

# Information-Based Complexity, Feedback and Dynamics in Convex Programming

Maxim Raginsky, *Member, IEEE*, and Alexander Rakhlin

**Abstract**—We study the intrinsic limitations of sequential convex optimization through the lens of feedback information theory. In the oracle model of optimization, an algorithm queries an *oracle* for noisy information about the unknown objective function and the goal is to (approximately) minimize every function in a given class using as few queries as possible. We show that, in order for a function to be optimized, the algorithm must be able to accumulate enough information about the objective. This, in turn, puts limits on the speed of optimization under specific assumptions on the oracle and the type of feedback. Our techniques are akin to the ones used in statistical literature to obtain minimax lower bounds on the risks of estimation procedures; the notable difference is that, unlike in the case of i.i.d. data, a sequential optimization algorithm can gather observations in a *controlled* manner, so that the amount of information at each step is allowed to change in time. In particular, we show that optimization algorithms often obey the law of diminishing returns: the signal-to-noise ratio drops as the optimization algorithm approaches the optimum. To underscore the generality of the tools, we use our approach to derive fundamental lower bounds for a certain active learning problem. Overall, the present work connects the intuitive notions of “information” in optimization, experimental design, estimation, and active learning to the quantitative notion of Shannon information.

**Index Terms**—Convex optimization, Fano’s inequality, feedback information theory, hypothesis testing with controlled observations, information-based complexity, information-theoretic converse, minimax lower bounds, sequential optimization algorithms, statistical estimation.

## I. INTRODUCTION

**M**ANY problems arising in such areas as communications and signal processing, control, machine learning, economics and many others require solving mathematical programs of the form

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

Manuscript received October 11, 2010; revised March 28, 2011; accepted March 28, 2011. Date of current version October 07, 2011. M. Raginsky was supported in part by the NSF under grant CCF-1017564. A. Rakhlin was supported by the NSF CAREER award DMS-0954737. A preliminary version of this work was presented at the 47th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September/October 2009.

M. Raginsky is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: m.raginsky@duke.edu).

A. Rakhlin is with the Department of Statistics, Wharton School of Business, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: rakhlin@wharton.upenn.edu).

Communicated by D. Palomar, Associate Editor for Detection and Estimation.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2011.2154375

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex objective function and  $X$  is a compact, convex subset of  $\mathbb{R}^n$ . Therefore, it is important to have a clear understanding of the *fundamental limits* on the efficiency of convex programming methods.

A systematic study of these fundamental limits was initiated in the 1970s by Nemirovski and Yudin [1]. 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 a given class to a desired accuracy. The results in [1] are very wide in scope and cover a variety of convex programming problems in Banach spaces; finite-dimensional versions are covered in [2] and [3].

For deterministic oracles, Nemirovski and Yudin derived lower bounds on the 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 [2]). 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, which expands upon our preliminary work [4], 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]–[7]. Namely, we reduce optimization to hypothesis testing with controlled observations and then relate the resulting probability of 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 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.

The idea of relating optimization to hypothesis testing is not new. For instance, Shapiro and Nemirovski [8] derive a lower bound on the information complexity of a certain class of 1-D linear optimization problems by reducing optimization to a binary hypothesis testing problem pertaining to the parameter of a Bernoulli random variable (the outcome of a coin toss). The reduction consists in showing that any good optimization algorithm can be converted into an accurate estimator of the coin bias based on repeated independent trials; then one can derive the lower bound on the information complexity (equivalently, the minimum necessary number of coin tosses) from the data processing inequality for divergence (or Fano’s inequality). This approach was recently extended to multidimensional optimization problems by Agarwal *et al.* [9], [10]. Like the present paper, their work uses information-theoretic methods to derive lower bounds on the oracle complexity of convex optimization and their results are qualitatively similar to some of ours. However, what sets our work apart from [8]–[10] is that we explicitly account for the controlled manner in which the algorithm interacts with the oracle. This, in turn, allows us to derive tight lower bounds on the rate of error decay for certain types of *infinite-step* descent algorithms, which is not possible with the reduction to coin tossing.

Sequential procedures have become increasingly popular in the field of machine learning, mostly due to the abundance of data and the resulting need to perform computation on-line. Convex optimization is not the only sequential setting being studied: recent research in machine learning has also focused on such scenarios as active learning, multi-armed bandits and experimental design, to name a few. In all these settings, one element is common: each additional “action” should provide additional “information” about some unknown quantity. Translating this intuitive notion of “information” into precise information-theoretic statements is often difficult. Our contribution consists in offering such a translation for convex optimization and closely related problems.

A. Notation

Given a continuous function  $f : X \rightarrow \mathbb{R}$  on a compact domain  $X \subset \mathbb{R}^n$ , we denote by  $f^*$  its minimum value over  $X$

$$f^* = \inf_{x \in X} f(x).$$

We will use several basic notions from nonsmooth convex analysis [11]. The *subdifferential* of  $f$  at  $x$ , denoted by  $\partial f(x)$ , is the set of all  $g \in \mathbb{R}^n$ , such that

$$f(y) \geq f(x) + g^\top(y - x), \quad \forall y \in \mathbb{R}^n.$$

Any such  $g$  is a *subgradient* of  $f$  at  $x$ . For a convex  $f$ , the subdifferential  $\partial f(x)$  is always nonempty. When  $|\partial f(x)| = 1$ , its only element is precisely the gradient  $\nabla f(x)$ . 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 defined as

$$D_X \triangleq \sup_{x, x' \in X} \|x - x'\|.$$

The  $n \times n$  identity matrix will be denoted by  $I_n$ .

All abstract spaces are assumed to be standard Borel (i.e., Borel subsets of a complete separable metric space) and will be equipped with their Borel  $\sigma$ -fields. If  $Z$  is such a space, then  $\mathcal{B}_Z$  will denote the corresponding  $\sigma$ -field. All functions between such spaces are assumed to be measurable. If  $Z_1$  and  $Z_2$  are two such spaces, then a *Markov kernel* [12], [13] from  $Z_1$  to  $Z_2$  is a mapping  $P : \mathcal{B}_{Z_2} \times Z_1 \rightarrow [0, 1]$ , such that for any  $z_1 \in Z_1$   $P(\cdot|z_1)$  is a probability measure on  $(Z_2, \mathcal{B}_{Z_2})$  and for any  $B \in \mathcal{B}_{Z_2}$   $P(B|\cdot)$  is a measurable function on  $Z_1$ .

We will use the standard notation  $P(dz_2|z_1)$  for such a kernel. We will work with the usual information-theoretic quantities, which are well-defined in standard Borel spaces [14]. Given two (Borel) probability measures  $\mathbb{P}$  and  $\mathbb{Q}$  on  $Z$ , their *divergence* is

$$D(\mathbb{P}||\mathbb{Q}) \triangleq \begin{cases} \int_Z \left( \log \frac{d\mathbb{P}}{d\mathbb{Q}} \right) d\mathbb{P}, & \text{if } \mathbb{P} \ll \mathbb{Q} \\ +\infty, & \text{otherwise} \end{cases}$$

where the notation  $\mathbb{P} \ll \mathbb{Q}$  means that  $\mathbb{P}$  is *absolutely continuous* w.r.t.  $\mathbb{Q}$ , i.e.,  $\mathbb{Q}(B) = 0$  for any  $B \in \mathcal{B}_Z$  implies that  $\mathbb{P}(B) = 0$  as well. If  $Z$  is a product space,  $Z = Z_1 \times Z_2$ , then the *conditional divergence* between two probability distributions  $\mathbb{P}$  and  $\mathbb{Q}$  on  $Z$  given  $\mathbb{P}_{Z_1}$  (the  $Z_1$ -marginal of  $\mathbb{P}$ ) is

$$\begin{aligned} & D(\mathbb{P}_{Z_2|Z_1} || \mathbb{Q}_{Z_2|Z_1} | \mathbb{P}_{Z_1}) \\ & \triangleq \int_{Z_1} \mathbb{P}_{Z_1}(dz_1) D(\mathbb{P}_{Z_2|Z_1}(\cdot|z_1) || \mathbb{Q}_{Z_2|Z_1}(\cdot|z_1)) \end{aligned} \quad (2)$$

where  $\mathbb{P}_{Z_2|Z_1}$  and  $\mathbb{Q}_{Z_2|Z_1}$  are any versions of the regular conditional probability distributions of  $Z_2$  given  $Z_1$  under  $\mathbb{P}$  and  $\mathbb{Q}$ , respectively. This definition extends in the obvious way to situations when  $Z_1$  or  $Z_2$  are themselves product spaces. Thus, if  $\mathbb{P}$  and  $\mathbb{Q}$  are two probability distributions for a random triple  $(Z_1, Z_2, Z_3)$  taking values in a product space  $Z = Z_1 \times Z_2 \times Z_3$ , such that, under  $\mathbb{Q}$ ,  $Z_2$  and  $Z_3$  are conditionally independent given  $Z_1$ , i.e.,  $\mathbb{Q}_{Z_3|Z_1, Z_2} = \mathbb{Q}_{Z_3|Z_1}$   $\mathbb{Q}$ -a.s., then we will write

$$\begin{aligned} & D(\mathbb{P}_{Z_3|Z_1, Z_2} || \mathbb{Q}_{Z_3|Z_1, Z_2} | \mathbb{P}_{Z_1, Z_2}) \\ & = D(\mathbb{P}_{Z_3|Z_1, Z_2} || \mathbb{Q}_{Z_3|Z_1} | \mathbb{P}_{Z_1, Z_2}). \end{aligned} \quad (3)$$

Given a random couple  $(Z_1, Z_2) \in Z$  with probability distribution  $\mathbb{P}$ , the *mutual information* between  $Z_1$  and  $Z_2$  is

$$I(Z_1; Z_2) \triangleq D(\mathbb{P} || \mathbb{P}_{Z_1} \otimes \mathbb{P}_{Z_2}) \equiv D(\mathbb{P}_{Z_2|Z_1} || \mathbb{P}_{Z_2} | \mathbb{P}_{Z_1}).$$

Given a random triple  $(Z_1, Z_2, Z_3) \in Z_1 \times Z_2 \times Z_3$ , the *conditional mutual information* between  $Z_2$  and  $Z_3$  given  $Z_1$  is

$$\begin{aligned} I(Z_2; Z_3 | Z_1) & \triangleq D(\mathbb{P}_{Z_2, Z_3 | Z_1} || \mathbb{P}_{Z_2 | Z_1} \otimes \mathbb{P}_{Z_3 | Z_1} | \mathbb{P}_{Z_1}) \\ & \equiv D(\mathbb{P}_{Z_3 | Z_1, Z_2} || \mathbb{P}_{Z_3 | Z_1} | \mathbb{P}_{Z_1, Z_2}) \end{aligned} \quad (4)$$

where (4) follows from Bayes’ rule and from (3). In other words, the conditional mutual information  $I(Z_2; Z_3 | Z_1)$  is given by the conditional divergence between the joint distribution of  $Z_1, Z_2, Z_3$  and the distribution under which  $Z_2$  and  $Z_3$  are conditionally independent given  $Z_1$ .

II. SEQUENTIAL OPTIMIZATION ALGORITHMS AND THEIR INFORMATION-BASED COMPLEXITY

The work of Nemirovski and Yudin [1] deals with fundamental limitations of sequential optimization algorithms in the

real-number model of computation. The basic setting is as follows. We have a class  $\mathcal{F}$  of convex functions  $f : X \rightarrow \mathbb{R}$  on some compact convex domain  $X \subset \mathbb{R}^n$ . We seek an “optimal” algorithm that would solve the optimization problem (1) with a given guarantee of accuracy regardless of which  $f \in \mathcal{F}$  were to be optimized. The algorithms of interest operate by repeatedly querying an *oracle* for information about the unknown objective  $f$  at appropriately selected points in  $X$  and then combining the accumulated information to form a solution. The notion of optimality of an algorithm pertains to the number of queries it makes before producing a solution, without regard to the *combinatorial complexity* of computing each query. In other words, we are interested in the *information-based complexity* (IBC) [15], [16] of convex optimization problems.

The theory of IBC is concerned with intrinsic difficulty of computational problems in terms of the minimum amount of information needed to solve every problem in a given class with a given guarantee of accuracy. The word “information” here does not refer to information in the sense of Shannon, but rather to what is known *a priori* about the problem being solved, as well as what an algorithm is allowed to *learn* during its operation. There are three aspects inherent in this notion of information—it is *partial*, *noisy* and *priced*. Let us explain informally what these three terms mean in the context of optimization by means of a simple example.

Let  $X = [0, 1]$  and consider the function class

$$\mathcal{F} = \left\{ f_\theta(x) \triangleq \frac{1}{2}|x - \theta|^2 : \theta \in X \right\}. \quad (5)$$

We wish to design an algorithm that minimizes every  $f = f_\theta \in \mathcal{F}$  to a given accuracy  $\varepsilon > 0$ . At the outset, the only *a priori* information available to the algorithm consists of the problem domain  $X$ , the function class  $\mathcal{F}$  and the desired accuracy  $\varepsilon$ . The algorithm is allowed to query the value and the derivative of  $f$  at any finite set of points  $\{x_1, \dots, x_T\} \subset X$  before arriving at a solution, which we denote by  $x_{T+1}$ . The queries are answered by an *oracle*, i.e., a (possibly stochastic) device that knows the function  $f$  (or, equivalently, the parameter  $\theta$ ) and responds to any query  $x \in X$  with  $Y(\theta, x, \omega) \in \mathbb{R}^2$ , where  $\omega$  is a random element from some probability space  $(\Omega, \mathcal{B}, \mathbb{P})$  that represents oracle noise. The random variable  $Y(\theta, x, \omega)$  is assumed to be a noisy observation of the pair  $(f_\theta(x), f'_\theta(x))$ . For concreteness, let us suppose that

$$\begin{aligned} Y(\theta, x, \omega) &= Y(\theta, x, (W, Z)) \\ &= (f_\theta(x) + W, f'_\theta(x) + Z) \\ &= \left( \frac{1}{2}|x - \theta|^2 + W, x - \theta + Z \right), \end{aligned} \quad (6)$$

where  $W$  and  $Z$  are an i.i.d. pair of  $\mathcal{N}(0, \sigma^2)$  random variables.

The interaction of the algorithm and the oracle takes place as follows. Let  $\{(W_t, Z_t)\}_{t=1}^\infty$  be an i.i.d. sequence. At time  $t = 1, 2, \dots$ , the algorithm computes the query  $X_t$  as a function of the past queries  $X_\tau, 1 \leq \tau \leq t - 1$  and the corresponding oracle responses  $Y_\tau = Y(\theta, X_\tau, (W_\tau, Z_\tau)), 1 \leq \tau \leq t - 1$ . At time  $t = 0$  the algorithm knows only that  $f \in \mathcal{F}$ ; this

represents the *a priori* information. At time  $t \geq 1$ , the algorithm acquires additional data  $(X^t, Y^t) = ((X_1, Y_1), \dots, (X_t, Y_t))$  and so can refine its *a priori* information. At every time step, the information is *partial* in the sense that there are (potentially infinitely) many functions consistent with it and it is also *noisy* due to the presence of the additive disturbances  $(W^t, Z^t)$ .

Formally, for the example outlined above, an algorithm that makes  $T$  queries (or a  $T$ -step algorithm) is a tuple  $\mathcal{A} = \{\mathcal{A}_t : X^{t-1} \times Y^{t-1} \rightarrow X\}_{t=1}^{T+1}$ , where  $Y = \mathbb{R}^2$ , so that, for  $1 \leq t \leq T$ ,  $X_t = \mathcal{A}_t(X^{t-1}, Y^{t-1})$  is the query at time  $t$  and  $X_{T+1} = \mathcal{A}_{T+1}(X^T, Y^T)$  is the solution. We assume that information is *priced* in the sense that the algorithm is charged some fixed cost  $c > 0$  for every query it makes. Thus, it is desired to keep the number of queries to a minimum. With this in mind, we can define the IBC for a given accuracy  $\varepsilon > 0$  as

$$\text{IBC}(\varepsilon) = \inf \left\{ T \geq 1 : \exists \mathcal{A} = \{\mathcal{A}_t\}_{t=1}^{T+1} \right. \\ \left. \text{s.t. } \sup_{\theta \in X} \mathbb{E} f_\theta(X_{T+1}) - f_\theta^* \leq \varepsilon \right\}$$

where the expectation is taken w.r.t. the noise process  $\{(W_t, Z_t)\}_{t=1}^\infty$ . For this particular problem it can be shown that

$$\text{IBC}(\varepsilon) = \begin{cases} 1, & \sigma^2 = 0, \varepsilon \in [0, 1/2) \\ \Theta\left(\frac{\sigma^2}{\varepsilon}\right), & \sigma^2 > 0, \varepsilon \in (0, 1/2) \\ 0, & \varepsilon \geq 1/2 \end{cases} \quad (7)$$

The first entry ( $\sigma^2 = 0, 0 \leq \varepsilon < 1/2$ ) follows because the algorithm can just query  $x_1 = 0$ , obtain the response  $y_1 = ((1/2)\theta^2, -\theta)$  and immediately compute  $x_2 = \theta$ ; the last entry ( $\sigma^2 > 0, \varepsilon \geq 1/2$ ) follows because the maximum value of any  $f_\theta \in \mathcal{F}$  on  $X$  is at most  $1/2$ . The intermediate regime ( $\sigma^2 > 0, \varepsilon \in (0, 1/2)$ ) is more involved. The main contribution of the present paper is a unified information-theoretic framework for deriving lower bounds on the IBC of arbitrary sequential algorithms for solving convex programming problems.

### A. Formal Definitions

The above discussion can be formalized as follows:

*Definition 1:* A *problem class* is a triple  $\mathcal{P} = (X, \mathcal{F}, \mathcal{O})$  consisting of the following objects:

- 1) a compact, convex *problem domain*  $X \subset \mathbb{R}^n$ ;
- 2) an *instance space*  $\mathcal{F}$ , which is a class of convex functions  $f : X \rightarrow \mathbb{R}$ ;
- 3) an *oracle*  $\mathcal{O} = (Y, P)$ , where  $Y$  is the *oracle information space* and  $P(dy|f, x)$ ,  $dy \in \mathcal{B}_Y$ ,  $f \in \mathcal{F}$ ,  $x \in X$ , is a Markov kernel.<sup>1</sup>

Some restrictions must be imposed in order to exclude oracles that are “too informative,” an extreme example being  $Y = \mathcal{F} \times X$  and  $P(dy|f, x) = \delta_{f,x}(dy)$ . One way to rule this out is to require the oracle in question to be *local* [1]:

<sup>1</sup>Recall that  $\mathcal{F}$  is a subset of  $C(X)$ , the space of all continuous real-valued functions on  $X$ . Equipped with the usual sup norm,  $C(X)$  is a separable Banach space, so a Markov kernel from  $\mathcal{F} \times X$  into  $Y$  is well-defined.

**Definition 2:** We say that an oracle  $\mathcal{O}$  is *local* if for every  $x \in \mathsf{X}$  and every pair  $f, f' \in \mathcal{F}$  such that  $f = f'$  in some open neighborhood of  $x$ , we have

$$P(dy|f, x) = P(dy|f', x), \quad \forall dy \in \mathcal{B}_Y.$$

It is easy to see that the oracle described right before the definition is not local. Indeed, fix a point  $x \in \mathsf{X}$  and consider any two functions  $f, f' \in \mathcal{F}$  that agree on some open neighborhood of  $x$ , but are not equal outside this neighborhood. Then  $P(dy|f, x) = \delta_{f, x}(dy)$ , but  $P(dy|f', x) = \delta_{f', x}(dy)$ , which violates locality. Most oracles encountered in practice are local (see, for instance, the examples in Section III).

To gain more insight into stochastic oracles, we can appeal to the basic structural result for Markov kernels: If  $Z_1$  and  $Z_2$  are standard Borel spaces, then any Markov kernel  $P(dz_2|z_1)$  from  $Z_1$  to  $Z_2$  can be realized in the form  $Z_2 = \Phi(z_1, W)$ , where  $W$  is a random variable uniformly distributed on  $[0, 1]$  and  $\Phi : Z_1 \times [0, 1] \rightarrow Z_2$  is a measurable mapping [13, Lemma 3.22]. Thus, for any stochastic oracle  $P(dy|f, x)$  we can find a *deterministic* oracle  $\psi : \mathcal{F} \times \mathsf{X} \rightarrow \mathsf{U}$  with some information space  $\mathsf{U}$  and a measurable mapping  $\Phi : \mathsf{U} \times [0, 1] \rightarrow \mathsf{Y}$ , such that  $P$  can be realized as

$$Y = \Phi(\psi(f, x), W) \quad (8)$$

with  $W$  as above. Thus,  $P$  will be local in the sense of Definition 2 whenever its “deterministic part”  $\psi$  is local.

Next, we make the notion of an optimization algorithm precise. In this paper, we deal only with deterministic algorithms, although all the results can be easily extended to cover randomized algorithms as well (cf. [1] for details):

**Definition 3:** A  $T$ -step algorithm for a given  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  is a sequence of mappings  $\mathcal{A} = \{\mathcal{A}_t : \mathsf{X}^{t-1} \times \mathsf{Y}^{t-1} \rightarrow \mathsf{X}\}_{t=1}^{T+1}$ . The set of all  $T$ -step algorithms for  $\mathcal{P}$  will be denoted by  $\mathfrak{A}_T(\mathcal{P})$ .

The interaction of any  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  with  $\mathcal{O}$ , shown in Fig. 1, 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  $t = 1, 2, \dots, T$ :
  - $\mathcal{A}$  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  $P(dY_t|f, X_t)$ .
- 3) At time  $t = T + 1$ ,  $\mathcal{A}$  outputs the candidate minimizer  $X_{T+1} = \mathcal{A}_{T+1}(X^T, Y^T)$ .

We can view the set-up of Fig. 1 as a discrete-time *stochastic dynamical system* with an unknown “parameter”  $f \in \mathcal{F}$ , input sequence  $\{X_t\}$  and output sequence  $\{Y_t\}$ . The objective is to drive the system as quickly as possible to an  $\varepsilon$ -minimizing state, i.e., any  $x \in \mathsf{X}$  such that  $f(x) - f^* < \varepsilon$ , for every  $f \in \mathcal{F}$ . We are interested in the *fundamental limits* on the speed with which this can be done. Defining the *error* of  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  on  $f \in \mathcal{F}$  by

$$\text{err}_{\mathcal{A}}(T, f) \triangleq f(X_{T+1}) - \inf_{x \in \mathsf{X}} f(x) = f(X_{T+1}) - f^*$$

we introduce the following definition:

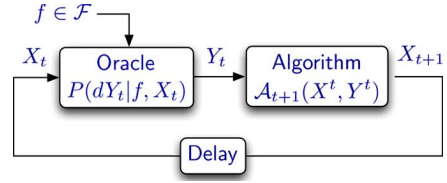


Fig. 1. Interaction of an algorithm  $\mathcal{A}$  and an oracle  $\mathcal{O}$ .

**Definition 4:** Fix a problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$ . For any  $r \geq 1$ ,  $\varepsilon > 0$  and  $\delta \in (0, 1)$ , we define the  $r$ th-order  $(\varepsilon, \delta)$ -complexity and the  $\varepsilon$ -complexity of  $\mathcal{P}$ , respectively, as

$$\begin{aligned} K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) &\triangleq \inf \left\{ T \geq 1 : \exists \mathcal{A} \in \mathfrak{A}_T(\mathcal{P}) \right. \\ &\quad \left. \text{s.t. } \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\mathcal{A}}^r(T, f) \geq \varepsilon) \leq \delta \right\} \\ K_{\mathcal{P}}^{(r)}(\varepsilon) &\triangleq \inf \left\{ T \geq 1 : \exists \mathcal{A} \in \mathfrak{A}_T(\mathcal{P}) \right. \\ &\quad \left. \text{s.t. } \sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}^r(T, f) < \varepsilon \right\}. \end{aligned}$$

When the underlying problem class  $\mathcal{P}$  is clear from context, we will write simply  $K^{(r)}(\varepsilon, \delta)$  and  $K^{(r)}(\varepsilon)$ . Moreover, when  $r = 1$  we will simply write  $K_{\mathcal{P}}(\cdot)$  or  $K(\cdot)$ .

The following is immediate from definitions (the proof is in Appendix B):

**Proposition 1:** For any  $\mathcal{P}$ ,  $r \geq 1$ ,  $\varepsilon > 0$ ,  $\delta \in (0, 1)$

$$K_{\mathcal{P}}^{(r)}(\varepsilon/\delta, \delta) \leq K_{\mathcal{P}}^{(r)}(\varepsilon).$$

The complexities  $K_{\mathcal{P}}^{(r)}(\varepsilon, \delta)$  and  $K_{\mathcal{P}}^{(r)}(\varepsilon)$  capture the intrinsic difficulty of sequential optimization over the problem class  $\mathcal{P}$  using any finite-step algorithm. However, most iterative optimization algorithms used in practice (such as stochastic gradient descent) are not run for a prescribed finite number of steps. Instead, they are run for however many steps are necessary until a desired accuracy is reached. Moreover, the error of the successive candidate minimizers produced by such an algorithm should decay monotonically with time. This observation motivates the following definitions:

**Definition 5:** A *weak infinite-step algorithm* for  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  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 set of all weak infinite-step algorithms for  $\mathcal{P}$  will be denoted by  $\mathfrak{A}_{\infty}(\mathcal{P})$ .

**Definition 6:** Given a problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}, \mathcal{O})$  and some  $r \geq 1$ , an algorithm  $\mathcal{A} \in \mathfrak{A}_{\infty}(\mathcal{P})$  is  $r$ -anytime if

$$\overline{\text{err}}_{\mathcal{A}}^{(r)}(t, \mathcal{F}) \triangleq \sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}^r(t, f) \rightarrow 0 \quad \text{as } t \rightarrow \infty. \quad (9)$$

We can now ask about fundamental limits on the rate of convergence in (9):

**Definition 7:** For any problem class  $\mathcal{P}$ , we define the  $r$ -anytime exponent as

$$\begin{aligned} \gamma_{\mathcal{P}}^{(r)} &\triangleq \sup \left\{ \gamma \geq 0 : \exists \mathcal{A} \in \mathfrak{A}_{\infty}(\mathcal{P}) \right. \\ &\quad \left. \text{s.t. } \limsup_{t \rightarrow \infty} t^{\gamma} \cdot \overline{\text{err}}_{\mathcal{A}}^{(r)}(t, \mathcal{F}) < \infty \right\}. \end{aligned}$$

According to the above definitions, the candidate minimizer  $X_{t+1}$  produced by a weak infinite-step algorithm  $\mathcal{A} \in \mathfrak{A}_\infty(\mathcal{P})$  after  $t$  queries is simultaneously the query at time  $t + 1$ . Many algorithms used in practice, such as stochastic gradient descent, are weak infinite-step algorithms. A more general class of algorithms, which we may call *strong* infinite-step algorithms, would also include strategies in which the process of issuing queries (i.e., gathering information about the objective) is separated from the process of generating candidate minimizers. Stochastic gradient descent with trajectory averaging [3], [17] is an example of such a strong algorithm. We do not consider strong infinite-step algorithms in this paper (except for a brief discussion in Appendix A), although their study is an interesting and important avenue for further research.

### III. EXAMPLES OF PROBLEM CLASSES AND PREVIEW OF SELECTED RESULTS

The following six examples show the variety of settings captured by our framework, ranging from “standard” optimization problems to such scenarios as parameter estimation, sequential experimental design and active learning.

*Example 1:* Given  $L > 0$ , let  $\mathcal{F}_{\text{Lip}}^L$  be the set of all convex functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  that are  $L$ -Lipschitz, i.e.,

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

Let  $\mathcal{Y} = \mathbb{R} \times \mathbb{R}^n$  and let  $P(dy|f, x)$  be a point mass concentrated at  $(f(x), g(x))$ , where, for each  $x \in \mathcal{X}$ ,  $g(x)$  is an arbitrary subgradient in  $\partial f(x)$ . This oracle provides noiseless *first-order information*. When  $L = 1$ , we will write  $\mathcal{F}_{\text{Lip}}$  instead of  $\mathcal{F}_{\text{Lip}}^1$ .

*Example 2:* Take  $\mathcal{F}_{\text{Lip}}^L$  as above, but now suppose that the oracle responds with

$$Y = (f(x) + W, g(x) + Z)$$

where  $W \in \mathbb{R}$  and  $Z \in \mathbb{R}^n$  are zero-mean random variables with finite second moments. Thus, any algorithm receives *noisy* first-order information and the oracle is local.

*Example 3:* Given  $\kappa > 0$ , let  $\mathcal{F}_{\text{SC}}^\kappa$  be the set of all differentiable functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  that are  $\kappa$ -strongly convex, i.e.,

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

As in the previous example, the oracle responds with

$$Y = (f(x) + W, g(x) + Z)$$

where  $W \in \mathbb{R}$  and  $Z \in \mathbb{R}^n$  are zero-mean random variables with finite second moments. When  $\kappa = 1$ , we will write  $\mathcal{F}_{\text{SC}}$  instead of  $\mathcal{F}_{\text{SC}}^\kappa$ .

*Example 4:* Fix a compact convex set  $\mathcal{X} \subset \mathbb{R}^n$  and a family of probability measures  $\{P_\theta : \theta \in \mathcal{X}\}$  on  $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ . Consider the class of convex functions

$$\mathcal{F} = \{f_\theta(x) : \theta \in \mathcal{X}\} \quad (10)$$

such that for every  $\theta \in \mathcal{X}$   $f_\theta(\theta) = \min_{x \in \mathcal{X}} f_\theta(x)$ . Consider also the oracle  $\mathcal{O} = (\mathcal{Y}, P)$ , defined by

$$P(dy|f_\theta, x) = P_\theta(dy), \quad \forall(\theta, x) \in \mathcal{X} \times \mathcal{X}. \quad (11)$$

This oracle ignores the query  $x$  and simply outputs a random element  $Y \sim P_\theta$ . The problem class  $(\mathcal{X}, \mathcal{F}, \mathcal{O})$  thus describes the statistical problem of *estimating the parameter of a probability distribution*. More generally, we can consider the function class

$$\mathcal{F} = \left\{ f_\theta(\cdot) = \mathbb{E}_\theta[F(\cdot, Y)] \equiv \int_{\mathcal{Y}} F(\cdot, y) P_\theta(dy) : \theta \in \mathcal{X} \right\} \quad (12)$$

where we assume that:

- for each fixed  $y \in \mathcal{Y}$ , the function  $x \mapsto F(x, y)$  is convex;
- $f_\theta(\theta) = \min_{x \in \mathcal{X}} f_\theta(x)$ .

The second condition says that  $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  is a *contrast function* [18]. Most classical problems in statistical inference, such as estimating the mean, the median, or the variance of a distribution, can be cast as minimizing a convex contrast function of the form (12). For instance, if  $\mathcal{X} \subset \mathbb{R}$ ,  $\mathcal{Y} = \mathbb{R}$ ,  $\mathbb{E}_\theta[Y] = \theta$  for each  $\theta \in \mathcal{X}$  and  $F(x, y) = (x - y)^2$ , then

$$f_\theta(x) = \mathbb{E}_\theta[(Y - x)^2]$$

with  $f_\theta(x) \geq f_\theta(\theta) \equiv \text{var}(P_\theta)$ , so we recover the problem of estimating the mean.

*Example 5:* As we have just seen, the queries are of no use in statistical estimation since the samples the statistician obtains depend only on the unknown parameter  $\theta$ . By contrast, the setting in which the statistician’s queries *do* affect the observations is known as *sequential experimental design* [19]–[21]. Consider the case when  $\mathcal{X} \subset \mathbb{R}^n$  is compact and convex, as in the above example. Suppose also that we have two families of probability measures on  $\mathcal{Y}$ ,  $\{Q_\theta : \theta \in \mathcal{X}\}$  and  $\{P_{\theta, x} : (\theta, x) \in \mathcal{X} \times \mathcal{X}\}$ . The function class is as in (12) but with  $Q_\theta$  replacing  $P_\theta$ , while the oracle now is defined by

$$P(dy|f_\theta, x) = P_{\theta, x}(dy), \quad \forall(\theta, x) \in \mathcal{X} \times \mathcal{X}.$$

Thus, the role of  $Q_\theta$  is to provide a measure of performance (or goodness-of-fit) of the final estimate of  $\theta$ , while  $P_{\theta, x}$  describes the experimental model (i.e., the relationship between the input  $X$  and the response  $Y$  given the parameter  $\theta$ ).

*Example 6:* Our last example is at the intersection of statistical learning theory and sequential experimental design. Let

$$\mathcal{X} = [0, 1], \quad \mathcal{Y} = \{-1, +1\}, \quad \mathcal{F} = \{f_\theta(x) = |x - \theta| : \theta \in \mathcal{X}\}.$$

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 [22] and will be treated in detail in Section VI.

We now briefly discuss some of the lower bounds that arise from the techniques introduced in the paper. First, Theorem 1 in Section V implies a general lower bound of the form

$$\Omega(n^\alpha \log(1/\varepsilon))$$

on the number of oracle calls required to  $\varepsilon$ -minimize every function in a given class, where the exponent  $\alpha > 0$  depends on the geometry of the problem domain  $X$  and on the complexity of the instance space  $\mathcal{F}$ . For convex Lipschitz functions and noiseless first-order oracles (Example 1), or more generally for stochastic oracles that are sufficiently “informative” in a sense we make precise, this lower bound holds with  $\alpha = 1$  (cf. the discussion right after Theorem 1). This lower bound is known to be optimal in the noiseless case [1] and in certain noisy scenarios when  $n = 1$  [23]; however, our techniques lead to a much more transparent proof of the bound.

For the noisy first-order oracle with zero-mean Gaussian noise of variance  $\sigma^2$ , we obtain lower bounds of the form

$$\Omega(\sigma^2 n^{\alpha_1} (1/\varepsilon)^{\alpha_2})$$

where the exponent  $\alpha_1$  depends, as before, on the geometry of  $X$ , on the complexity of  $\mathcal{F}$ , as well as on whether the oracle supplies full first-order information (function value and subgradient) or just the subgradient. The exponent  $\alpha_2$  depends on the details of the function class  $\mathcal{F}$ . More specifically:

- for  $\mathcal{F}_{\text{Lip}}$  (Example 2), we have  $\alpha_2 = 2$  (Theorem 2 in Section V);
- for  $\mathcal{F}_{\text{SC}}$  (Example 3), we have  $\alpha_2 = 1$  (Theorem 3 in Section V).

The corresponding result for convex Lipschitz functions in  $n = 1$  can be found in [1], [8], yet we obtain the optimal dependence on  $n$  for higher dimensions. Our lower bound for strongly convex functions seems to be new; in particular, Nemirovski and Yudin [1] only consider the noiseless case, while Agarwal *et al.* [9], [10] consider noisy first-order oracles, but with a different oracle model, which does not allow additive noise due to a coin-tossing construction. Ignoring the dependence on the dimension, we also obtain the error decay rate  $\Omega(\sigma^2/t)$  for  $\mathcal{F}_{\text{SC}}$  when we restrict ourselves to anytime infinite-step algorithms (Theorem 5 in Section VI). To the best of our knowledge, such analysis does not appear anywhere else in the literature. The bounds of (7) essentially capture the fundamental limits of strongly convex programming in one dimension and can be easily deduced using our techniques (a sketch of the derivation is given in Section IV-A). We also derive new (and tighter) lower bounds on anytime algorithms for minimizing higher-order polynomials under a second-moment error criterion (Theorems 6 and 7 in Section VI).

Apart from “standard” optimization problems, our framework seamlessly captures several statistical problems with an optimization flavor. In particular, in Section V-D we look at information-based complexity of statistical estimation and sequential experimental design (Examples 4 and 5, respectively). Here we do not aim at obtaining tight rates for specific settings of interest, but rather show the connections to the techniques employed in statistics. Finally, we show in Section VI-C that our methodology leads to a particularly easy derivation of a

lower bound for the active learning problem of Example 6. This bound was previously obtained in [22] using a much more involved argument relying on a careful construction of a “difficult” subset of functions.

Overall, our main contributions are the development of a general framework that captures many diverse settings with optimization flavor, as well as a novel analysis that takes into account the effect of feedback upon the dynamics of the interaction between the algorithm and the oracle.

#### IV. SETTING THE STAGE: OPTIMIZATION VERSUS HYPOTHESIS TESTING WITH FEEDBACK

We now lay down the foundations of 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.

Conceptually, our techniques are akin to the ones used in statistical literature to obtain minimax lower bounds on the risks of estimation procedures [5]–[7]. The main idea is this. Given a problem class  $\mathcal{P} = (X, \mathcal{F}, \mathcal{O})$ , we construct a “difficult” finite subclass  $\mathcal{F}' = \{f_0, \dots, f_{N-1}\} \subset \mathcal{F}$ , such that the functions in it are nearly indistinguishable from one another based on the information supplied by the oracle in response to any possible query and yet they are sufficiently far apart from one another, so that a candidate approximate minimizer for any one of them fails to minimize all the remaining functions to the same accuracy. Once such a class is constructed, we consider a fictitious situation in which Nature selects an element of  $\mathcal{F}'$  *uniformly at random*. Then for every  $T$ -step algorithm  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  we can construct a probability space  $(\Omega, \mathcal{B}, \mathbb{P})$  with the following random variables defined on it:

- $M \in \{0, \dots, N - 1\}$ , which encodes the random choice of a problem instance in  $\mathcal{F}'$ ;
- $X^{T+1} \in X^{T+1}$ , where  $X^T$  are the queries issued by  $\mathcal{A}$  and  $X_{T+1}$  is the candidate minimizer;
- $Y^T \in Y^T$  are the responses of  $\mathcal{O}$  to the queries issued by  $\mathcal{A}$ .

These variables describe the interaction between Nature, the algorithm and the oracle and thus have the causal ordering

$$M, X_1, Y_1, \dots, X_t, Y_t, \dots, X_T, Y_T, X_{T+1}$$

where,  $\mathbb{P}$ -almost surely

$$\begin{aligned} \mathbb{P}(M = i) &= \frac{1}{N} \\ \mathbb{P}(X_t \in A | M, X^{t-1}, Y^{t-1}) &= \mathbf{1}_{\{\mathcal{A}_t(X^{t-1}, Y^{t-1}) \in A\}} \\ \mathbb{P}(Y_t \in B | M, X^t, Y^{t-1}) &= P(B | f_M, X_t) \end{aligned} \quad (13)$$

for all  $i \in \{0, \dots, N - 1\}$ ,  $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$ . In other words,  $M \rightarrow (X^{t-1}, Y^{t-1}) \rightarrow X_t$  and  $(X^{t-1}, Y^{t-1}) \rightarrow (M, X_t) \rightarrow Y_t$  are Markov chains for every  $t$ .

The reason for such punctilious bookkeeping is that now we can relate the problem faced by  $\mathcal{A}$  to sequential hypothesis testing with feedback, as defined by Burnashev [24]. We can think of  $M$  as encoding the choice of one of  $N$  equiprobable hypotheses. At each time  $t$ , the algorithm issues a query  $X_t$

and receives an observation  $Y_t$  which is stochastically related to  $X_t$  and  $M$  via the kernel  $\mathbb{P}(dY_t|M, X_t) = P(dY_t|f_M, X_t)$ . The current query may depend only on the past queries and observations. At time  $T+1$ , the algorithm produces a candidate minimizer,  $X_{T+1}$ . As we will shortly demonstrate, we can use the information available to  $\mathcal{A}$  at time  $T+1$  to construct an estimate  $\hat{M}_T$  of the true hypothesis  $M$ .<sup>2</sup> Once this is done, we can analyze the *mutual information*  $I(M; \hat{M}_T)$ , which is well-defined because we have specified  $\mathbb{P}$ . In particular, the analysis hinges on the following observations. Suppose that  $\mathcal{A}$  is such that for some  $r \geq 1$ ,  $\varepsilon > 0$  and  $\delta \in (0, 1)$  we have

$$\Pr\left(\text{err}_{\mathcal{A}}^r(T, f) \geq \varepsilon\right) \leq \delta, \quad \forall f \in \mathcal{F}$$

where the probability is w.r.t. the randomness in the oracle's responses. Then, first of all

$$\mathbb{P}\left(\text{err}_{\mathcal{A}}^r(T, f_M) \geq \varepsilon\right) \leq \delta.$$

We will use this fact, together with the “geometric” distinguishability of the functions  $\{f_i\}$ , to show that  $\mathbb{P}(\hat{M}_T \neq M) \leq \delta$  and, as a consequence, that there exists some  $\Psi_1(r, \varepsilon, \delta) > 0$ , such that

$$I(M; \hat{M}_T) \geq \Psi_1(r, \varepsilon, \delta). \quad (14)$$

In other words, a good algorithm should be able to obtain a nontrivial amount of information about the hypothesis  $M$ . On the other hand, by the data processing inequality,  $I(M; \hat{M}_T) \leq I(M; X^T, Y^T)$  and we will use statistical indistinguishability of  $\{f_i\}$ , as well as the structure of the oracle, to obtain an upper bound of the form

$$I(M; \hat{M}_T) \leq T\Psi_2(r, \varepsilon) \quad (15)$$

with some  $\Psi_2(r, \varepsilon) < +\infty$ . The two bounds are then combined to yield

$$T \geq \frac{\Psi_1(r, \varepsilon, \delta)}{\Psi_2(r, \varepsilon)} \implies K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \geq \frac{\Psi_1(r, \varepsilon, \delta)}{\Psi_2(r, \varepsilon)}. \quad (16)$$

### A. An Illustrative Example

To illustrate our method in action, we will sketch the derivation of the nontrivial part of the lower bound in (7), i.e., when  $\varepsilon \in (0, 1/2)$ . Let

$$x_0^* = \begin{cases} 1/2 - \sqrt{2\varepsilon}, & \varepsilon \in (0, 1/8) \\ 0, & \varepsilon \in [1/8, 1/2) \end{cases}$$

$$x_1^* = \begin{cases} 1/2 + \sqrt{2\varepsilon}, & \varepsilon \in (0, 1/8) \\ 1, & \varepsilon \in [1/8, 1/2). \end{cases}$$

It is easy to see that  $x_0^*, x_1^* \in [0, 1]$ . Consider two functions

$$f_m(x) = \frac{1}{2}(x - x_m^*)^2, \quad m \in \{0, 1\}.$$

<sup>2</sup>It is important to keep in mind that the hypothesis testing set-up is purely fictitious — indeed,  $\mathcal{A}$  may or may not know that the problem instances are drawn at random among  $\{f_0, \dots, f_{N-1}\}$ , rather than arbitrarily from the entire instance space  $\mathcal{F}$ . The point is, though, that the average performance of  $\mathcal{A}$  on  $\mathcal{F}'$  cannot be better than its worst-case performance on  $\mathcal{F}$ . In statistical terms, the minimax risk of  $\mathcal{A}$  over  $\mathcal{F}$  is bounded below by the Bayes risk over any subset of  $\mathcal{F}$ .

A simple calculation shows that for any  $x \in [0, 1]$  such that

$$f_0(x) - f_0^* = f_0(x) = \frac{1}{2}(x - x_0^*)^2 < \varepsilon$$

we must have  $f_1(x) - f_1^* = f_1(x) > \varepsilon$  and the same holds with the roles of  $f_0$  and  $f_1$  reversed. Thus, any  $\varepsilon$ -minimizer of  $f_0$  fails to  $\varepsilon$ -minimize  $f_1$  and vice versa.

On the other hand, the probability distribution of the output of the first-order Gaussian oracle (6) for any query  $x \in [0, 1]$  when  $M = 0$  is very close to its  $M = 1$  counterpart. Indeed, letting  $Y \in \mathbb{R}^2$  denote the output of the oracle, we have

$$\mathbb{P}_{Y|M=m, X=x} = \mathcal{N}\left(\frac{1}{2}(x - x_m^*)^2, \sigma^2\right) \otimes \mathcal{N}(x - x_m^*, \sigma^2).$$

Then it is not hard to show that, for  $m \in \{0, 1\}$

$$D(\mathbb{P}_{Y|M=m, X=x} \parallel \mathbb{P}_{Y|M=1-m, X=x}) = O\left(\frac{\varepsilon}{\sigma^2}\right). \quad (17)$$

In other words, the functions  $f_0$  and  $f_1$  are nearly indistinguishable from one another based on the outcome of a single query.

Now suppose that Nature selects an index  $M \in \{0, 1\}$  uniformly at random. Consider a  $T$ -step algorithm that  $\varepsilon$ -minimizes every function in the class  $\mathcal{F}$  defined in (5) with probability at least  $1 - \delta$ , where  $\delta \in (0, 1/2)$ . Let  $r = 1$ . Then Lemma 1 in Section IV-C can be used to show that the lower bound (14) holds with

$$\Psi_1(1, \varepsilon, \delta) = \log 2 - h_2(\delta) > 0$$

where  $h_2(\delta) \triangleq -\delta \log \delta - (1-\delta) \log(1-\delta)$  is the binary entropy function. On the other hand, using Lemmas 2 and 4, as well as (17), we can show that the upper bound(15) holds with

$$\Psi_2(1, \varepsilon) = \max_{m \in \{0, 1\}} \max_{x \in [0, 1]} D(\mathbb{P}_{Y|M=m, X=x} \parallel \mathbb{P}_{Y|M=1-m, X=x}) = O\left(\frac{\varepsilon}{\sigma^2}\right).$$

Hence, according to (16), any  $T$ -step algorithm that  $\varepsilon$ -minimizes every function in the class  $\mathcal{F}$  of (5) with probability at least  $1 - \delta$  must satisfy

$$T = \Omega\left(\frac{\sigma^2(\log 2 - h_2(\delta))}{\varepsilon}\right).$$

From this and from Proposition 1, we can obtain the lower bound  $\Omega(\sigma^2/\varepsilon)$  of (7). The matching upper bound  $O(\sigma^2/\varepsilon)$  is achieved by stochastic gradient descent [3].

### B. Reduction to Hypothesis Testing With Feedback

We now develop our information-theoretic methodology in the general setting of Section II-A.

Let us fix a problem class  $\mathcal{P} = (\mathcal{X}, \mathcal{F}, \mathcal{O})$ . To set up our analysis, we first endow the instance space  $\mathcal{F}$  with a “distance”  $d(\cdot, \cdot)$  that has the following property: for any  $x \in \mathcal{X}$  and any  $\varepsilon > 0$

$$d(f, g) \geq 2\varepsilon \text{ and } f(x) < f^* + \varepsilon \implies g(x) > g^* + \varepsilon. \quad (18)$$

In other words, an  $\varepsilon$ -minimizer of a function cannot simultaneously be an  $\varepsilon$ -minimizer of a distant function. It is easy to

construct a  $d$  satisfying (18) for any particular class  $\mathcal{F}$  of continuous functions, although such a  $d$  need not be a metric. For example, if we consider the class

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

for some  $\Theta \subset X$ , then  $d(f_\theta, f_{\theta'}) = \|\theta - \theta'\|$  satisfies (18). Indeed,  $\|\theta - \theta'\| \geq 2\varepsilon$  and  $\|x - \theta\| < \varepsilon$  imply  $\|x - \theta'\| > \varepsilon$  by the triangle inequality. For a general  $\mathcal{F}$ , we can also define

$$d(f, g) \triangleq \inf_{x \in X} [f(x) + g(x)] - [f^* + g^*]$$

the distance-like function introduced in [9], [10]. This definition coincides with  $d(f_\theta, f_{\theta'}) = \|\theta - \theta'\|$  for the parametric set  $\mathcal{F}_\Theta$ ; however, (18) is the most general requirement. Note that we will often implicitly restrict our consideration to a subclass of  $\mathcal{F}$  and define an appropriate  $d$  on that subclass.

Let us fix the exponent  $r \geq 1$  and 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  $2\varepsilon^{1/r}$  apart in  $d(\cdot, \cdot)$ . Given any  $T \in \mathbb{N}$  and an algorithm  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$ , we can now construct the probability space  $(\Omega, \mathcal{B}, \mathbb{P})$ , as described in the introduction to this section. Given  $X_{T+1}$ , the output of  $\mathcal{A}$ , we can define the “estimator”

$$\hat{M}_T(X^T, Y^T) \triangleq \arg \min_{i=0, \dots, N-1} [f_i(X_{T+1}) - f_i^*] \quad (19)$$

which simply selects that function in  $\mathcal{F}'$  for which the error of  $X_{T+1}$  is the smallest. Since  $X_{T+1}$  is  $\sigma(X^T, Y^T)$ -measurable, the estimator  $\hat{M}_T$  is indeed a function only of the information available to  $\mathcal{A}$  after time  $T$ .

### C. Information Bounds

The main object of interest will be the mutual information  $I(M; \hat{M}_T)$ . We first show that any “good”  $T$ -step algorithm obtains a nonzero amount of information about  $M$  at the end of its operation:

*Lemma 1:* Fix some  $r \geq 1$ ,  $\delta \in (0, 1/2)$  and  $\varepsilon > 0$ . Suppose  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  attains

$$\sup_{f \in \mathcal{F}} \Pr(\text{err}_{\mathcal{A}}^r(T, f) \geq \varepsilon) \leq \delta. \quad (20)$$

Let  $\mathcal{F}' \subset \mathcal{F}$  be a finite set  $\{f_0, \dots, f_{N-1}\}$  of functions, such that

$$d(f_i, f_j) \geq 2\varepsilon^{1/r}, \quad \forall i \neq j.$$

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 (19) satisfies the bound

$$I(M; \hat{M}_T) \geq (1 - \delta) \log N - \log 2 > 0. \quad (21)$$

If  $N = 2$ , then

$$I(M; \hat{M}_T) \geq \log 2 - h_2(\delta) > 0 \quad (22)$$

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. Define, for each  $i$ , the event

$$E_i \triangleq \{\text{err}_{\mathcal{A}}^r(T, f_i) \geq \varepsilon\}.$$

We first show that the event  $\{\hat{M}_T \neq i\}$  implies  $E_i$ . Indeed, if  $E_i$  does not occur, then from the fact that  $d(f_i, f_j) \geq 2\varepsilon^{1/r}$  for all  $j \neq i$  and from (18) we deduce that

$$f_j(X_{T+1}) - f_j^* > \varepsilon^{1/r} > f_i(X_{T+1}) - f_i^*, \quad \forall j \neq i$$

so it must be the case that  $\hat{M}_T = i$ . Therefore

$$\begin{aligned} \delta &\geq \max_{i=0, \dots, N-1} \mathbb{P}(E_i | M = i) \\ &\geq \max_{i=0, \dots, N-1} \mathbb{P}(\hat{M}_T \neq i | M = i) \\ &\geq \mathbb{P}(\hat{M}_T \neq M). \end{aligned}$$

Now suppose that  $N > 4$ . Then we can invoke the following version of Fano’s inequality [25]

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

Rearranging, we get (21). When  $N = 2$ , we use a stronger form of Fano’s inequality (see, e.g., Section 2.10 in [26])

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

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 (22). ■

On the other hand, the amount of information  $I(M; \hat{M}_T)$  cannot be too large:

*Lemma 2:* Any estimator  $\hat{M} : X^T \times Y^T \rightarrow \{0, \dots, N-1\}$  [and, in particular, the estimator  $\hat{M}_T$  defined in (19)] satisfies

$$I(M; \hat{M}) \leq \sum_{t=1}^T I(M; Y_t | X^t, Y^{t-1}). \quad (23)$$

*Remark 2:* The terms  $I(M; Y_t | X^t, Y^{t-1})$  have analogues in the literature on information-theoretic experimental design (see, e.g., [19], [20]). In that context, they represent the average reduction of uncertainty about the unknown variable  $M$  after observing the *experimental outcome*  $Y_t$  based on the *design point*  $X_t = \mathcal{A}_t(X^{t-1}, Y^{t-1})$ .

*Proof:* We have

$$I(M; \hat{M}) \leq I(M; X^T, Y^T) \quad (24)$$

$$= \sum_{t=1}^T I(M; X_t, Y_t | X^{t-1}, Y^{t-1}) \quad (25)$$

$$= \sum_{t=1}^T [I(M; X_t | X^{t-1}, Y^{t-1}) + I(M; Y_t | X^t, Y^{t-1})] \quad (26)$$

$$= \sum_{t=1}^T I(M; Y_t | X^t, Y^{t-1}) \quad (27)$$



where (24) is a consequence of the data processing inequality; (25) and (26) use the chain rule; and (27) uses the fact that  $M \rightarrow (X^{t-1}, Y^{t-1}) \rightarrow X_t$  is a Markov chain. ■

#### D. Refinement of the Upper Bounds

Lemmas 1 and 2 are the two main elements of our approach. In order to apply them, we need to get a handle on the conditional mutual information terms on the right-hand side of (23). The following two lemmas, whose proofs can be found in Appendix B, give us just the right tools for that:

*Lemma 3:* Consider any estimator  $\hat{M} : \mathcal{X}^T \times \mathcal{Y}^T \rightarrow \{0, \dots, N-1\}$ . Then, considering any realization of the oracle  $\mathcal{O}$  in the form (8), we have the bound

$$I(M; \hat{M}) \leq \sum_{t=1}^T I(U_t; Y_t)$$

where  $U_t = \psi(f_M, X_t)$  is the output of the “deterministic part” of the oracle.

*Lemma 4:* Consider any estimator  $\hat{M} : \mathcal{X}^T \times \mathcal{Y}^T \rightarrow \{0, \dots, N-1\}$ . For any sequence of conditional probability measures  $\{\mathbb{Q}_{Y_t|X^t, Y^{t-1}}\}_{t=1}^T$  on  $(\Omega, \mathcal{B})$  satisfying the conditions

$$\mathbb{P}_{Y_t|X^t, Y^{t-1}} \ll \mathbb{Q}_{Y_t|X^t, Y^{t-1}}, \quad t = 1, \dots, T \quad (28)$$

we have the bound

$$I(M; \hat{M}) \leq \sum_{t=1}^T D(\mathbb{P}_{Y_t|M, X^t, Y^{t-1}} \| \mathbb{Q}_{Y_t|X^t, Y^{t-1}} | \mathbb{P}_{M, X^t, Y^{t-1}}) \quad (29)$$

*Remark 3:* By hypothesis on the behavior of the oracle,  $(X^{t-1}, Y^{t-1}) \rightarrow (M, X_t) \rightarrow Y_t$  is a Markov chain. Hence,  $\mathbb{P}_{Y_t|M, X^t, Y^{t-1}}$  in (29) can be replaced with  $\mathbb{P}_{Y_t|M, X_t}$ .

The key to using Lemma 4 is in the judicious choice of the “comparison” measures  $\mathbb{Q}_{Y_t|X^t, Y^{t-1}}$ . In particular, we will use two different strategies of choosing the  $\mathbb{Q}$ ’s, which in turn lead to two different types of bounds:

- **Information Radius (IR) bound**—This bound is useful for analyzing arbitrary finite-time algorithms. For each  $t$ , take  $\mathbb{Q}_{Y_t|X^t, Y^{t-1}} = \mathbb{Q}_{Y_t|X_t}$  to be the mixture

$$\mathbb{Q}_{Y_t|X_t} = \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{P}_{Y_t|M=i, X_t}.$$

Then, letting  $M'$  denote an independent copy of  $M$  and noting that  $\mathbb{Q}_{Y_t|X_t} = \mathbb{E}_{M'} \mathbb{P}_{Y_t|M', X_t}$ , we obtain

$$\begin{aligned} I(M; \hat{M}) &= \sum_{t=1}^T D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{E}_{M'} \mathbb{P}_{Y_t|M', X_t} | \mathbb{P}_{M, X_t}) \\ &\leq \sum_{t=1}^T \mathbb{E}_{M'} D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{P}_{Y_t|M', X_t} | \mathbb{P}_{M, X_t}) \quad (30) \end{aligned}$$

$$\begin{aligned} &= \sum_{t=1}^T \mathbb{E}_{M, X_t} \mathbb{E}_{M'} D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{P}_{Y_t|M', X_t}) \\ &\leq T \max_{i,j} \sup_{x \in \mathcal{X}} D(\mathbb{P}_{Y|M=i, X=x} \| \mathbb{P}_{Y|M=j, X=x}) \quad (31) \end{aligned}$$

where (30) follows from Jensen’s inequality and convexity of the divergence. The use of the term “information radius” is inspired by an analogous concept in the theory of information-based complexity [16]: the divergence  $D(\mathbb{P}_{Y|M=i, X=x} \| \mathbb{P}_{Y|M=j, X=x})$  quantifies how close, in a statistical sense, the oracle’s responses are for a given query point  $x \in \mathcal{X}$  and a given pair  $i, j$ . Viewing the random variable  $Y \sim \mathbb{P}_{Y|M=i, X=x}$  as (stochastic, noisy) *information* about the function  $f_i$  at the point  $x$ , we can interpret the quantity multiplying  $T$  in (31) as a measure of *ambiguity* of this information. We use IR bounds in Section V.

- **Lyapunov Function (LF) bound**—This bound is useful for analyzing anytime algorithms. It relies on the idea that, with certain types of problem classes, the oracle responds with “pure noise” whenever the query point happens to hit upon a minimizer. In other words, there exists a probability measure  $Q^*$  on  $\mathcal{Y}$ , such that

$$P(dy|f, x) = Q^*(dy) \quad \text{if } x \in \arg \min_x f$$

where  $\arg \min_x f \triangleq \{x \in \mathcal{X} : f(x) = f^*\}$ . Moreover,

for an anytime algorithm it is often the case that the conditional divergence  $D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{Q}_{Y_t}^* | \mathbb{P}_{M, X_t})$ , where  $\mathbb{Q}_{Y_t}^*$  is an independent copy of  $Q^*$ , decreases with  $t$  and, hence, can be thought of as a Lyapunov function for the problem at hand (in fact, Lyapunov functions of the divergence type have been used before to analyze the convergence of specific stochastic optimization algorithms [27]). This leads to the natural choice of  $\mathbb{Q}_{Y_t|X^t, Y^{t-1}} = \mathbb{Q}_{Y_t}^*$  and to the bound

$$I(M; \hat{M}) \leq \sum_{t=1}^T D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{Q}_{Y_t}^* | \mathbb{P}_{M, X_t}). \quad (32)$$

We apply the LF bounds in Section VI to the study of anytime algorithms.

We should point out that the use of an auxiliary measure  $\mathbb{Q}$  has been pioneered by Yang and Barron [6] (with later refinements by Yang [28]) in the usual setting of statistical estimation from i.i.d. data; there, the relevant bound was of the form

$$\begin{aligned} I(M; Y^T) &\leq D(\mathbb{P}_{Y^T|M} \| \mathbb{Q}_{Y^T} | \mathbb{P}_M) \\ &\leq \max_m D(\mathbb{P}_{Y^T|M=m} \| \mathbb{Q}_{Y^T}) \end{aligned}$$

(cf. [6, p.1571]). Similarly, the “symmetrization trick” involving an independent copy of  $M$  and an application of Jensen’s inequality, as in (30) above, is used often in the statistics literature (cf. [5] and references therein). Our innovation consists in first performing a sequential decomposition of the mutual information  $I(M; X^T, Y^T)$ , carefully taking into account all the Markov structures that arise due to the causality constraints that must be obeyed by the algorithm and then choosing an appropriate auxiliary measure for each time  $t = 1, \dots, T$ .

#### V. LOWER BOUNDS FOR ARBITRARY ALGORITHMS

We now apply the lemmas of the preceding section to the problem of deriving lower bounds on the information-based complexity of several problem classes. These bounds hold for *arbitrary* finite-step or infinite-step algorithms.

A. A General Information-Theoretic Lower Bound

Our first bound applies to *any* problem class. However, this generality comes at a price: the bound is nontrivial (i.e., tight) only in certain cases.

*Theorem 1:* Consider a problem class  $\mathcal{P} = (X, \mathcal{F}, \mathcal{O})$ , with any realization of the oracle  $\mathcal{O}$  in the form (8). Given any  $\eta > 0$ , define the *packing number*

$$N(\mathcal{F}, d, \eta) \triangleq \max \left\{ N \geq 1 : \exists f_0, \dots, f_{N-1} \in \mathcal{F} \right. \\ \left. \text{s.t. } d(f_i, f_j) \geq 2\eta, \forall i \neq j \right\}.$$

Then, for any  $r \geq 1$ , any  $\varepsilon$  such that  $N(\mathcal{F}, d, \varepsilon^{1/r}) > 4$  and any  $\delta \in (0, 1/2)$ , the following bounds hold:

$$K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \geq \frac{1}{C^*} \left[ (1 - \delta) \log N(\mathcal{F}, d, \varepsilon^{1/r}) - \log 2 \right] \quad (33)$$

$$K_{\mathcal{P}}^{(r)}(\varepsilon) \geq \frac{1}{C^*} \left[ \frac{2}{3} \log N(\mathcal{F}, d, (3\varepsilon)^{1/r}) - \log 2 \right] \quad (34)$$

with

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

where the supremum is over all random variables  $U$  taking values in  $\mathcal{U}_{X, \mathcal{F}} = \psi(\mathcal{F}, X)$  and the mutual information is between  $U$  and  $Y = \Phi(U, W)$ , cf. (8).

*Remark 4:* The number  $C^*$  is the *Shannon capacity* of the random transformation  $Y = \Phi(U, W)$  when its input is constrained to lie in the information space of the deterministic oracle  $\psi$ . When  $C^* = \infty$ , the bounds (33) and (34) simply say that  $K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \geq 0$  and  $K_{\mathcal{P}}^{(r)}(\varepsilon) \geq 0$ .

*Proof:* Let  $\mathcal{F}_{\varepsilon}^{(r)} = \{f_0, \dots, f_{N-1}\} \subset \mathcal{F}$ ,  $N = N(\mathcal{F}, d, \varepsilon^{1/r})$ , be a maximal packing set in  $\mathcal{F}$ . Given  $\delta \in (0, 1/2)$ , consider any  $T$  and any algorithm  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  such that  $\Pr(\text{err}_{\mathcal{A}}^r(T, f) \geq \varepsilon) \leq \delta$ . Then we can apply Lemma 1 to get

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

On the other hand, from Lemma 3 and the definition of  $C^*$

$$I(M; \hat{M}_T) \leq \sum_{t=1}^T I(U_t; Y_t) \leq TC^*.$$

Combining these two bounds, we get (33), while (34) follows after applying Proposition 1 with  $\delta = 1/3$ . ■

As an example, let  $X = B_{\infty}^n$  and  $\mathcal{F} = \mathcal{F}_{\text{Lip}}$  (cf. Example 1). Consider the case  $r = 1$ . Let  $\Lambda_{\varepsilon}$  be a maximal  $2\varepsilon$ -packing of  $X$  in  $\ell_2$ . A simple volume counting argument shows that

$$|\Lambda_{\varepsilon}| \geq v_n^{-1} (1/\varepsilon)^n$$

where  $v_n = \text{vol}(B_2^n)$ . Consider the subclass of  $\mathcal{F}_{\text{Lip}}$  consisting of all functions of the form  $f_{\theta}(x) = \|x - \theta\|$ ,  $\theta \in \Lambda_{\varepsilon}$ . Then for any two distinct functions  $f_{\theta}, f_{\theta'}$  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 noisy oracle with  $C^* < +\infty$

$$K_{\mathcal{P}}(\varepsilon) = \Omega \left( n \log \frac{1}{\varepsilon} \right).$$

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* [1], [2]. 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 [22], [23], [29].

B. First-Order Oracles With Gaussian Noise

If the oracle provides *noisy* first-order information, the above logarithmic lower bound can be tightened significantly. We now present lower bounds for two problem classes—convex Lipschitz functions (cf. Example 2) and strongly convex functions (cf. Example 3)—when the oracle supplies first-order information corrupted by additive white Gaussian noise. This is an oracle that, for a function  $f$  and a query point  $x$ , responds with

$$Y = (f(x) + W, g(x) + Z) \quad (35)$$

where, as before,  $g(x)$  is an arbitrary subgradient in  $\partial f(x)$  and  $W \sim \mathcal{N}(0, \sigma^2)$ ,  $Z \sim \mathcal{N}(0, \sigma^2 I_n)$  are mutually independent. We will refer to this oracle as the *first-order Gaussian (FOG) oracle*. We will also consider the *subgradient-only Gaussian (SOG) oracle*  $Y = g(x) + Z$ . For simplicity, we will assume that the algorithm knows the structure of the deterministic selector mapping  $(f, x) \mapsto g(x) \in \partial f(x)$ , since this knowledge can only help. We will see that the  $\varepsilon$ -complexities of these problem classes have *polynomial* dependence on  $1/\varepsilon$ , but differ in their dependence on the problem dimension  $n$ . Special cases of these results for linear functions for  $n = 1$  and  $r = 1$  can be found, for example, in [8]. The exponent of  $1/\varepsilon$  in the bounds will, generally, depend on the smoothness of functions in  $\mathcal{F}$ . We remark that noise variance for each coordinate of the subgradient is a constant  $\sigma^2$ , implying that the expected squared  $\ell_2$  norm of the noisy subgradient scales linearly with  $n$ . We shall keep this in mind when considering achievability of the lower bounds by specific algorithms. It is also straightforward to treat the case when  $Z \sim \mathcal{N}(0, (\sigma^2/n)I_n)$ , i.e., when  $\sigma^2$  bounds the *total* (as opposed to per-coordinate) noise variance. We begin by particularizing the IR bound (31) to the oracles under consideration (the proof is given in Appendix B):

*Lemma 5 (IR Bounds for Gaussian Oracles):*

$$I(M; \hat{M}) \leq \begin{cases} \frac{T}{2\sigma^2} \max_{i,j} \sup_{x \in X} \left\{ [f_i(x) - f_j(x)]^2 \right. \\ \left. + \|g_i(x) - g_j(x)\|^2 \right\} & (\text{FOG}) \\ \frac{T}{2\sigma^2} \max_{i,j} \sup_{x \in X} \|g_i(x) - g_j(x)\|^2 & (\text{SOG}). \end{cases} \quad (36)$$

We can now address the complexity of minimizing Lipschitz convex functions over a compact domain (cf. Example 2):

*Theorem 2:* Consider the problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}_{\text{Lip}}, \mathcal{O})$  with a Gaussian oracle, where  $\mathsf{X} \subset \mathbb{R}^n$  with  $n \geq 16$ . Define

$$s_{\mathsf{X}} \triangleq \max \{s \geq 0 : sB_{\infty}^n \subseteq \mathsf{X}\}.$$

Then for any  $r \geq 1$ ,  $\varepsilon \leq (s_{\mathsf{X}}\sqrt{n/8})^r$  and  $\delta \in (0, 1/2)$ , the following bounds hold:

$$K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \geq \begin{cases} \frac{((1-\delta)n-8)ns_{\mathsf{X}}^2 \log 2}{128(ns_{\mathsf{X}}^2+1)} \cdot \frac{\sigma^2}{\varepsilon^{2/r}} & (FOG) \\ \frac{((1-\delta)n-8)ns_{\mathsf{X}}^2 \log 2}{128} \cdot \frac{\sigma^2}{\varepsilon^{2/r}} & (SOG). \end{cases} \quad (37)$$

*Proof:* By the Varshamov-Gilbert bound (see, e.g., Lemma 2.9 in [7]), 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. In other words, there exists a subset  $\{\xi_0, \dots, \xi_{N-1}\}$  of the vertices of  $B_{\infty}^n$  with  $N > 2^{n/8}$ , such that

$$\|\xi_i - \xi_j\|^2 = 4 \sum_{k=1}^n \mathbf{1}_{\{\xi_{i,k} \neq \xi_{j,k}\}} \geq \frac{n}{2}, \quad \forall i \neq j. \quad (38)$$

Define the functions

$$f_i(x) \triangleq \frac{\varepsilon^{1/r}}{s_{\mathsf{X}}} \sqrt{\frac{8}{n}} \|x - s_{\mathsf{X}}\xi_i\|, \quad i = 0, 1, \dots, N-1.$$

Since  $\varepsilon \leq (s_{\mathsf{X}}\sqrt{n/8})^r$  and  $\{s_{\mathsf{X}}\xi_i\}_{i=0}^{N-1} \subset \mathsf{X}$ , these functions lie in  $\mathcal{F}_{\text{Lip}}$  and each  $f_i$  is uniquely minimized at  $x_i^* = s_{\mathsf{X}}\xi_i$  with  $f_i^* = 0$ . Moreover, upon defining

$$d(f_i, f_j) \triangleq \varepsilon^{1/r} \sqrt{\frac{8}{n}} \|\xi_i - \xi_j\|$$

it follows from (38) and the triangle inequality that  $d(f_i, f_j) \geq 2\varepsilon^{1/r}$  for all  $i \neq j$  and that this  $d$  satisfies the condition (18) on the set  $\{f_i\}_{i=0}^{N-1}$ . Hence, if there exist some  $T \in \mathbb{N}$  and an algorithm  $\mathcal{A} \in \mathcal{A}_T(\mathcal{P})$  that attains (20), we can apply Lemma 1 to obtain

$$\begin{aligned} I(M; \hat{M}_T) &\geq (1-\delta) \log N - \log 2 \\ &> \frac{(1-\delta)n-8}{8} \log 2. \end{aligned} \quad (39)$$

Next we will bound  $I(M; \hat{M}_T)$  from above. For any  $x \in \mathsf{X}$  and any pair  $i, j$ , we have

$$\begin{aligned} |f_i(x) - f_j(x)|^2 &\leq \frac{8\varepsilon^{2/r}}{n} \|\xi_i - \xi_j\|^2 \\ &= \frac{8\varepsilon^{2/r}}{n} \sum_{k=1}^n (\xi_{i,k} - \xi_{j,k})^2 \\ &= \frac{32\varepsilon^{2/r}}{n} \sum_{k=1}^n \mathbf{1}_{\{\xi_{i,k} \neq \xi_{j,k}\}} \\ &\leq 32\varepsilon^{2/r} \end{aligned}$$

and any subgradient of  $f_i$  at  $x$  has  $\ell_2$  norm not exceeding  $(\varepsilon^{1/r}/s_{\mathsf{X}})\sqrt{8/n}$ . Hence, applying Lemma 5, we get

$$I(M; \hat{M}_T) \leq \begin{cases} \frac{16T\varepsilon^{2/r}}{\sigma^2} \left(1 + \frac{1}{ns_{\mathsf{X}}^2}\right) & (FOG) \\ \frac{16T\varepsilon^{2/r}}{n\sigma^2 s_{\mathsf{X}}^2} & (SOG). \end{cases} \quad (40)$$

Combining (39) and (40) and rearranging, we get (37).  $\blacksquare$

Upper bounds on stochastic gradient descent—an algorithm which only uses the subgradient information—for  $r = 1$  are of the form  $O(G^2 D_{\mathsf{X}}^2 / \varepsilon^2)$ , where  $G^2$  is an upper bound on the expected squared norm of the noisy gradient [3]. 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 SOG oracle:

*Corollary 1:* Under the conditions of Theorem 2, we have

$$\begin{aligned} \mathsf{X} = \rho B_{\infty}^n &\implies K_{\mathcal{P}}^{(r)}(\varepsilon) = \begin{cases} \Omega\left(\frac{n^2\rho^2}{1+n\rho^2} \cdot \frac{\sigma^2}{\varepsilon^{2/r}}\right) & (FOG) \\ \Omega\left(n^2\rho^2 \cdot \frac{\sigma^2}{\varepsilon^{2/r}}\right) & (SOG) \end{cases} \\ \mathsf{X} = \rho B_2^n &\implies K_{\mathcal{P}}^{(r)}(\varepsilon) = \begin{cases} \Omega\left(\frac{n\rho^2}{1+\rho^2} \cdot \frac{\sigma^2}{\varepsilon^{2/r}}\right) & (FOG) \\ \Omega\left(n\rho^2 \cdot \frac{\sigma^2}{\varepsilon^{2/r}}\right) & (SOG). \end{cases} \end{aligned}$$

When the functions in  $\mathcal{F}$  are *strongly convex* (cf. Example 3), rather than convex Lipschitz, the complexity of optimization will decrease:

*Theorem 3:* Consider the problem class  $\mathcal{P} = (\mathsf{X}, \mathcal{F}_{\text{SC}}, \mathcal{O})$  with a first-order or gradient-only Gaussian oracle, where  $\mathsf{X} \subset \mathbb{R}^n$  with  $n \geq 16$ . Then for any  $r \geq 1$ ,  $\varepsilon \leq (ns_{\mathsf{X}}^2/16)^r$ ,  $\delta \in (0, 1/2)$ , the following bounds hold:

$$K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \geq \begin{cases} \frac{((1-\delta)n-8) \log 2}{256(D_{\mathsf{X}}^2+1)} \cdot \frac{\sigma^2}{\varepsilon^{1/r}}, & (FOG) \\ \frac{((1-\delta)n-8) \log 2}{256} \cdot \frac{\sigma^2}{\varepsilon^{1/r}}, & (SOG). \end{cases} \quad (41)$$

*Proof:* Given  $n$ , construct the set  $\{\xi_0, \dots, \xi_{N-1}\} \subset \{-1, +1\}^n$  as in the proof of Theorem 2 and define the functions

$$f_i(x) \triangleq \frac{1}{2} \left\| x - \sqrt{\frac{16\varepsilon^{1/r}}{n}} \xi_i \right\|^2, \quad i = 0, 1, \dots, N-1.$$

Since  $\varepsilon \leq (ns_{\mathsf{X}}^2/16)^r$

$$\left\{ x_i^* \triangleq \sqrt{\frac{16\varepsilon^{1/r}}{n}} \xi_i \right\}_{i=0}^{N-1} \subset \sqrt{\frac{16\varepsilon^{1/r}}{n}} B_{\infty}^n \subseteq s_{\mathsf{X}} B_{\infty}^n \subseteq \mathsf{X}.$$

Thus, the functions  $f_i$  lie in  $\mathcal{F}_{\text{SC}}$  and each  $f_i$  is uniquely minimized by  $x_i^*$  with  $f_i^* = 0$ . Moreover, upon defining

$$d(f_i, f_j) \triangleq \frac{4\varepsilon^{1/r}}{n} \|\xi_i - \xi_j\|^2$$

it follows from (38) and the triangle inequality that  $d(f_i, f_j) \geq 2\varepsilon^{1/r}$  for all  $i \neq j$  and that this  $d$  satisfies (18) on the set  $\{f_i\}_{i=0}^{N-1}$ . Hence, if there is some  $T \in \mathbb{N}$  and some  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  satisfying (20), we can apply Lemma 1 to obtain (39).

Now we will derive an upper bound on  $I(M; \hat{M}_T)$ . For any  $x \in X$  and any pair  $i, j$ , we have

$$\begin{aligned} &|f_i(x) - f_j(x)| \\ &= \frac{1}{2} \left( \|x - x_i^*\| + \|x - x_j^*\| \cdot \left| \|x - x_i^*\| - \|x - x_j^*\| \right| \right) \\ &\leq D_X \|x_i^* - x_j^*\| \end{aligned}$$

where the last step uses the definition of  $D_X$  and the triangle inequality. Thus

$$|f_i(x) - f_j(x)|^2 \leq D_X^2 \|x_i^* - x_j^*\|^2 \leq 64D_X^2 \varepsilon^{1/r}.$$

Also

$$\|\nabla f_i(x) - \nabla f_j(x)\|^2 = \|x_i^* - x_j^*\|^2 \leq 64\varepsilon^{1/r}.$$

Therefore, applying Lemma 5, we get

$$I(M; \hat{M}_T) \leq \begin{cases} \frac{32(D_X^2 + 1)T\varepsilon^{1/r}}{32T\varepsilon^{1/r} \sigma^2} & (FOG) \\ \frac{\sigma^2}{\sigma^2} & (SOG). \end{cases} \quad (42)$$

Combining (39) and (42), we get (41). ■

*Corollary 2:* Under the conditions of Theorem 3, we have

$$\begin{aligned} X = \rho B_\infty^n &\implies K_{\mathcal{P}}^{(r)}(\varepsilon) = \begin{cases} \Omega\left(\frac{n}{n\rho^2 + 1} \cdot \frac{\sigma^2}{\varepsilon^{1/r}}\right) & (FOG) \\ \Omega\left(n \cdot \frac{\sigma^2}{\varepsilon^{1/r}}\right) & (SOG) \end{cases} \\ X = \rho B_2^n &\implies K_{\mathcal{P}}^{(r)}(\varepsilon) = \begin{cases} \Omega\left(\frac{n}{\rho^2 + 1} \cdot \frac{\sigma^2}{\varepsilon^{1/r}}\right) & (FOG) \\ \Omega\left(n \cdot \frac{\sigma^2}{\varepsilon^{1/r}}\right) & (SOG). \end{cases} \end{aligned}$$

### C. Noisy Oracles Satisfying a Moment Bound

Our information-theoretic technique can be used to give a simpler derivation of the lower bounds obtained by Nemirovski and Yudin [1, Ch.5], for Lipschitz convex functions and noisy first-order oracles satisfying a certain moment constraint.

Let  $X = B_\infty^n$  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.,  $\mathbb{E}[V^0|f, x] = f(x)$ ,  $\mathbb{E}[V^1|f, x] \in \partial f(x)$ ,  $\forall f \in \mathcal{F}, x \in X$ .
- (C2) There exist constants  $\alpha > 1, L > 0$ , such that

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

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

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

*Theorem 4:* There exists an oracle  $\mathcal{O} \in \Pi(\alpha, L)$ , such that the corresponding problem class  $\mathcal{P} = (X, \mathcal{F}_{\text{Lip}}, \mathcal{O})$  satisfies

$$K_{\mathcal{P}}(\varepsilon, \delta) \geq \frac{\log 2 - h_2(\delta)}{c \log 2} \varepsilon^{-\alpha/(\alpha-1)} \quad (43)$$

for all  $\varepsilon \in (0, \min\{L/2^{1/\alpha}, 1\}), \delta \in (0, 1/2)$  with some  $c = c(\alpha, L) > 0$ .

*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 following noisy oracle defined by Nemirovski and Yudin [1, p.198]. Choose a constant  $c > 0$  such that  $c^{1-\alpha} < \min\{L^\alpha/2, 1\}$  and let  $p_{\varepsilon, \alpha} \triangleq c\varepsilon^{\alpha/(\alpha-1)}$ . On the set  $\mathcal{F} \setminus \{f_0, f_1\}$ , this oracle acts noiselessly, while on the set  $\{f_0, f_1\}$  it acts as follows: given  $f_i, i \in \{0, 1\}$  and  $x \in X$ , it outputs

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

It is an easy exercise to show that this oracle belongs to  $\Pi(\alpha, L)$ ; moreover, on the set  $\{f_0, f_1\}$  this oracle can be realized in the form (8) with  $\psi(f_i, x) = (f_i(x), \nabla f_i(x)), i \in \{0, 1\}$ .

Consider an algorithm  $\mathcal{A}$  that achieves (20) with  $r = 1$ . Let  $U_t = \psi(f_{M_t}, X_t)$ . Then  $I(U_t; Y_t | X^t, Y^{t-1}) \leq I(U_t; Y_t | X_t)$  because  $(X^t, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$  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

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

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

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

From Lemma 1, we have  $I(M; \hat{M}_T) \geq \log 2 - h_2(\delta)$ . Combining these bounds and rearranging, we get (43). ■

The statement of Theorem 4 should be interpreted in the following sense (cf. also [1]): given  $X$  and  $\mathcal{F}$  as above

$$\sup_{\mathcal{O} \in \Pi(\alpha, L)} K_{(X, \mathcal{F}_{\text{Lip}}, \mathcal{O})}(\varepsilon) = \Omega(\varepsilon^{-\alpha/(\alpha-1)}).$$

Thus, we have a lower bound which is robust relative to  $\Pi(\alpha, L)$ . However, this bound is sharp only for  $\alpha \in (1, 2]$  [1], [9], [10]: for  $\alpha \geq 2$ , the correct bound is  $\Omega(1/\varepsilon^2)$ . This can be easily seen from the results of the preceding section on the Gaussian first-order oracle.

### D. Statistical Estimation and Sequential Experimental Design

Finally, let us see how the problems of parametric statistical estimation (Example 4) and experimental design (Example 5) can be viewed through the lens of optimization.

Let us consider statistical estimation first. We will use the notation of Example 4. Typically, one considers the setting in which the statistician gets  $T$  i.i.d. samples  $Y_1, \dots, Y_T$  drawn from some  $P_\theta$ , where  $\theta$  is unknown. The quantity of interest is the *minimax risk*

$$R_T^* \triangleq \inf_{\hat{\theta}_T} \sup_{\theta \in X} \mathbb{E}_\theta \left( f_\theta(\hat{\theta}_T) - f_\theta(\theta) \right)^r$$

where the infimum is over all measurable estimators  $\hat{\theta}_T : \mathcal{Y}^T \rightarrow \mathcal{X}$  and  $f_\theta$  is an element of an appropriate class  $\mathcal{F}$  of loss functions, such as (10) or (12).

Now, the output of any such estimator can be viewed as the final result of some algorithm  $\mathcal{A} \in \mathcal{A}_T(\mathcal{P})$ . Thus, we simply follow our general recipe and isolate a finite subset  $\mathcal{F}' = \{f_{\theta_0}, \dots, f_{\theta_{N-1}}\}$  such that, with a suitably defined “distance”  $d(\cdot, \cdot)$  that satisfies (18), we have

$$d(f_{\theta_i}, f_{\theta_j}) \geq 2\varepsilon^{1/r}, \quad \forall i \neq j.$$

Suppose that we can arrange things in such a way that the cardinality of such an  $\mathcal{F}'$  is independent of  $\varepsilon$ , but may still depend on  $n$  and  $r$ :  $N = N(r, n)$  (this is possible in many cases, cf. the proofs of Theorems 2 and 3). Then we simply apply the IR bound to get

$$K_{\mathcal{P}}^{(r)}(\varepsilon) = \Omega\left(\frac{\log N(r, n)}{\max_{i,j} D(P_{\theta_i} \| P_{\theta_j})}\right). \quad (44)$$

Assuming, as is often the case, that

$$\max_{i,j} D(P_{\theta_i} \| P_{\theta_j}) = C(r, n)\varepsilon^{\gamma/r}$$

for some  $C(r, n) > 0, \gamma > 0$ , we can invert (44) and obtain the minimax lower bound

$$R_T^* = \Omega\left(\left(\frac{\log N(r, n)}{C(r, n)}\right)^{r/\gamma} T^{-r/\gamma}\right).$$

Similar considerations apply to sequential experimental design as well (Example 5), except there we have a *design strategy*  $\{\hat{X}_t : \mathcal{Y}^{t-1} \rightarrow \mathcal{X}\}_{t=1}^T$  and the estimator  $\hat{\theta}_T : \mathcal{Y}^T \rightarrow \mathcal{X}$ , where at time  $t = 1, \dots, T$  we choose the design point  $X_t = \hat{X}_t(Y^{t-1})$  and obtain a sample  $Y_t \sim P_{\theta, X_t}$  and then at time  $T+1$  we process all the samples to get the estimate  $X_{T+1} = \hat{\theta}_T(Y^T)$ . The minimax risk is then

$$R_T^* = \inf_{\hat{X}^T, \hat{\theta}_T} \sup_{\theta \in \mathcal{X}} \mathbb{E} \left( f_\theta(\hat{\theta}_T) - f_\theta(\theta) \right)^r.$$

The connection to optimization is even more apparent than in the estimation setting and the IR bound technique yields

$$K_{\mathcal{P}}^{(r)}(\varepsilon) = \Omega\left(\frac{\log N(r, n)}{\max_{i,j} \sup_{x \in \mathcal{X}} D(P_{\theta_i, x} \| P_{\theta_j, x})}\right).$$

Just as before, this bound can be inverted to get a lower bound on the minimax risk. Note, however, that what we have is a lower bound on the number of the design points needed to guarantee that the minimax risk is below  $\varepsilon$ .

## VI. LOWER BOUNDS FOR ANYTIME ALGORITHMS

Conceptually, our use of the IR bounds is akin to the methods used in statistics to obtain minimax lower bounds through local

entropy estimates and a device like the Assouad lemma (cf. [5] and references therein). In both cases, in order to get the right rates it is essential to arrange things so that the size of the “packing” set  $\mathcal{F}'$  is independent of  $\varepsilon$ . However, one drawback of the IR bounds is that they do not take into account the *dynamics* of the algorithm, pertaining to the manner in which its expected error evolves with time. Instead, we must use uniform, worst-case bounds on the uncertainty remaining after each successive oracle call. However, it could be argued on practical grounds that the only optimization algorithms that are of any value are the ones whose performance gradually and monotonically improves with time, as more and more queries are issued—that is, anytime algorithms. In this section, we show that the LF bounds can be used to track the evolution of the mutual information over time. As a consequence, we will be able to derive upper bounds on the anytime exponent for certain problem classes.

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. We will also show that, in some cases, insisting on the anytime property may mean that the algorithm will take *longer* to get to the point after which its expected error drops below some desired level. This seemingly strange conclusion reflects the fact that, without placing any restrictions on the algorithm’s trajectory, we are allowing “bizarre” (and not very practical) strategies that wander around the problem domain for a while, gathering information without much regard to how close they are to a minimizer and then—boom!—produce an excellent solution. With such algorithms, it is certainly no surprise that they may hit upon a good solution more quickly than an “honest” anytime algorithm that must proceed incrementally and inexorably towards a minimizer.

In contrast to the local technique based on the IR bounds, our use of the LF bounds in this section can be thought of as a *global technique* [6], [28], [30]. The main idea is as follows. Suppose we have an anytime algorithm whose worst-case expected errors decay at some rate  $\varepsilon_t \rightarrow 0$ . Then, for each  $T$ , we consider an  $\varepsilon_T$ -packing of the problem domain (with respect to a suitable metric, typically just the usual Euclidean norm  $\|\cdot\|$ ), which will induce a packing of the function class  $\mathcal{F}$ . This packing will be of size  $\Omega(\varepsilon_T^{-n})$ . Since the algorithm does well on every function in this large packing set. Thus, if the objective function is drawn uniformly at random from this set, then combining the lower bound of Lemma 1 with the LF upper bound will result in a relation of the form

$$n \log \left( \frac{1}{\varepsilon_T} \right) \preceq \sum_{t=1}^T \varepsilon_t^\gamma$$

where  $\gamma > 0$  depends on the smoothness of the functions in  $\mathcal{F}$ . This relation must hold for all but finitely many values of  $T$ . The

optimal rate is then derived by balancing the entropy  $n \log \varepsilon_T^{-1}$  and the sum of diminishing mutual information terms.

### A. Strongly Convex Functions

We first consider the case of strongly convex functions with Lipschitz-continuous gradients (an often made assumption [3], [27]). Given  $X$ , let  $\mathcal{F}_{\kappa,L}$  denote the set of all functions  $f : X \rightarrow \mathbb{R}$  that satisfy the following conditions:

- Each  $f \in \mathcal{F}_{\kappa,L}$  is  $\kappa$ -strongly convex (cf. Example 3).
- For each  $f$ , the mapping  $x \mapsto \nabla f(x)$  is  $L$ -Lipschitz

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

Consider the noisy first-order oracle  $Y = (f(x) + W, \nabla f(x) + Z)$  as in (35). 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  $\nabla f_i(x_i^*) = 0$ , where  $x_i^*$  is the (unique) minimizer of  $f_i$  on  $X$ . Let  $M$  denote the uniform random variable on  $\{0, \dots, N-1\}$ . Then we have the following (the proof is in Appendix B):

*Lemma 6:* At every time  $t = 1, 2, \dots$ , any algorithm  $\mathcal{A} \in \mathfrak{A}_\infty(\mathcal{P})$  satisfies

$$I(M; Y_t | X^t, Y^{t-1}) \leq \frac{(L/\kappa)^2 (D_X^2 + 1)}{\sigma^2} \max_i \mathbb{E} \text{err}_{\mathcal{A}}(t, f_i). \quad (45)$$

Moreover, if  $\mathcal{A}$  is 1-anytime (cf. Definition 6), then

$$I(M; Y_t | X^t, Y^{t-1}) \leq \frac{(L/\kappa)^2}{\sigma^2} (D_X^2 + 1) \cdot \overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F}_{\kappa,L}) \xrightarrow{t \rightarrow \infty} 0. \quad (46)$$

Thus, 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. In other words, anytime algorithms for strongly convex programming obey a *law of diminishing returns*. 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 6 in conjunction with the information bounds of Section V, we establish the following upper bound on the anytime exponent of strongly convex programming problems:

*Theorem 5:* Consider the problem class  $\mathcal{P} = (X, \mathcal{F}, \mathcal{O})$  with  $X = B_\infty^n$ ,  $\mathcal{F} = \mathcal{F}_{\kappa,L}$  with  $\kappa = 1$  and  $L \geq 1$  and the Gaussian first-order oracle (35). Then  $\gamma_{\mathcal{P}} \leq 1$ . In other words, on this problem class,  $O(t^{-1})$  is the optimal error decay rate for all 1-anytime algorithms whose errors decay polynomially with  $t$ .

*Proof:* Consider any algorithm  $\mathcal{A} \in \mathfrak{A}_\infty(\mathcal{P})$  whose worst-case errors  $\varepsilon_t \triangleq \overline{\text{err}}_{\mathcal{A}}(t, \mathcal{F})$  satisfy

$$\limsup_{t \rightarrow \infty} t^\gamma \varepsilon_t < \infty$$

for some  $\gamma \geq 0$ . In other words,  $\varepsilon_t = O(t^{-\gamma})$ . By Markov's inequality, we have, for every  $T$

$$\sup_{f \in \mathcal{F}} \Pr \left( \text{err}_{\mathcal{A}}(T, f) \geq 3\varepsilon_T \right) \leq \frac{\sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}(T, f)}{3\varepsilon_T} \leq \frac{1}{3}.$$

Let us fix some  $T$ , let  $\Lambda_T = \{\theta_0, \dots, \theta_{N-1}\}$  denote a maximal  $2\sqrt{3\varepsilon_T}$ -packing set in  $X$  (w.r.t.  $\|\cdot\|$ ) and define

$$f_i(x) \triangleq \frac{1}{2} \|x - \theta_i\|^2, \quad i = 0, \dots, N-1.$$

By volume counting,  $N \geq v_n^{-1} (1/3\varepsilon_T)^{n/2}$ . We also have  $d(f_i, f_j) = \frac{1}{2} \|\theta_i - \theta_j\|^2 \geq 6\varepsilon_T$ . By Lemma 1

$$I(M; \hat{M}_T) \geq \frac{n}{3} \log \left( \frac{1}{\varepsilon_T} \right) + c_n \quad (47)$$

where  $c_n = \frac{1}{3} \log \left( \frac{1}{3^n 8 v_n^2} \right)$ . On the other hand, applying Lemma 6, we obtain

$$I(M; \hat{M}_T) \leq \frac{n+1}{\sigma^2} \sum_{t=1}^T \varepsilon_t. \quad (48)$$

Combining (47) and (48), we see that the sequence  $\{\varepsilon_t\}$  must satisfy the following inequalities:

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

where  $c'_n = \sigma^2 c_n / (n+1)$ . From Lemma C.1, we, therefore, conclude that there exists an infinite subsequence of times  $1 \leq t_1 < t_2 < \dots$ , such that  $\varepsilon_{t_j} \geq c t_j^{-1}$  for some  $c > 0$ . Since  $\varepsilon_t = O(t^{-\gamma})$  by hypothesis, we must have  $\gamma \leq 1$ . ■

The bound  $\Omega(t^{-1})$  is tight and can be achieved by stochastic gradient descent [3]. Note that the methods of Section V can be used to explicitly identify the dependence of the lower bound on the problem dimension  $n$ .

### B. Comparison of IR and LF Bounds

A natural question is whether LF bounds for anytime algorithms provide tighter lower bounds when compared to IR bounds without the anytime assumption. This is indeed the case, as we demonstrate through an example. Interestingly, the difference is not present for linear and quadratic functions, but appears for higher degree polynomials. Consider a simple set-up with  $X = [-1, 1]$

$$\mathcal{F} = \mathcal{F}_m = \{f_\theta(x) = (x - \theta)^m : \theta \in X\}$$

for some even  $m \in \mathbb{N}$  and the noisy first-order oracle

$$Y = (f(x) + W, f'(x) + Z)$$

where  $W, Z$  are mutually independent  $\mathcal{N}(0, \sigma^2)$  random variables.

*Theorem 6:* On this problem class, for any  $\varepsilon \leq 1$  and any  $\delta \in (0, 1/2)$  we have

$$K_{\mathcal{P}}^{(2)}(\varepsilon) = \Omega \left( \frac{\sigma^2}{4^m m^4 \varepsilon^{1/m}} \right). \quad (49)$$

*Proof:* Let us define

$$d(f_\theta, f_{\theta'}) \triangleq 2^{1-m} (\theta - \theta')^m, \quad \forall \theta, \theta' \in X. \quad (50)$$

By convexity of  $x \mapsto x^m$  (recall that  $m$  is even), we have for any  $x, \theta, \theta' \in X$

$$\begin{aligned} (\theta - \theta')^m &= 2^m \left( \frac{\theta - x}{2} + \frac{x - \theta'}{2} \right)^m \\ &\leq 2^{m-1} [(x - \theta)^m + (x - \theta')^m]. \end{aligned}$$

Hence, the  $d(\cdot, \cdot)$  defined in (50) satisfies (18). In particular, if we fix the functions

$$f_0(x) = (x - \varepsilon^{1/2m})^m, \text{ and } f_1(x) = (x + \varepsilon^{1/2m})^m$$

then  $d(f_0, f_1) = 2\varepsilon^{1/2}$ . Let  $M$  have uniform distribution on  $\{0, 1\}$ . Consider now any algorithm  $\mathcal{A} \in \mathfrak{A}_T(\mathcal{P})$  that attains  $\text{err}_{\mathcal{A}}^2(T, f_\theta) < \varepsilon$  with probability at least  $1 - \delta$  for every  $\theta \in X$ . Applying Lemma 1, we obtain

$$I(M; \hat{M}_T) \geq \log 2 - h_2(\delta). \quad (51)$$

On the other hand, from Lemma 5, we have

$$I(M; \hat{M}_T) \leq \frac{1}{2\sigma^2} \sup_{x \in X} \{ [f_0(x) - f_1(x)]^2 + [f'_0(x) - f'_1(x)]^2 \}.$$

From convexity

$$\begin{aligned} [f_0(x) - f_1(x)]^2 &= [(x - \varepsilon^{1/2m})^m - (x + \varepsilon^{1/2m})^m]^2 \\ &\leq m^2 4^m \varepsilon^{1/m}. \end{aligned}$$

Likewise, applying the mean-value theorem to the function  $x \mapsto x^{m-1}$ , we get

$$\begin{aligned} [f'_0(x) - f'_1(x)]^2 &= m^2 \left[ (x - \varepsilon^{1/2m})^{m-1} - (x + \varepsilon^{1/2m})^{m-1} \right]^2 \\ &\leq [m(m-1)]^2 2^{2(m-1)} \varepsilon^{1/m}. \end{aligned}$$

Thus, we obtain the upper bound

$$I(M; \hat{M}_T) = O\left(\frac{T 4^m m^4 \varepsilon^{1/m}}{\sigma^2}\right). \quad (52)$$

Combining (51) and (52), we obtain (49).  $\blacksquare$

*Theorem 7:* Consider the same problem class. Then  $\gamma_{\mathcal{P}}^{(2)} \leq \frac{m}{m-1}$ .

*Proof:* As before, consider any algorithm  $\mathcal{A} \in \mathfrak{A}_\infty(\mathcal{P})$  whose worst-case errors  $\varepsilon_t \triangleq \overline{\text{err}}_{\mathcal{A}}^{(2)}(t, \mathcal{F})$  satisfy

$$\limsup_{t \rightarrow \infty} t^\gamma \varepsilon_t < \infty$$

for some  $\gamma \geq 0$ . Applying the same argument based on Markov's inequality as in the proof of Theorem 5, we see that, for any  $T$ ,  $\mathcal{A}$  attains  $\text{err}_{\mathcal{A}}^2(T, f_\theta) < 3\varepsilon_T$  with probability at least  $2/3$  on every  $f_\theta \in \mathcal{F}$ . Given  $T$ , let  $\Lambda_T = \{\theta_0, \dots, \theta_{N-1}\}$  denote the largest finite subset of  $X = [-1, 1]$ , such that

$$|\theta_i - \theta_j| \geq 2(3\varepsilon_T)^{1/2m}, \quad \forall i \neq j.$$

A simple counting argument shows that  $N \geq \left(\frac{1}{3\varepsilon_T}\right)^{1/2m}$ . Moreover, the functions  $f_i(x) \triangleq (x - \theta_i)^m, i = 0, \dots, N-1$ ,

satisfy  $d(f_i, f_j) \geq 2\sqrt{3\varepsilon_T}$  for  $i \neq j$ , where  $d(\cdot, \cdot)$  is defined in (50).

By Lemma 1

$$I(M; \hat{M}_T) \geq \frac{1}{3m} \log\left(\frac{1}{\varepsilon_T}\right) + c_m$$

where  $c_m = \frac{1}{3} \log\left(\frac{1}{31^{1/m} 8}\right)$ . We will now combine this lower bound with an appropriate LF bound. Let  $Q^*$  denote the bivariate normal distribution  $\mathcal{N}(0, \sigma^2 I_2)$ . Then, for every  $i = 0, \dots, N-1$  we have

$$Y^{(i)} = (f_i(\theta_i) + W, f'_i(\theta_i) + Z) = (W, Z) \sim Q^*.$$

Hence, applying the LF bound with  $\mathbb{Q}_{Y_i}^* = Q^*$  as in the proof of Lemma 6, we obtain

$$\begin{aligned} I(M; \hat{M}_T) &\leq \frac{1}{2\sigma^2} \sum_{t=1}^T \mathbb{E} \{ f_M^2(X_t) + [f'_M(X_t)]^2 \} \\ &= \frac{1}{2\sigma^2} \sum_{t=1}^T \mathbb{E} \left\{ (X_t - \theta_M)^{2m} + m^2 (X_t - \theta_M)^{2(m-1)} \right\} \\ &= \frac{1}{2\sigma^2} \sum_{t=1}^T \mathbb{E} \left\{ [f_M(X_t) - f_M^*]^2 + m^2 [f_M(X_t) - f_M^*]^{\frac{2(m-1)}{m}} \right\} \\ &\leq \frac{1}{2\sigma^2} \sum_{t=1}^T \left\{ \mathbb{E} [f_M(X_t) - f_M^*]^2 + m^2 (\mathbb{E} [f_M(X_t) - f_M^*]^2)^{\frac{m-1}{m}} \right\} \\ &\leq \frac{m^2 + 1}{2\sigma^2} \sum_{t=1}^T (\mathbb{E} [f_M(X_t) - f_M^*]^2)^{\frac{m-1}{m}} \\ &= \frac{m^2 + 1}{2\sigma^2} \sum_{t=1}^T (\mathbb{E} \text{err}_{\mathcal{A}}^2(t, f_M))^{\frac{m-1}{m}} \\ &\leq \frac{m^2 + 1}{2\sigma^2} \sum_{t=1}^T \varepsilon_t^{\frac{m-1}{m}} \end{aligned} \quad (53)$$

where in (53) we have used the concavity of the function  $u \mapsto u^{(m-1)/m}$ . Therefore, we conclude that the sequence  $\{\varepsilon_t\}$  must satisfy

$$\frac{2\sigma^2}{3(m^3 + m^2)} \log\left(\frac{1}{\varepsilon_T}\right) + \frac{2\sigma^2 c_m}{m^2 + 1} \leq \sum_{t=1}^T \varepsilon_t^{\frac{m-1}{m}}$$

for all sufficiently large  $T$ . Applying Lemma C.1, we conclude that there exists an infinite subsequence of times  $t_1 < t_2 < \dots$ , such that  $\varepsilon_{t_j} \geq c t_j^{-m/(m-1)}$  for some constant  $c > 0$ . Since  $\varepsilon_t = O(t^{-\gamma})$  by hypothesis, we must have  $\gamma \leq \frac{m}{m-1}$ .  $\blacksquare$

For  $m = 2$ , the two results indicate the same order of complexity,  $T \succeq \varepsilon^{-1/2}$ ; however, for  $m = 4$  and larger, the bounds differ, giving  $T \succeq \varepsilon^{-1/m}$  for arbitrary algorithms and  $T \succeq \varepsilon^{(1-m)/m}$  for anytime algorithms, which is larger. We conclude that, in general, the LF bounding technique leads to tighter bounds for optimization algorithms which actually converge monotonically to the optimal solution.

C. 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 [22]. In general, active learning is more difficult than (convex) optimization. However, for the case below, we can apply the tools developed in this paper. The reason for including this example is twofold: first, to show that problems beyond convex optimization can be attacked with our information-theoretic method and second to exhibit a problem with a noise model more complicated than those encountered so far in the paper.

The active learning problem is stated as follows. We have a pair  $(X, Z)$  of jointly distributed random variables  $X \in \mathbb{X} = [0, 1]$  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  $G^*(x) \triangleq \mathbf{1}_{\{\eta(x) \geq 1/2\}}$  for this problem [31] 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

$$c|x - \theta|^{\kappa-1} \leq |\eta(x) - 1/2| \leq C|x - \theta|^{\kappa-1} \quad (54)$$

where the first inequality (known as the Tsybakov noise condition [32]) 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  $X_t \in \mathbb{X}$ , 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  $P_{Z|X}(\cdot|X_t)$ . 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  $G^*$

$$R(G_{X_t}) - R(G^*) = \int_{[X_t, 1] \Delta [0, 1]} |2\eta(x) - 1| dx \quad (55)$$

where  $\Delta$  denotes symmetric difference between sets. (The risk of a classifier  $G : \mathbb{X} \rightarrow \{0, 1\}$  is defined as  $R(G) \triangleq \Pr(G(X) \neq Z)$  and the Bayes risk is  $R(G^*) \triangleq \inf_G R(G)$ [31].)

Castro and Nowak [22] have shown that 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 6. Let  $\mathbb{X}$  and  $\mathcal{F}$  be as described there and associate to each  $P_{Z|X} \in \Pi(\kappa, c, C)$  a noisy oracle with  $Y = \{-1, +1\}$  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 8:* 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 excess risk is  $\Omega(2^{-6C^2t})$ .<sup>3</sup>

*Proof:* For each  $\theta \in [0, 1]$ , find some  $P_{Z|X}^\theta \in \Pi(\kappa, c, C)$ , such that the inequalities in (54) hold for all values of  $x \in \mathbb{X}$ . 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 (55) and (54) we get

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

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

$$\mathbb{E}[R(G_{X_t}) - R(G_\theta)] \geq (2c/\kappa) \mathbb{E}|X_t - \theta|^\kappa. \quad (56)$$

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

$$\mathbb{E}|X_t - \theta|^\kappa \leq \kappa r_t / 2c, \quad t = 1, 2, \dots \quad (57)$$

Let  $\varepsilon_t \triangleq (3\kappa r_t / 2c)^{1/\kappa}$ . Then using (57) and Markov's inequality, we see that for this strategy we must have

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

In other words, this active learning strategy gives rise to an optimization algorithm  $\mathcal{A}$  for the problem class  $\mathcal{P} = (\mathbb{X}, \mathcal{F}, \mathcal{O})$ , 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  $\Pr(\text{err}_{\mathcal{A}}(t, f) \geq \varepsilon_t) \leq 1/3$  for all  $t \geq T_0$ .

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

$$I(M; \hat{M}_T) \geq \frac{2}{3} \log \left( \frac{1}{\varepsilon_T} \right) - \frac{5}{3} \log 2. \quad (58)$$

<sup>3</sup>The exponent in this lower bound is not tight, since there exists a specific strategy that achieves the excess risk of  $O(2^{-c^2 t \log e/2})$  when  $\kappa = 1$  [22].



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 the LF bound with  $Q^*$  the uniform distribution on  $\{-1, +1\}$ . Then

$$\begin{aligned} I(M, X_t; Y_t) &\leq D(\mathbb{P}_{Y_t|M, X_t} \| \mathbb{Q}_{Y_t}^* | \mathbb{P}_{M, X_t}) \\ &\leq 4\mathbb{E}_{M, X_t} \left\{ (\mathbb{P}(Y_t = 1 | M, X_t) - 1/2)^2 \right\} \\ &= 4\mathbb{E}_{M, X_t} \left\{ |\eta_M(X_t) - 1/2|^2 \right\} \\ &\leq 4C^2 \mathbb{E}_{M, X_t} |X_t - \theta_M|^{2(\kappa-1)} \end{aligned} \quad (59)$$

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 (54). Suppose first that  $\kappa > 1$ . 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 (57) and Lemma 2, we can bound the mutual information  $I(M; \hat{M}_T)$  as

$$I(M; \hat{M}_T) \leq 4C^2 \sum_{t=1}^T \left( \frac{\kappa T t}{2C} \right)^{\frac{2(\kappa-1)}{\kappa}} = \frac{4C^2}{3^{\frac{2\kappa-2}{\kappa}}} \sum_{t=1}^T \varepsilon_t^{2(\kappa-1)}. \quad (60)$$

Combining (58) and (60), 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  $T \geq T_0$ . Lemma C.1 then states 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 (59) we have  $I(M, X_t; Y_t) \leq 4C^2$  for all  $t$ . This, together with (58), 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  $\varepsilon_T = \Omega(2^{-6C^2T})$  and  $r_T = \Omega(2^{-6C^2T})$ . ■

## VII. CONCLUDING REMARKS

Sequential optimization algorithms operating in the presence of uncertainty must be able to accumulate information in order to reduce uncertainty. As we have shown in this paper, there

are fundamental limitations on the rate at which this uncertainty can be reduced, depending on the richness of the class of objective functions faced by the algorithm, the noisiness and the structure of the oracle that supplies information to the algorithm and the manner in which the algorithm may approach the optimum (i.e., monotonically or not). In order to derive these fundamental limitations, we have developed a comprehensive information-theoretic machinery that makes use of the fact (which we have proved) that the problem of sequential optimization is, in a certain sense, at least as hard as hypothesis testing with feedback (or with controlled observations). This observation then leads to quantitative estimates that relate the minimum number of oracle queries needed to achieve a given level of accuracy to the overall reduction of uncertainty about the objective function being optimized. The latter is measured by the *mutual information* between the random choice of the objective and the history of algorithm's queries and oracle's responses. Carefully taking into account all the Markovian structures that are imposed by the sequential and the adaptive nature of the algorithm, we can obtain different upper bounds on this mutual information.

Using this machinery, we have derived tight lower bounds in several settings in optimization, both for arbitrary and for anytime optimization algorithms (in some cases improving upon existing results) and beyond, e.g., for experimental design and active learning. One promising direction for future work is to consider algorithms with *query costs*, i.e., when issuing each query incurs a cost that may depend on the query and the goal is to balance the total cost of querying with the final optimization error. Recent work by Naghshvar and Javidi [33] considers a hypothesis testing problem of this kind by relating it to optimal stopping for a Markov decision process and the techniques developed in that work may be useful for deriving information-theoretic lower bounds for optimization problems with query costs.

## APPENDIX A

### FINITE-STEP VERSUS STRONG INFINITE-STEP ALGORITHMS

As we pointed out in Section II, our definition of an infinite-step algorithm is somewhat restrictive, as it allows only the algorithms that use their most recently computed candidate minimizer as the next query. The following definition removes this restriction:

*Definition A.1:* A strong infinite-step algorithm for a problem class  $\mathcal{P} = (\mathcal{X}, \mathcal{F}, \mathcal{O})$  is a sequence of mappings  $\tilde{\mathcal{A}} = \{\tilde{\mathcal{A}}_t : \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \rightarrow \mathcal{X} \times \mathcal{X}\}_{t=1}^\infty$ . The set of all infinite-step algorithms for  $\mathcal{P}$  will be denoted by  $\tilde{\mathcal{A}}_\infty(\mathcal{P})$ .

The interaction of any  $\tilde{\mathcal{A}} \in \tilde{\mathcal{A}}_\infty(\mathcal{P})$  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  $\tilde{\mathcal{A}}$ .
- 2) At each time  $t = 1, 2, \dots$ :
  - $\tilde{\mathcal{A}}$  computes

$$(X_t, \hat{X}_t) = \tilde{\mathcal{A}}_t(X^{t-1}, Y^{t-1})$$

where  $X_\tau$  and  $\hat{X}_\tau$  are, respectively, the query and the candidate minimizer at time  $\tau$ .

- $\mathcal{O}$  responds with a random element  $Y_t \in \mathcal{Y}$  according to  $P(dY_t | f, X_t)$ .

In other words, both  $\hat{X}_t$ , the candidate minimizer at time  $t$  and  $X_t$ , the query at time  $t$ , are computed on the basis of all currently available data, i.e.,  $(X^{t-1}, Y^{t-1})$ , yet the algorithm has more freedom, since at time  $t + 1$  it can query the oracle with an arbitrary point, rather than just  $\hat{X}_t$ . The error of  $\tilde{\mathcal{A}}$  on  $f \in \mathcal{F}$  at time  $t$  is given by

$$\text{err}_{\tilde{\mathcal{A}}}(t, f) = f(\hat{X}_t) - \inf_{x \in \mathcal{X}} f(x) = f(\hat{X}_t) - f^*.$$

*Definition A.2:* Fix a problem class  $\mathcal{P} = (\mathcal{X}, \mathcal{F}, \mathcal{O})$ . For any  $r \geq 1$ ,  $\varepsilon > 0$  and  $\delta \in (0, 1)$ , we define the  $r$ th-order infinite-step  $(\varepsilon, \delta)$ -complexity and the  $\varepsilon$ -complexity of  $\mathcal{P}$ , respectively, as

$$\begin{aligned} K_{\mathcal{P}}^{(r), \infty}(\varepsilon, \delta) &\triangleq \inf \left\{ T \geq 1 : \exists \tilde{\mathcal{A}} \in \tilde{\mathcal{A}}_{\infty}(\mathcal{P}) \right. \\ &\quad \left. \text{s.t. } \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\tilde{\mathcal{A}}}^r(t, f) \geq \varepsilon) \leq \delta, \forall t > T \right\} \\ K_{\mathcal{P}}^{(r), \infty}(\varepsilon) &\triangleq \inf \left\{ T \geq 1 : \exists \tilde{\mathcal{A}} \in \tilde{\mathcal{A}}_{\infty}(\mathcal{P}) \right. \\ &\quad \left. \text{s.t. } \sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\tilde{\mathcal{A}}}^r(t, f) < \varepsilon, \forall t > T \right\}. \end{aligned}$$

It turns out that these notions of complexity are equivalent to the ones introduced earlier:

*Proposition A.1:* For any problem class  $\mathcal{P}$  and all  $r \geq 1$ ,  $\varepsilon > 0$ ,  $\delta \in (0, 1)$ , we

$$K_{\mathcal{P}}^{(r), \infty}(\varepsilon, \delta) = K_{\mathcal{P}}^{(r)}(\varepsilon, \delta) \quad (\text{A.1})$$

$$K_{\mathcal{P}}^{(r), \infty}(\varepsilon) = K_{\mathcal{P}}^{(r)}(\varepsilon). \quad (\text{A.2})$$

*Proof:* We only prove (A.1), since the proof of (A.2) is similar. Likewise, we will only consider the  $r = 1$  case.

First we prove that  $K_{\mathcal{P}}^{\infty}(\varepsilon, \delta) \leq K_{\mathcal{P}}(\varepsilon, \delta)$ . We can assume that  $K_{\mathcal{P}}(\varepsilon, \delta) < \infty$ , for otherwise the inequality holds *a fortiori*. Given  $\varepsilon$  and  $\delta$ , consider any  $T$  for which there exists some  $T$ -step algorithm  $\mathcal{A} \in \mathcal{A}_T(\mathcal{P})$ , such that

$$\sup_{f \in \mathcal{F}} \Pr(\text{err}_{\mathcal{A}}(T, f) \geq \varepsilon) \leq \delta.$$

Given  $\mathcal{A}$ , we can construct a strong infinite-step algorithm  $\tilde{\mathcal{A}} \in \tilde{\mathcal{A}}_{\infty}(\mathcal{P})$  as follows. Choose an arbitrary  $T$ -tuple  $(\hat{x}_1, \dots, \hat{x}_T) \in \mathcal{X}^T$  and let

$$\begin{aligned} \tilde{\mathcal{A}}_t(x^{t-1}, y^{t-1}) &= \begin{cases} (\mathcal{A}_t(x^{t-1}, y^{t-1}), \hat{x}_t), & t = 1, \dots, T \\ (\mathcal{A}_{T+1}(x^T, y^T), \mathcal{A}_{T+1}(x^T, y^T)), & t > T. \end{cases} \end{aligned}$$

Then it's clear that for any  $t > T$

$$\begin{aligned} \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\tilde{\mathcal{A}}}(t, f) \geq \varepsilon) &= \sup_{f \in \mathcal{F}} \Pr(f(\hat{X}_t) - f^* \geq \varepsilon) \\ &= \sup_{f \in \mathcal{F}} \Pr(f(\mathcal{A}_{T+1}(X^T, Y^T)) - f^* \geq \varepsilon) \\ &= \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\mathcal{A}}(T, f) \geq \varepsilon) \\ &\leq \delta. \end{aligned}$$

Hence,  $K_{\mathcal{P}}^{\infty}(\varepsilon, \delta) \leq K_{\mathcal{P}}(\varepsilon, \delta)$ .

Next, we prove  $K_{\mathcal{P}}(\varepsilon, \delta) \leq K_{\mathcal{P}}^{\infty}(\varepsilon, \delta)$ . Again, we can assume that  $K_{\mathcal{P}}^{\infty}(\varepsilon, \delta) < \infty$ . Consider an algorithm  $\tilde{\mathcal{A}} \in \tilde{\mathcal{A}}_{\infty}(\mathcal{P})$ , such that

$$\sup_{t > T} \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\tilde{\mathcal{A}}}(t, f) \geq \varepsilon) \leq \delta \quad (\text{A.3})$$

for some  $T$ . Let  $\Pi_1$  and  $\Pi_2$  denote the two coordinate projection mappings from  $\mathcal{X} \times \mathcal{X}$  onto  $\mathcal{X}$ , i.e.,  $\Pi_1(x, x') = x$  and  $\Pi_2(x, x') = x'$  and define  $\mathcal{A} \in \mathcal{A}_T(\mathcal{P})$  by setting

$$\mathcal{A}_t = \begin{cases} \Pi_1 \circ \tilde{\mathcal{A}}_t, & t = 1, \dots, T \\ \Pi_2 \circ \tilde{\mathcal{A}}_{T+1}, & t = T + 1. \end{cases}$$

Then from (A.3)

$$\begin{aligned} \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\mathcal{A}}(T, f) \geq \varepsilon) &= \sup_{f \in \mathcal{F}} \Pr(f(X_{T+1}) - f^* \geq \varepsilon) \\ &= \sup_{f \in \mathcal{F}} \Pr(f(\mathcal{A}_{T+1}(X^T, Y^T)) - f^* \geq \varepsilon) \\ &= \sup_{f \in \mathcal{F}} \Pr(\text{err}_{\tilde{\mathcal{A}}}(T + 1, f) \geq \varepsilon) \\ &\leq \delta \end{aligned}$$

which implies  $K_{\mathcal{P}}(\varepsilon, \delta) \leq K_{\mathcal{P}}^{\infty}(\varepsilon, \delta)$ . ■

## APPENDIX B MISCELLANEOUS PROOFS

*A) Proof of Proposition 1:* Given  $\varepsilon$  and  $\mathcal{P}$ , consider any  $T$  for which there exists some algorithm  $\mathcal{A} \in \mathcal{A}_T(\mathcal{P})$  that satisfies  $\sup_{f \in \mathcal{F}} \mathbb{E} \text{err}_{\mathcal{A}}^r(T, f) \leq \varepsilon$ . Then Markov's inequality gives

$$\Pr(\text{err}_{\mathcal{A}}^r(T, f) \geq \varepsilon/\delta) \leq \frac{\mathbb{E} \text{err}_{\mathcal{A}}^r(T, f)}{\varepsilon/\delta} \leq \delta, \quad \forall f \in \mathcal{F}.$$

Hence,  $T \geq K_{\mathcal{P}}^{(r)}(\varepsilon/\delta, \delta)$ . Taking the infimum over all such  $T$ , we arrive at the proof.

*B) Proof of Lemma 3:* First, we modify the construction of the probability space  $(\Omega, \mathcal{B}, \mathbb{P})$  in Section IV by introducing the random variables  $U^T \in \mathcal{U}^T$  that describe the responses of the ‘‘clean’’ (deterministic) oracle  $\psi : \mathcal{F} \times \mathcal{X} \rightarrow \mathcal{U}$  to the queries  $X^T$ . The relevant causal ordering is

$$M, X_1, U_1, Y_1, \dots, X_t, U_t, Y_t, \dots, X_T, U_T, Y_T, X_{T+1}$$

where,  $\mathbb{P}$ -almost surely, we have (13) and

$$\begin{aligned} \mathbb{P}(X_t \in A | M, X^{t-1}, U^{t-1}, Y^{t-1}) &= \mathbf{1}_{\{\mathcal{A}_t(X^{t-1}, Y^{t-1}) \in A\}} \\ \mathbb{P}(U_t \in C | M, X^t, U^{t-1}, Y^{t-1}) &= \mathbf{1}_{\{\psi(f_M, X_t) \in C\}} \\ \mathbb{P}(Y_t \in B | M, X^t, U^t, Y^{t-1}) &= Q(B | U_t) \end{aligned}$$

for all  $A \in \mathcal{B}_X, B \in \mathcal{B}_Y, C \in \mathcal{B}_U$ . That is,

$$\begin{aligned} (M, U^{t-1}) &\rightarrow (X^{t-1}, Y^{t-1}) \rightarrow X_t, \\ (X^{t-1}, U^{t-1}, Y^{t-1}) &\rightarrow (M, X_t) \rightarrow U_t \end{aligned}$$

and  $(X^{t-1}, U^{t-1}, Y^{t-1}) \rightarrow (M, X_t) \rightarrow U_t$  and

$$(M, X^t, U^{t-1}, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$$

are Markov chains for each  $t$ . Then we can write

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

because  $U_t$  is completely determined by  $M$  and  $X_t$  via  $U_t = \psi(f_M, X_t)$ . Moreover

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

where the first step is by the chain rule and the second step is due to the fact that  $M \rightarrow (U_t, X^t, Y^{t-1}) \rightarrow Y_t$  is a Markov chain. This follows by applying the weak union and the decomposition properties of conditional independence [34, p.11] to the Markov chain  $(M, X^t, U^{t-1}, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$ . By the same token,  $(X^t, Y^{t-1}) \rightarrow U_t \rightarrow Y_t$  is also a Markov chain, so we have  $I(U_t; Y_t | X^t, Y^{t-1}) \leq I(U_t; Y_t)$ .

C) *Proof of Lemma 4:* Let us fix some  $t$  and consider the conditional mutual information term  $I(M; Y_t | X^t, Y^{t-1})$  in the summation in Lemma 2

$$\begin{aligned} I(M; Y_t | X^t, Y^{t-1}) &= D(\mathbb{P}_{Y_t | M, X^t, Y^{t-1}} \| \mathbb{P}_{Y_t | X^t, Y^{t-1}} | \mathbb{P}_{M, X^t, Y^{t-1}}) \quad (\text{B.4}) \\ &= \mathbb{E} \left[ \log \frac{d\mathbb{P}_{Y_t | M, X^t, Y^{t-1}}}{d\mathbb{P}_{Y_t | X^t, Y^{t-1}}} \right] \\ &= \mathbb{E} \left[ \log \frac{d\mathbb{P}_{Y_t | M, X^t, Y^{t-1}}}{d\mathbb{Q}_{Y_t | X^t, Y^{t-1}}} \right] - \mathbb{E} \left[ \log \frac{d\mathbb{P}_{Y_t | X^t, Y^{t-1}}}{d\mathbb{Q}_{Y_t | X^t, Y^{t-1}}} \right] \quad (\text{B.5}) \end{aligned}$$

$$\begin{aligned} &= D(\mathbb{P}_{Y_t | M, X^t, Y^{t-1}} \| \mathbb{Q}_{Y_t | X^t, Y^{t-1}} | \mathbb{P}_{M, X^t, Y^{t-1}}) \\ &\quad - D(\mathbb{P}_{Y_t | X^t, Y^{t-1}} \| \mathbb{Q}_{Y_t | X^t, Y^{t-1}} | \mathbb{P}_{X^t, Y^{t-1}}) \quad (\text{B.6}) \end{aligned}$$

$$\leq D(\mathbb{P}_{Y_t | M, X^t, Y^{t-1}} \| \mathbb{Q}_{Y_t | X^t, Y^{t-1}} | \mathbb{P}_{M, X^t, Y^{t-1}}) \quad (\text{B.7})$$

where (B.4) follows from (4), (B.5) and (B.6) are justified by virtue of (28), while (B.7) follows from the fact that the divergence is nonnegative.

D) *Proof of Lemma 5:* The random variables  $V^0 = f_M(X) + W$  and  $V^1 = g_M(X) + Z$  are conditionally independent given  $M = i$  and  $X = x$

$$\mathbb{P}_{Y | M=i, X=x} = \mathbb{P}_{V^0 | M=i, X=x} \otimes \mathbb{P}_{V^1 | M=i, X=x}$$

where

$$\begin{aligned} \mathbb{P}_{V^0 | M=i, X=x} &= \mathcal{N}(f_i(x), \sigma^2) \\ \mathbb{P}_{V^1 | M=i, X=x} &= \mathcal{N}(g_i(x), \sigma^2 I_n). \end{aligned}$$

Therefore

$$\begin{aligned} &D(\mathbb{P}_{Y | M=i, X=x} \| \mathbb{P}_{Y | M=j, X=x}) \\ &= D(\mathbb{P}_{V^0 | M=i, X=x} \| \mathbb{P}_{V^0 | M=j, X=x}) \\ &\quad + D(\mathbb{P}_{V^1 | M=i, X=x} \| \mathbb{P}_{V^1 | M=j, X=x}) \\ &= D(\mathcal{N}(f_i(x), \sigma^2) \| \mathcal{N}(f_j(x), \sigma^2)) \\ &\quad + D(\mathcal{N}(g_i(x), \sigma^2 I_n) \| \mathcal{N}(g_j(x), \sigma^2 I_n)) \\ &= \frac{1}{2\sigma^2} \left\{ [f_i(x) - f_j(x)]^2 + \|g_i(x) - g_j(x)\|^2 \right\}. \end{aligned}$$

Plugging this into (31), we get (36).

E) *Proof of Lemma 6:* Let  $Q^*$  denote the product normal distribution  $\mathcal{N}(c^*, \sigma^2) \otimes \mathcal{N}(0, \sigma^2 I_n)$ . Observe that, for every  $i = 0, 1, \dots, N-1$

$$Y^{(i)} = (f_i(x_i^*) + W, \nabla f_i(x_i^*) + Z) = (c^* + W, Z) \sim Q^*.$$

Let  $X_t$  denote the query of  $\mathcal{A}$  at time  $t$  and let  $Y_t$  be the corresponding oracle response. Then

$$\mathbb{P}_{Y_t | M=i, X_t=x_t} = \mathcal{N}(f_i(x_t), \sigma^2) \otimes \mathcal{N}(\nabla f_i(x_t), \sigma^2 I_n).$$

Hence, applying the LF bound (32) with  $\mathbb{Q}_{Y_t}^* = Q^*$ , we can write

$$\begin{aligned} I(M; Y_t | X^t, Y^{t-1}) &\leq \frac{1}{2\sigma^2} \max_i \mathbb{E} \left\{ [f_i(X_t) - c^*]^2 + \|\nabla f_i(X_t)\|^2 \right\}. \quad (\text{B.8}) \end{aligned}$$

We now relate the right-hand side of (B.8) to the performance of  $\mathcal{A}$ . First of all, by convexity of  $f_i$

$$\begin{aligned} f_i(X_t) - c^* &= f_i(X_t) - f_i(x_i^*) \\ &\leq \nabla f_i(X_t)^\top (X_t - x_i^*) \\ &\leq \|\nabla f_i(X_t)\| \|X_t - x_i^*\| \\ &\leq LD_X \|X_t - x_i^*\| \quad (\text{B.9}) \end{aligned}$$

where in the last step we have used the fact that

$$\|\nabla f_i(X_t)\| = \|\nabla f_i(X_t) - \nabla f_i(x_i^*)\| \leq L \|X_t - x_i^*\| \leq LD_X.$$

On the other hand, from strong convexity we have that

$$f_i(X_t) \geq c^* + (\kappa^2/2) \|X_t - x_i^*\|^2. \quad (\text{B.10})$$

Combining (B.9) and (B.10), we, therefore, obtain

$$\begin{aligned} [f_i(X_t) - c^*]^2 &\leq 2D_X^2 (L/\kappa)^2 [f_i(X_t) - c^*] \\ &= 2D_X^2 (L/\kappa)^2 \text{err}_{\mathcal{A}}(t, f_i). \quad (\text{B.11}) \end{aligned}$$

Moreover, because  $\nabla f_i(x_i^*) = 0$ , we can write

$$\begin{aligned} \|\nabla f_i(X_t)\|^2 &= \|\nabla f_i(X_t) - \nabla f_i(x_i^*)\|^2 \\ &\leq L^2 \|X_t - x_i^*\|^2 \\ &\leq 2(L/\kappa)^2 \text{err}_{\mathcal{A}}(t, f_i). \quad (\text{B.12}) \end{aligned}$$

Substituting (B.11) and (B.12) into (B.8), we get (45). Equation (46) is immediate from definitions.

## APPENDIX C

### LEMMA ON FUNCTIONAL RECURRENCES

*Lemma C.1:* Suppose that  $\{\varepsilon_t\}$  is a sequence of nonnegative reals satisfying

$$K \log \left( \frac{1}{\varepsilon_T} \right) - L \leq \sum_{t=1}^T \varepsilon_t^\alpha, \quad \forall T$$

for some  $K, L, \alpha > 0$ . Then there exists some constant  $0 < c < (K/\alpha)^{1/\alpha}$ , such that  $\varepsilon_t \geq ct^{-1/\alpha}$  for infinitely many values of  $t$ .

*Proof:* The proof is by contradiction. Suppose first that  $t^{1/\alpha}\varepsilon_t < c$  for all  $t$ . Then

$$K \log \frac{T^{1/\alpha}}{c} - L < c^\alpha \sum_{t=1}^T \frac{1}{t} \leq c^\alpha (\log T + 1), \forall T.$$

Rearranging, we get

$$\left(\frac{K}{\alpha} - c^\alpha\right) \log T \leq K \log c + L + c^\alpha, \quad \forall T.$$

Since  $K/\alpha - c^\alpha > 0$ , this implies that  $\log T$  is bounded for all large and positive  $T$ , which is impossible. Hence, there exists some set  $S \subseteq \mathbb{N}$ , such that  $\varepsilon_t \geq ct^{-1/\alpha}$  for all  $t \in S$ . We now show that  $S$  must necessarily be countably infinite. Suppose, to the contrary, that it's finite. Then there is some  $T_0$ , such that  $\varepsilon_t < ct^{-1/\alpha}$  for all  $t \geq T_0$ . In that case, for  $T > T_0$ , we can write

$$\begin{aligned} K \log \frac{T^{1/\alpha}}{c} - L &\leq c^\alpha \left( \sum_{t=1}^{T_0} \varepsilon_t^\alpha + \sum_{t=T_0+1}^T \frac{1}{t} \right) \\ &\equiv c^\alpha \left( K(T_0, \alpha) + \sum_{t=T_0+1}^T \frac{1}{t} \right) \\ &\leq c^\alpha (K(T_0, \alpha) + \log T - \log T_0). \end{aligned}$$

Rearranging, we see that the inequality

$$\begin{aligned} \left(\frac{K}{\alpha} - c^\alpha\right) \log T \\ \leq K \log c + L + c^\alpha (K(T_0, \alpha) - \log T_0) \end{aligned}$$

must hold for all  $T > T_0$ . Since  $K/\alpha > c^\alpha$  by hypothesis, this implies that  $\log T$  is bounded for  $T > T_0$ , which is, again, impossible. Thus,  $\varepsilon_t \geq ct^{-1/\alpha}$  for infinitely many values of  $t$ .

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their probing questions and numerous suggestions, which have greatly improved the paper. In particular, we would like to thank one reviewer for suggesting the definition of a strong infinite-step algorithm.

REFERENCES

[1] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Hoboken, NJ: Wiley, 1983.  
 [2] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Boston, MA: Kluwer, 2004.  
 [3] 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.  
 [4] M. Raginsky and A. Rakhlin, "Information complexity of black-box convex optimization: A new look via feedback information theory," in *Proc. 47th Annu. Allerton Conf. Communication, Control and Computing*, Sep./Oct. 2009, pp. 803–810.  
 [5] B. Yu, "Assouad, Fano and Le Cam," in *Festschrift for Lucien Le Cam*. New York: Springer, 1997, pp. 423–435.

[6] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.  
 [7] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York: Springer, 2009.  
 [8] A. Shapiro and A. Nemirovski, "On complexity of stochastic programming," in *Continuous Optimization: Trends and Applications*, V. Jeyakumar and A. M. Rubinov, Eds. New York: Springer, 2005, pp. 111–144.  
 [9] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, "Information-theoretic lower bounds on the oracle complexity of convex optimization," *Advances in Neural Information Processing Systems 22*. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., pp. 1–9, 2009.  
 [10] A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright, Information-Theoretic Lower Bounds on the Oracle Complexity of Convex Optimization, Sep. 2010 [Online]. Available: <http://arxiv.org/abs/1009.0571>  
 [11] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. New York: Springer, 2001.  
 [12] A. N. Shiryaev and V. G. Spokoiny, *Statistical Experiments and Decisions: Asymptotic Theory*. Singapore: World Scientific, 2000.  
 [13] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. New York: Springer, 2002.  
 [14] R. M. Gray, *Entropy and Information Theory*. New York: Springer, 1991.  
 [15] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*. Boston, MA: Academic, 1988.  
 [16] L. Plaskota, *Noisy Information and Computational Complexity*. Cambridge, U.K.: Cambridge Univ. Press, 1996.  
 [17] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, Jul. 1992.  
 [18] J. Pfanzagl, "On the measurability and consistency of minimum contrast estimates," *Metrika*, vol. 14, no. 1, pp. 249–272, Dec. 1969.  
 [19] V. V. Fedorov, *Theory of Optimal Experiments*. New York: Academic, 1972.  
 [20] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neur. Comput.*, vol. 4, pp. 590–604, 1992.  
 [21] L. Paninski, "Asymptotic theory of information-theoretic experimental design," *Neural Comput.*, vol. 17, pp. 1480–1507, 2005.  
 [22] R. Castro and R. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.  
 [23] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problem. Inf. Transm.*, vol. 10, pp. 223–231, 1974.  
 [24] M. V. Burnashev, "Sequential discrimination of hypotheses with control of observations," *Math. USSR Izv.*, vol. 15, no. 3, pp. 419–440, 1980.  
 [25] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1247–1251, Jul. 1994.  
 [26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.  
 [27] M. L. Vil'k and S. V. Shil'man, "Convergence and optimality of implementable adaptation algorithms (informational approach)," *Problems Inf. Transm.*, vol. 20, pp. 314–326, 1985.  
 [28] Y. Yang, "How powerful can any regression learning procedure be?," in *Proc. 11th Int. Conf. Artif. Intell. Statist. (AISTATS)*, M. Meila and X. Shen, Eds., San Juan, PR, Mar. 2007.  
 [29] R. D. Nowak, The Geometry of Generalized Binary Search, Oct. 2009 [Online]. Available: <http://arxiv.org/abs/0910.4397>  
 [30] A. Guntuboyina, "Lower bounds for the minimax risk using  $f$ -divergences and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, Apr. 2011.  
 [31] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.  
 [32] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Ann. Statist.*, vol. 32, no. 1, pp. 135–166, 2004.  
 [33] M. Naghshvar and T. Javidi, "Active M-ary sequential hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, Jun. 2010, pp. 1623–1627.  
 [34] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

**Maxim Raginsky** (S'99–M'00) received the B.S. and M.S. degrees in 2000 and the Ph.D. degree in 2002 from Northwestern University, Evanston, IL, all in electrical engineering. From 2002 to 2004 he was a Postdoctoral Researcher at the Center for Photonic Communication and Computing at Northwestern University, where he pursued work on quantum cryptography and quantum communication and information theory. From 2004 to 2007, he was a Beckman Foundation Postdoctoral Fellow at the University of Illinois in Urbana-Champaign, where he carried out research on information theory, statistical learning and computational neuroscience. Since September 2007, he has been with Duke University, where he is now Assistant Research Professor of Electrical and Computer Engineering. His interests include statistical signal processing, information theory, statistical learning and nonparametric estimation. He is particularly interested in problems that combine the communication, signal processing and machine learning components in a novel and nontrivial way, as well as in the theory and practice of robust statistical inference with limited information.

**Alexander Rakhlin** received the B.A. degrees in mathematics and computer science from Cornell University in 2000 and Ph.D. in computational neuroscience from MIT in 2006. From 2006 to 2009, he was a postdoctoral fellow at the Department of Electrical Engineering and Computer Sciences, UC Berkeley. Since 2009, he has been on the faculty in the Department of Statistics, University of Pennsylvania. He has been a co-director of Penn Research in Machine Learning (PRiML) since 2010. His interests include machine learning, statistics, optimization, and game theory.