $\mathcal{M}_T$ is allowed to make T queries to an oracle with information function $\phi(\cdot\,; g_{\alpha^*})$, thereby obtaining the information sequence

$$\phi(x_1^T; g_\alpha^*) := \{\phi(x_t; g_\alpha^*), t = 1, 2, \ldots, T\}.$$

Our next lemma shows that if the method $\mathcal{M}_T$ achieves a low minimax error over the class $\mathcal{G}(\delta)$, then one can use its output to construct a hypothesis test that returns the true parameter α^* at least 2/3 of the time. (In this statement, we recall the definition (2) of the minimax error in optimization.)

Lemma 2. *Suppose that based on the data $\phi(x_1^T; g_\alpha^*)$, there exists a method $\mathcal{M}_T$ that achieves a minimax error satisfying*

$$\mathbb{E}\big[\epsilon_T(\mathcal{M}_T, \mathcal{G}(\delta), \mathbb{S}, \phi)\big] \leq \frac{\psi(\delta)}{9}. \tag{20}$$

Based on such a method $\mathcal{M}_T$, one can construct a hypothesis test $\widehat{\alpha} : \phi(x_1^T; g_\alpha^*) \to \mathcal{V}$ such that $\max_{\alpha^*\in\mathcal{V}} \mathbb{P}_\phi[\widehat{\alpha} \neq \alpha^*] \leq \frac{1}{3}$.

Proof. Given a method $\mathcal{M}_T$ that satisfies the bound (20), we construct an estimator $\widehat{\alpha}(\mathcal{M}_T)$ of the true vertex α^* as follows. If there exists some $\alpha \in \mathcal{V}$ such that $g_\alpha(x_T) - g_\alpha(x_\alpha) \leq \frac{\psi(\delta)}{3}$ then we set $\widehat{\alpha}(\mathcal{M}_T)$ equal to α. If no such α exists, then we choose $\widehat{\alpha}(\mathcal{M}_T)$ uniformly at random from $\mathcal{V}$.

From Lemma 1, there can exist only one such $\alpha \in \mathcal{V}$ that satisfies this inequality. Consequently, using Markov's inequality, we have $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \mathbb{P}_\phi\big[\epsilon_T(\mathcal{M}_T, g_{\alpha^*}, \mathbb{S}, \phi) \geq \psi(\delta)/3\big] \leq \frac{1}{3}$. Maximizing over α^* completes the proof. □

We have thus shown that having a low minimax optimization error over $\mathcal{G}(\delta)$ implies that the vertex $\alpha^* \in \mathcal{V}$ can be identified most of the time.

4.1.3 Oracle answers and coin tosses

We now describe stochastic first order oracles ϕ for which the samples $\phi(x_1^T; g_\alpha)$ can be related to coin tosses. In particular, we associate a coin with each dimension $i \in \{1, 2, \ldots, d\}$, and consider the set of coin bias vectors lying in the set

$$\Theta(\delta) = \big\{(1/2 + \alpha_1\delta, \ldots, 1/2 + \alpha_d\delta) \mid \alpha \in \mathcal{V}\big\}, \tag{21}$$

Given a particular function $g_\alpha \in \mathcal{G}(\delta)$—or equivalently, vertex $\alpha \in \mathcal{V}$—we consider two oracles ϕ that present noisy value and gradient samples from g_α, as summarized in the following.

Oracle A: 1-dimensional unbiased gradients

(a) Pick an index $i \in \{1, \ldots, d\}$ uniformly at random.

(b) Draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i\delta$.

(c) For the given input $x \in \mathbb{S}$, return the value $\widehat{g}_{\alpha,A}(x)$ and a sub-gradient $\widehat{z}_{\alpha,A}(x) \in \partial\widehat{g}_{\alpha,A}(x)$ of the function

$$\widehat{g}_{\alpha,A} := c\big[b_i f_i^+ + (1 - b_i) f_i^-\big].$$

By construction, the function value and gradients returned by Oracle A are unbiased estimates of those of g_α. In particular, since each co-ordinate i is chosen with probability $1/d$, we have

$$\mathbb{E}\big[\widehat{g}_{\alpha,A}(x)\big] = \frac{c}{d}\sum_{i=1}^{d}\big[b_i f_i^+(x) + (1 - b_i) f_i^-(x)\big] \;=\; g_\alpha(x),$$

with a similar relation for the gradient. Furthermore, as long as the base functions f_i^+ and f_i^- have gradients bounded by 1, we have $\mathbb{E}[\|\widehat{z}_{\alpha,A}(x)\|_p] \leq c$ for all $p \in [1, \infty]$.

Parts of proofs are based on an oracle reduces function values and gradients that are *d-dimensional* in nature.

Oracle B: d-dimensional unbiased gradients

(a) For $i = 1, \ldots, d$, draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_i \delta$.

(b) For the given input $x \in \mathbb{S}$, return the value $\widehat{g}_{\alpha,B}(x)$ and a sub-gradient $\widehat{z}_{\alpha,B}(x) \in \partial \widehat{g}_{\alpha,B}(x)$ of the function

$$\widehat{g}_{\alpha,B} := \frac{c}{d} \sum_{i=1}^{d} \big[b_i f_i^+ + (1 - b_i) f_i^- \big].$$

As with Oracle A, this oracle returns unbiased estimates of the function values and gradients. We also note that if the derivatives of f_i^+ and f_i^- are upper bounded in absolute value by 1, we have

$$\|\nabla \widehat{g}_{\alpha,B}(x)\|_p^2 = \frac{c^2}{d^2} \Big(\sum_{i=1}^{d} \big| b_i \partial f_i^+(x) + (1 - b_i) \partial f_i^-(x) \big|^p \Big)^{2/p} \leq c^2 d^{2/p - 2}. \tag{22}$$

In our later uses of Oracles A and B, we will choose the pre-factor c appropriately so as to produce the desired Lipschitz constants.

4.1.4 Lower bounds on coin-tossing

Finally, we use information-theoretic methods to lower bound the probability of correctly estimating the true parameter $\alpha^* \in \mathcal{V}$ in our model. At each round of either Oracle A or Oracle B, we can consider a set of d coin tosses, with an associated vector $\theta^* = (\frac{1}{2} + \alpha_1^* \delta, \ldots, \frac{1}{2} + \alpha_d^* \delta)$ of parameters. At any round, the output of Oracle A can (at most) reveal the instantiation $b_i \in \{0, 1\}$ of a randomly chosen index, whereas Oracle B can at most reveal the entire vector $(b_1, b_2, \ldots, b_d)$. Our goal is to lower bound the probability of estimating the true parameter α^*, based on a sequence of length T.

Lemma 3. *Suppose that the Bernoulli parameter vector α^* is chosen uniformly at random from the packing set $\mathcal{V}$, and suppose that the outcome of $\ell \leq d$ coins chosen uniformly at random is revealed at each round $t = 1, \ldots, T$. Then for any $\delta \in (0, 1/4]$, any hypothesis test $\widehat{\alpha}$ satisfies*

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{16 \ell T \delta^2 + \log 2}{\frac{d}{2} \log(2/\sqrt{e})}, \tag{23}$$

where the probability is taken over both randomness in the oracle and the choice of α^*.

Note that we will apply the lower bound (23) with $\ell = 1$ in the case of Oracle A, and $\ell = d$ in the case of Oracle B.

Proof. For each time $t = 1, 2, \ldots, T$, let U_t denote the randomly chosen subset of size ℓ, and let $Y_t \in \{-1, 0, 1\}^d$ be a random vector with entries

$$Y_{t,i} = \begin{cases} X_{t,i} & \text{if } i \in U_t, \text{ and} \\ -1 & \text{if } i \notin U_t. \end{cases}$$

By Fano's inequality [9], we have the lower bound

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{I(\{(U_t, Y_t\}_{t=1}^T; \alpha^*) + \log 2}{\log |\mathcal{V}|},$$

where $I(\{(U_t, Y_t\}_{t=1}^T; \alpha^*)$ denotes the mutual information between the sequence $\{(U_t, Y_t)\}_{t=1}^T$ and the random parameter vector α^*. As discussed earlier, we are guaranteed that $\log |\mathcal{V}| \geq \frac{d}{2} \log(2/\sqrt{e})$. Consequently, in order to prove the lower bound (23), it suffices to establish the upper bound $I(\{U_t, Y_t\}_{t=1}^T; \alpha^*) \leq 16T \, \ell \, \delta^2$.

By the independent and identically distributed nature of the sampling model, we have

$$I(((U_1, Y_1), \ldots, (U_T, Y_T)); \alpha^*) = \sum_{t=1}^{T} I((U_t, Y_t); \alpha^*) \; = \; T \; I((U_1, Y_1); \alpha^*),$$

so that it suffices to upper bound the mutual information for a single round. To simplify notation, from here onwards we write (Y, U) to mean the pair (Y_1, U_1); the remainder of our proof is devoted to establishing that $I(Y; U) \leq 16 \, \ell \, \delta^2$,

By chain rule for mutual information [9], we have

$$I((U, Y); \alpha^*) = I(Y; \alpha^* \mid U) + I(\alpha^*; U). \tag{24}$$

Since the subset U is chosen independently of α^*, we have $I(\alpha^*; U) = 0$, and so it suffices to upper bound the first term. By definition of conditional mutual information [9], we have

$$I(Y; \alpha^* \mid U) = \mathbb{E}_U \big[D(\mathbb{P}_{Y|\alpha^*, U} \, \| \, \mathbb{P}_{Y|U}) \big]$$

Since α has a uniform distribution over $\mathcal{V}$, we have $\mathbb{P}_{Y|U} = \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} \mathbb{P}_{Y|\alpha, U}$, so that by convexity of the Kullback-Leibler divergence,

$$D(\mathbb{P}_{Y|\alpha^*, U} \, \| \, \mathbb{P}_{Y|U}) \leq \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D(\mathbb{P}_{Y|\alpha^*, U} \, \| \, \mathbb{P}_{Y|\alpha, U}). \tag{25}$$

Now for any pair $\alpha^*, \alpha \in \mathcal{V}$, the KL divergence $D(\mathbb{P}_{Y|\alpha^*, U} \, \| \, \mathbb{P}_{Y|\alpha, U})$ can be at most the KL divergence between ℓ independent pairs of Bernoulli variates with parameters $\frac{1}{2} + \delta$ and $\frac{1}{2} - \delta$. Letting $D(\delta)$ denote the Kullback-Leibler divergence between a single Bernoulli pair, a little calculation yields

$$\begin{aligned} D(\delta) &= (\frac{1}{2} + \delta) \log \frac{\frac{1}{2} + \delta}{\frac{1}{2} - \delta} + (\frac{1}{2} - \delta) \log \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta} \\ &= 2\delta \log \big(1 + \frac{4\delta}{1 - 2\delta}\big) \\ &\leq \frac{8\delta^2}{1 - 2\delta}. \end{aligned}$$

Consequently, as long as $\delta \leq 1/4$, we have $D(\delta) \leq 16\delta^2$. Returning to the bound (25), we conclude that $D(\mathbb{P}_{Y|\alpha^*, U} \, \| \, \mathbb{P}_{Y|U}) \leq 16\ell\delta^2$. Taking averages over U, we obtain the bound $I(Y; \alpha^* \mid U) \leq 16 \, \ell \, \delta^2$, and applying the decomposition (24) yields $I((U, Y); \alpha^*) \leq 16\ell\delta^2$, thereby completing the proof. □

Equipped with these tools, we are now prepared to prove our main results.

4.2 Proof of Theorem 1

We begin with oracle complexity for bounded Lipschitz functions, as stated in Theorem 1. We first prove the result for the set $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$.

Part (a)—Proof for $p \in [1, 2]$: Consider Oracle A that returns the quantities $(\widehat{g}_{\alpha,A}(x), \widehat{z}_{\alpha,A}(x))$. By definition of the oracle, each round reveals only at most one coin flip, meaning that we can apply Lemma 3 with $\ell = 1$, thereby obtaining the lower bound

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\frac{16T\delta^2 + \log 2}{d \log(2/\sqrt{e})}. \tag{26}$$

We now seek an upper bound $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha]$ using Lemma 2. In order to do so, we need to specify the base functions (f_i^+, f_i^-) involved. For $i = 1, \ldots, d$, we define

$$f_i^+(x) := \big|x(i) + \frac{1}{2}\big|, \quad \text{and} \quad f_i^-(x) := \big|x(i) - \frac{1}{2}\big|. \tag{27}$$

Given that $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$, we see that the minimizers of g_α are contained in S. Also, both the functions are 1-Lipschitz in the ℓ_1-norm. By the construction (15), we are guaranteed that for any subgradient of g_α, we have

$$\|z_\alpha(x)\|_p \leq 2c \qquad \text{for all } p \geq 1.$$

Therefore, in order to ensure that g_α is L-Lipschitz in the dual ℓ_q-norm, it suffices to set $c = L/2$.

Let us now lower bound the discrepancy function (17). We first observe that each function g_α is minimized over the set $\mathbb{B}_\infty\big(\frac{1}{2}\big)$ at the vector $x_\alpha := -\alpha/2$, at which point it achieves its minimum value

$$\min_{x \in \mathbb{B}_\infty(\frac{1}{2})} g_\alpha(x) = \frac{cd}{2} - cd\delta.$$

Furthermore, we note that for any $\alpha \neq \beta$, we have

$$\begin{aligned}
g_\alpha(x) + g_\beta(x) &= \frac{c}{d}\sum_{i=1}^d \Big[\big(\frac{1}{2} + \alpha_i + \frac{1}{2} + \beta_i\big) f_i^+(x) + \big(\frac{1}{2} - \alpha_i + \frac{1}{2} - \beta_i\big) f_i^-(x)\Big] \\
&= \frac{c}{d}\sum_{i=1}^d \Big[\big(1 + \alpha_i + \beta_i\big) f_i^+(x) + \big(1 - \alpha_i - \beta_i\big) f_i^-(x)\Big] \\
&= \frac{c}{d}\sum_{i=1}^d \Big[\big(f_i^+(x) + f_i^-(x)\big)\mathbb{I}(\alpha_i \neq \beta_i) + \big((1 + 2\alpha_i) f_i^+(x) + (1 - 2\alpha_i) f_i^-(x)\big)\mathbb{I}(\alpha_i = \beta_i)\Big].
\end{aligned}$$

When $\alpha_i = \beta_i$ then $x_\alpha(i) = x_\beta(i) = -\alpha_i/2$, so that this co-ordinate does not make a contribution to the discrepancy function $\rho(g_\alpha, g_\beta)$. On the other hand, when $\alpha_i \neq \beta_i$, we have

$$f_i^+(x) + f_i^-(x) = \big|x(i) + \frac{1}{2}\big| + \big|x(i) - \frac{1}{2}\big| \geq 1 \quad \text{for all } x \in \mathbb{R}.$$

Consequently, any such co-ordinate yields a contribution of $2c\delta$ to the discrepancy. Recalling our packing set (14) with $d/4$ separation in Hamming norm, we conclude that for any distinct $\alpha \neq \beta$ within our packing set,

$$\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\,\Delta_H(\alpha, \beta) \geq \frac{c\delta}{2},$$

so that by definition of ψ, we have established the lower bound $\psi(\delta) \geq \frac{c\delta}{2}$.

Setting the target error $\epsilon := \frac{c\delta}{18} < 1/2$, we observe that this choice ensures that $\epsilon < \frac{\psi(\delta)}{9}$. Consequently, we may apply Lemma 2 to obtain the upper bound $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. Combining this upper bound with the lower bound (26) yields the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16T\delta^2 + \log 2}{d\log(2/\sqrt{e})}.$$

Recalling that $c = \frac{L}{2}$, making the substitution $\delta = \frac{18\epsilon}{c} = \frac{36\epsilon}{L}$, and performing some algebra yields

$$T = \Omega\big(\frac{L^2 d}{\epsilon^2}\big) \quad \text{for all } d \geq 11.$$

Combined with Theorem 5.3.1 of NY [4], we conclude that this lower bound holds for all dimensions d.

Part (b)—Proof for $p > 2$: The preceding proof based on Oracle A is also valid for $p > 2$, but yields a relatively weak result. Here we show how the use of Oracle B yields the stronger claim stated in Theorem 1(b). When using this oracle, all d coin tosses at each round are revealed, so that Lemma 3 with $\ell = d$ yields the lower bound

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\frac{16\,T\,d\,\delta^2 + \log 2}{d\log(2/\sqrt{e})}. \tag{28}$$

We now seek an upper bound on $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha]$. As before, we use the set $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$, and the previous definitions (27) of $f_i^+(x)$ and $f_i^-(x)$. From our earlier analysis (in particular, equation (22)), the the Lipschitz constant of $g_\alpha(x)$ is at most $cd^{1/p-1}$, so that setting $c = Ld^{1-1/p}$ yields functions that are Lipschitz with parameter L.

As before, for any distinct pair $\alpha, \beta \in \mathcal{V}$, we have the lower bound

$$\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\,\Delta_H(\alpha,\beta) \;\geq\; \frac{c\delta}{2},$$

so that $\psi(\delta) \geq \frac{c\delta}{2}$. Consequently, if we set the target error $\epsilon := \frac{c\delta}{18} < 1/2$, then we are guaranteed that $\epsilon < \frac{\psi(\delta)}{9}$, as is required for applying Lemma 2. Application of this lemma yields the upper bound $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$. Combined with the lower bound (28), we obtain the inequality

$$\frac{1}{3} \geq 1 - 2\frac{16\,d\,T\,\delta^2 + \log 2}{d\log(2/\sqrt{e})}.$$

Substituting $\delta = 18\epsilon/c$ yields the scaling $\epsilon = \mathcal{O}(\frac{c}{\sqrt{T}})$ for all $d \geq 11$. Recalling that $c = Ld^{1-1/p}$, we obtain the bound (10). Combining this with Theorem 5.3.1 of NY [4] gives the claim for all dimensions.

We have thus completed the proof of Theorem 1 in the special case $\mathbb{S} = \mathbb{B}_\infty(\frac{1}{2})$. In order to prove the general claims, which scale with r when $B_\infty(r) \subseteq \mathbb{S}$, we note that our preceding proof required only that $\mathbb{S} \supseteq \mathbb{B}_\infty(\frac{1}{2})$ so that the minimizing points $x_\alpha = -\alpha/2 \in \mathbb{S}$ for all α. In the general case, we define our base functions to be

$$f_i^+(x) = \big|x(i) + \frac{r}{2}\big|, \quad \text{and} \quad f_i^-(x) = \big|x(i) - \frac{r}{2}\big|.$$

With this choice, the functions $g_\alpha(x)$ are minimized at $x_\alpha = -r\alpha/2$, and $\inf_{x\in\mathbb{S}} g_\alpha(x) = cd/2 - cr\delta$. Mimicking the previous steps with $r = 1/2$, we obtain the lower bound

$$\rho(g_\alpha, \beta) \geq \frac{cr\delta}{2} \quad \forall \alpha \neq \beta \in \mathcal{V}.$$

The rest of the proof above did not depend on $\mathbb{S}$, so that we again obtain the lower bound $T = \Omega\big(\frac{d}{\delta^2}\big)$. In this case, the difference in ρ computation means that $\epsilon = \frac{L\delta}{36r}$, from which the general claim follows.

4.3 Proof of Theorem 2

We now turn to the proof of lower bounds on the oracle complexity of the class of strongly convex functions from Definition 3.

Part (a)—Proof for $p = 1$: Let us first consider the special case $\mathbb{S} = \mathbb{B}_\infty(r)$. We start by setting $\gamma^2 = Ld^{-1/p}/r = L/d$, where we have recalled that $p = 1$. In this case, we define the base class

$$f_i^+(x) = \frac{1}{2}\big(x(i) + \frac{r}{2}\big)^2, \quad \text{and} \quad f_i^-(x) = \frac{1}{2}\big(x(i) - \frac{r}{2}\big)^2,$$

for $i = 1, \ldots, d$. With this choice, the functions g_α are strongly convex with respect to the Euclidean norm with coefficient $\gamma = c/d$. Further, each f_i^+, f_i^- is cr-Lipschitz, since it depends on only one co-ordinate. Thus the functions g_α are also cr-Lipschitz with respect to the $\|\cdot\|_1$-norm as before. We consider Oracle A that returns one-dimensional gradients, so that $\mathbb{E}\|\widehat{g}_\alpha(x)\|^2 \leq c^2$. Once again, we need to lower bound the discrepancy $\rho(g_\alpha, g_\beta)$ so as to complete the proof. In this case, the functions $g_\alpha(x)$ are minimized at $x = -\alpha r\delta$ and

$$\min_{x\in\mathbb{B}_\infty(r)} g_\alpha(x) = \frac{cr^2}{2}\left(\frac{1}{4} - \delta^2\right).$$

Now some calculation as in the proof of Theorem 1(a) shows that $\rho(g_\alpha, g_\beta) = \frac{c\delta^2 r^2}{d}\,\Delta_H(\alpha, \beta)$ for all $\alpha \neq \beta$. The remainder of the proof is identical to Theorem 1(a), and we obtain the result by substituting $c = L/r$ and $\gamma^2 = L/dr$. To obtain the result for arbitrary values of $\gamma^2 \leq 4L/dr$, we set

$$f_i^+(x) = r\theta|x(i) + r| + (1-\theta)\left(x(i) + \frac{r}{2}\right)^2, \quad \text{and} \quad f_i^-(x) = r\theta|x(i) - r| + (1-\theta)\left(x(i) - \frac{r}{2}\right)^2.$$

The resulting functions g_α are cr-Lipschitz and $c(1-\theta)/d$-strongly convex. Hence $\gamma^2 = L(1-\theta)/dr$. Hence we can set $\theta \in [0, 1]$ to match the desired value of strong convexity. A little calculation as before shows that for this family of functions,

$$\rho(g_\alpha, g_\beta) = \frac{\delta^2 r^2 c}{4(1-\theta)}, \quad \forall \alpha \neq \beta \in \mathcal{V}.$$

We set $c = L/r$ to match the Lipschitz constant, and substitute $\theta = 1 - \gamma^2 dr/L$. Mimicking the rest of the proof above with these quantities gives

$$T = \Omega\left(\frac{L^2}{\gamma^2\epsilon}\right).$$

Finally, we recall that this proof assumed explicitly that $\mathbb{S} = \mathbb{B}_\infty(r)$. Of course if $\mathbb{S}$ is strictly larger than $\mathbb{B}_\infty(r)$, then the complexity of optimization over $\mathbb{S}$ is only greater than that over $\mathbb{B}_\infty(r)$ and the claim of the theorem follows.

Part (a)—Proof for $p > 2$: We start by setting $\gamma^2 = Ld^{-1/p}/r$. We use the same base class as before, but now switch to Oracle B that returns d-dimensional gradients. As before, we have

$$\mathbb{E}\|\nabla \widehat{g}_{\alpha,A}(x)\|^2 \leq c^2 d^{2/p-2} r^2.$$

We set $c = Ld^{1-1/p}/r$ so as to obtain functions that are L-Lipschitz. In this case, the strong convexity modulus is

$$\gamma^2 = c/d = \frac{Ld^{-1/p}}{r}$$

as desired. Also $\rho(g_\alpha, g_\beta) = \frac{2c\delta^2 r^2}{d}\ \Delta_H(\alpha, \beta)$ for all $\alpha \neq \beta$ as before. The remainder of the proof is identical to Theorem 1(b), and we obtain the result by substituting $c = Ld^{1-1/p}/r$ and $\gamma^2 = Ld^{-1/p}/r$. Finally, the result for arbitrary settings of γ^2 is obtained via the same argument as the proof of part (a).

4.4 Proof of Theorem 3

We begin by constructing an appropriate subset of $\mathcal{F}_{\mathrm{sp}}(k)$ over which the Fano method can be applied. Let $\mathcal{V}(k) := \{\alpha^1, \ldots, \alpha^M\}$ be a set of vectors, such that each $\alpha^j \in \{-1, 0, +1\}^d$ satisfies

$$\|\alpha^j\|_0 = k \quad \text{for all } j = 1, \ldots, M, \text{ and} \qquad \Delta_H(\alpha^j, \alpha^\ell) \geq \frac{k}{2} \quad \text{for all } j \neq \ell.$$

It can be shown that there exists such a packing set with $|\mathcal{V}(k)| \geq \exp\left(\frac{k}{2}\log\frac{d-k}{k/2}\right)$ elements (e.g., see Lemma 5 in Raskutti et al. [14]).

For any $\alpha \in \mathcal{V}(k)$, we define the function

$$g_\alpha(x) := c\left[\sum_{i=1}^d \left\{\left(\frac{1}{2} + \alpha_i\delta\right)\left|x(i) + \frac{1}{2}\right| + \left(\frac{1}{2} - \alpha_i\delta\right)\left|x(i) - \frac{1}{2}\right|\right\} + \delta\sum_{i=1}^d |x(i)|\right]. \tag{29}$$

In this definition, the quantity $c > 0$ is a pre-factor to be chosen later, and $\delta \in (0, \frac{1}{4}]$ is a given error tolerance. Observe that each function $g_\alpha \in \mathcal{G}(\delta; k)$ is convex, and Lipschitz with parameter c with respect to the $\|\cdot\|_\infty$ norm.

Central to the remainder of the proof is the function class $\mathcal{G}(\delta; k) := \{g_\alpha,\ \alpha \in \mathcal{V}(k)\}$. In particular, we need to control the discrepancy $\psi(\delta; k) := \psi(\mathcal{G}(\delta; k))$ for this class. The following result, proven in Appendix A, provides a suitable lower bound:

Lemma 4. *We have*

$$\psi(\delta; k) \ = \ \inf_{\alpha \neq \beta \in \mathcal{V}(k)} \rho(g_\alpha, g_\beta) \geq \frac{ck\delta}{4}. \tag{30}$$

Using Lemma 4, we may complete the proof of Theorem 3. Define the base functions

$$f_i^+(x) := \big|x(i) + \frac{1}{2}\big| + \frac{\delta}{2}|x(i)|, \quad \text{and} \quad f_i^-(x) := \big|x(i) - \frac{1}{2}\big| + \frac{\delta}{2}|x(i)|.$$

Consider Oracle B, which returns d-dimensional gradients based on the function

$$\widehat{g}_{\alpha,B}(x) = \frac{c}{d}\sum_{i=1}^{d}\big[b_i f_i^+(x) + (1-b_i) f_i^-(x)\big],$$

where $\{b_i\}$ are Bernoulli variables. By construction, the function $\widehat{g}_{\alpha,B}$ is at most $3c$-Lipschitz in ℓ_∞ norm, so that setting $c = \frac{L}{3}$ yields an L-Lipschitz function.

Our next step is to use Fano's inequality [9] to lower bound the probability of error in the multiway testing problem associated with this stochastic oracle, following an argument similar to (but somewhat simpler than) the proof of Lemma 3. Fano's inequality yields the lower bound

$$\mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq 1 - \frac{\frac{1}{\binom{|\mathcal{V}|}{2}}\sum_{\alpha\neq\beta} D(\mathbb{P}_\alpha \,\|\mathbb{P}_\beta) + \log 2}{\log|\mathcal{V}|}. \tag{31}$$

(As in the proof of Lemma 3, we have used convexity of mutual information [9] to bound it by the average of the pairwise KL divergences.) By construction, any two parameters $\alpha, \beta \in \mathcal{V}$ differ in at most $2k$ places, and the remaining entries are all zeroes in both vectors. The proof of Lemma 3 shows that for $\delta \in [0, \frac{1}{4}]$, each of these $2k$ places makes a contribution of at most $16\delta^2$. Recalling that we have T samples, we conclude that $D(\mathbb{P}_\alpha \,\|\mathbb{P}_\beta) \leq 32kT\delta^2$. Substituting this upper bound into the Fano lower bound (31) and recalling that the cardinality of $\mathcal{V}$ is at least $\exp\big(\frac{k}{2}\log\frac{d-k}{k/2}\big)$, we obtain

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq 1 - 2\left(\frac{32kT\delta^2 + \log 2}{\frac{k}{2}\log\frac{d-k}{k/2}}\right) \tag{32}$$

By Lemma 4 and our choice $c = L/3$, we have

$$\psi(\delta) \geq \frac{ck\delta}{4} \;=\; \frac{Lk\delta}{12},$$

Therefore, if we aim for the target error $\epsilon = \frac{Lk\delta}{108}$, then we are guaranteed that $\epsilon \leq \frac{\psi(\delta)}{9}$, as is required for the application of Lemma 2. This lemma implies that $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq 1/3$, which when combined with the earlier bound (32) yields

$$\frac{1}{3} \geq 1 - 2\left(\frac{32kT\delta^2 + \log 2}{\frac{k}{2}\log\frac{d-k}{k/2}}\right).$$

Rearranging yields the lower bound

$$T = \Omega\left(\frac{\log\frac{d-k}{k/2}}{\delta^2}\right) \;=\; \Omega\left(L^2\, k^2\, \frac{\log\frac{d-k}{k/2}}{\epsilon^2}\right),$$

where the second step uses the relation $\delta = \frac{108\epsilon}{Lk}$. As long as $k \leq \lfloor d/2 \rfloor$, we have $\log\frac{d-k}{k/2} = \Theta(\log\frac{d}{k})$, which completes the proof.

5 Discussion

In this paper, we have studied the complexity of convex optimization within the stochastic first-order oracle model. We derived lower bounds for various function classes, including convex functions, strongly convex functions, and convex functions with sparse optima. As we discussed, our lower bounds are sharp in general, since there are matching upper bounds achieved by known algorithms, among them stochastic gradient descent and stochastic mirror descent. Our bounds also reveal various dimension-dependent and geometric aspects of the stochastic oracle complexity of convex optimization. An interesting aspect of our proof technique is the features in common statistical minimax theory. In particular, our proofs are based on constructing packing sets, defined with respect to a semi-metric that measures how the degree of separation between the optima of different functions. We then leveraged information-theoretic techniques, in particular Fano's inequality and its variants, in order to establish lower bounds.

There are various directions for future research. It would be interesting to consider the effect of memory constraints on the complexity of convex optimization, or to derive lower bounds for problems of distributed optimization. We suspect that the proof techniques developed in this paper may be useful for studying these related problems.

Acknowledgements

AA and PLB gratefully acknowledge partial support from NSF awards DMS-0707060 and DMS-0830410 and DARPA-HR0011-08-2-0002. AA was also supported in part by a Microsoft Research Fellowship. MJW and PR were partially supported by funding from the National Science Foundation (DMS-0605165, and DMS-0907632). In addition, MJW received funding from the Air Force Office of Scientific Research (AFOSR-09NL184). We also thank the anonymous reviewers at NIPS 2009 for the helpful suggestions.

A Proof of Lemma 4

Recalling the definition (17) of the discrepancy ρ, we need to compute the quantities $\inf_{x \in \mathbb{S}} g_\alpha(x)$ and $\inf_{x \in \mathbb{S}}\{g_\alpha(x)+g_\beta(x)\}$. Beginning with the former quantity, first observe that for any $x \in \mathbb{B}_\infty(\frac{1}{2})$, we have

$$\left[\frac{1}{2} + \alpha_i\delta\right] \left|x(i) + \frac{1}{2}\right| + \left[\frac{1}{2} - \alpha_i\delta\right] \left|x(i) - \frac{1}{2}\right| = \frac{1}{2} + 2\alpha_i\delta x(i). \tag{33}$$

We now consider one of the individual terms arising in the definition (15) of the function g_α. Using the relation (33), it can be written as

$$\begin{aligned}\left(\frac{1}{2} + \alpha_i\delta\right) f_i^+(x) + \left(\frac{1}{2} - \alpha_i\delta\right) f_i^-(x) &= \left(\frac{1}{2} + \alpha_i\delta\right)\left|x(i) + \frac{1}{2}\right| + \left(\frac{1}{2} - \alpha_i\delta\right)\left|x(i) - \frac{1}{2}\right| + \delta|x(i)| \\ &= \begin{cases} \frac{1}{2} + (2\alpha_i + 1)\delta x(i) & \text{if } x(i) \geq 0 \\ \frac{1}{2} + (2\alpha_i - 1)\delta x(i) & \text{if } x(i) \leq 0 \end{cases}\end{aligned}$$

From this representation, we see that whenever $\alpha_i \neq 0$, then the i^{th} term in the summation defining g_α minimized at $x(i) = -\alpha_i/2$, at which point it takes on its minimum value $1/2 - \delta/2$. On the

other hand, for any term with $\alpha_i = 0$, the function is minimized at $x(i) = 0$ with associated minimum value of $1/2$. Combining these two facts shows that the vector $-\frac{\alpha}{2}$ is an element of the set $\arg\min_{x \in \mathbb{S}} g_\alpha(x)$, and moreover that

$$\inf_{x \in \mathbb{S}} g_\alpha(x) = c\big(\frac{d}{2} - k\frac{\delta}{2}\big). \tag{34}$$

We now turn to the computation of $\inf_{x \in \mathbb{S}}\{g_\alpha(x) + g_\beta(x)\}$. From the relation (33) and the definitions of g_α and g_β, some algebra yields

$$\inf_{x \in \mathbb{S}} \big\{g_\alpha(x) + g_\beta(x)\big\} = c \inf_{x \in \mathbb{S}} \sum_{i=1}^{d} \Big\{1 + 2\delta\big[(\alpha_i + \beta_i)x(i) + |x(i)|\big]\Big\}. \tag{35}$$

Let us consider the minimizer of the i^{th} term in this summation. First, suppose that $\alpha_i \neq \beta_i$, in which case there are two cases to consider.

- If $\alpha_i \neq \beta_i$ and neither α_i nor β_i is zero, then we must have $\alpha_i + \beta_i = 0$, so that the minimum value of 1 is achieved at $x(i) = 0$.
- Otherwise, suppose that $\alpha_i \neq 0$ and $\beta_i = 0$. In this case, we see from Equation (35) that it is equivalent to minimize $1 + \alpha_i x(i) + |x(i)|$. Setting $x(i) = -\alpha_i$ achieves the minimum value of 1.

In the remaining two cases, we have $\alpha_i = \beta_i$.

- If $\alpha_i = \beta_i \neq 0$, then the component is minimized at $x(i) = -\alpha_i/2$ and the minimum value along the component is $1 - 2\delta$.
- If $\alpha_i = \beta_i = 0$, then the minimum value is 1, achieved at $x(i) = 0$.

Consequently, accumulating all of these individual cases into a single expression, we obtain

$$\inf_{x \in \mathbb{S}} \big\{g_\alpha(x) + g_\beta(x)\big\} = c\,\big(d - 2\delta \sum_{i=1}^{d} \mathbb{I}[\alpha_i = \beta_i \neq 0]\big). \tag{36}$$

Finally, combining equations (34) and (36) in the definition of ρ, we find that

$$\begin{aligned}
\rho(g_\alpha, g_\beta) &= c\big[d - \delta \sum_{i=1}^{d} \mathbb{I}[\alpha_i = \beta_i \neq 0] - 2\big(\frac{d}{2} - \frac{k\delta}{2}\big)\big] \\
&= \frac{c\delta}{2}\left[2k - \sum_{i=1}^{d} \mathbb{I}[\alpha_i = \beta_i \neq 0]\right] \\
&= \frac{c}{2}\delta\Delta_H(\alpha, \beta),
\end{aligned}$$

where the second equality follows since α and β have exactly k non-zero elements each. Finally, since $\mathcal{V}$ is an $k/2$-packing set in Hamming distance, we have $\Delta_H(\alpha, \beta) \geq k/2$, which completes the proof.

B Upper bounds via mirror descent

This appendix is devoted to background on the family of mirror descent methods. We first describe the basic form of the algorithm and some known convergence results, before showing that different forms of mirror descent provide matching upper bounds for several of the lower bounds established in this paper, as discussed in the main text.

B.1 Background on mirror descent

Mirror descent is a generalization of (projected) stochastic gradient descent, first introduced by Nemirovski and Yudin [4]; here we follow a more recent presentation of it due to Beck and Teboulle [15]. For a given norm $\|\cdot\|$, let $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a differentiable function that is 1-strongly convex with respect to $\|\cdot\|$, meaning that

$$\Phi(y) \geq \Phi(x) + \langle \nabla\Phi(x),\, y - x\rangle + \frac{1}{2}\|y - x\|^2.$$

We assume that Φ is a function of Legendre type [16, 17], which implies that the conjugate dual Φ^* is differentiable on its domain with $\nabla\Phi^* = \big(\nabla\Phi\big)^{-1}$. For a given proximal function, we let D_Φ be the Bregman divergence induced by Φ, given by

$$D_\Phi(x, y) := \Phi(x) - \Phi(y) - \langle \nabla\Phi(y),\, x - y\rangle. \tag{37}$$

With this set-up, we can now describe the mirror descent algorithm based on the proximal function Φ for minimizing a convex function f over a convex set $\mathbb{S}$ contained within the domain of Φ. Starting with an arbitrary initial $x_0 \in \mathbb{S}$, it generates a sequence $\{x_t\}_{t=0}^\infty$ contained within $\mathbb{S}$ via the updates

$$x_{t+1} = \arg\min_{x \in \mathbb{S}} \big\{\eta_t \langle x,\, \nabla f(x_t)\rangle + D_\Phi(x, x_t)\big\}, \tag{38}$$

where $\eta_t > 0$ is a stepsize.

A special case of this algorithm is obtained by choosing the proximal function $\Phi(x) = \frac{1}{2}\|x\|_2^2$, which is 1-strongly convex with respect to the Euclidean norm. The associated Bregman divergence $D_\Phi(x, y) = \frac{1}{2}\|x - y\|_2^2$ is simply the Euclidean norm, so that the updates (38) correspond to a standard projected gradient descent method. If one receives only an unbiased estimate of the gradient $\nabla f(x_t)$, then this algorithm corresponds to a form of projected stochastic gradient descent. Moreover, other choices of the proximal function lead to different stochastic algorithms, as discussed below.

In order to state convergence rates for this algorithm, convexity and Lipschitz assumptions about the functions f are required. Following the set-up used in our lower bound analysis, we assume that $\|\nabla\widehat{z}(x_t)\|_* \leq L$ for all $x \in \mathbb{S}$, where $\|v\|_* := \sup_{\|x\| \leq 1}\langle x,\, v\rangle$ is the dual norm defined by $\|\cdot\|$. Given stochastic mirror descent based on unbiased estimates of the gradient, Beck and Teboulle [15] showed that with the initialization $x_0 = \arg\min_{x \in \mathbb{S}} \Phi(x)$ and stepsizes $\eta_t = 1/\sqrt{t}$, the optimization error of the sequence $\{x_t\}$ is bounded as

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[f(x_t) - f(x^*)\big] \leq L\sqrt{\frac{D_\Phi(x^*, x_1)}{T}} \;\leq\; L\sqrt{\frac{\Phi(x^*)}{T}} \tag{39}$$

Note that this averaged convergence is a little different from the convergence of x_T discussed in our lower bounds. In order to relate the two quantities, observe that by Jensen's inequality

$$\mathbb{E}\left[f\left(\frac{\sum_{t=1}^{T} x_t}{T}\right)\right] \leq \frac{1}{T}\mathbb{E}\big[f(x_t)\big].$$

Consequently, based on mirror descent for $T-1$ rounds, we may set $x_T = \frac{1}{T-1}\sum_{t=1}^{T-1} x_t$ so as to obtain the same convergence bounds up to constant factors. In the following discussion, we assume this choice of x_T for comparing the mirror descent upper bounds to our lower bounds.

B.2 Matching upper bounds

Now consider the form of mirror descent obtained by choosing the proximal function

$$\Phi_a(x) := \frac{1}{(a-1)}\|x\|_a^2 \qquad \text{for } 1 < a \leq 2. \tag{40}$$

Note that this proximal function is 1-strongly convex with respect to the ℓ_a-norm, meaning that

$$\frac{1}{(a-1)}\|x\|_a^2 \geq \frac{1}{(a-1)}\|y\|_a^2 + \left(\nabla \frac{1}{(a-1)}\|x\|_a^2\right)^T (x-y) + \frac{1}{2}\|x-y\|_a^2.$$

Upper bounds for Theorem 1: For this case, we use mirror descent based on the proximal function Φ_a with $a = q$. Under the condition $\|x^*\|_\infty \leq 1$, a condition which holds in our lower bounds, we obtain

$$\|x^*\|_q \leq \|x^*\|_\infty \, d^{1/q} \; = \; d^{1/q},$$

which implies that $\Phi_q(x^*) = \mathcal{O}(d^{2/q})$. Under the conditions of Theorem 1, we have $\|\nabla f(x_t)\|_p \leq L$ where $p = 1 - 1/q$ defines the dual norm. Note that the condition $1 < q \leq 2$ implies that $p \geq 2$. Substituting this in the upper bound (39) yields

$$\mathbb{E}\big[f(x_T) - f(x^*)\big] = \mathcal{O}\Big(L\,\sqrt{d^{2/q}/T}\Big) \; = \; \mathcal{O}\Big(L d^{1-1/p}\sqrt{\frac{1}{T}}\Big),$$

which matches the lower bound from Theorem 1(b). Moreover, setting $q = p = 2$ corresponds to stochastic gradient descent, and yields an upper bound to match the lower bound of Theorem 1(a).

Upper bounds for Theorem 3: In order to recover matching upper bounds in this case, we use the function Φ_a from equation (40) with $a = \frac{2\log d}{2\log d - 1}$. In this case, the resulting upper bound (39) on the convergence rate takes the form

$$\mathbb{E}\Big[f(x_T) - f(x^*)\Big] \; = \; \mathcal{O}\Big(L\sqrt{\frac{\|x^*\|_a^2}{2(a-1)T}}\Big) = \mathcal{O}\Big(L\sqrt{\frac{\|x^*\|_a^2 \log d}{T}}\Big), \tag{41}$$

since $\frac{1}{a-1} = 2\log d - 1$. Based on the conditions of Theorem 3, we are guaranteed that x^* is k-sparse, with every component bounded by 1 in absolute value, so that $\|x^*\|_a^2 \leq k^{2/a} \leq k^2$, where the final inequality follows since $a > 1$. Substituting this upper bound back into Equation (41) yields

$$\mathbb{E}\big[f(x_T) - f(x^*)\big] = \mathcal{O}\Big(L\sqrt{\frac{k^2 \log d}{T}}\Big).$$

Note that whenever $k = \mathcal{O}(d^{1-\delta})$ for some $\delta > 0$, then we have $\log d = \Theta(\log \frac{d}{k})$, in which case this upper bound matches the lower bound from Theorem 3 up to constant factors, as claimed.

References

[1] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.

[2] D. Bertsekas, *Nonlinear programming*. Belmont, MA: Athena Scientific, 1995.

[3] Y. Nesterov, *Introductory lectures on convex optimization: Basic course*. Kluwer Academic Publishers, 2004.

[4] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. John Wiley UK/USA, 1983.

[5] A. S. Nemirovski, "Efficient methods in convex programming," *Lecture notes*. [Online]. Available: http://www2.isye.gatech.edu/~nemirovs/OPTI_LectureNotes.pdf

[6] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, "Information-theoretic lower bounds on the oracle complexity of convex optimization," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1–9.

[7] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Mach. Learn.*, vol. 69, no. 2-3, pp. 169–192, 2007.

[8] E. Hazan and S. Kale, "An optimal algorithm for stochastic strongly-convex optimization," 2010. [Online]. Available: http://arxiv.org/PS_cache/arxiv/pdf/1006/1006.2425v1.pdf

[9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[10] R. Z. Has'minskii, "A lower bound on the risks of nonparametric estimates of densities in the uniform metric," *Theory Prob. Appl.*, vol. 23, pp. 794–798, 1978.

[11] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrsch. verw. Gebiete*, vol. 65, pp. 181–327, 1983.

[12] B. Yu, "Assouad, Fano and Le Cam," in *Festschrift in Honor of L. Le Cam on his 70th Birthday*. Springer-Verlag, 1993.

[13] J. Matousek, *Lectures on discrete geometry*. New York: Springer-Verlag, 2002.

[14] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," 2009. [Online]. Available: http://arxiv.org/abs/0910.2042

[15] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167 – 175, 2003.

[16] G. Rockafellar, *Convex Analysis*. Princeton: Princeton University Press, 1970.

[17] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms.* New York: Springer-Verlag, 1993, vol. 1.