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

Alekh Agarwal
Computer Science Division
UC Berkeley
alekh@cs.berkeley.edu

Peter Bartlett
Computer Science Division
Department of Statistics
UC Berkeley
bartlett@cs.berkeley.edu

Pradeep Ravikumar
Department of Computer Sciences
UT Austin
pradeepr@cs.utexas.edu

Martin J. Wainwright
Department of EECS, and
Department of Statistics
UC Berkeley
wainwrig@eecs.berkeley.edu

Abstract

Despite a large literature on upper bounds on complexity of convex optimization, relatively less attention has been paid to the fundamental hardness of these problems. Given the extensive use of convex optimization in machine learning and statistics, gaining a 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. We also discuss implications of these results for the understanding the inherent complexity of large-scale learning and estimation problems.

1 Introduction

Convex optimization forms the backbone of many algorithms for statistical learning and estimation. In large-scale learning problems, in which the problem dimension and/or data are large, it is essential to exploit bounded computational resources in a (near)-optimal manner. For such problems, understanding the computational complexity of convex optimization is a key issue.

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 alogrithms, among them gradient descent, mirror descent, interior point programming, and stochastic gradient descent, to name a few. We refer the reader to standard texts on optimization (e.g., [4, 1, 10]) 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 [8] (hereafter referred to as NY). One obstacle to a classical complexity-theoretic analysis, as the authors observed, was 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 [8], we refer the reader to Nesterov [10] or the lecture notes by Nemirovski [7].

In this paper, we consider the computational complexity of stochastic convex optimization in 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 [8] 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. Our proofs exploit a new notion of the discrepancy between two functions that appears to be natural for optimization problems. They are based on 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.

2 Background and problem formulation

In this section, we introduce background on the oracle model of complexity for convex optimization, and then define the oracles considered in this paper.

2.1 Convex optimization in the oracle model

Convex optimization is the task of minimizing a convex function f over a convex set $S \subseteq \mathbb{R}^d$. Assuming that the minimum is achieved, it corresponds to computing an element x_f^* that achieves the minimum—that is, $x_f^* \in \arg\min_{x \in S} f(x)$. An *optimization method* is any procedure that solves this task, typically by repeatedly selecting values from 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 [8], based on the oracle model of optimization. More precisely, an *oracle* is a (possibly random) function $\phi : S \mapsto \mathcal{I}$ that answers any query $x \in S$ 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 k^{th} order, the answer to a query x_t consists of $f(x_t)$ and the first k 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.

Given some number of rounds T, an optimization method $\mathcal{M}$ designed to approximately minimize the convex function f over the convex set S proceeds as follows: at any given round $t = 1, \ldots, T$, the method $\mathcal{M}$ queries at $x_t \in S$, 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

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

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

2.2 Minimax error

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 $\epsilon(\mathcal{M}, f, S, \phi)$ is also a random variable. Accordingly, for the case of stochastic oracles, we measure the accuracy in terms of the expected value $\mathbb{E}_\phi[\epsilon(\mathcal{M}, f, S, \phi)]$, where the expectation is taken over the oracle randomness.

Given a class of functions $\mathcal{F}$, and the class $\mathbb{M}_T$ of optimization methods making T oracle queries, we can define the minimax error

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

Note that this definition depends on the optimization set S. In order to obtain uniform bounds, we define $\mathbb{S} := \{S \subseteq \mathbb{R}^d : S \text{ convex}, \|x-y\|_\infty \leq 1 \text{ for } x, y \in S\}$, and consider the worst-case average error over all $S \in \mathbb{S}$, given by

$$\epsilon^*(\mathcal{F}, \phi) := \sup_{S \in \mathbb{S}} \epsilon^*(\mathcal{F}, S, \phi). \tag{3}$$

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

It is worth noting that oracle complexity measures only the number of queries to the oracle—for instance, the number of (approximate) function or gradient evaluations. However, it does not track computational cost within each component of the oracle query (e.g., the actual flop count associated with evaluating the gradient).

2.3 Types of Oracle

In this paper we study the class of stochastic first order oracles, which we will denote simply by $\mathcal{O}$. For this class of oracles, the information set $\mathcal{I}$ consists of pairs of noisy function and gradient evaluations; consequently, any oracle ϕ in this class can be written as

$$\phi(x, f) = (\widehat{f}(x), \widehat{g}(x)), \tag{4}$$

where $\widehat{f}(x)$ and $\widehat{g}(x)$ are random variables that are unbiased as estimators of the function and gradient values respectively (i.e., $\mathbb{E}\widehat{f}(x) = f(x)$ and $\mathbb{E}\widehat{g}(x) = \nabla f(x)$). Moreover, we assume that both $\widehat{f}(x)$ and $\widehat{g}(x)$ have variances bounded by one. When the gradient is not defined at x, the notation $\nabla f(x)$ should be understood to mean any arbitrary subgradient at x. Recall that a subgradient of a convex function f is any vector $v \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + v^\top (y - x).$$

Stochastic gradient methods are popular examples of algorithms for such oracles.

Notation: For the convenience of the reader, we collect here some notation used throughout the paper. We use x_1^t to refer to the sequence $(x_1, \ldots, x_t)$. We refer to the i-th coordinate of any vector $x \in \mathbb{R}^d$ as $x(i)$. For a convex set S, the radius of the largest inscribed and smallest circumscribing ℓ_∞ balls are denoted as r_∞ and R_∞ resp. For a convex function f, its minimizer over a set S will be denoted as x_f^* when S is obvious from context. We will often use the notation x_α^* to denote the minimizer of f_α if α is an index variable over a class. For two distributions p and q, $\mathrm{KL}(p||q)$ refers to the Kullback Leibler divergence between the distributions. The notation $\mathbb{I}(A)$ is the 0-1 valued indicator random variable of the set (equivalently event) A. For two vectors $\alpha, \beta \in \{-1, +1\}^d$, we define the Hamming distance $\Delta_H(\alpha, \beta) := \sum_{i=1}^d \mathbb{I}[\alpha_i \neq \beta_i]$.

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. In particular, we provide some tight lower bounds on the complexity of stochastic oracle optimization. We begin by analyzing the minimax oracle complexity of optimization for the class of *convex Lipschitz functions.* Recall that a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^d$ and $\lambda \in (0, 1)$, we have the inequality $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. For some constant $L > 0$, we say that the function f is L-Lipschitz on S if $|f(x) - f(y)| \leq L\|x - y\|_\infty$ for all $x, y \in S$.

Before stating the results, we note that scaling the Lipschitz constant scales minimax optimization error linearly. Hence, to keep our results scale-free, we consider 1-Lipschitz functions only. As the diameter of S is also bounded by 1, this automatically enforces that $|f(x)| \leq 1, \; \forall x \in S$.

Theorem 1. *Let $\mathcal{F}^C$ be the class of all bounded convex 1-Lipschitz functions on $\mathbb{R}^d$. Then there is a constant c (independent of d) such that*

$$\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}^C, \phi) \geq c\sqrt{\frac{d}{T}}. \tag{5}$$

Remarks: This lower bound is tight in the minimax sense, since the method of stochastic gradient descent attains a matching upper bound for all stochastic first order oracles for any convex set S (see Chapter 5 of NY [8]). Also, even though this lower bound requires the oracle to have only bounded variance, we will use an oracle based on Bernoulli random variables, which has all moments bounded. As a result there is no hope to get faster rates in a simple way by assuming bounds on higher moments for the oracle. This is in interesting contrast to the case of having less than 2 bounded moments where we get slower rates (again, see Chapter 5 of NY [8]).

The above lower bound is obtained by considering the worst case over all convex sets. However, we expect optimization over a smaller convex set to be easier than over a large set. Indeed, we can easily obtain a corollary of Theorem 1 that quantifies this intuition.

Corollary 1. *Let $\mathcal{F}^C$ be the class of all bounded convex 1-Lipschitz functions on $\mathbb{R}^d$. Let S be a convex set such that it contains an ℓ_∞ ball of radius r_∞ and is contained in an ℓ_∞ ball of radius R_∞. Then there is a universal constant c such that,*

$$\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}^C, S, \phi) \geq c \frac{r_\infty}{R_\infty} \sqrt{\frac{d}{T}}. \tag{6}$$

Remark: The ratio $\frac{r_\infty}{R_\infty}$ is also common in results of [8], and is called the asphericity of S. As a particular application of above corollary, consider S to be the unit ℓ_2 ball. Then $r_\infty = \frac{1}{\sqrt{d}}$, and $R_\infty = 1$. which gives a dimension independent lower bound. This lower bound for the case of the ℓ_2 ball is indeed tight, and is recovered by the stochastic gradient descent algorithm [8].

Just as optimization over simpler sets gets easier, optimization over simple function classes should be easier too. A natural function class that has been studied extensively in the context of better upper bounds is that of strongly convex functions. For any given norm $\|\cdot\|$ on S, a function f is *strongly convex with coefficient* κ means that $f(x) \geq f(y) + \nabla f(y)^T(x-y) + \frac{\kappa}{2}\|x-y\|^2$ for all $x, y \in S$. For this class of functions, we obtain a smaller lower bound on the minimax oracle complexity of optimization.

Theorem 2. *Let $\mathcal{F}^S$ be the class of all bounded strongly convex and 1-Lipschitz functions on $\mathbb{R}^d$. Then there is a universal constant c such that,*

$$\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}^S, \phi) \geq c \frac{d}{T}. \tag{7}$$

Once again there is a matching upper bound using stochastic gradient descent for example, when the strong convexity is with respect to the ℓ_2 norm. The corollary depending on the geometry of S follows again.

Corollary 2. *Let $\mathcal{F}^S$ be the class of all bounded convex 1-Lipschitz functions on $\mathbb{R}^d$. Let S be a convex set such that it contains an ℓ_∞ ball of radius r_∞ and is contained in an ℓ_∞ ball of radius R_∞. Then there is a universal constant c such that* $\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}^S, S, \phi) \geq c\left(\frac{r_\infty}{R_\infty}\right)^2 \frac{d}{T}$.

In comparison, Nemirovski and Yudin [8] obtained a lower bound scaling as $\Omega\left(\frac{1}{\sqrt{T}}\right)$ for the class $\mathcal{F}^C$. Their bound applies only to the class $\mathcal{F}^C$, and does not provide any dimension dependence, as opposed to the bounds provided here. Obtaining the correct dependence yields tight minimax results, and allows us to highlight the dependence of bounds on the geometry of the set S. Our proofs are information-theoretic in nature. We characterize the hardness of optimization in terms of a relatively easy to compute complexity measure. As a result, our technique provides tight lower bounds for smaller function classes like strongly convex functions rather easily. Indeed, we will also state a result for general function classes.

3.1 An application to statistical estimation

We now describe a simple application of the results developed above to obtain results on the oracle complexity of statistical estimation, where the typical setup is the following: given a convex loss function ℓ, a class of functions $\mathcal{F}$ indexed by a d-dimensional parameter θ so that $\mathcal{F} = \{f_\theta : \theta \in$

$\mathbb{R}^d\}$, find a function $f \in \mathcal{F}$ such that $\mathbb{E}\ell(f) - \inf_{f\in\mathcal{F}} \mathbb{E}\ell(f) \leq \epsilon$. If the distribution were known, this is exactly the problem of computing the ϵ-accurate optimizer of a convex function, assuming the function class $\mathcal{F}$ is convex. Even though we do not have the distribution in practice, we typically are provided with i.i.d. samples from it, which can be used to obtain unbiased estimates of the value and gradients of the risk functional $\mathbb{E}\ell(f)$ for any given f. If indeed the computational model of the estimator were restricted to querying these values and gradients, then the lower bounds in the previous sections would apply. Our bounds, then allow us to deduce the oracle complexity of statistical estimation problems in this realistic model. In particular, a case of interest is when we fix a convex loss function ℓ and consider the worst oracle complexity over all possible distributions under which expectation is taken. From our bounds, it is straightforward to deduce:

- For the absolute loss $\ell(f(x), y) = |f(x) - y|$, the oracle complexity of ϵ-accurate estimation over all possible distributions is $\Omega\left(d/\epsilon^2\right)$.
- For the quadratic loss $\ell(f(x), y) = (f(x) - y)^2$, the oracle complexity of ϵ-accurate estimation over all possible distributions is $\Omega\left(d/\epsilon\right)$.

We can use such an analysis to determine the limits of statistical estimation under computational constraints. Several authors have recently considered this problem [3, 9], and provided upper bounds for particular algorithms. In contrast, our results provide *algorithm-independent* lower bounds on the complexity of statistical estimation within the oracle model. An interesting direction for future work is to broader the oracle model so as to more accurately reflect the computational trade-offs in learning and estimation problems, for instance by allowing a method to pay a higher price to query an oracle with lower variance.

4 Proofs of results

We now turn to the proofs of our main results, beginning with a high-level outline of the main ideas common to our proofs.

4.1 High-level outline

Our main idea is to embed the problem of estimating the parameter of a Bernoulli vector (alternatively, the biases of d coins) into a convex optimization problem. We start with an appropriately chosen subset of the vertices of a d-dimensional hypercube each of which corresponds to some value of the Bernoulli vector. For any given function class, we then construct a "difficult" subclass of functions parameterized by these hypercube vertices. We then show that being able to optimize any function in this subclass requires estimating its hypercube vertex, that is, the corresponding biases of the d coins. But the only information for this estimation would be from the coin toss outcomes revealed by the oracle in T queries. With this set-up, we are able to apply the Fano lower bound for statistical estimation, as has been done in past work on nonparametric estimation (e.g., [5, 2, 11]).

In more detail, the proofs of Theorems 1 and 2 are both based on a common set of steps, which we describe here.

Step I: 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 on the space $\{-1, +1\}^d$, we choose $\mathcal{V}$ to be a $d/4$-packing of this hypercube. That is, $\mathcal{V}$ is a subset of the hypercube such that for all $\alpha, \beta \in \mathcal{V}$, the Hamming distance satisfies $\Delta_H(\alpha, \beta) \geq d/4$. By standard arguments [6], we can construct such a packing set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq (2/\sqrt{e})^{d/2}$.

We then let $\mathcal{G}_{\text{base}} = \{f_i^+, f_i^-, \, i = 1, \ldots, d\}$ denote some base set of $2d$ functions (to be chosen depending on the problem at hand). Given the packing set $\mathcal{V}$ and some parameter $\delta \in [0, 1/4]$, we define a larger class (with a total of $|\mathcal{V}|$ functions) via $\mathcal{G}(\delta) := \{g_\alpha, \ \alpha \in \mathcal{V}\}$, where each function $g_\alpha \in \mathcal{G}(\delta)$ has the form

$$g_\alpha(x) = \frac{1}{d} \sum_{i=1}^{d} \big\{(1/2 + \alpha_i \delta) f_i^+(x) + (1/2 - \alpha_i \delta)\, f_i^-(x)\big\}. \tag{8}$$

In our proofs, the subclasses $\mathcal{G}_{\text{base}}$ and $\mathcal{G}(\delta)$ are chosen such that $\mathcal{G}(\delta) \subseteq \mathcal{F}$, the functions f_i^+, f_i^- are bounded over the convex set S with a Lipschitz constant independent of dimension d, and the minimizers x_β of g_β over $\mathbb{R}^d$ are contained in S for all $\beta \in \mathcal{V}$. We demonstrate specific choices in the proofs of Theorems 1 and 2.

Step II: Optimizing well is equivalent to function identification. In this step, we show that if a method can optimize over the subclass $\mathcal{G}(\delta)$ up to a certain tolerance $\psi(\mathcal{G}(\delta))$, 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

$$\rho(f, g) = \inf_{x \in S} \left[f(x) + g(x) - f(x_f^*) - g(x_g^*) \right]. \tag{9}$$

The discrepancy measure is non-negative, symmetric in its arguments,[1] and satisfies $\rho(f, g) = 0$ if and only if $x_f^* = x_g^*$, so that we may refer to it as a semimetric.

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

$$\psi(\mathcal{G}(\delta)) = \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta), \tag{10}$$

which we also denote 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 $\mathcal{G}(\delta)$. Note that x_α^* denotes a minimizing argument of the function g_α.

Lemma 1. *For any $\widetilde{x} \in S$, there can be at most one function $g_\alpha \in \mathcal{G}(\delta)$ for which $g_\alpha(\widetilde{x}) - g_\alpha(x_\alpha^*) \leq \frac{\psi(\delta)}{3}$.*

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

Proof. For a given $\widetilde{x} \in S$, 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 (10), for any $\beta \in \mathcal{V},\ \beta \neq \alpha$, we have

$$\psi(\delta) \quad \leq \quad g_\alpha(\widetilde{x}) - g_\alpha(x_\alpha^*) + g_\beta(\widetilde{x}) - g_\beta(x_\beta^*) \ \leq\ \psi(\delta)/3 + g_\beta(\widetilde{x}) - g_\beta(x_\beta^*),$$

which implies that $g_\beta(\widetilde{x}) - g_\beta(x_\beta^*) \geq 2\psi(\delta)/3$, from which the claim follows.

□

Suppose that we choose some function $g_{\alpha^*} \in \mathcal{G}(\delta)$, and some method $\mathcal{M}_T$ is allowed to make T queries to an oracle with information function $\phi(\cdot, g_{\alpha^*})$. Our next lemma shows that in this set-up, if the method $\mathcal{M}_T$ can optimize well over the class $\mathcal{G}(\delta)$, then it must be capable of determining the true function g_{α^*}. Recall the definition (2) of the minimax error in optimization:

Lemma 2. *Suppose that some method $\mathcal{M}_T$ has minimax optimization error upper bounded as*

$$\mathbb{E}\big[\epsilon^*(\mathcal{M}_T, \mathcal{G}(\delta), S, \phi)\big] \leq \frac{\psi(\delta)}{9}. \tag{11}$$

Then the method $\mathcal{M}_T$ can construct an estimator $\widehat{\alpha}(\mathcal{M}_T)$ such that $\max\limits_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \frac{1}{3}$.

Proof. Given a method $\mathcal{M}_T$ that satisfies the bound (11), 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 $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha^*] \leq \mathbb{P}_\phi\big[\epsilon^*(\mathcal{M}_T, g_{\alpha^*}, S, \phi) \geq \psi(\delta)/3\big] \ \leq\ \frac{1}{3}$. 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.

[1] However, it fails to satisfy the triangle inequality and so is not a metric.

Step III: Oracle answers and coin tosses. We now demonstrate a stochastic first order oracle ϕ for which the samples $\{\phi(x_1, g_\alpha), \ldots, \phi(x_T, g_\alpha)\}$ 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

$$\Theta(\delta) = \big\{(1/2 + \alpha_1 \delta, \ldots, 1/2 + \alpha_d \delta) \;\mid\; \alpha \in \mathcal{V}\big\}, \tag{12}$$

Given a particular function $g_\alpha \in \mathcal{G}(\delta)$ (or equivalently, vertex $\alpha \in \mathcal{V}$), we consider the oracle ϕ that presents noisy value and gradient samples from g_α according to the following prescription:

- Pick an index $i_t \in \{1, \ldots, d\}$ uniformly at random.
- Draw $b_{i_t} \in \{0, 1\}$ according to a Bernoulli distribution with parameter $1/2 + \alpha_{i_t} \delta$.
- Return the value and sub-gradient of the function

$$\widehat{g}_\alpha(x) = b_{i_t} f_{i_t}^{+}(x) + (1 - b_{i_t}) f_{i_t}^{-}(x).$$

By construction, the function value and gradient samples are unbiased estimates of those of g_α; moreover, the variance of the effective "noise" is bounded independently of d as long as the Lipschitz constant is independent of d since the function values and gradients are bounded on S.

Step IV: Lower bounds on coin-tossing Finally, we use information-theoretic methods to lower bound the probability of correctly estimating the true vertex $\alpha^* \in \mathcal{V}$ in our model.

Lemma 3. *Given an arbitrary vertex $\alpha^* \in \mathcal{V}$, suppose that we toss a set of d coins with bias $\theta^* = (\frac{1}{2} + \alpha_1^* \delta, \ldots, \frac{1}{2} + \alpha_2^* \delta)$ a total of T times, but that the outcome of only one coin chosen uniformly at random is revealed at every round. Then for all $\delta \leq 1/4$, any estimator $\widehat{\alpha}$ satisfies*

$$\inf_{\widehat{\alpha}} \max_{\alpha^* \in \mathcal{V}} \mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq \left\{1 - \frac{16T\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}\right\}.$$

Proof. Denote the Bernoulli distribution for the i-th coin by P_{θ_i}. Let $Y_t \in \{1, \ldots, d\}$ be the variable indicating the coin revealed at time T, and let $X_t \in \{0, 1\}$ denote its outcome. With some abuse of notation, we also denote the distribution of (X_t, Y_t) by P_θ, and that of the entire data $\{(X_t, Y_t)\}_{t=1}^{T}$ by P_θ^T. Note that $P_\theta(i, b) = \frac{1}{d} P_{\theta_i}(b)$. We now apply a version of Fano's lemma [11] to the set of distributions P_θ^T for $\theta \in \Theta(\delta)$. In particular, using the proof of Lemma 3 in [11] we get:

$$\mathrm{KL}(P_\theta^T || P_{\theta'}^T) \;\leq\; b,\ \forall \theta, \theta' \in \Theta(\delta) \;\Rightarrow\; \inf_{\widehat{\theta}} \max_{\theta \in \Theta(\delta)} \mathbb{P}_\theta[\widehat{\theta} \neq \theta] \geq \left(1 - \frac{b + \log 2}{\log |\Theta|}\right). \tag{13}$$

In our case, we upper bound b as follows:

$$b = \mathrm{KL}(P_\theta^T || P_{\theta'}^T) = \sum_{t=1}^{T} \mathrm{KL}(P_\theta(X_t, Y_t) || P_{\theta'}(X_t, Y_t)) \;=\; \frac{1}{d} \sum_{t=1}^{T} \sum_{i=1}^{d} \mathrm{KL}(P_{\theta_i}(X_t) || P_{\theta_i'}(X_t)).$$

Each term $\mathrm{KL}(P_{\theta_i}(X_t) || P_{\theta_i'}(X_t))$ is at most the KL divergence $g(\delta)$ between Bernoulli variates with parameters $1/2 + \delta$ and $1/2 - \delta$. A little calculation shows that

$$g(\delta) \quad = \quad 2\delta \log\left(1 + \frac{4\delta}{1 - 2\delta}\right) \;\leq\; \frac{8\delta^2}{1 - 2\delta},$$

which is less than $16\delta^2$ as long as $\delta \leq 1/4$. Consequently, we conclude that $b \leq 16T\delta^2$. Also, we note that $\mathbb{P}[\widehat{\alpha} \neq \alpha^*] = \mathbb{P}_\theta[\widehat{\theta} \neq \theta^*]$. Substituting these values and the size of $\mathcal{V}$ into (13) yields the claim. □

4.2 Proofs of main results

We are now in a position to prove our main theorems.

Proof of Theorem 1: By the construction of our oracle, it is clear that, at each round, only one coin is revealed to the method $\mathcal{M}_T$. Thus Lemma 3 applies to the estimator $\widehat{\alpha}(\mathcal{M}_T)$:

$$\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \geq \left(1 - 2\frac{16T\delta^2 + \log 2}{d\log(2/\sqrt{e})}\right). \tag{14}$$

In order to obtain an upper bound on $\mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha]$ using Lemma 2, we need to identify the subclass $\mathcal{G}_{\mathrm{base}}$ of $\mathcal{F}^C$. For $i = 1, \ldots, d$, define:

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

We take S to be the ℓ_∞ ball of radius $1/2$. It is clear then that the minimizers of g_α are contained in S. Also, the functions f_i^+, f_i^- are bounded in $[0, 1]$ and 1-Lipschitz in the ∞-norm, giving the same properties for each function g_α. Finally, we note that $\rho(g_\alpha, g_\beta) = \frac{2\delta}{d}\, \Delta_H(\alpha, \beta) \geq \frac{\delta}{2}$ for $\alpha \neq \beta \in \mathcal{V}$. Setting $\epsilon = \delta/18 < 1/2$, we obtain $\epsilon^*(\mathcal{F}^C, \phi) \leq \epsilon = \frac{\delta}{18} = \frac{\psi(\delta)}{9}$. Then by Lemma 2, we have $\mathbb{P}_\phi[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}$ which, when combined with equation (14), yields $\frac{1}{3} \geq \big(1 - 2\frac{16T\delta^2 + \log 2}{d\log(2/\sqrt{e})}\big)$. Substituting $\delta = 18\epsilon$ yields $T = \Omega\big(\frac{d}{\epsilon^2}\big)$ for all $d \geq 11$. Combining this with Theorem 5.3.1 of NY [8] gives $T = \Omega\left(\frac{d}{\epsilon^2}\right)$ for all d.

To prove Corollary 1, we note that the proof of Theorem 1 required $r_\infty \geq \frac{1}{2}$. If not, it is easy to see that the computation of ρ on $\mathcal{G}(\delta)$ scales by r_∞. Further, if the set is contained in a ball of radius R_∞, then we need to scale the function with $\frac{1}{R_\infty}$ to keep the function values bounded. Taking both these dependences into account gives the desired result.

Proof of Theorem 2: In this case, we define the base class

$$f_i^+(x) = \big(x(i) + 1/2\big)^2, \quad \text{and} \quad f_i^-(x) = \big(x(i) - 1/2\big)^2, \qquad \text{for } i = 1, \ldots, d.$$

Then the functions g_α are strongly convex w.r.t. the Euclidean norm with coefficient $\kappa = 1/d$. Some calculation shows that $\rho(g_\alpha, g_\beta) = \frac{2\delta^2}{d}\, \Delta_H(\alpha, \beta)$ for all $\alpha \neq \beta$. The remainder of the proof is identical to Theorem 1.

The reader might suspect that the dimension dependence in our lower bound for strongly convex functions is not tight, due to the dependence of κ on the dimension d. However, this is the largest possible value of κ under the assumptions of the theorem.

4.3 A general result

Armed with the greater understanding from these proofs, we can now state a general result for any function class $\mathcal{F}$. The proof is similar to that of earlier results.

Theorem 3. *For any function class $\mathcal{F} \subseteq \mathcal{F}^C$, suppose a given base set of functions $\mathcal{G}_{\mathrm{base}}$ yields the subclass $\mathcal{G}(\delta)$ and the measure ψ as defined in* (10)*, and this function is monotone increasing. Then there exists a universal constant c such that* $\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}^{\mathcal{S}}, \phi) \geq c\, \psi\Big(\sqrt{\frac{\log|\mathcal{G}(\delta)|}{T}}\Big)$.

Acknowledgements We gratefully acknowledge the support of the NSF under award DMS-0830410 and of DARPA under award HR0011-08-2-0002. Alekh is supported in part by MSR PhD Fellowship.

References

[1] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

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

[3] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *NIPS*. 2008.

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

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

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

[7] A. S. Nemirovski. Efficient methods in convex programming. *Lecture notes*.

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

[9] S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *ICML*, 2008.

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

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