Query Complexity of Derivative-Free Optimization

Kevin G. Jamieson
University of Wisconsin
Madison, WI 53706, USA
kgjamieson@wisc.edu

Robert D. Nowak
University of Wisconsin
Madison, WI 53706, USA
nowak@engr.wisc.edu

Benjamin Recht
University of Wisconsin
Madison, WI 53706, USA
brecht@cs.wisc.edu

Abstract

This paper provides lower bounds on the convergence rate of Derivative Free Optimization (DFO) with noisy function evaluations, exposing a fundamental and unavoidable gap between the performance of algorithms with access to gradients and those with access to only function evaluations. However, there are situations in which DFO is unavoidable, and for such situations we propose a new DFO algorithm that is proved to be near optimal for the class of strongly convex objective functions. A distinctive feature of the algorithm is that it uses only Boolean-valued function comparisons, rather than function evaluations. This makes the algorithm useful in an even wider range of applications, such as optimization based on paired comparisons from human subjects, for example. We also show that regardless of whether DFO is based on noisy function evaluations or Boolean-valued function comparisons, the convergence rate is the same.

1 Introduction

Optimizing large-scale complex systems often requires the tuning of many parameters. With training data or simulations one can evaluate the relative merit, or incurred loss, of different parameter settings, but it may be unclear how each parameter influences the overall objective function. In such cases, derivatives of the objective function with respect to the parameters are unavailable. Thus, we have seen a resurgence of interest in Derivative Free Optimization (DFO) [1, 2, 3, 4, 5, 6, 7, 8]. When function evaluations are noiseless, DFO methods can achieve the same rates of convergence as noiseless gradient methods up to a small factor depending on a low-order polynomial of the dimension [9, 5, 10]. This leads one to wonder if the same equivalence can be extended to the case when function evaluations and gradients are noisy.

Sadly, this paper proves otherwise. We show that when function evaluations are noisy, the optimization error of *any* DFO is $\Omega(\sqrt{1/T})$, where T is the number of evaluations. This lower bound holds even for strongly convex functions. In contrast, noisy gradient methods exhibit $\Theta(1/T)$ error scaling for strongly convex functions [9, 11]. A consequence of our theory is that finite differencing cannot achieve the rates of gradient methods when the function evaluations are noisy.

On the positive side, we also present a new derivative-free algorithm that achieves this lower bound with near optimal dimension dependence. Moreover, the algorithm uses only boolean comparisons of function values, not actual function values. This makes the algorithm applicable to situations in which the optimization is only able to probably correctly decide if the value of one configuration is better than the value of another. This is especially interesting in optimization based on human subject feedback, where paired comparisons are often used instead of numerical scoring. The convergence rate of the new algorithm is optimal in terms of T and near-optimal in terms of its dependence on the ambient dimension. Surprisingly, our lower bounds show that this new algorithm that uses only function comparisons achieves the same rate in terms of T as any algorithm that has access to function evaluations.

2 Problem formulation and background

We now formalize the notation and conventions for our analysis of DFO. A function f is *strongly convex with constant* τ on a convex set $\mathcal{B} \subset \mathbb{R}^d$ if there exists a constant $\tau > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\tau}{2}||x - y||^2$$

for all $x, y \in \mathcal{B}$. The gradient of f, if it exists, denoted ∇f, is *Lipschitz with constant* L if $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for some $L > 0$. The class of strongly convex functions with Lipschitz gradients defined on a nonempty, convex set $\mathcal{B} \subset \mathbb{R}^n$ which take their minimum in $\mathcal{B}$ with parameters τ and L is denoted by $\mathcal{F}_{\tau,L,\mathcal{B}}$.

The problem we consider is minimizing a function $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$. The function f is not explicitly known. An optimization procedure may only query the function in one of the following two ways.

Function Evaluation Oracle: For any point $x \in \mathcal{B}$ an optimization procedure can observe

$$E_f(x) = f(x) + w$$

where $w \in \mathbb{R}$ is a random variable with $\mathbb{E}[w] = 0$ and $\mathbb{E}[w^2] = \sigma^2$.

Function Comparison Oracle: For any pair of points $x, y \in \mathcal{B}$ an optimization procedure can observe a binary random variable $C_f(x, y)$ satisfying

$$\mathbb{P}\left(C_f(x, y) = \text{sign}\{f(y) - f(x)\}\right) \geq \frac{1}{2} + \min\left\{\delta_0, \mu|f(y) - f(x)|^{\kappa-1}\right\} \tag{1}$$

for some $0 < \delta_0 \leq 1/2$, $\mu > 0$ and $\kappa \geq 1$. When $\kappa = 1$, without loss of generality assume $\mu \leq \delta_0 \leq 1/2$. Note $\kappa = 1$ implies that the comparison oracle is correct with a probability that is greater than 1/2 and independent of x, y. If $\kappa > 1$, then the oracle's reliability decreases as the difference between $f(x)$ and $f(y)$ decreases.

To illustrate how the function comparison oracle and function evaluation oracles relate to each other, suppose $C_f(x, y) = \text{sign}\{E_f(y) - E_f(x)\}$ where $E_f(x)$ is a function evaluation oracle with additive noise w. If w is Gaussian distributed with mean zero and variance σ^2 then $\kappa = 2$ and $\mu \geq \left(4\pi\sigma^2 e\right)^{-1/2}$ (see supplementary materials). In fact, this choice of w corresponds to Thurston's law of comparative judgment which is a popular model for outcomes of pairwise comparisons from human subjects [12]. If w is a "spikier" distribution such as a two-sided Gamma distribution with shape parameter in the range of $(0, 1]$ then all values of $\kappa \in (1, 2]$ can be realized (see supplementary materials).

Interest in the function comparison oracle is motivated by certain popular derivative-free optimization procedures that use only comparisons of function evaluations (e.g. [7]) and by optimization problems involving human subjects making paired comparisons (for instance, getting fitted for prescription lenses or a hearing aid where unknown parameters specific to each person are tuned with the familiar queries "better or worse?"). Pairwise comparisons have also been suggested as a novel way to tune web-search algorithms [13]. Pairwise comparison strategies have previously been analyzed in the finite setting where the task is to identify the best alternative among a finite set of alternatives (sometimes referred to as the dueling-bandit problem) [13, 14]. The function comparison oracle presented in this work and its analysis are novel. The main contributions of this work and new art are as follows (i) lower bounds for the function evaluation oracle in the presence of measurement noise (ii) lower bounds for the function comparison oracle in the presence of noise and (iii) an algorithm for the function comparison oracle, which can also be applied to the function evaluation oracle setting, that nearly matches both the lower bounds of (i) and (ii).

We prove our lower bounds for strongly convex functions with Lipschitz gradients defined on a compact, convex set $\mathcal{B}$, and because these problems are a subset of those involving all convex functions (and have non-empty intersection with problems where f is merely Lipschitz), the lower bound also applies to these larger classes. While there are known theoretical results for DFO in the noiseless setting [15, 5, 10], to the best of our knowledge we are the first to characterize lower bounds for DFO in the stochastic setting. Moreover, we believe we are the first to show a novel upper bound for stochastic DFO using a function comparison oracle (which also applies to the function evaluation oracle). However, there are algorithms with upper bounds on the rates of convergence for stochastic

DFO with the function evaluation oracle [15, 16]. We discuss the relevant results in the next section following the lower bounds .

While there remains many open problems in stochastic DFO (see Section 6), rates of convergence with a stochastic gradient oracle are well known and were first lower bounded by Nemirovski and Yudin [15]. These classic results were recently tightened to show a dependence on the dimension of the problem [17]. And then tightened again to show a better dependence on the noise [11] which matches the upper bound achieved by stochastic gradient descent [9]. The aim of this work is to start filling in the knowledge gaps of stochastic DFO so that it is as well understood as the stochastic gradient oracle. Our bounds are based on simple techniques borrowed from the statistical learning literature that use natural functions and oracles in the same spirit of [11].

3 Main results

The results below are presented with simplifying constants that encompass many factors to aid in exposition. Explicit constants are given in the proofs in Sections 4 and 5. Throughout, we denote the minimizer of f as x_f^*. The expectation in the bounds is with respect to the noise in the oracle queries and (possible) optimization algorithm randomization.

3.1 Query complexity of the function comparison oracle

Theorem 1. *For every $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ let C_f be a function comparison oracle with parameters (κ, μ, δ_0). Then for $n \geq 8$ and sufficiently large T*

$$\inf_{\widehat{x}_T} \sup_{f \in \mathcal{F}_{\tau,L,\mathcal{B}}} \mathbb{E}\left[f(\widehat{x}_T) - f(x_f^*)\right] \geq \begin{cases} c_1 \exp\left\{-c_2 \frac{T}{n}\right\} & \text{if } \kappa = 1 \\ c_3 \left(\frac{n}{T}\right)^{\frac{1}{2(\kappa-1)}} & \text{if } \kappa > 1 \end{cases}$$

where the infimum is over the collection of all possible estimators of x_f^* using at most T queries to a function comparison oracle and the supremum is taken with respect to all problems in $\mathcal{F}_{\tau,L,\mathcal{B}}$ and function comparison oracles with parameters (κ, μ, δ_0). The constants c_1, c_2, c_3 depend the oracle and function class parameters, as well as the geometry of $\mathcal{B}$, but are independent of T and n.

For upper bounds we propose a specific algorithm based on coordinate-descent in Section 5 and prove the following theorem for the case of unconstrained optimization, that is, $\mathcal{B} = \mathbb{R}^n$.

Theorem 2. *For every $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ with $\mathcal{B} = \mathbb{R}^n$ let C_f be a function comparison oracle with parameters (κ, μ, δ_0). Then there exists a coordinate-descent algorithm that is adaptive to unknown $\kappa \geq 1$ that outputs an estimate $\widehat{x}_T$ after T function comparison queries such that with probability $1 - \delta$*

$$\sup_{f \in \mathcal{F}_{\tau,L,B}} \mathbb{E}\left[f(\widehat{x}_T) - f(x_f^*)\right] \leq \begin{cases} c_1 \exp\left\{-c_2 \sqrt{\frac{T}{n}}\right\} & \text{if } \kappa = 1 \\ c_3 n \left(\frac{n}{T}\right)^{\frac{1}{2(\kappa-1)}} & \text{if } \kappa > 1 \end{cases}$$

where c_1, c_2, c_3 depend the oracle and function class parameters as well as T,n, and $1/\delta$, but only poly-logarithmically.

3.2 Query complexity of the function evaluation oracle

Theorem 3. *For every $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ let E_f be a function evaluation oracle with variance σ^2. Then for $n \geq 8$ and sufficiently large T*

$$\inf_{\widehat{x}_T} \sup_{f \in \mathcal{F}_{\tau,L,\mathcal{B}}} \mathbb{E}\left[f(\widehat{x}_T) - f(x_f^*)\right] \geq c \left(\frac{n\sigma^2}{T}\right)^{\frac{1}{2}}$$

where the infimum is taken with respect to the collection of all possible estimators of x_f^* using just T queries to a function evaluation oracle and the supremum is taken with respect to all problems in $\mathcal{F}_{\tau,L,\mathcal{B}}$ and function evaluation oracles with variance σ^2. The constant c depends on the oracle and function class parameters, as well as the geometry of $\mathcal{B}$, but is independent of T and n.

Because a function evaluation oracle can always be turned into a function comparison oracle (see discussion above), the algorithm and upper bound in Theorem 2 with $\kappa = 2$ applies to many typical function evaluation oracles (e.g. additive Gaussian noise), yielding an upper bound of $\left(n^3\sigma^2/T\right)^{1/2}$ ignoring constants and log factors. This matches the rate of convergence as a function of T and σ^2, but has worse dependence on the dimension n.

Alternatively, under a less restrictive setting, Nemirovski and Yudin proposed two algorithms for the class of convex, Lipschitz functions that obtain rates of $n^{1/2}/T^{1/4}$ and $p(n)/T^{1/2}$, respectively, where $p(n)$ was left as an unspecified polynomial of n [15]. While focusing on stochastic DFO with bandit feedback, Agarwal *et. al.* built on the ideas developed in [15] to obtain a result that they point out implies a convergence rate of $n^{16}/T^{1/2}$ in the optimization setting considered here [16]. Whether or not these rates can be improved to those obtained under the more restrictive function classes of above is an open question.

A related but fundamentally different problem that is somewhat related with the setting considered in this paper is described as online (or stochastic) convex optimization with multi-point feedback [18, 5, 19]. Essentially, this setting allows the algorithm to probe the value of the function f plus noise at multiple locations where the noise changes at each time step, but each set of samples at each time experiences the *same* noise. Because the noise model of that work is incompatible with the one considered here, no comparisons should be made between the two.

4 Lower Bounds

The lower bounds in Theorems 1 and 3 are proved using a general minimax bound [20, Thm. 2.5]. Our proofs are most related to the approach developed in [21] for active learning, which like optimization involves a Markovian sampling process. Roughly speaking, the lower bounds are established by considering a simple case of the optimization problem in which the global minimum is known a priori to belong to a finite set. Since the simple case is "easier" than the original optimization, the minimum number of queries required for a desired level of accuracy in this case yields a lower bound for the original problem.

The following theorem is used to prove the bounds. In the terms of the theorem, f is a function to be minimized and P_f is the probability model governing the noise associated with queries when f is the true function.

Theorem 4. *[20, Thm. 2.5] Consider a class of functions $\mathcal{F}$ and an associated family of probability measures $\{P_f\}_{f\in\mathcal{F}}$. Let $M \geq 2$ be an integer and $f_0, f_1, \ldots, f_M$ be functions in $\mathcal{F}$. Let $d(\cdot,\cdot) : \mathcal{F}\times\mathcal{F} \to \mathbb{R}$ be a semi-distance and assume that:*

1. *$d(f_i, f_j) \geq 2s > 0$, for all $0 \leq i < j \leq M$,*
2. *$\frac{1}{M}\sum_{j=1}^{M} KL(P_i||P_0) \leq a \log M$,*

where the Kullback-Leibler divergence $KL(P_i||P_0) := \int \log \frac{dP_i}{dP_0} dP_i$ is assumed to be well-defined (i.e., P_0 is a dominating measure) and $0 < a < 1/8$. Then

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{P}(d(\widehat{f}, f) \geq s) \;\geq\; \inf_{\widehat{f}} \max_{f\in\{f_0,\ldots,f_M\}} \mathbb{P}(d(\widehat{f}, f) \geq s) \;\geq\; \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2a - 2\sqrt{\frac{a}{\log M}}\right) > 0\,,$$

where the infimum is taken over all possible estimators based on a sample from P_f.

We are concerned with the functions in the class $\mathcal{F} := \mathcal{F}_{\tau,L,\mathcal{B}}$. The volume of $\mathcal{B}$ will affect only constant factors in our bounds, so we will simply denote the class of functions by $\mathcal{F}$ and refer explicitly to $\mathcal{B}$ only when necessary. Let $x_f := \arg\min_x f(x)$, for all $f \in \mathcal{F}$. The semi-distance we use is $d(f, g) := \|x_f - x_g\|$, for all $f, g \in \mathcal{F}$. Note that each point in $\mathcal{B}$ can be specified by one of many $f \in \mathcal{F}$. So the problem of selecting an f is equivalent to selecting a point $x \in \mathcal{B}$. Indeed, the semi-distance defines a collection of equivalence classes in $\mathcal{F}$ (i.e., all functions having a minimum at $x \in \mathcal{B}$ are equivalent). For every $f \in \mathcal{F}$ we have $\inf_{g\in\mathcal{F}} f(x_g) = \inf_{x\in\mathcal{B}} f(x)$, which is a useful identity to keep in mind.

We now construct the functions $f_0, f_1, \ldots, f_M$ that will be used for our proofs. Let $\Omega = \{-1, 1\}^n$ so that each $\omega \in \Omega$ is a vertex of the d-dimensional hypercube. Let $\mathcal{V} \subset \Omega$ with cardinality $|\mathcal{V}| \geq 2^{n/8}$

such that for all $\omega \neq \omega' \in \mathcal{V}$, we have $\rho(\omega, \omega') \geq n/8$ where $\rho(\cdot, \cdot)$ is the Hamming distance. It is known that such a set exists by the Varshamov-Gilbert bound [20, Lemma 2.9]. Denote the elements of $\mathcal{V}$ by $\omega_0, \omega_1, \ldots, \omega_M$. Next we state some elementary bounds on the functions that will be used in our analysis.

Lemma 1. *For $\epsilon > 0$ define the set $\mathcal{B} \subset \mathbb{R}^n$ to be the ℓ_∞ ball of radius ϵ and define the functions on $\mathcal{B}$: $f_i(x) := \frac{\tau}{2}||x - \epsilon\omega_i||^2$, for $i = 0, \ldots, M$, $\omega_i \in \mathcal{V}$, and $x_i := \arg\min_x f_i(x) = \epsilon\omega_i$. Then for all $0 \leq i < j \leq M$ and $x \in \mathcal{B}$ the functions $f_i(x)$ satisfy*

1. *f_i is strongly convex-τ with Lipschitz-τ gradients and $x_i \in \mathcal{B}$*
2. *$||x_i - x_j|| \geq \epsilon\sqrt{\frac{n}{2}}$*
3. *$|f_i(x) - f_j(x)| \leq 2\tau n\epsilon^2$.*

We are now ready to prove Theorems 1 and 3. Each proof uses the functions $f_0, \ldots, f_M$ a bit differently, and since the noise model is also different in each case, the KL divergence is bounded differently in each proof. We use the fact that if X and Y are random variables distributed according to Bernoulli distributions P_X and P_Y with parameters $1/2 + \mu$ and $1/2 - \mu$, then $\mathrm{KL}(P_X||P_Y) \leq 4\mu^2/(1/2 - \mu)$. Also, if $X \sim \mathcal{N}(\mu_X, \sigma^2) =: P_X$ and $Y \sim \mathcal{N}(\mu_Y, \sigma^2) =: P_y$ then $\mathrm{KL}(P_X||P_Y) = \frac{1}{2\sigma^2}||\mu_X - \mu_Y||^2$.

4.1 Proof of Theorem 1

First we will obtain the bound for the case $\kappa > 1$. Let the comparison oracle satisfy

$$\mathbb{P}\left(C_{f_i}(x, y) = \mathrm{sign}\{f_i(y) - f_i(x)\}\right) \;=\; \frac{1}{2} + \min\left\{\mu|f_i(y) - f_i(x)|^{\kappa-1}, \delta_0\right\}.$$

In words, $C_{f_i}(x, y)$ is correct with probability as large as the right-hand-side of above and is monotonic increasing in $f_i(y) - f_i(x)$. Let $\{x_k, y_k\}_{k=1}^T$ be a sequence of T pairs in $\mathcal{B}$ and let $\{C_{f_i}(x_k, y_k)\}_{k=1}^T$ be the corresponding sequence of noisy comparisons. We allow the sequence $\{x_k, y_k\}_{k=1}^T$ to be generated in any way subject to the Markovian assumption that $C_{f_i}(x_k, y_k)$ given (x_k, y_k) is conditionally independent of $\{x_i, y_i\}_{i<k}$. For $i = 0, \ldots, M$, and $\ell = 1, \ldots, T$ let $P_{i,\ell}$ denote the joint probability distribution of $\{x_k, y_k, C_{f_i}(x_k, y_k)\}_{k=1}^\ell$, let $Q_{i,\ell}$ denote the conditional distribution of $C_{f_i}(x_\ell, y_\ell)$ given (x_ℓ, y_ℓ), and let S_ℓ denote the conditional distribution of (x_ℓ, y_ℓ) given $\{x_k, y_k, C_{f_i}(x_k, y_k)\}_{k=1}^{\ell-1}$. Note that S_ℓ is only a function of the underlying optimization algorithm and does not depend on i.

$$\mathrm{KL}(P_{i,T}||P_{j,T}) = \mathbb{E}_{P_{i,T}}\left[\log\frac{P_{i,T}}{P_{j,T}}\right] = \mathbb{E}_{P_{i,T}}\left[\log\frac{\prod_{\ell=1}^T Q_{i,\ell}S_\ell}{\prod_{\ell=1}^T Q_{j,\ell}S_\ell}\right] = \mathbb{E}_{P_{i,T}}\left[\log\frac{\prod_{\ell=1}^T Q_{i,\ell}}{\prod_{\ell=1}^T Q_{j,\ell}}\right]$$
$$= \sum_{\ell=1}^T \mathbb{E}_{P_{i,T}}\left[\mathbb{E}_{P_{i,T}}\left[\log\frac{Q_{i,\ell}}{Q_{j,\ell}}\middle|\{x_k, y_k\}_{k=1}^T\right]\right] \leq T\sup_{x_1,y_1\in\mathcal{B}}\mathbb{E}_{P_{i,1}}\left[\mathbb{E}_{P_{i,1}}\left[\log\frac{Q_{i,1}}{Q_{j,1}}\middle|x_1, y_1\right]\right]$$

By the second claim of Lemma 1, $|f_i(x) - f_j(x)| \leq 2\tau n\epsilon^2$, and therefore the bound above is less than or equal to the KL divergence between the Bernoulli distributions with parameters $\frac{1}{2} \pm \mu\left(2\tau n\epsilon^2\right)^{(\kappa-1)}$, yielding the bound

$$\mathrm{KL}(P_{i,T}|P_{j,T}) \leq \frac{4T\mu^2\left(2\tau n\epsilon^2\right)^{2(\kappa-1)}}{1/2 - \mu\left(2\tau n\epsilon^2\right)^{(\kappa-1)}} \;\leq 16T\mu^2\left(2\tau n\epsilon^2\right)^{2(\kappa-1)}$$

provided ϵ is sufficiently small. We also assume ϵ (or, equivalently, $\mathcal{B}$) is sufficiently small so that $|f_i(x) - f_j(x)|^{\kappa-1} \leq \delta_0$. We are now ready to apply Theorem 4. Recalling that $M \geq 2^{n/8}$, we want to choose ϵ such that

$$\mathrm{KL}(P_{i,T}|P_{j,T}) \leq 16T\mu^2\left(2\tau n\epsilon^2\right)^{2(\kappa-1)} \leq a\frac{n}{8}\log(2) \leq a\log M$$

with an a small enough so that we can apply the theorem. By setting $a = 1/16$ and equating the two sides of the equation we have $\epsilon = \epsilon_T := \frac{1}{2\sqrt{n}}\left(\frac{2}{\tau}\right)^{1/2}\left(\frac{n\log(2)}{2048\mu^2 T}\right)^{\frac{1}{4(\kappa-1)}}$ (note that this also implies a

sequence of sets $\mathcal{B}_T$ by the definition of the functions in Lemma 1). Thus, the semi-distance satisfies

$$d(f_j, f_i) = ||x_j - x_i|| \geq \sqrt{n/2}\epsilon_T \geq \frac{1}{2\sqrt{2}} \left(\frac{2}{\tau}\right)^{1/2} \left(\frac{n \log(2)}{2048\mu^2 T}\right)^{\frac{1}{4(\kappa-1)}} =: 2s_T \,.$$

Applying Theorem 4 we have

$$\begin{aligned}\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{P}(\|x_{\widehat{f}} - x_f\| \geq s_T) &\geq \inf_{\widehat{f}} \max_{i\in\{0,\ldots,M\}} \mathbb{P}(\|x_{\widehat{f}} - x_i\| \geq s_T) = \inf_{\widehat{f}} \max_{i\in\{0,\ldots,M\}} \mathbb{P}(d(\widehat{f}, f_i) \geq s_T) \\ &\geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2a - 2\sqrt{\frac{a}{\log M}}\right) > 1/7\,,\end{aligned}$$

where the final inequality holds since $M \geq 2$ and $a = 1/16$. Strong convexity implies that $f(x) - f(x_f) \geq \frac{\tau}{2}||x - x_f||^2$ for all $f \in \mathcal{F}$ and $x \in \mathcal{B}$. Therefore

$$\begin{aligned}\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{P}\left(f(x_{\widehat{f}}) - f(x_f) \geq \frac{\tau}{2}s_T^2\right) &\geq \inf_{\widehat{f}} \max_{i\in\{0,\ldots,M\}} \mathbb{P}\left(f_i(x_{\widehat{f}}) - f_i(x_i) \geq \frac{\tau}{2}s_T^2\right) \\ &\geq \inf_{\widehat{f}} \max_{i\in\{0,\ldots,M\}} \mathbb{P}\left(\frac{\tau}{2}\|x_{\widehat{f}} - x_i\|^2 \geq \frac{\tau}{2}s_T^2\right) \\ &= \inf_{\widehat{f}} \max_{i\in\{0,\ldots,M\}} \mathbb{P}\left(\|x_{\widehat{f}} - x_i\| \geq s_T\right) > 1/7\,.\end{aligned}$$

Finally, applying Markov's inequality we have

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}\left[f(x_{\widehat{f}}) - f(x_f)\right] \geq \frac{1}{7}\left(\frac{1}{32}\right)\left(\frac{n\log(2)}{2048\mu^2 T}.\right)^{\frac{1}{2(\kappa-1)}}$$

4.2 Proof of Theorem 1 for $\kappa = 1$

To handle the case when $\kappa = 1$ we use functions of the same form, but the construction is slightly different. Let ℓ be a positive integer and let $M = \ell^n$. Let $\{\xi_i\}_{i=1}^M$ be a set of uniformly space points in $\mathcal{B}$ which we define to be the unit cube in $\mathbb{R}^n$, so that $\|\xi_i - \xi_j\| \geq \ell^{-1}$ for all $i \neq j$. Define $f_i(x) := ||x - \xi_i||^2$, $i = 1, \ldots, M$. Let $s := \frac{1}{2\ell}$ so that $d(f_i, f_j) := ||x_i^* - x_j^*|| \geq 2s$. Because $\kappa = 1$, we have $\mathbb{P}\left(C_{f_i}(x, y) = \text{sign}\{f_i(y) - f_i(x)\}\right) \geq \mu$ for some $\mu > 0$, all $i \in \{1, \ldots, M\}$, and all $x, y \in \mathcal{B}$. We bound $\text{KL}(P_{i,T}||P_{j,T})$ in exactly the same way as we bounded it in Section 4.1 except that now we have $C_{f_i}(x_k, y_k) \sim$ Bernoulli$(\frac{1}{2} + \mu)$ and $C_{f_j}(x_k, y_k) \sim$ Bernoulli$(\frac{1}{2} - \mu)$. It then follows that if we wish to apply the theorem, we want to choose s so that

$$\text{KL}(P_{i,T}|P_{j,T}) \leq 2T\mu^2/(1/2 - \mu) \leq a \log M = an \log\left(\tfrac{1}{2s}\right)$$

for some $a < 1/8$. Using the same sequence of steps as in Section 4.1 we have

$$\inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}\left[f(x_{\widehat{f}}) - f(x_f)\right] \geq \frac{1}{7}\left(\frac{1}{2}\right)^2 \exp\left\{-\frac{128T\mu^2}{n(1/2-\mu)}\right\}.$$

4.3 Proof of Theorem 3

Let f_i for all $i = 0, \ldots, M$ be the functions considered in Lemma 1. Recall that the evaluation oracle is defined to be $E_f(x) := f(x) + w$, where w is a random variable (independent of all other random variables under consideration) with $\mathbb{E}[w] = 0$ and $\mathbb{E}[w^2] = \sigma^2 > 0$. Let $\{x_k\}_{k=1}^n$ be a sequence of points in $\mathcal{B} \subset \mathbb{R}^n$ and let $\{E_f(x_k)\}_{k=1}^T$ denote the corresponding sequence of noisy evaluations of $f \in \mathcal{F}$. For $\ell = 1, \ldots, T$ let $P_{i,\ell}$ denote the joint probability distribution of $\{x_k, E_{f_i}(x_k)\}_{k=1}^\ell$, let $Q_{i,\ell}$ denote the conditional distribution of $E_{f_i}(x_k)$ given x_k, and let S_ℓ denote the conditional distribution of x_ℓ given $\{x_k, E_f(x_k)\}_{k=1}^{\ell-1}$. S_ℓ is a function of the underlying optimization algorithm and does not depend on i. We can now bound the KL divergence between any two hypotheses as in Section 4.1:

$$\text{KL}(P_{i,T}||P_{j,T}) \leq T \sup_{x_1\in\mathcal{B}} \mathbb{E}_{P_{i,1}}\left[\mathbb{E}_{P_{i,1}}\left[\log\frac{Q_{i,1}}{Q_{j,1}}\middle| x_1\right]\right] .$$

To compute a bound, let us assume that w is Gaussian distributed. Then

$$\begin{aligned}\mathrm{KL}(P_{i,T}||P_{j,T}) &\leq T \sup_{z\in\mathcal{B}} \mathrm{KL}\left(\mathcal{N}(f_i(z),\sigma^2)||\mathcal{N}(f_j(z),\sigma^2)\right) \\ &= \frac{T}{2\sigma^2}\sup_{z\in\mathcal{B}}|f_i(z)-f_j(z)|^2 \leq \frac{T}{2\sigma^2}\left(2\tau n\epsilon^2\right)^2\end{aligned}$$

by the third claim of Lemma 1. We then repeat the same procedure as in Section 4.1 to attain

$$\inf_{\widehat{f}}\sup_{f\in\mathcal{F}}\mathbb{E}\left[f(x_{\widehat{f}})-f(x_f)\right] \geq \frac{1}{7}\left(\frac{1}{32}\right)\left(\frac{n\sigma^2\log(2)}{64T}\right)^{\frac{1}{2}}.$$

5 Upper bounds

The algorithm that achieves the upper bound using a pairwise comparison oracle is a combination of standard techniques and methods from the convex optimization and statistical learning literature. The algorithm is explained in full detail in the supplementary materials, and is summarized as follows. At each iteration the algorithm picks a coordinate uniformly at random from the n possible dimensions and then performs an approximate line search. By exploiting the fact that the function is strongly convex with Lipschitz gradients, one guarantees using standard arguments that the approximate line search makes a sufficient decrease in the objective function value in expectation [22, Ch.9.3]. If the pairwise comparison oracle made no errors then the approximate line search is accomplished by a binary-search-like scheme, essentially a golden section line-search algorithm [23]. However, when responses from the oracle are only probably correct we make the line-search robust to errors by repeating the same query until we can be confident about the true, uncorrupted direction of the pairwise comparison using a standard procedure from the active learning literature [24] (a similar technique was also implemented for the bandit setting of derivate-free optimization [8]). Because the analysis of each component is either known or elementary, we only sketch the proof here and leave the details to the supplementary materials.

5.1 Coordinate descent

Given a candidate solution x_k after $k \geq 0$ iterations, the algorithm defines a search direction $d_k = \mathbf{e}_i$ where i is chosen uniformly at random from the possible n dimensions and $\mathbf{e}_i$ is a vector of all zeros except for a one in the ith coordinate. We note that while we only analyze the case where the search direction d_k is a coordinate direction, an analysis with the same result can be obtained with d_k chosen uniformly from the unit sphere. Given d_k, a line search is then performed to find an $\alpha_k \in \mathbb{R}$ such that $f(x_{k+1}) - f(x_k)$ is sufficiently small where $x_{k+1} = x_k + \alpha_k d_k$. In fact, as we will see in the next section, for some input parameter $\eta > 0$, the line search is guaranteed to return an α_k such that $|\alpha_k - \alpha^*| \leq \eta$ where $\alpha_* = \min_{\alpha\in\mathbb{R}} f(x_k + d_k\alpha^*)$. Using the fact that the gradients of f are Lipschitz (L) we have

$$f(x_k+\alpha_k d_k) - f(x_k+\alpha^* d_k) \leq \frac{L}{2}||(\alpha_k-\alpha^*)d_k||^2 = \frac{L}{2}|\alpha_k-\alpha^*|^2 \leq \frac{L}{2}\eta^2.$$

If we define $\hat{\alpha_k} = -\frac{\langle\nabla f(x_k), d_k\rangle}{L}$ then we have

$$\begin{aligned}f(x_k+\alpha_k d_k) - f(x_k) &\leq f(x_k+\alpha^* d_k) - f(x_k) + \frac{L}{2}\eta^2 \\ &\leq f(x_k+\hat{\alpha}_k d_k) - f(x_k) + \frac{L}{2}\eta^2 \leq -\frac{\langle\nabla f(x_k), d_k\rangle^2}{2L} + \frac{L}{2}\eta^2\end{aligned}$$

where the last line follows from applying the fact that the gradients are Lipschitz (L). Arranging the bound and taking the expectation with respect to d_k we get

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] - \tfrac{L}{2}\eta^2 \leq \mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\mathbb{E}\left[||\nabla f(x_k)||^2\right]}{2nL} \leq \mathbb{E}\left[f(x_k) - f(x^*)\right]\left(1 - \tfrac{\tau}{4nL}\right)$$

where the second inequality follows from the fact that f is strongly convex (τ). If we define $\rho_k := \mathbb{E}\left[f(x_k) - f(x^*)\right]$ then we equivalently have

$$\rho_{k+1} - \frac{2nL^2\eta^2}{\tau} \leq \left(1 - \frac{\tau}{4nL}\right)\left(\rho_k - \frac{2nL^2\eta^2}{\tau}\right) \leq \left(1 - \frac{\tau}{4nL}\right)^k\left(\rho_0 - \frac{2nL^2\eta^2}{\tau}\right)$$

which leads to the following result.

Theorem 5. *Let $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ with $\mathcal{B} = \mathbb{R}^n$. For any $\eta > 0$ assume the line search returns an α_k that is within η of the optimal after at most $T_\ell(\eta)$ queries from the pairwise comparison oracle. If x_K is an estimate of $x^* = \arg\min_x f(x)$ after requesting no more than K pairwise comparisons, then*

$$\sup_f \mathbb{E}[f(x_K) - f(x_*)] \leq \tfrac{4nL^2\eta^2}{\tau} \qquad \textit{whenever} \qquad K \geq \tfrac{4nL}{\tau} \log\left(\tfrac{f(x_0)-f(x^*)}{\eta^2 2nL^2/\tau}\right) T_\ell(\eta)$$

where the expectation is with respect to the random choice of d_k at each iteration.

This implies that if we wish $\sup_f \mathbb{E}[f(x_K) - f(x_*)] \leq \epsilon$ it suffices to take $\eta = \sqrt{\frac{\epsilon\tau}{4nL^2}}$ so that at most $\frac{4nL}{\tau} \log\left(\frac{f(x_0)-f(x^*)}{\epsilon/2}\right) T_\ell\left(\sqrt{\frac{\epsilon\tau}{4nL^2}}\right)$ pairwise comparisons are requested.

5.2 Line search

This section is concerned with minimizing a function $f(x_k+\alpha_k d_k)$ over some $\alpha_k \in \mathbb{R}$. In particular, we wish to find an $\alpha_k \in \mathbb{R}$ such that $|\alpha_k - \alpha^*| \leq \eta$ where $\alpha_* = \min_{\alpha\in\mathbb{R}} f(x_k + d_k\alpha^*)$. First assume that the function comparison oracle makes no errors. The line search operates by maintaining a pair of boundary points α^+, α^- such that if at some iterate we have $\alpha^* \in [\alpha^-, \alpha^+]$ then at the next iterate, we are guaranteed that α^* is still contained inside the boundary points but $|\alpha^+ - \alpha^-| \leftarrow \frac{1}{2}|\alpha^+ - \alpha^-|$. An initial set of boundary points $\alpha^+ > 0$ and $\alpha^- < 0$ are found using simple binary search. Thus, regardless of how far away or close α^* is, we converge to it exponentially fast. Exploiting the fact that f is strongly convex (τ) with Lipschitz (L) gradients we can bound how far away or close α^* is from our initial iterate.

Theorem 6. *Let $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ with $\mathcal{B} = \mathbb{R}^n$ and let C_f be a function comparison oracle that makes no errors. Let $x \in \mathbb{R}^n$ be an initial position and let $d \in \mathbb{R}^n$ be a search direction with $||d|| = 1$. If α_K is an estimate of $\alpha^* = \arg\min_\alpha f(x + d\alpha)$ that is output from the line search after requesting no more than K pairwise comparisons, then for any $\eta > 0$*

$$|\alpha_K - \alpha^*| \leq \eta \qquad \textit{whenever} \qquad K \geq 2\log_2\left(\tfrac{256L(f(x)-f(x+d\,\alpha^*))}{\tau^2\eta^2}\right).$$

5.3 Making the line search robust to errors

Now assume that the responses from the pairwise comparison oracle are only probably correct in accordance with the model introduced above. Essentially, the robust procedure runs the line search as if the oracle made no errors except that each time a comparison is needed, the oracle is repeatedly queried until we can be confident about the true direction of the comparison. This strategy applied to active learning is well known because of its simplicity and its ability to adapt to unknown noise conditions [24]. However, we mention that when used in this way, this sampling procedure is known to be sub-optimal so in practice, one may want to implement a more efficient approach like that of [21]. Nevertheless, we have the following lemma.

Lemma 2. *[24] For any $x, y \in \mathcal{B}$ with $\mathbb{P}\left(C_f(x,y) = sign\{f(y) - f(x)\}\right) = p$, with probability at least $1-\delta$ the coin-tossing algorithm of [24] correctly identifies the sign of $\mathbb{E}\left[C_f(x,y)\right]$ and requests no more than $\frac{\log(2/\delta)}{4|1/2-p|^2}\log_2\left(\frac{\log(2/\delta)}{4|1/2-p|^2}\right)$ pairwise comparisons.*

It would be convenient if we could simply apply the result of Lemma 2 to our line search procedure. Unfortunately, if we do this there is no guarantee that $|f(y) - f(x)|$ is bounded below so for the case when $\kappa > 1$, it would be impossible to lower bound $|1/2 - p|$ in the lemma. To account for this, we will sample at multiple locations per iteration as opposed to just two in the noiseless algorithm to ensure that we can always lower bound $|1/2 - p|$. Intuitively, strong convexity ensures that f cannot be arbitrarily flat so for any three equally spaced points x, y, z on the line d_k, if $f(x)$ is equal to $f(y)$, then it follows that the absolute difference between $f(x)$ and $f(z)$ must be bounded away from zero. Applying this idea and union bounding over the total number of times one must call the coin-tossing algorithm, one finds that with probability at least $1-\delta$, the total number of calls to the pairwise comparison oracle over the course of the whole algorithm does not exceed $\widetilde{O}\left(\frac{nL}{\tau}\left(\frac{n}{\epsilon}\right)^{2(\kappa-1)}\log^2\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)\log(n/\delta)\right)$. By finding a $T > 0$ that satisfies this bound for any ϵ we see that this is equivalent to a rate of $O\left(n\log(n/\delta)\left(\frac{n}{T}\right)^{\frac{1}{2(\kappa-1)}}\right)$ for $\kappa > 1$ and $O\left(\exp\left\{-c\sqrt{\frac{T}{n\log(n/\delta)}}\right\}\right)$ for $\kappa = 1$, ignoring polylog factors.

References

[1] T. Eitrich and B. Lang. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of computational and applied mathematics*, 196(2):425–436, 2006.

[2] R. Oeuvray and M. Bierlaire. A new derivative-free algorithm for the medical image registration problem. *International Journal of Modelling and Simulation*, 27(2):115–124, 2007.

[3] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to derivative-free optimization*, volume 8. Society for Industrial Mathematics, 2009.

[4] Warren B. Powell and Ilya O. Ryzhov. *Optimal Learning*. John Wiley and Sons, 2012.

[5] Y. Nesterov. Random gradient-free minimization of convex functions. *CORE Discussion Papers*, 2011.

[6] N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *Arxiv preprint arXiv:0912.3995*, 2009.

[7] R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[8] A. Agarwal, D.P. Foster, D. Hsu, S.M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *Arxiv preprint arXiv:1107.1744*, 2011.

[9] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574, 2009.

[10] V. Protasov. Algorithms for approximate calculation of the minimum of a convex function from its values. *Mathematical Notes*, 59:69–74, 1996. 10.1007/BF02312467.

[11] M. Raginsky and A. Rakhlin. Information-based complexity, feedback, and dynamics in convex programming. *Information Theory, IEEE Transactions on*, (99):1–1, 2011.

[12] L.L. Thurstone. A law of comparative judgment. *Psychological Review; Psychological Review*, 34(4):273, 1927.

[13] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 2012.

[14] K.G. Jamieson and R.D. Nowak. Active ranking using pairwise comparisons. *Neural Information Processing Systems (NIPS)*, 2011.

[15] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[16] A. Agarwal, D.P. Foster, D. Hsu, S.M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *Arxiv preprint arXiv:1107.1744*, 2011.

[17] A. Agarwal, P.L. Bartlett, P. Ravikumar, and M.J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, (99):1–1, 2010.

[18] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory (COLT)*, 2010.

[19] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. 2012.

[20] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[21] R.M. Castro and R.D. Nowak. Minimax bounds for active learning. *Information Theory, IEEE Transactions on*, 54(5):2339–2353, 2008.

[22] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

[23] R.P. Brent. *Algorithms for minimization without derivatives*. Dover Pubns, 2002.

[24] M. Kääriäinen. Active learning in the non-realizable case. In *Algorithmic Learning Theory*, pages 63–77. Springer, 2006.