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

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

Peter L. Bartlett
Computer Science Division
Department of Statistics
UC Berkeley
peter@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

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. Recent years have seen a surge in optimization methods tailored to sparse optimization problems. In this paper, we study the complexity of stochastic convex optimization in an oracle model of computation, when the objective is optimized at a sparse vector in a high dimensional space. Our result is matched by an appropriately tuned method of mirror descent, establishing the minimiax optimality of the result.

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. High-dimensional data presents a challenge to convex optimization methods since the minimax complexity of deterministic as well as stochastic convex optimization is known to scale with the dimension of the space [1, 2, 3].

A large body of recent work has focused on developing optimization algorithms specifically tailored to high-dimensional problems [4, 5, 6]. A common assumption throughout this line of work is that the objective function is optimized at a sparse point. Under this assumption, the aforementioned approaches enjoy a mild logarithmic scaling with the dimension of the space, making the methods suitable for high-dimensional problems.

It is natural to ask what is the smallest possible number of queries with which an algorithm might be able to optimize a function under such a sparsity assumption. In this paper we establish a lower bound on the complexity of sparse convex optimization in a stochastic first order oracle model, first introduced by Nemirovski and Yudin [1] (hereafter referred to as NY). Our result matches the best known upper bound attained by an appropriately tuned method of mirror descent [1, 7], as well as the other methods mentioned earlier. This establishes the minimax optimality of the above methods for solving sparse optimization problems.

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 $x \in \mathbb{R}^d$, 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.

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 $\mathcal{F}$ which is optimized at a point with only few non-zero coordinates, 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 [1], and measure computational labor based on the oracle model of optimization. In particular we focus on the stochastic first-order oracle. Such an oracle takes a query $x \in \mathbb{S}$ and returns a tuple $\phi(x, f) = (\widehat{f}(x), \widehat{z}(x))$ such that

$$\mathbb{E}[\widehat{f}(x)] = f(x), \qquad \mathbb{E}[\widehat{z}(x)] \in \partial f(x), \quad \text{and} \quad \mathbb{E}\big[\, \|\widehat{z}(x)\|_p^2 \,\big] \leq \sigma^2. \tag{1}$$

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 sub-gradient $z \in \partial f(x)$. When f is actually differentiable, then $\widehat{z}(x)$ is an unbiased estimate of the gradient $\nabla f(x)$. The third condition in equation (1) controls the "noisiness" of the sub-gradient estimates in terms of the ℓ_p-norm.

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. We remark here that all the methods for high-dimensional problems mentioned earlier fit this general template. For a given oracle function ϕ (which defines the distribution of the random variables), 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

$$\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi) := f(x_T) - \min_{x \in \mathbb{S}} f(x) \;=\; f(x_T) - f(x_f^*), \tag{2}$$

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. 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

$$\epsilon_T^*(\mathcal{F}, \mathbb{S}; \phi) := \inf_{\mathcal{M} \in \mathbb{M}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon_T(\mathcal{M}, f, \mathbb{S}, \phi)]. \tag{3}$$

So as to ease the notation, when the oracle ϕ is clear from the context, we simply write $\epsilon_T^*(\mathcal{F}, \mathbb{S})$.

We now turn to defining the function class $\mathcal{F}$ of interest precisely. In all cases, we consider real-valued convex functions defined over some convex set $\mathbb{S}$. For a vector $x \in \mathbb{R}^d$, we use $\|x\|_0$ to denote the number of non-zero elements in x. We now define a class of Lipschitz functions with sparse minimizers.

Definition 1. *For a convex set $\mathbb{S} \subset \mathbb{R}^d$ and positive integer $k \leq \lfloor d/2 \rfloor$, we define $\mathcal{F}_{\text{sp}}(k, \mathbb{S}, L)$ to be the class of all convex functions $f : \mathbb{S} \mapsto \mathbb{R}$ such that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad \textit{for all } x, y \in \mathbb{S}, \tag{4}$$

and $\exists x^* \in \arg\min_{x \in \mathbb{S}} f(x)$ satisfying $\|x^*\|_0 \leq k$.

Note that the Lipschitz condition is equivalent to assuming that $\|z\|_\infty \leq L$ for all $z \in \partial f(x)$ and for all $x \in \mathbb{S}$. The Lipschitz condition assumed here is different from our previous work[3], where we assumed a bound on the ℓ_1-norm of the gradients. In the context of high-dimensional optimization, however, the ℓ_∞ bound is most natural. Indeed for a high-dimensional vector, assuming a small bound on the ℓ_1-norm implies that any single coordinate is getting smaller in the limit as the number of dimensions grows, making the problem easier in some sense. On the contrary, for some of the optimization problems encountered in this setup such as sparse linear regression with Gaussian noise (see e.g. [8]), the ℓ_2 norm of the gradient grows as $\sqrt{d}$ with high probability. Hence we assume only an ℓ_∞ bound on the gradients to match the assumptions of the problems that are frequently encountered in high-dimensional setups. We frequently use the shorthand notation $\mathcal{F}_{\text{sp}}(k)$ when the set $\mathbb{S}$ and parameters L are clear from context. In words, the set $\mathcal{F}_{\text{sp}}(k)$ consists of all convex functions that are L-Lipschitz in the ℓ_∞-norm, and have at least one k-sparse optimizer.

3 Oracle complexity for convex, Lipschitz functions with sparse optima

With the setup of stochastic convex optimization in place, we are now in a position to state the main result of this paper, and to discuss some of its consequences. For the remainder of this paper, we set the oracle variance bound σ to be the same as the Lipschitz constant L in our results. The following theorem provides a lower bound on the complexity of optimization functions from the class $\mathcal{F}_{\text{sp}}$.

Theorem 1. *Let $\mathcal{F}_{\text{sp}}$ be the class of all convex functions that are L-Lipschitz with respect to the ℓ_∞ norm and have a k-sparse optimizer. Let $\mathbb{S}$ be any convex set containing a unit ball in ℓ_∞ norm. Then for $k \leq \lfloor \frac{d}{2} \rfloor$, the oracle complexity satisfies the lower bound*

$$\sup_{\phi \in \mathbb{O}_{p,L}} \epsilon^*(\mathcal{F}_{\text{sp}}, \phi) = \Omega\left(L\sqrt{\frac{k^2 \log \frac{d}{k}}{T}}\right). \tag{5}$$

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 the method of mirror descent based on $\|\cdot\|_{1+\varepsilon}$ norm with $\varepsilon = 2 \log d/(2 \log d - 1)$. Then it can be shown (see e.g. Chapter 11 of [9]) 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 (5) up to constant factors whenever $k = \mathcal{O}(d^{1-\delta})$. To the best of our knowledge, Theorem 1 provides the first tight lower bound on the oracle complexity of sparse optimization. The result also establishes the computational optimality of methods achieving a matching upper bound, since we have shown that no other method can achieve a better convergence rate in the worst case.

3.1 Proof of main result

We now turn to the proof of our main result. We start by fixing the convex set $\mathbb{S}$ to be the unit ball in ℓ_∞ norm. Since the complexity of optimizing over a large set is only larger than that over any subset, this also implies the general result for any set containing the unit ℓ_∞ ball.

Our proof follows the technique from our previous work [3]. In particular, we define a hard subclass of functions embedded in $\mathcal{F}_{\text{sp}}$. We index this subclass by a set of appropriately chosen vectors in $\{-1, 0, +1\}^d$. 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. [10]).

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

$$g_\alpha(x) := c\left[\sum_{i=1}^{d}\left\{\left(\frac{1}{2}+\alpha_i\delta\right)\left|x(i)+\frac{1}{2}\right| + \left(\frac{1}{2}-\alpha_i\delta\right)\left|x(i)-\frac{1}{2}\right|\right\} + \delta\sum_{i=1}^{d}|x(i)|\right]. \tag{6}$$

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. We use the notation $\mathcal{G}(\delta, k)$ to refer to the class of functions $\{g_\alpha : \alpha \in \mathcal{V}(k)\}$. 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. Thus $g_\alpha \in \mathcal{F}_{\text{sp}}(k, \mathbb{S}, c)$.

For ease of notation in the future, we will also define base functions f_i^+, f_i^- for $i = 1, \ldots, d$ that constitute the functions g_α. We define

$$f_i^+ := d\left(\left|x(i)+\frac{1}{2}\right| + \frac{\delta}{2}|x(i)|\right), \quad \text{and} \quad f_i^-(x) := d\left(\left|x(i)-\frac{1}{2}\right| + \frac{\delta}{2}|x(i)|\right). \tag{7}$$

It is easily seen that

$$g_\alpha(x) = \frac{c}{d}\sum_{i=1}^{d}\left[\left(\frac{1}{2}+\alpha_1\delta\right)f_i^+(x) + \left(\frac{1}{2}-\alpha_i\delta\right)f_i^-(x)\right].$$

The next key ingredient is to show that for $\alpha \neq \beta$, g_α and g_β are different so that any optimization method needs to be able to distinguish between them based on the stochastic gradient samples. Following previous work, we show this by demonstrating a large ρ-separation between g_α and g_β. Recall that for two convex functions, the semi-metric ρ is defined in Agarwal et al [3] as:

$$\rho(f, g) := \inf_{x \in \mathbb{S}}\left[f(x) + g(x) - f(x_f^*) - g(x_g^*)\right]. \tag{8}$$

Here x_f^* and x_g^* are the minimizers of f and g respectively. We show that any two elements of $\mathcal{G}(\delta, k)$ are well-separated by analyzing the quantity

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

Understanding the scaling of ψ with δ was critical to the proofs in our previous work [3], as it was shown that any method attaining a minimax error (3) smaller than $\psi(\delta)/9$ is capable of solving an estimation problem of identifying the true function g_α based on the oracle's responses. Our next lemma provides a lower bound on $\psi(\delta)$ for the class $\mathcal{G}(\delta, k)$.

Lemma 1. *For g_α defined as in Equation 6, for $\alpha \neq \beta \in \mathcal{V}(k)$ we have*

$$\psi(\delta; k) \;=\; \inf_{\alpha \neq \beta \in \mathcal{V}(k)} \rho(g_\alpha, g_\beta) \geq \frac{ck\delta}{4}. \tag{9}$$

The proof of this lemma is omitted due to lack of space and can be found in the full draft [11]. Let $\mathcal{M}_T$ be any optimization method that takes T rounds and produces a point $x_T \in \mathbb{S}$ such that the minimax optimization error of x_T is smaller than $ck\delta/36$. Then the above lemma, combined with Lemmas 1 and 2 of Agarwal et al [3] allows us to conclude that $\mathcal{M}_T$ reliably estimates the parameter α used by the oracle. That is, the method implicitly constructs an estimator $\widehat{\alpha}(\mathcal{M}_T)$ such that

$$\max_{\alpha \in \mathcal{V}} \mathbb{P}[\widehat{\alpha}(\mathcal{M}_T) \neq \alpha] \leq \frac{1}{3}, \tag{10}$$

where α indexes the function used by the oracle.

In order to provide a lower bound on the number of queries needed by any method, we would like to show a lower bound on the probability of error in this estimation problem, based on T noisy

samples received from the oracle. To do this, we use a particular choice of oracle and then show a lower bound on the error probability for this oracle using Fano's inequality [12, 13].

We start by defining a stochastic first-order oracle that meets the conditions of Equation 1. It turns out that the oracle from our previous work [3] that provides 1-dimensional stochastic gradient samples is unsuited for this problem. This is because here we require our gradients to be bounded in ℓ_∞ norm while the previous work assumed the gradients (and their variance) to be bounded in ℓ_1 norm. As a result, in our setup we can provide d-dimensional gradients instead, with each co-ordinate being large which allows us to inject a lot more noise in the problem.

We associate a coin with each coordinate of the problem and consider the set of coin bias vectors lying in the set $\{(1/2 + \alpha_1\delta), \ldots, (1/2 + \alpha_d\delta) \ : \ \alpha \in \mathcal{V}(k)\}$. Given a particular function $g_\alpha \in \mathcal{G}(\delta, k)$, we present noisy value and gradient samples according to the following prescription:

- For each $i = 1, \ldots, d$, draw $b_i \in \{0, 1\}$ according to a Bernoulli distribution with bias $1/2+\alpha_i\delta$.
- Return the value and sub-gradient of the function:

$$\widehat{g}(x) = \frac{c}{d}\sum_{i=1}^{d}[b_i f_i^+(x) + (1 - b_i)f_i^-(x)].$$

It is easy to see that $\widehat{g}$ defined above satisfies conditions (1) if $c = L/3$. As the final ingredient, we would like to lower bound the probability of error in estimating the bias vectors of the coins, based on tosses received from the oracle above. We adapt a version of Fano's inequality from the work of Yu [12]. Let $\mathbb{P}_\alpha$ be the distribution of the function values and gradients generated by our stochastic first-order oracle when using the function g_α. Then Lemma 3 of Yu [12] allows us to conclude that

$$\inf_{\widehat{\alpha}} \sup_{\alpha \in \mathcal{V}(k)} \mathbb{P}_\alpha[\widehat{\alpha} \neq \alpha] \geq 1 - \frac{b + \log 2}{\log |\mathcal{V}(k)|},$$

where b is an upper bound on the Kullback-Leibler divergence between $\mathbb{P}_\alpha$ and $\mathbb{P}_\beta$ for $\alpha \neq \beta \in \mathcal{V}(k)$. It can be shown using some algebra that in our case, $b = 32kT\delta^2$ suffices. Recalling that $|\mathcal{V}(k)| \geq \exp\left(\frac{k}{2}\log\frac{d-k}{k/2}\right)$, we get

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

Detailed proof for this fact can be found in Lemma 3 of the long version [11]. Combining this with the upper bound (10) yields the lower bound

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

To complete the proof, we observe that we want to obtain a lower bound on T in terms of the minimax optimization error. Since the upper bound on error probability applies to all methods with optimization error no more than $ck\delta/36 = Lk\delta/108$, we get the relation $\epsilon = Lk\delta/108$ where ϵ is the optimization error of $\mathcal{M}_T$. Substituting this back in the previous lower bounds gives

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

which completes the statement of the theorem for $\mathbb{S}$ being the unit ℓ_∞ ball. The general result follows from observing that enlarging the convex set only increases the oracle complexity of optimization.

References

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

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

[3] 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.

[4] Y. Nesterov, "Gradient methods for minimizing composite objective function," Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Tech. Rep. 76, 2007.

[5] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 2116–2124.

[6] J. Duchi and Y. Singer, "Efficient learning using forward-backward splitting," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 495–503.

[7] 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.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. New York: Springer Series in Statistics, 2001.

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[10] 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

[11] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, "Information-theoretic lower bounds on the oracle complexity of convex optimization," *ArXiv e-prints*, 2010. [Online]. Available: http://arxiv.org/abs/1009.0571v1

[12] B. Yu, "Assouad, Fano and Le Cam," in *Festschrift for Lucien Le Cam*. Berlin: Springer-Verlag, 1997, pp. 423–435.

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