



University of Pennsylvania  
ScholarlyCommons

---

Statistics Papers

Wharton Faculty Research

---


2009

# Information Complexity of Black-Box Convex Optimization: A New Look via Feedback Information Theory

Maxim Raginsky

Alexander Rakhlin  
*University of Pennsylvania*

Follow this and additional works at: [http://repository.upenn.edu/statistics\\_papers](http://repository.upenn.edu/statistics_papers)

 Part of the [Computer Sciences Commons](#), and the [Statistics and Probability Commons](#)

---

## Recommended Citation

Raginsky, M., & Rakhlin, A. (2009). Information Complexity of Black-Box Convex Optimization: A New Look via Feedback Information Theory. *47th Annual Allerton Conference on, Monticello, IL*, 803-510. <http://dx.doi.org/10.1109/ALLERTON.2009.5394945>

This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/statistics\\_papers/397](http://repository.upenn.edu/statistics_papers/397)  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Information Complexity of Black-Box Convex Optimization: A New Look via Feedback Information Theory

## **Abstract**

This paper revisits information complexity of black-box convex optimization, first studied in the seminal work of Nemirovski and Yudin, from the perspective of feedback information theory. These days, large-scale convex programming arises in a variety of applications, and it is important to refine our understanding of its fundamental limitations. The goal of black-box convex optimization is to minimize an unknown convex objective function from a given class over a compact, convex domain using an iterative scheme that generates approximate solutions by querying an oracle for local information about the function being optimized. The information complexity of a given problem class is defined as the smallest number of queries needed to minimize every function in the class to some desired accuracy. We present a simple information-theoretic approach that not only recovers many of the results of Nemirovski and Yudin, but also gives some new bounds pertaining to optimal rates at which iterative convex optimization schemes approach the solution. As a bonus, we give a particularly simple derivation of the minimax lower bound for a certain active learning problem on the unit interval.

## **Keywords**

convex programming, information theory, learning (artificial intelligence), active learning problem, black box convex optimization, convex programming, feedback information theory, information complexity, minimax lower bound, feedback, history, information theory, iterative methods, large-scale systems, minimax techniques, signal processing, signal processing algorithms, statistics, stochastic resonance

## **Disciplines**

Computer Sciences | Statistics and Probability

# Information Complexity of Black-Box Convex Optimization: A New Look via Feedback Information Theory

Maxim Raginsky<sup>†</sup> and Alexander Rakhlin<sup>‡</sup>

**Abstract**—This paper revisits information complexity of black-box convex optimization, first studied in the seminal work of Nemirovski and Yudin, from the perspective of feedback information theory. These days, large-scale convex programming arises in a variety of applications, and it is important to refine our understanding of its fundamental limitations. The goal of black-box convex optimization is to minimize an unknown convex objective function from a given class over a compact, convex domain using an iterative scheme that generates approximate solutions by querying an oracle for local information about the function being optimized. The information complexity of a given problem class is defined as the smallest number of queries needed to minimize every function in the class to some desired accuracy. We present a simple information-theoretic approach that not only recovers many of the results of Nemirovski and Yudin, but also gives some new bounds pertaining to optimal rates at which iterative convex optimization schemes approach the solution. As a bonus, we give a particularly simple derivation of the minimax lower bound for a certain active learning problem on the unit interval.

## I. INTRODUCTION

Convex optimization problems of the form

$$\min\{f(x) : x \in X\}, \quad (1)$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex objective function and  $X$  is a compact, convex subset of  $\mathbb{R}^n$ , arise in such areas as communications and signal processing, control, machine learning, economics, and many others. For this reason, it is important to have a clear understanding of the *fundamental limits* on the efficiency of convex programming methods.<sup>1</sup>

A systematic study of these fundamental limits was initiated in the 1970's by Nemirovski and Yudin [2]. In their framework, an optimization algorithm is a sequential procedure that repeatedly queries a black-box *oracle* for information about the function being optimized, each query depending on the past information. The oracle may be deterministic (for example, giving the value of the function and its derivatives up to some order at any point) or stochastic. This leads to the notion of *information-based complexity*, i.e., the smallest number of oracle calls needed to minimize any function in the class to a desired accuracy. The results in [2] are very wide in scope and cover a variety of convex programming problems in Banach spaces; finite-dimensional versions are covered in [3] and [4].

<sup>†</sup>Department of Electrical and Computer Engineering, Duke University, Durham, NC. E-mail: m.raginsky@duke.edu

<sup>‡</sup>Department of Statistics, University of Pennsylvania, Philadelphia, PA. E-mail: rakhlin@wharton.upenn.edu

<sup>1</sup>A related work is independently undertaken by Agarwal et al [1].

For deterministic oracles, Nemirovski and Yudin derived lower bounds on information complexity of convex programming using a “counterfactual” argument: given any algorithm that purports to optimize all functions in some class  $\mathcal{F}$  to some degree of accuracy  $\varepsilon$  using at most  $T$  oracle calls, one explicitly constructs, for a particular history of queries and oracle responses, a function in  $\mathcal{F}$  which is consistent with this history, and yet cannot be  $\varepsilon$ -minimized by the algorithm using fewer than  $T$  oracle calls (see also [3]). A similar approach was also used for stochastic oracles.

Proper application of this *method of resisting oracles* requires a lot of ingenuity. In particular, the stochastic case involves fairly contrived noise models, unlikely to be encountered in practice. In this paper, we will show that the same (and many other) lower bounds can be derived using a much simpler information-theoretic technique reminiscent of the way one proves minimax lower bounds in statistics [5], [6]. Namely, we reduce optimization to statistical estimation and then relate the probability of estimation error to information complexity using Fano's inequality and a series of mutual information bounds. These bounds highlight the role of *feedback* in choosing the next query based on the past observations. One notable feature of our approach is that it does not require constructing particularly “strange” functions or noise models. Moreover, we derive a “law of diminishing returns” for a wide class of convex optimization schemes which says that the decay of optimization error is offset by the decay of the rate at which the algorithm can reduce its uncertainty about the objective function.

**Notation.** Given a function  $f : X \rightarrow \mathbb{R}$ , where  $X \subset \mathbb{R}^n$  is compact and convex, we denote by  $f^*$  its minimum value over  $X$ :  $f^* = \inf_{x \in X} f(x)$ . The *subdifferential* of  $f$  at  $x$ , denoted by  $\partial f(x)$ , is the set of all  $g \in \mathbb{R}^n$ , such that  $f(y) \geq f(x) + g^\top(y - x)$ ,  $\forall y \in \mathbb{R}^n$ . Any such  $g$  is a *subgradient* of  $f$  at  $x$ . When  $|\partial f(x)| = 1$ , its only element is precisely the gradient  $\nabla f(x)$ . Abusing notation, we write  $\nabla f(x)$  for an arbitrary subgradient of  $f$  at  $x$  (which always exists for a convex  $f$ ). By  $\|x\|_p$  we denote the  $\ell_p$  norm of  $x \in \mathbb{R}^n$ ; the  $\ell_2$  norm will also be denoted by  $\|\cdot\|$ . By  $B_p^n$  we denote the unit ball in  $\mathbb{R}^n$  in the  $\ell_p$  norm. The  $\ell_2$ -diameter of  $X$  is denoted by  $D_X$ . All spaces are assumed to be Borel measurable and equipped with appropriate  $\sigma$ -fields. If  $Z$  is such a space, then  $\mathcal{B}_Z$  will denote the  $\sigma$ -field. All functions between such spaces are likewise assumed to be measurable.

## II. CONVEX OPTIMIZATION WITH ORACLES

In the query model studied in [2] and here, we must solve (1), where  $f$  comes from some class  $\mathcal{F}$  and is initially

unknown. Any procedure we use gathers information about  $f$  by querying an oracle with points in  $\mathsf{X}$  subject to certain causality constraints. More precisely, we have the following:

**Definition 1.** A problem class is a triple  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  consisting of the following objects: (i) a compact, convex problem domain  $\mathsf{X} \subset \mathbb{R}^n$ ; (ii) an instance space  $\mathcal{F}$ , which is a class of convex functions  $f : \mathsf{X} \rightarrow \mathbb{R}$ ; and (iii) an oracle  $\mathcal{O} = (\mathsf{Y}, P)$ , where  $\mathsf{Y}$  is the oracle information space and  $P(dy|f, x)$ ,  $dy \in \mathcal{B}_{\mathsf{Y}}$ ,  $f \in \mathcal{F}$ ,  $x \in \mathsf{X}$ , is a Markov kernel.

**Definition 2.** An oracle  $\mathcal{O} = (\mathsf{Y}, P)$  is oblivious if there exist a deterministic map  $\psi : \mathcal{F} \times \mathsf{X} \rightarrow \mathsf{U}$  into some space  $\mathsf{U}$  and a Markov kernel  $Q(dy|u)$ ,  $dy \in \mathcal{B}_{\mathsf{Y}}$ ,  $u \in \mathsf{U}$ , such that

$$P(dy|f, x) = Q(dy|\psi(f, x)), \quad dy \in \mathcal{B}_{\mathsf{Y}}, f \in \mathcal{F}, x \in \mathsf{X}.$$

Otherwise,  $\mathcal{O}$  will be called nonoblivious.

**Example 1.** Let  $\mathcal{F}_{\text{Lip}}$  be the set of all convex functions  $f : \mathsf{X} \rightarrow \mathbb{R}$  that are 1-Lipschitz, i.e.,  $|f(x) - f(y)| \leq \|x - y\|$ ,  $\forall x, y \in \mathsf{X}$ . Let  $\mathsf{Y} = \mathbb{R} \times \mathbb{R}^n$  and let  $P(dy|f, x)$  be a point mass concentrated at  $(f(x), \nabla f(x))$  for some  $\nabla f(x) \in \partial f(x)$ . This oracle provides noiseless first-order information.

**Example 2.** Take  $\mathcal{F}_{\text{Lip}}$  as above, but now suppose that the oracle responds with  $Y = (f(x) + W, \nabla f(x) + Z)$ , where  $W \in \mathbb{R}$  and  $Z \in \mathbb{R}^n$  are zero-mean random variables with bounded variances. Thus, any algorithm receives noisy first-order information, and the oracle is oblivious.

**Example 3.** As an example of a problem class with a nonoblivious oracle, let  $\mathsf{X} = [0, 1]$ ,  $\mathsf{Y} = \{-1, +1\}$ ,  $\mathcal{F} = \{f_{\theta}(x) = |x - \theta| : \theta \in \mathsf{X}\}$ . To define the oracle, suppose that there exist some  $0 < c, C < 1/2$  and  $\kappa \in [1, \infty)$ , such that

$$c|x - \theta|^{\kappa-1} \leq |P(Y = 1|f_{\theta}, x) - 1/2| \leq C|x - \theta|^{\kappa-1},$$

where the first inequality holds for all  $x$  in a sufficiently small neighborhood of  $\theta$ . This oracle provides a noisy subgradient of  $f_{\theta}$  at  $x$ , and the amount of noise depends on the distance between  $x$  and  $\theta$ . This problem class is related to *active learning* of a threshold function on the unit interval [7], and will be treated in detail in Section IV.

An *algorithm* for a given  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  is a sequence of mappings  $\mathcal{A} = \{\mathcal{A}_t : \mathsf{X}^{t-1} \times \mathsf{Y}^{t-1} \rightarrow \mathsf{X}\}_{t=1}^{\infty}$ . The interaction of  $\mathcal{A}$  with  $\mathcal{O}$  is described recursively as follows:

- 1) At time  $t = 0$ , a problem instance  $f \in \mathcal{F}$  is selected by Nature and revealed to  $\mathcal{O}$ , but not to  $\mathcal{A}$ .
- 2) At each time  $t = 1, 2, \dots$ :
  - $\mathcal{A}$  queries  $\mathcal{O}$  with  $x_t = \mathcal{A}_t(x^{t-1}, y^{t-1})$ , where  $(x_{\tau}, y_{\tau}) \in \mathsf{X} \times \mathsf{Y}$  is the algorithm's query and the oracle's response at time  $\tau \leq t - 1$ .
  - $\mathcal{O}$  responds with a random element  $y_t \in \mathsf{Y}$  according to  $P(dy_t|f, x_t)$ .

The *error* of  $\mathcal{A}$  on  $f$  after  $T$  steps of operation is given by

$$\text{err}_{\mathcal{A}}(T, f) \triangleq f(x_T) - \min_{x \in \mathsf{X}} f(x) = f(x_T) - f^*.$$

Given  $\varepsilon > 0$ ,  $t \geq 1$ , and  $f \in \mathcal{F}$ , let us define the event

$$E_t^{\varepsilon}(\mathcal{A}, f) \triangleq \{\text{err}_{\mathcal{A}}(t, f) \geq \varepsilon\}.$$

**Definition 3.** For any  $\varepsilon > 0$  and  $\delta \in (0, 1)$ , the  $(\varepsilon, \delta)$ -computing time of  $\mathcal{A}$  w.r.t.  $\mathcal{P}$ , denoted by  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon, \delta)$ , is

$$T_{\mathcal{A}, \mathcal{P}}(\varepsilon, \delta) \triangleq \sup_{f \in \mathcal{F}} \inf \left\{ \tau \geq 1 : \forall t \geq \tau, \Pr(E_t^{\varepsilon}(\mathcal{A}, f)) \leq \delta \right\}.$$

The  $\varepsilon$ -computing time, denoted by  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon)$ , is

$$T_{\mathcal{A}, \mathcal{P}}(\varepsilon) \triangleq \sup_{f \in \mathcal{F}} \inf \left\{ \tau \geq 1 : \forall t \geq \tau, \mathbb{E} \text{err}_{\mathcal{A}}(t, f) < \varepsilon \right\}.$$

When the underlying problem class  $\mathcal{P}$  is clear from context, we will write simply  $T_{\mathcal{A}}(\varepsilon, \delta)$  and  $T_{\mathcal{A}}(\varepsilon)$ .

We remark that our framework encompasses statistical estimation with  $L_2$  loss considered in [5]. To sketch the reduction, consider the collection of densities  $\{p_{\theta} : \theta \in \Theta\}$ , where  $\Theta$  is a subset of a Hilbert space. The non-oblivious oracle response  $y_t$  is defined as a random sample from  $p_{\theta}$ , ignoring the query point  $x_t \in \Theta$ . The value of feedback has thus been nullified. In contrast, it is precisely the sequential nature of stochastic optimization and the diminishing value of feedback at each step that distinguish this work from the lower bounds based on the entire sample.

### III. LOWER BOUNDS FOR ARBITRARY ALGORITHMS

We now describe our information-theoretic method for determining lower bounds on the information complexity of convex programming. The basic strategy is to show that the minimum number of oracle queries is constrained by the average rate at which each new query can reduce the algorithm's uncertainty about the function being optimized.

#### A. Reduction to statistical estimation

Consider a problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  and suppose that the instance space  $\mathcal{F}$  can be endowed with a "distance"  $d(\cdot, \cdot)$  with the following property: for any  $x \in \mathsf{X}$ ,

$$d(f, g) \geq 2\varepsilon \text{ and } f(x) < f^* + \varepsilon \Rightarrow g(x) > g^* + \varepsilon. \quad (2)$$

In other words, an  $\varepsilon$ -minimizer of a function cannot simultaneously be an  $\varepsilon$ -minimizer of a distant function. Note that, similarly to [5],  $d$  need not satisfy properties of a metric. It is easy to show that  $d$  satisfying (2) exists for any class  $\mathcal{F}$  of convex functions. For example, if we consider the class

$$\mathcal{F}_{\Theta} \triangleq \{f_{\theta}(x) = \|x - \theta\| : \theta \in \Theta\}$$

for some  $\Theta \subset \mathsf{X}$ , then  $d(f_{\theta}, f_{\theta'}) = \|\theta - \theta'\|$  satisfies (2). Now consider any finite  $\mathcal{F}' = \{f_0, \dots, f_{N-1}\} \subset \mathcal{F}$ , such that any two distinct  $f_i, f_j \in \mathcal{F}'$  are at least  $2\varepsilon$  apart in  $d(\cdot, \cdot)$ . Given the history  $(X^T, Y^T)$  of queries and oracle answers up to time  $T$ , let us define the estimator

$$\hat{M}_T(X^T, Y^T) \triangleq \arg \min_{m=0, \dots, N-1} [f_m(X_T) - f_m^*]. \quad (3)$$

**Lemma 1.** Fix some  $\delta \in (0, 1/2)$  and  $\varepsilon > 0$ . Consider any algorithm  $\mathcal{A}$  with  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon, \delta) = T$ . Let  $M$  be uniformly distributed on  $\{0, 1, \dots, N-1\}$ , and suppose that  $\mathcal{A}$  is fed

with the random problem instance  $f_M \in \mathcal{F}'$ . If  $N > 4$ , then the estimator  $\hat{M}_T$  defined in (3) satisfies the bound

$$I(M; \hat{M}_T) \geq (1 - \delta) \log N - \log 2 > 0. \quad (4)$$

If  $N = 2$ , then

$$I(M; \hat{M}_T) \geq \log 2 - h_2(\delta) > 0, \quad (5)$$

where  $h_2(\delta) \triangleq -\delta \log \delta - (1 - \delta) \log(1 - \delta)$  is the binary entropy function.

**Remark 1.** In the sequel, we will consider only the cases when the set  $\mathcal{F}'$  is either “rich”, so that  $N \gg 4$ , or has only two elements, so  $N = 2$ .

*Proof.* Consider an algorithm  $\mathcal{A}$  with the claimed properties. If  $\mathcal{A}$  operates on any  $f_m \in \mathcal{F}'$ , then the event  $E_T^e(\mathcal{A}, f_m)$  will occur with probability at most  $\delta$ . From (2), we must have  $\hat{M}_T = m$  on the complement of  $E_T^e(\mathcal{A}, f_m)$ . Therefore,

$$\begin{aligned} \delta &\geq \max_{m=0, \dots, N-1} \Pr(E_T^e(\mathcal{A}, f_m)) \\ &\geq \max_{m=0, \dots, N-1} \Pr(\hat{M}_T \neq m) \\ &\geq \Pr(\hat{M}_T \neq M) \end{aligned}$$

Suppose first that  $N > 4$ . Then we can invoke the following version of Fano’s inequality [8]:

$$\Pr(\hat{M}_T \neq M) \geq 1 - \frac{I(M; \hat{M}_T) + \log 2}{\log N}.$$

Rearranging, we get (4). When  $N = 2$ , we use a stronger form of Fano’s inequality (see, e.g., Section 2.10 in [9]):

$$h_2(\Pr(\hat{M} \neq M)) \geq \log 2 - I(M; \hat{M}_T).$$

Since  $\delta \mapsto h_2(\delta)$  is monotone increasing on  $[0, 1/2]$ , we get  $h_2(\delta) \geq \log 2 - I(M; \hat{M}_T)$ . Rearranging, we get (5).  $\square$

## B. Information bounds

Lemma 1 gives a lower bound on  $I(M; \hat{M}_T)$ . This lower bound will be combined with the following upper bounds:

**Lemma 2.** Any estimator  $\hat{M} : \mathcal{X}^T \times \mathcal{Y}^T \rightarrow \{0, \dots, N - 1\}$  [and, in particular, the estimator  $\hat{M}_T$  defined in (3)] satisfies

$$I(M; \hat{M}) \leq \sum_{t=1}^T I(M; Y_t | X^t, Y^{t-1}).$$

Suppose now that the oracle is oblivious (refer to Definition 2). Then each term in the above summation simplifies:

$$I(M; Y_t | X^t, Y^{t-1}) = I(U_t; Y_t | X^t, Y^{t-1}) \leq I(U_t; Y_t),$$

where  $U_t = \psi(f_M, X_t)$ . Furthermore,

$$I(U_t; Y_t) \leq C^* \triangleq \sup_{U \in \mathcal{U}_{\mathcal{X}, \mathcal{F}}} I(U; Y),$$

where the supremum is over all random variables  $U$  taking values in  $\mathcal{U}_{\mathcal{X}, \mathcal{F}} = \psi(\mathcal{F}, \mathcal{X})$ , and the mutual information is between  $U$  and  $Y$  related via the Markov kernel  $Q(dy|u)$ .

**Remark 2.** The last bound of the lemma is nontrivial only if the number  $C^*$  is finite. This number is the Shannon capacity

of the noisy channel induced by  $Q$ .

*Proof.* We begin by writing the following:

$$I(M; \hat{M}) \leq I(M; X^T, Y^T) \quad (6)$$

$$= \sum_{t=1}^T I(M; X_t, Y_t | X^{t-1}, Y^{t-1}) \quad (7)$$

$$= \sum_{t=1}^T [I(M; X_t | X^{t-1}, Y^{t-1}) + I(M; Y_t | X^t, Y^{t-1})] \quad (8)$$

$$= \sum_{t=1}^T I(M; Y_t | X^t, Y^{t-1}), \quad (9)$$

where (6) is a consequence of the data processing inequality; (7) and (8) use the chain rule; and (9) uses the fact that  $M \rightarrow (X^{t-1}, Y^{t-1}) \rightarrow X_t$  is a Markov chain. Moreover, for an oblivious oracle we can write  $I(M; Y_t | X^t, Y^{t-1}) = I(M, U_t; Y_t | X^t, Y^{t-1})$  because, given  $X_t$  and  $M$ ,  $U_t$  is completely determined via  $U_t = \psi(f_M, X_t)$ . Therefore,

$$\begin{aligned} &I(M, U_t; Y_t | X^t, Y^{t-1}) \\ &= I(U_t; Y_t | X^t, Y^{t-1}) + I(M; Y_t | U_t, X^t, Y^{t-1}) \\ &= I(U_t; Y_t | X^t, Y^{t-1}), \end{aligned}$$

where the first step is by the chain rule and the second step is due to the fact that, for an oblivious oracle,  $M \rightarrow U_t \rightarrow Y_t$  is a Markov chain, conditionally on  $(X^t, Y^{t-1})$ . Since  $(X^t, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$  is also a Markov chain, we have

$$I(U_t; Y_t | X^t, Y^{t-1}) \leq I(U_t; Y_t) \leq C^*,$$

and the lemma is proved.  $\square$

These bounds can be particularized to specific oracles. For example, consider a noisy oblivious first-order oracle

$$Y = (f(x) + W, \nabla f(x) + Z),$$

where  $W \in \mathbb{R}$  and  $Z \in \mathbb{R}^n$  are zero mean and mutually independent. For concreteness, we will assume that  $W \sim N(0, \sigma^2)$  and  $Z \sim N(0, \sigma^2 I_n)$ , where  $I_n$  is the  $n \times n$  identity matrix. Then we have the following bound:

**Lemma 3.** For the above noisy first order oracle, we have

$$I(U_t; Y_t) \leq \frac{1}{2\sigma^2} \{ \text{var } f_M(X_t) + \mathbb{E} \|\nabla f_M(X_t)\|^2 \}. \quad (10)$$

If all  $f_m \in \mathcal{F}'$  have the same minimum value  $c^*$ , then

$$I(U_t; Y_t) \leq \frac{1}{2\sigma^2} \{ \mathbb{E} [(f_M(X_t) - c^*)^2] + \mathbb{E} \|\nabla f_M(X_t)\|^2 \} \quad (11)$$

Finally, if the oracle only supplies the noisy value of the subgradient,  $Y = \nabla f(x) + Z$ , then we will have

$$I(U_t; Y_t) \leq \frac{1}{2\sigma^2} \mathbb{E} \|\nabla f_M(X_t)\|^2. \quad (12)$$

*Proof.* Let us denote by  $U_t = (f_t^M, \nabla_t^M)$  the noiseless first-order information  $f_t^M = f_M(X_t)$  and  $\nabla_t^M = \nabla f_M(X_t)$ , and by  $Y_t = (V_t^0, V_t^1)$  the noisy observation of  $U_t$ :  $V_t^0 = f_t^M + W_t$ ,  $V_t^1 = \nabla_t^M + Z_t$ . By the independence of  $W_t$  and

$Z_t$ , we have

$$I(U_t; Y_t) \leq I(f_t^M; V_t^0) + I(\nabla_t^M; V_t^1).$$

We will separately bound  $I(f_t^M; V_t^0)$  and  $I(\nabla_t^M; V_t^1)$ , using the fact that mutual information  $I(A; B)$  between any two random variables  $A$  and  $B$  can be written as

$$I(A; B) = D(P_{B|A} \| P_{B'} | P_A) - D(P_B \| P_{B'}), \quad (13)$$

where  $B'$  is any random variable such that  $P_B \ll P_{B'}$ . For  $I(f_t^M; V_t^0)$ , use (13) with  $A = f_t^M$ ,  $B = f_t^M + W_t$ , and  $B' = c + W_t$ , where  $c$  is an arbitrary constant. Then

$$\begin{aligned} I(f_t^M; V_t^0) &\leq \min_{c \in \mathbb{R}} \mathbb{E} [D(N(f_t^M, \sigma^2) \| N(c, \sigma^2))] \\ &= \frac{1}{2\sigma^2} \min_{c \in \mathbb{R}} \mathbb{E} [(f_M(X_t) - c)^2] \\ &= \frac{1}{2\sigma^2} \text{var } f_M(X_t). \end{aligned}$$

Similarly, for  $I(\nabla_t^M; V_t^1)$  use (13) with  $A = \nabla_t^M$ ,  $B = \nabla_t^M + Z_t$ , and  $B' = Z_t$ . Then

$$\begin{aligned} I(\nabla_t^M; V_t^1) &\leq \mathbb{E} [D(N(\nabla_t^M, \sigma^2 I_d) \| N(0, \sigma^2 I_d))] \\ &= \frac{1}{2\sigma^2} \mathbb{E} \|\nabla f_M(X_t)\|^2. \end{aligned}$$

In both cases, we have used the well-known formula for the divergence between two normal distributions. Adding up the two estimates, we get (10). To obtain (11), use  $c = c^*$  to overbound  $\text{var } f_M(X_t)$  by  $\mathbb{E}[(f_M(X_t) - c^*)^2]$ . The proof of (12) is similar to that of (10).  $\square$

From now on, we will adhere to the following notation:

- $\mathcal{F}_\Theta$ , for any  $\Theta \subseteq \mathbb{X}$ , is the parametric class  $\{f_\theta(x) = \|x - \theta\| : \theta \in \Theta\}$
- $M$  is the uniformly distributed random variable describing the choice of a problem instance from a given set  $\{f_0, \dots, f_{N-1}\}$
- $\hat{M}_T$  is the estimator defined in (3)
- $(X_t, Y_t)$  is the query/answer pair at time  $t$
- $U_t$ , when the oracle is oblivious, is the deterministic response at time  $t$ :  $U_t = \psi(f_M, X_t)$
- $W, W_t \sim N(0, \sigma^2)$  and  $Z, Z_t \sim N(0, \sigma^2 I_n)$ , always
- $V_t^0$  and  $V_t^1$  are noisy versions of the function value  $f_M(X_t)$  and the subgradient  $\nabla f_M(X_t)$  at time  $t$

### C. A general information-theoretic lower bound

We now give a general information-theoretic lower bound for any problem class and any oblivious oracle, provided the Shannon capacity  $C^*$  of its noisy channel  $Q$  is finite.

**Theorem 1.** Consider a problem class  $\mathcal{P} = (\mathbb{X}, \mathcal{F}, \mathcal{O})$  with an oblivious oracle. Given  $\varepsilon > 0$ , define the packing number

$$\begin{aligned} N(\mathcal{F}, d, \varepsilon) &\triangleq \max \left\{ N \geq 1 : \right. \\ &\quad \left. \exists f_0, \dots, f_{N-1} \in \mathcal{F} : d(f_i, f_j) \geq 2\varepsilon, \forall i \neq j \right\}. \end{aligned}$$

Then, for any  $\varepsilon$  such that  $N(\mathcal{F}, d, \varepsilon) > 4$  and any  $\delta \in$

$(0, 1/2)$ , the following bounds hold for any algorithm  $\mathcal{A}$ :

$$T_{\mathcal{A}, \mathcal{P}}(\varepsilon, \delta) \geq \frac{1}{C^*} [(1 - \delta) \log N(\mathcal{F}, d, \varepsilon) - \log 2]; \quad (14)$$

$$T_{\mathcal{A}, \mathcal{P}}(\varepsilon) \geq \frac{1}{C^*} \left[ \frac{2}{3} \log N(\mathcal{F}, d, 3\varepsilon) - \log 2 \right]. \quad (15)$$

*Proof.* Let  $\mathcal{F}_\varepsilon = \{f_0, \dots, f_{N-1}\} \subset \mathcal{F}$ ,  $N = N(\mathcal{F}, d, \varepsilon)$ , be a maximal packing set in  $\mathcal{F}$ . Given  $\delta \in (0, 1/2)$  and an algorithm  $\mathcal{A}$  with  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon, \delta) = T$ , apply Lemma 1 to get

$$I(M; \hat{M}_T) \geq (1 - \delta) \log N - \log 2.$$

By Lemma 2,  $I(M; \hat{M}_T) \leq TC^*$ . Combining these two bounds, we get (14). Now, if  $\mathcal{A}$  satisfies  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon) = T$  for some  $\varepsilon > 0$ , then by Markov's inequality it will also satisfy

$$\sup_{f \in \mathcal{F}} \Pr(E_t^{3\varepsilon}(\mathcal{A}, f)) \leq \frac{\sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}(t, f)}{3\varepsilon} < \frac{1}{3}$$

for all  $t \geq T$ . Thus,  $T_{\mathcal{A}, \mathcal{P}}(3\varepsilon, 1/3) \leq T$ , and applying the same argument as above we get (15).  $\square$

**Example 4.** Let  $\mathbb{X} = B_\infty^n$  and  $\mathcal{F} = \mathcal{F}_{\text{Lip}}$ . Let  $\Lambda_\varepsilon$  be a maximal  $2\varepsilon$ -packing of  $\mathbb{X}$  in  $\ell_2$ . A simple volume counting argument shows that  $|\Lambda_\varepsilon| \geq v_n^{-1}(1/\varepsilon)^n$ , where  $v_n = \text{vol}(B_2^n)$ . Then for any two distinct functions  $f_\theta, f_{\theta'} \in \mathcal{F}_{\Lambda_\varepsilon}$  we will have  $d(f_\theta, f_{\theta'}) = \|\theta - \theta'\| \geq 2\varepsilon$ , so  $N(\mathcal{F}_{\text{Lip}}, d, \varepsilon) \geq v_n^{-1}(1/\varepsilon)^n$ . Theorem 1 then gives the following lower bound for any algorithm  $\mathcal{A}$  and any oblivious oracle with  $C^* < +\infty$ :  $T_{\mathcal{A}}(\varepsilon) = \Omega(n \log(1/\varepsilon))$ . For noiseless first-order oracles, the same lower bound follows from a binary search argument, and can be achieved using the (computationally infeasible) *method of centers of gravity* [2], [3]. In order to achieve this bound with a *noisy* oracle, an algorithm must pose queries that reduce the uncertainty by an amount that is independent of  $\varepsilon$ . This is possible with certain kinds of oracles, and will be treated in the full version of this paper.

### D. Lipschitz convex functions and noisy first-order oracles

If the oracle provides noisy first-order information, the logarithmic lower bound of Example 4 can be tightened significantly. We exhibit several applications of Lemmas 1-3 to  $\mathcal{F} = \mathcal{F}_{\text{Lip}}$  and an oracle that supplies first-order information corrupted by additive white Gaussian noise. The  $\varepsilon$ -computing times in these examples have *quadratic* dependence on  $1/\varepsilon$  but differ in their dependence on the dimension. Special cases of these results for linear functions in  $n = 1$  can be found, for example, in [10]. It should be pointed out that rates other than  $\Omega(\varepsilon^{-2})$  are possible due to 1) non-Gaussian noise or 2) different rates, depending on the smoothness of functions in  $\mathcal{F}$ , at which the information  $I(M; Y_t | X^t, Y^{t-1})$  is reduced as  $X_t$  approaches a minimizer.

In what follows, we will distinguish two types of oracles: the *gradient-only* oracle provides the gradient information, while *first-order* oracle provides both the gradient and the function value. We have the following general bounds:

**Theorem 2.** Consider a problem class  $\mathcal{P} = (\mathbb{X}, \mathcal{F}_{\text{Lip}}, \mathcal{O})$  with an oblivious first-order or gradient-only noisy oracle. Let  $N$  be the size of a maximal  $(D_X/c)$ -packing of  $\mathbb{X}$  in  $\ell_2$  for

some  $c \geq 1$ , and assume  $N > 4$ . Then, for any  $\varepsilon \leq D_X/2c$  and any  $\delta \in (0, 1/2)$ , the following bounds hold for any algorithm  $\mathcal{A}$ :

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{[(1-\delta)\log N - \log 2]\sigma^2}{2c^2\varepsilon^2} \cdot \frac{D_X^2}{D_X^2 + 1}$$

for the first-order noisy oracle, and

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{[(1-\delta)\log N - \log 2]\sigma^2}{2c^2\varepsilon^2} \cdot D_X^2$$

for the gradient-only noisy oracle.

**Remark 3.** Upper bounds on stochastic gradient descent – an algorithm which only uses the gradient information – are of the form  $O(G^2 D_X^2/\varepsilon^2)$ , where  $G^2$  is an upper bound on the expected squared norm of the noisy gradient [4]. As we show below, this is matched by our lower bounds. Indeed,  $G^2 \propto n\sigma^2$  for the additive Gaussian noise with variance  $\sigma^2$ . For the unit sphere we thus obtain  $\Omega(n\sigma^2/\varepsilon^2)$ ; for the unit hypercube we obtain  $\Omega(n^2\sigma^2/\varepsilon^2)$  for the gradient-only oracle.

**Remark 4.** The bound on  $f_M(X_t)$  in the proof below can be tightened: the information given by the *value* of the function falls below the information given by the *gradient* once the query point  $X_t$  is  $1/D_X$ -close to the minimum. It is not clear if this indicates a faster (in terms of  $n$ ) initial speed of optimization for the hypercube if the function value is used. Analysis which considers the dynamics of the process is carried out in Section IV.

*Proof of Theorem 2.* Set  $\gamma = \frac{2c\varepsilon}{D_X}$ . Let  $\Theta = \{\theta_0, \dots, \theta_{N-1}\}$  be a maximal  $(D_X/c)$ -packing set of  $\mathsf{X}$  and define  $\mathcal{F}' = \gamma\mathcal{F}_\Theta = \{f_m = \gamma f_{\theta_m} : \theta_m \in \Theta\}$ . Clearly,  $d(\gamma f_\theta, \gamma f_{\theta'}) \triangleq \gamma\|\theta - \theta'\|$  satisfies (2), and  $d(f_m, f_{m'}) \geq 2\varepsilon$  for all  $f_m, f_{m'} \in \mathcal{F}'$ . Note that  $f_m(x) = \gamma\|x - \theta_m\| \leq 2c\varepsilon$  and  $\|\nabla f_m(x)\| \leq \gamma = \frac{2c\varepsilon}{D_X}$  for any  $f_m \in \mathcal{F}'$  and  $x \in \mathsf{X}$  (the last bound holds with equality when  $x \neq \theta_m$ ). Applying Lemma 3, we get

$$I(U_t; Y_t) \leq \frac{2c\varepsilon^2}{\sigma^2} (1 + D_X^{-2})$$

for first-order oracle. The term  $1 + D_X^{-2}$  drops down to  $D_X^{-2}$  for the gradient-only oracle. Combining Lemmas 1 and 2,

$$(1-\delta)\log N - \log 2 \leq \frac{2Tc^2\varepsilon^2}{\sigma^2} (1 + D_X^{-2})$$

for the first-order oracle and, again,  $1 + D_X^{-2}$  becomes  $D_X^{-2}$  for gradient-only oracles. Rearranging yields the result.  $\square$

**Corollary 1.** Suppose  $n \geq 16$ , and  $\mathsf{X}$  contains a hypercube  $sB_\infty^n$ . Then for any  $\delta \in (0, 1/2)$  and any  $\varepsilon \leq s\sqrt{n}/8$ , any algorithm  $\mathcal{A}$  satisfies

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{\log 2 \cdot \sigma^2 s^2}{256\varepsilon^2} \cdot \frac{n[n(1-\delta) - 8]}{s^2 n + 1}$$

for the first-order oracle. For the gradient-only oracle,

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{\log 2 \cdot \sigma^2 s^2}{256\varepsilon^2} \cdot n[n(1-\delta) - 8]$$

*Proof.* By the Varshamov–Gilbert bound (Lemma 2.9 in [6]),

there exists an  $n/8$ -packing of size  $N > 2^{n/8} \geq 4$  of the binary cube  $\{-1, +1\}^n$  in the Hamming distance. This packing gives an  $(s\sqrt{n}/4)$ -packing of the scaled hypercube  $sB_\infty^n$  in  $\ell_2$ . Using Theorem 2 with  $D_X \geq s\sqrt{n}$  and  $c = 4$  yields the result.  $\square$

**Corollary 2.** Suppose  $n \geq 16$  and  $\mathsf{X}$  contains a Euclidean ball  $sB_2^n$ . Then for any  $\delta \in (0, 1/2)$  and any  $\varepsilon \leq s/8$ , any algorithm  $\mathcal{A}$  satisfies

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{\log 2 \cdot \sigma^2 s^2}{256\varepsilon^2} \cdot \frac{[n(1-\delta) - 8]}{s^2 + 1}$$

for the first-order oracle. For the gradient-only oracle,

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{\log 2 \cdot \sigma^2 s^2}{256\varepsilon^2} \cdot [n(1-\delta) - 8]$$

Corollary 2 follows immediately from Corollary 1 by noting that  $\frac{s}{\sqrt{n}}B_\infty^n \subset sB_2^n$ .

### E. Noisy oracles satisfying a moment bound

We close this section by showing how our information-theoretic technique can be used to recover the lower bounds derived by Nemirovski and Yudin [2, Ch. 5] for Lipschitz convex functions and noisy first-order oracles satisfying a certain moment constraint.

Let  $\mathsf{X} = B_\infty^n$  and  $\mathcal{F} = \mathcal{F}_{\text{Lip}}$ , and consider the class of all noisy first-order oracles whose output  $Y = (V^0, V^1) \in \mathbb{R} \times \mathbb{R}^n$  satisfies the following two conditions:

- (C1) It is unbiased, i.e.,  $\mathbb{E}[V^0|f, x] = f(x)$ ,  $\mathbb{E}[V^1|f, x] \in \partial f(x)$ ,  $\forall f \in \mathcal{F}, x \in \mathsf{X}$ .
- (C2) There exist constants  $r > 1, L > 0$ , such that

$$\mathbb{E}[|V^0 - f(x)|^r | f, x] \leq L^r, \quad \mathbb{E}[\|V^1\|^r | f, x] \leq L^r$$

for all  $f \in \mathcal{F}, x \in \mathsf{X}$ .

We will denote the class of all such oracles by  $\Pi(r, L)$ .

**Theorem 3.** There exists an oracle  $\mathcal{O} \in \Pi(r, L)$ , such that any algorithm  $\mathcal{A}$  operating on the corresponding problem class satisfies

$$T_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{\log 2 - h_2(\delta)}{c \log 2} \varepsilon^{-r/(r-1)} \quad (16)$$

for all  $\varepsilon \in (0, 1], \delta \in (0, 1/2)$  with some  $c = c(r, L) > 0$ .

*Proof.* Define two functions  $f_0(x) = -\xi^\top x$  and  $f_1(x) = \xi^\top x$ , where  $\xi \in \mathbb{R}^n$  has all coordinates equal to  $\varepsilon/n$ , and consider the noisy oracle defined by Nemirovski and Yudin [2, p. 198]. Choose a constant  $c > 0$  such that  $c^{(1-r)/r} < \min\{L, 1\}$ , and let  $p_{\varepsilon, r} \triangleq c\varepsilon^{r/(r-1)}$ . On the set  $\mathcal{F} \setminus \{f_0, f_1\}$ , our oracle acts noiselessly, while on the set  $\{f_0, f_1\}$  it acts as follows: given  $f_m, m \in \{0, 1\}$ , and  $x \in \mathsf{X}$ , it outputs

$$Y = \begin{cases} (0, 0), & \text{with probability } p_{\varepsilon, r} \\ p_{\varepsilon, r}^{-1}(f_m(x), \nabla f_m(x)), & \text{with probability } 1 - p_{\varepsilon, r}. \end{cases}$$

It is an easy exercise to show that this oracle belongs to  $\Pi(r, L)$ . Moreover, on the set  $\{f_0, f_1\}$  this oracle is oblivious because, given  $f_m$  and  $x$ , its output is a noisy version of  $(f_m(x), \nabla f_m(x))$ .

Consider an algorithm  $\mathcal{A}$  such that  $T_{\mathcal{A}}(\varepsilon, \delta) = T$ . Then  $I(U_t; Y_t | X^t, Y^{t-1}) \leq I(U_t; Y_t | X_t)$  because, given  $X_t, (X^{t-1}, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$  is a Markov chain. Now, given  $X_t = x_t$ ,  $U_t$  can take only two values, namely  $(-\xi^\top x_t, -\xi)$  or  $(\xi^\top x_t, \xi)$ . Thus,  $H(U_t | X_t) \leq \log 2$ . Moreover, since the mutual information  $I(A; B | C)$  is convex in  $P_{B|A, C}$ , we have

$$I(U_t; Y_t | X_t) \leq p_{\varepsilon, r} H(U_t | X_t) \leq p_{\varepsilon, r} \log 2.$$

Summing over the  $T$  rounds and using Lemma 2, we get

$$I(M; \hat{M}_T) \leq \sum_{t=1}^T I(U_t; Y_t | X_t) \leq T c \varepsilon^{r/(r-1)} \log 2.$$

From Lemma 1, we have  $I(M; \hat{M}_T) \geq \log 2 - h_2(\delta)$ . Combining these bounds and rearranging, we get (16).  $\square$

The statement of Theorem 3 should be interpreted in the following sense (cf. also [2]): given  $\mathsf{X}$  and  $\mathcal{F}$  as above, any algorithm  $\mathcal{A}$  will satisfy  $\sup_{\mathcal{O} \in \Pi(r, L)} T_{\mathcal{A}}(\varepsilon) = \Omega(\varepsilon^{-r/(r-1)})$ . Thus, we have a lower bound on the information complexity of any algorithm which is robust relative to  $\Pi(r, L)$ .

#### IV. LOWER BOUNDS FOR ANYTIME ALGORITHMS

In deriving the lower bounds of Section III, we have been following a certain recipe: given an algorithm that requires  $T$  oracle calls to  $\varepsilon$ -minimize every function in some class of interest with probability at least  $1 - \delta$ , we obtained a lower bound on  $T$  using a chain of inequalities of the form  $\phi_1(\varepsilon, \delta) \leq I(M; \hat{M}_T) \leq T \phi_2(\varepsilon)$ , where  $I(M; \hat{M}_T)$ , roughly speaking, is the average amount of information the algorithm can extract, after  $T$  oracle calls, about an unknown function drawn at random from some set of cardinality  $N = N(\varepsilon)$ . This gave us tight lower bounds of the form  $T \geq \phi_1(\varepsilon, \delta) / \phi_2(\varepsilon)$  for a variety of problem classes.

However, one aspect of this approach is somewhat unsatisfying. In bounding the mutual information  $I(M; \hat{M}_T)$ , we have not taken into account the *dynamics* of the algorithm, pertaining to the manner in which its expected error evolves with time. Instead, we have settled for uniform, worst-case bounds on the uncertainty remaining after each successive oracle call. In this section, we describe another technique that tracks the evolution of the mutual information over time and can be used to derive lower bounds for algorithms whose expected errors are known *a priori* to decay with time. We will call any such algorithm *anytime*.

We will show that the amount of information extracted by an anytime algorithm at each time step obeys a *law of diminishing returns*: as the queries  $X_t$  approach the minimizer, the rate at which the algorithm can reduce its uncertainty about the objective function slows down. Moreover, assuming that the worst-case expected error of such an algorithm decays polynomially with time, we will obtain lower bounds on the rate of this decay.

Let us briefly draw parallels to the work of Yang and Barron [5]. The authors showed that optimal rates of estimation are determined by a certain *critical separation*  $\varepsilon_T$ , which balances  $T\varepsilon_T^2$  and the metric entropy at resolution  $\varepsilon_T$ . The technique of Section III is similar in nature. In the present

section, however, we extend this idea by carefully tracking the diminishing information. The optimal rate is then given by the critical separation  $\varepsilon_T$  which balances the entropy  $\log N$  and the sum of diminishing mutual information terms.

First, some notation. Given an algorithm  $\mathcal{A}$  for a problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$ , let us denote by  $\overline{\text{err}}_{\mathcal{A}}(t, f)$  the worst-case expected error of  $\mathcal{A}$  at time  $t$ :

$$\overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F}) \triangleq \sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}(t, f).$$

**Definition 4.** An algorithm  $\mathcal{A}$  will be called anytime if  $\lim_{t \rightarrow \infty} \overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F}) = 0$ .

##### A. Strongly convex functions

We first consider the case of *strongly convex* functions. Given  $\mathsf{X}$ , let  $\mathcal{F}_{\kappa, L}$  denote the set of all functions  $f : \mathsf{X} \rightarrow \mathbb{R}$  that satisfy the following conditions:

- Each  $f \in \mathcal{F}_{\kappa, L}$  is *strongly convex* with parameter  $\kappa$ :

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{\kappa^2}{2} \|x - y\|^2, \quad \forall x, y \in \mathsf{X}.$$

- For each  $f$ , the mapping  $x \mapsto \nabla f(x)$  is  $L$ -Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathsf{X}.$$

Consider the noisy first-order oracle  $Y = (f(x) + W, \nabla f(x) + Z)$ , and suppose that there exists an algorithm  $\mathcal{A}$  whose worst-case errors decay at a given rate  $\{\varepsilon_t\}_{t=1}^\infty$ :

$$\overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F}) = \varepsilon_t, \quad t = 1, 2, \dots \quad (17)$$

Let  $\{f_0, \dots, f_{N-1}\} \subset \mathcal{F}_{\kappa, L}$  be a finite set of functions, such that  $f_0^* = \dots = f_{N-1}^* = c^*$  and  $\nabla f_m(x_m^*) = 0$ , where  $x_m^*$  is the (unique) minimizer of  $f_m$  on  $\mathsf{X}$ . Then we have:

**Lemma 4.** At every time  $t = 1, 2, \dots$ , any algorithm  $\mathcal{A}$  such that (17) holds also satisfies

$$I(U_t; Y_t) \leq \frac{(L/\kappa)^2}{\sigma^2} (D_{\mathsf{X}}^2 + 1) \varepsilon_t. \quad (18)$$

*Proof.* By Lemma 3,

$$I(U_t; Y_t) \leq \frac{1}{2\sigma^2} \left\{ \mathbb{E}[(f_M(X_t) - c^*)^2] + \mathbb{E}\|\nabla f_M(X_t)\|^2 \right\}. \quad (19)$$

The fact that  $\nabla f_M(x_M^*) = 0$  and the Lipschitz condition on the gradient imply that  $\|\nabla f_M(x)\| \leq LD_{\mathsf{X}}$  for all  $x \in \mathsf{X}$ . By convexity of  $f_M$ ,

$$\begin{aligned} f_M(X_t) - f_M^* &\leq \nabla f(X_t)^\top (X_t - x_M^*) \\ &\leq \|\nabla f(X_t)\| \|X_t - x_M^*\| \leq LD_{\mathsf{X}} \|X_t - x_M^*\|. \end{aligned}$$

On the other hand, from strong convexity we have that  $f_M(X_t) \geq f_M^* + (\kappa^2/2) \|X_t - x_M^*\|^2$ , which, together with (17), gives  $\mathbb{E}\|X_t - x_M^*\|^2 \leq 2\varepsilon_t/\kappa^2$ . Therefore, we can write

$$\mathbb{E}[(f_M(X_t) - c^*)^2] \leq 2D_{\mathsf{X}}^2 (L/\kappa)^2 \varepsilon_t. \quad (20)$$

Moreover, because  $\nabla f_M(x_M^*) = 0$ , we can write

$$\begin{aligned} \mathbb{E}\|\nabla f_M(X_t)\|^2 &= \mathbb{E}\|\nabla f_M(X_t) - \nabla f_M(x_M^*)\|^2 \\ &\leq L^2 \mathbb{E}\|X_t - x_M^*\|^2 \leq 2(L/\kappa)^2 \varepsilon_t. \end{aligned} \quad (21)$$



Substituting (20) and (21) into (19), we get (18).  $\square$

The lemma says that the decay of the expected error in minimizing a strongly convex function is accompanied by the decay of the average information gain, and, moreover, the two quantities decay at the same rate. For this reason, we call this the *law of diminishing returns* for strongly convex programming. Evidently, this phenomenon is due to the fact that, as the algorithm zeroes in on the minimizer, the signal-to-noise ratio keeps dropping because the mean-square error and the mean-square norm of the gradient both decrease as  $O(\varepsilon_t)$ . Using Lemma 4 in conjunction with the information bounds of Section III, we can prove the following:

**Theorem 4.** *Let  $\mathsf{X} = B_\infty^n$  and  $\mathcal{F} = \mathcal{F}_{\kappa,L}$  with  $\kappa = 1$  and  $L \geq 1$ . Suppose there exists an anytime algorithm  $\mathcal{A}$  that satisfies  $\overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F}_{\kappa,L}) = O(t^{-\gamma})$  for some  $\gamma > 0$ . Then  $\gamma \leq 1$ . In other words,  $O(t^{-1})$  is the optimal error decay rate for all anytime algorithms for this problem whose errors decay polynomially with  $t$ ; equivalently,  $T_{\mathcal{A}}(\varepsilon) = O(\varepsilon^{-1})$  is the optimal decay rate of the  $\varepsilon$ -computing time.*

*Proof.* By the anytime property, given  $\{\varepsilon_t\}$ , there exists some  $T_0$  such that  $t = T_{\mathcal{A},\mathcal{P}}(\varepsilon_t), \forall t \geq T_0$ . By Markov's inequality, we must have  $T_{\mathcal{A},\mathcal{P}}(3\varepsilon_t, 1/3) \leq t$  for all  $t \geq T_0$ .

Thus, for each  $T \geq T_0$  let  $\Lambda_T = \{\theta_0, \dots, \theta_{N-1}\}$  denote a maximal  $2\sqrt{3\varepsilon_T}$ -packing set in  $\mathsf{X}$  (w.r.t.  $\|\cdot\|$ ), and define

$$f_m(x) \triangleq (1/2)\|x - \theta_m\|^2, \quad m = 0, \dots, N-1.$$

By volume counting,  $N \geq v_n^{-1}(1/3\varepsilon_T)^{n/2}$ . We also have  $d(f_m, f_{m'}) = \frac{1}{2}\|\theta_m - \theta_{m'}\|^2 \geq 6\varepsilon_T$ . By Lemma 1,

$$I(M; \hat{M}_T) \geq \frac{n}{3} \log\left(\frac{1}{\varepsilon_T}\right) + c_n, \quad (22)$$

where  $c_n = (1/3) \log(1/3^n 8v_n^2)$ . On the other hand, applying Lemmas 2 and 4, we obtain

$$I(M; \hat{M}_T) \leq \frac{n+1}{\sigma^2} \sum_{t=1}^T \varepsilon_t. \quad (23)$$

Combining (22) and (23), we see that the sequence  $\{\varepsilon_t\}$  must satisfy the following inequalities:

$$\frac{\sigma^2 n}{3(n+1)} \log\left(\frac{1}{\varepsilon_T}\right) + c'_n \leq \sum_{t=1}^T \varepsilon_t, \quad \forall T \geq T_0 \quad (24)$$

where  $c'_n = \sigma^2 c_n / (n+1)$ . It can be shown that (24) implies the existence of an infinite subsequence of times  $1 \leq t_1 < t_2 < \dots$ , such that  $\varepsilon_{t_j} = \Omega(t_j^{-1})$ . Since  $\varepsilon_t = O(t^{-\gamma})$  by hypothesis, we must have  $\gamma \leq 1$ .  $\square$

The bound  $\Omega(t^{-1})$  is tight and can be achieved by stochastic gradient descent [4]. Note that the methods of Section III can be used to explicitly identify the dependence of the lower bound on the problem dimension  $n$ .

## B. Active learning

Our technique for analyzing anytime optimization algorithms can also be used to give a particularly simple

derivation of the minimax lower bound for active learning of a threshold function on the unit interval [7]. In a very sketchy form, the active learning problem is stated as follows. We have a pair  $(X, Z)$  of jointly distributed random variables  $X \in \mathsf{X} = [0, 1]$  and  $Z \in \{0, 1\}$ , where the marginal distribution  $P_X$  is uniform on  $[0, 1]$ , while the conditional distribution  $P_{Z|X}$  is unknown. We do, however, have some prior knowledge about  $P_{Z|X}$ . Define  $\eta(x) \triangleq \mathbb{E}[Z|X=x]$ . Then we assume the following:

- There exists some  $\theta \in [0, 1]$ , such that  $\eta(x) < 1/2$  for  $x < \theta$  and  $\eta(x) \geq 1/2$  otherwise. In other words, the *Bayes classifier*  $G^*$  for this problem is of the form  $G^*(x) = G_\theta(x) = \mathbf{1}_{\{x \geq \theta\}}$ .
- For some  $0 < c < C < 1/2$  and  $\kappa \in [1, \infty)$ , we have

$$c|x - \theta|^{\kappa-1} \leq |\eta(x) - 1/2| \leq C|x - \theta|^{\kappa-1}, \quad (25)$$

where the first inequality holds for all  $x$  in a sufficiently small neighborhood of  $\theta$ .

Let  $\Pi(\kappa, c, C)$  denote the class of all conditional probability distributions  $P_{Z|X}$  satisfying these two conditions. We wish to determine the unknown *threshold*  $\theta$  using an *active strategy*: at time  $t$ , we request a label  $z_t \in \{0, 1\}$  at a point  $x_t \in \mathsf{X}$ , chosen as a function of the history  $(x^{t-1}, z^{t-1})$ . Given our query  $x_t$ , the label  $z_t$  is generated at random according to  $P_{Z|X=x_t}$ . At time  $t$ , the candidate classifier is  $G_{x_t}(x) = \mathbf{1}_{\{x \geq x_t\}}$ . The performance of the strategy after  $t$  time steps is measured by the *excess risk* relative to  $G^*$ :

$$R(G_{x_t}) - R(G^*) = \int_{[x_t, 1] \Delta [\theta, 1]} |2\eta(x) - 1| dx, \quad (26)$$

where  $\Delta$  denotes symmetric difference between sets. [The risk of a classifier  $G : x \mapsto \{-1, +1\}$  is defined as  $R(G) \triangleq \Pr(G(X) \neq Z)$ , and the Bayes risk is  $R(G^*) \triangleq \inf_G R(G)$ .]

Castro and Nowak [7] have shown any active strategy will have excess risks of  $\Omega(t^{-\kappa/(2\kappa-2)})$ , and gave an explicit scheme that achieves the rate  $O(t^{-\kappa/(2\kappa-2)})$ . Their proof of the lower bound relies on an intricate construction of two distributions  $P_{Z|X}^{(1)}, P_{Z|X}^{(2)} \in \Pi(\kappa, c, C)$  that are close in a statistical sense, but far apart in the sense of their Bayes risks. We now show that the same lower bound can be derived using our machinery without any careful function tuning. To that end, we will cast this problem in the optimization setting, as alluded to in Example 3. Let  $\mathsf{X}$  and  $\mathcal{F}$  be as described there, and associate to each  $P_{Z|X} \in \Pi(\kappa, c, C)$  a noisy nonoblivious oracle with  $\mathsf{Y} = \{-1, +1\}$  and  $P(Y = 1|f, x) = P(Y = 1|\theta, x) = \eta(x)$ . With this correspondence in place, we can now prove the following:

**Theorem 5.** *Let  $\kappa \in (1, 2]$ . Suppose that there exists an active learning strategy satisfying*

$$\sup_{P_{Z|X} \in \Pi(\kappa, c, C)} \mathbb{E}[R(G_{X_t}) - R(G^*)] = O(t^{-\gamma})$$

*for some  $\gamma > 0$ . Then  $\gamma \leq \kappa/(2\kappa-2)$ . Thus,  $O(t^{-\kappa/(2\kappa-2)})$  is the optimal decay rate for all active learning strategies whose excess risks decay as  $\text{Poly}(t^{-1})$ . If  $\kappa = 1$ , then the optimal lower bound on the excess risk is  $\Omega(2^{-t})$ .*

*Proof.* For each  $\theta \in [0, 1]$ , find some  $P_{Z|X}^\theta \in \Pi(\kappa, c, C)$ , such that the inequalities in (25) hold for *all* values of  $x \in \mathbf{X}$ . Given a candidate classifier  $G_{X_t}$ , consider the excess risk  $R(G_{X_t}) - R(G_\theta)$ . Assume for now that  $\theta > X_t$ . Then from (26) and (25) we get

$$R(G_{X_t}) - R(G_\theta) \geq 2c \int_{X_t}^\theta (\theta - x)^{\kappa-1} dx = \frac{2c}{\kappa} (\theta - X_t)^\kappa.$$

The case  $X_t < \theta$  is similar. Thus, the expected excess risk of any strategy at time  $t$  can be bounded as

$$\mathbb{E}[R(G_{X_t}) - R(G_\theta)] \geq (2c/\kappa) \mathbb{E}|X_t - \theta|^\kappa. \quad (27)$$

Now suppose we have a learning strategy whose worst-case excess risks decay at a prescribed rate  $\{r_t\}$ :

$$\sup_{P_{Z|X} \in \Pi(\kappa, c, C)} \mathbb{E}[R(G_{X_t}) - R(G^*)] = r_t, \quad t = 1, 2, \dots$$

Then from this and (27) we have that, for every  $P_{Z|X}^\theta$ , this strategy satisfies

$$\mathbb{E}|X_t - \theta|^\kappa \leq \kappa r_t / 2c, \quad t = 1, 2, \dots \quad (28)$$

Let  $\varepsilon_t \triangleq (3\kappa r_t / 2c)^{1/\kappa}$ . Then using (28) and Markov's inequality, we see that for this strategy we must have

$$\sup_{\theta \in [0, 1]} \Pr(|X_t - \theta| \geq \varepsilon_t | \theta) \leq 1/3, \quad \forall t = 1, 2, \dots \quad (29)$$

In other words, this active learning strategy gives rise to an *optimization algorithm*  $\mathcal{A}$  for the problem class  $\mathcal{P} = (\mathbf{X}, \mathcal{F}, \mathcal{O})$ , where  $\mathcal{O}$  is specified by  $P(Y = 1 | \theta, x) = \mathbb{E}_\theta[Z | X = x]$ , and there exists some  $T_0 \geq 1$ , such that  $T_{\mathcal{A}, \mathcal{P}}(\varepsilon_t, 1/3) \leq t, \forall t \geq T_0$ .

Now for each  $T \geq T_0$  let  $\Lambda_T = \{\theta_0, \dots, \theta_{N-1}\}$  be a maximal  $2\varepsilon_T$ -packing of  $[0, 1]$ . Simple counting shows that  $N \geq 1/2\varepsilon_T$ . Consider the set  $\mathcal{F}' = \{f_m = f_{\theta_m} : \theta \in \Lambda_T\} \subset \mathcal{F}$ , and denote  $\eta_m(x) \triangleq \mathbb{E}_{\theta_m}[Z | X = x]$ . Then, in our usual notation, we have from Lemma 1 that

$$I(M; \hat{M}_T) \geq \frac{2}{3} \log \left( \frac{1}{\varepsilon_T} \right) - \frac{5}{3} \log 2. \quad (30)$$

Next we apply Lemma 2. To that end, let us inspect the terms  $I(M; Y_t | X^t, Y^{t-1})$ :

$$\begin{aligned} I(M; Y_t | X^t, Y^{t-1}) &= I(M, X_t; Y_t | X^{t-1}, Y^{t-1}) - I(X_t; Y_t | X^{t-1}, Y^{t-1}) \\ &\leq I(M, X_t; Y_t | X^{t-1}, Y^{t-1}) \leq I(M, X_t; Y_t), \end{aligned}$$

where the first step uses the chain rule, the second is because mutual information is nonnegative, and the third is because  $(X^{t-1}, Y^{t-1}) \rightarrow (M, X_t) \rightarrow Y_t$  is a Markov chain. Now we use (13) with  $A = (M, X_t)$ ,  $B = Y_t$ , and  $B'$  uniformly distributed on  $\{-1, +1\}$ . Then

$$\begin{aligned} I(M, X_t; Y_t) &\leq D(P_{Y_t | M, X_t} \| P_{B'} | P_{M, X_t}) \\ &\leq 4\mathbb{E}_{M, X_t} \{ (\Pr[Y = 1 | M, X_t] - 1/2)^2 \} \\ &= 4\mathbb{E}_{M, X_t} \{ |\eta_M(X_t) - 1/2|^2 \} \\ &\leq 4C^2 \mathbb{E}_{M, X_t} |X_t - \theta_M|^{2(\kappa-1)}, \end{aligned} \quad (31)$$

where in the second step we used the fact that

$$d(p || 1/2) \triangleq p \log 2p + (1-p) \log [2(1-p)] \leq 4(p-1/2)^2$$

for all  $p \in [0, 1]$ , and in the last step we used (25). Suppose first that  $\kappa \in (1, 2]$ . Because  $\kappa \leq 2$ , the function  $x \mapsto x^{2(\kappa-2)/\kappa}$  is concave, and we can write

$$\mathbb{E}|X_t - \theta_M|^{2(\kappa-1)} \leq (\mathbb{E}|X_t - \theta_M|^\kappa)^{2(\kappa-1)/\kappa}.$$

Using this in conjunction with (28) and Lemma 2, we can bound the mutual information  $I(M; \hat{M}_T)$  as

$$I(M; \hat{M}_T) \leq 4C^2 \sum_{t=1}^T \left( \frac{\kappa r_t}{2c} \right)^{\frac{2(\kappa-1)}{\kappa}} = \frac{4C^2}{3^{\frac{2\kappa-2}{\kappa}}} \sum_{t=1}^T \varepsilon_t^{2(\kappa-1)}. \quad (32)$$

Combining (30) and (32), we have

$$\frac{3^{(\kappa-2)/\kappa}}{2C^2} \log \left( \frac{1}{\varepsilon_T} \right) - \frac{5 \cdot 3^{(\kappa-2)/\kappa}}{4C^2} \log 2 \leq \sum_{t=1}^T \varepsilon_t^{2(\kappa-1)}.$$

An inequality like this must hold for all  $T \geq T_0$ . From this it can be shown that there exists an infinite subsequence of times  $1 \leq t_1 < t_2 < \dots$ , such that  $\varepsilon_{t_j} = \Omega \left( t_j^{-1/(2\kappa-2)} \right)$ , or, equivalently, that  $r_{t_j} = \Omega \left( t_j^{-\kappa/(2\kappa-2)} \right)$ . Since by hypothesis  $r_t = O(t^{-\gamma})$ , we must have  $\gamma \leq \kappa/(2\kappa-2)$ .

When  $\kappa = 1$ , from (31) we have  $I(M, X_t; Y_t) \leq 4C^2$  for all  $t$ . This, together with (30), gives

$$\frac{1}{6C^2} \log \left( \frac{1}{\varepsilon_T} \right) - \frac{5}{12} \log 2 \leq T, \quad \forall T \geq T_0.$$

which gives  $\varepsilon_T = \Omega(2^{-T})$  and  $r_T = \Omega(2^{-T})$ .  $\square$

## REFERENCES

- [1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, "Information-theoretic lower bounds on the oracle complexity of convex optimization," in *NIPS*, 2009, to appear.
- [2] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [3] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Kluwer, 2004.
- [4] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.
- [5] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.
- [6] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, 2009.
- [7] R. Castro and R. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.
- [8] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1247–1251, July 1994.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
- [10] A. Shapiro and A. Nemirovski, "On complexity of stochastic programming," in *Continuous Optimization: Trends and Applications*. Springer, 2005, pp. 111–144.