6 Conclusion

This paper presented lower bounds on the performance of derivative-free optimization for (i) an oracle that provides noisy function evaluations and (ii) an oracle that provides probably correct boolean comparisons between function evaluations. Our results were proven for the class of strongly convex functions but because this class is a subset of all, possibly non-convex functions, our lower bounds hold for much larger classes as well. Under both oracle models we showed that the expected error decays like $\Omega\left((n/T)^{1/2}\right)$. Furthermore, for the class of strongly convex functions with Lipschitz gradients, we proposed an algorithm that achieves a rate of $\widetilde{O}\left(n(n/T)^{1/2}\right)$ for both oracle models which shows that the lower bounds are tight with respect to the dependence on the number of iterations T and no more than a factor of n off in terms of the dimension.

A number of open questions still remain. In particular, one would like to resolve the gap between the lower and upper bounds with respect to the dependence on the dimension. Due to real world constraints, it is also desirable to extend the pairwise comparison algorithm to operate under the conditions of constrained optimization where $\mathcal{B}$ is a convex, proper subset of $\mathbb{R}^d$. Also, while the analysis of our algorithm relies heavily on the assumption that the function is strongly convex with Lipschitz gradients, it is unclear whether these assumptions are necessary to achieve the same rates of convergence. Developing a practical algorithm that achieves our lower bounds and does not suffer from these limiting assumptions would be a significant contribution.

A Bounds on (κ, μ, δ_0) for some distributions

In this section we relate the function evaluation oracle to the function comparison oracle for some common distributions. That is, if $E_f(x) = f(x) + w$ for some random variable w, we lower bound the probability $\eta(y, x) := \mathbb{P}(\text{sign}\{E_f(y) - E_f(x)\} = \text{sign}\{f(y) - f(x)\})$ in terms of the parameterization of (1).

Lemma 3. *Let w be a Gaussian random variable with mean zero and variance σ^2. Then* $\eta(y, x) \geq \frac{1}{2} + \min\left\{\frac{1}{\sqrt{2\pi e}}, \frac{1}{\sqrt{4\pi\sigma^2 e}}|f(y) - f(x)|\right\}$.

Proof. Notice that $\eta(y, x) = \mathbb{P}(Z + |f(y) - f(x)|/\sqrt{2\sigma^2} \geq 0)$ where Z is a standard normal. The result follows by lower bounding the density of Z by $\frac{1}{\sqrt{2\pi e}}\mathbf{1}\{|Z| \leq 1\}$ and integrating where $\mathbf{1}\{\cdot\}$ is equal to one when its arguments are true and zero otherwise. □

We say w is a 2-sided gamma distributed random variable if its density is given by $\frac{\beta^\alpha}{2\Gamma(\alpha)}|x|^{\alpha-1}e^{-\beta|x|}$ for $x \in [-\infty, \infty]$ and $\alpha, \beta > 0$. Note that this distribution is unimodal only for $\alpha \in (0, 1]$ and is equal to a Laplace distribution for $\alpha = 1$. This distribution has variance $\sigma^2 = \alpha/\beta^2$.

Lemma 4. *Let w be a 2-sided gamma distributed random variable with parameters $\alpha \in (0, 1]$ and $\beta > 0$. Then* $\eta(y, x) \geq \frac{1}{2} + \min\left\{\frac{1}{4\alpha^2\Gamma(\alpha)^2}\left(\frac{\alpha}{e}\right)^{2\alpha}, \frac{(\beta/2e)^{2\alpha}}{4\alpha^2\Gamma(\alpha)^2}|f(y) - f(x)|^{2\alpha}\right\}$.

Proof. Let $E_f(y) = f(y) + w$ and $E_f(x) = f(x) + w'$ where w and w' are i.i.d. 2-sided gamma distributed random variables. If we lower bound $e^{-\beta|x|}$ with $e^{-\alpha}\mathbf{1}\{|x| \leq \alpha/\beta\}$ and integrate we find that $\mathbb{P}(-t/2 \leq w \leq 0) \geq \min\left\{\frac{1}{2\alpha\Gamma(\alpha)}\left(\frac{\alpha}{e}\right)^\alpha, \frac{(\beta/e)^\alpha}{2\alpha\Gamma(\alpha)}(t/2)^\alpha\right\}$. And by the symmetry and independence of w and w' we have $\mathbb{P}(-t \leq w - w') \geq \frac{1}{2} + \mathbb{P}(-t/2 \leq w \leq 0)\mathbb{P}(-t/2 \leq w \leq 0)$. □

While the bound in the lemma immediately above can be shown to be loose, these two lemmas are sufficient to show that the entire range of $\kappa \in (1, 2]$ is possible.

B Upper Bounds - Extended

The algorithm that achieves the upper bound using a pairwise comparison oracle is a combination of a few standard techniques and methods pulled from the convex optimization and statistical learning literature. The algorithm can be 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 that is known in the literature as the 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].

B.1 Coordinate descent algorithm

n-dimensional Pairwise comparison algorithm
Input: $x_0 \in \mathbb{R}^n$, $\eta \geq 0$
For k=0,1,2,...
  Choose $d_k = \mathbf{e}_i$ for $i \in \{1, \ldots, n\}$ chosen uniformly at random
  Obtain α_k from a line-search such that
    $|\alpha_k - \alpha^*| \leq \eta$ where $\alpha^* = \arg\min_\alpha f(x_k + \alpha d_k)$
  $x_{k+1} = x_k + \alpha_k d_k$
end

Figure 1: Algorithm to minimize a convex function in d dimensions. Here $\mathbf{e}_i$ is understood to be a vector of all zeros with a one in the ith position.

Theorem 7. *Let $f \in \mathcal{F}_{\tau,L,\mathcal{B}}$ with $\mathcal{B} = \mathbb{R}^n$. For any $\eta > 0$ assume the line search in the algorithm of Figure 1 requires 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*

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

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

Proof. First note that $||d_k|| = 1$ for all k with probability 1. Because the gradients of f are Lipschitz (L) we have from Taylor's theorem

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), \alpha_k d_k \rangle + \frac{\alpha_k^2 L}{2}.$$

Note that the right-hand-side is convex in α_k and is minimized by

$$\hat{\alpha_k} = -\frac{\langle \nabla f(x_k), d_k \rangle}{L}.$$

However, recalling how α_k is chosen, if $\alpha^* = \arg\min_\alpha f(x_k + \alpha d_k)$ then we have

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

This implies

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

Taking the expectation with respect to d_k, we have

$$\begin{aligned}\mathbb{E}\left[f(x_{k+1})\right] &\leq \mathbb{E}\left[f(x_k)\right] - \mathbb{E}\left[\frac{\langle \nabla f(x_k), d_k\rangle^2}{2L}\right] + \frac{L}{2}\eta^2 \\ &= \mathbb{E}\left[f(x_k)\right] - \mathbb{E}\left[\mathbb{E}\left[\frac{\langle \nabla f(x_k), d_k\rangle^2}{2L}\middle| d_0, \ldots, d_{k-1}\right]\right] + \frac{L}{2}\eta^2 \\ &= \mathbb{E}\left[f(x_k)\right] - \mathbb{E}\left[\frac{||\nabla f(x_k)||^2}{2nL}\right] + \frac{L}{2}\eta^2\end{aligned}$$

where we applied the law of iterated expectation. Let $x^* = \arg\min_x f(x)$ and note that x^* is a unique minimizer by strong convexity (τ). Using the previous calculation we have

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

where the second inequality follows from

$$\begin{aligned}\left(f(x_k) - f(x^*)\right)^2 \leq& \left(\langle \nabla f(x_k), x_k - x^*\rangle\right)^2 \\ \leq& ||\nabla f(x_k)||^2 ||x_k - x^*||^2 \leq ||\nabla f(x_k)||^2 \left(\frac{\tau}{2}\right)^{-1}\left(f(x_k) - f(x^*)\right).\end{aligned}$$

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 completes the proof. □

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.

B.2 Line search

This section is concerned with minimizing a function $f(x_k + \alpha d_k)$ over some $\alpha \in \mathbb{R}$. Because we are minimizing over a single variable, α, we will restart the indexing at 0 such that the line search algorithm produces a sequence $\alpha_0, \alpha_1, \ldots, \alpha_{K'}$. This indexing should not be confused with the indexing of the iterates $x_1, x_2, \ldots, x_K$. We will first present an algorithm that assumes the pairwise comparison oracle makes no errors and then extend the algorithm to account for the noise model introduced in Section 2.

Consider the algorithm of Figure 2. At each iteration, one is guaranteed to eliminate at least $1/2$ the search space at each iteration such that at least $1/4$ the search space is discarded for every pairwise comparison that is requested. However, with a slight modification to the algorithm, one can guarantee a greater fraction of removal (see the golden section line-search algorithm). We use this sub-optimal version for simplicity because it will help provide intuition for how the robust version of the algorithm works.

Theorem 8. *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 algorithm of Figure 2 after requesting no more than K pairwise comparisons, then for any $\eta > 0$*

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

Proof. First note that if α_K is output from the algorithm, we have $\frac{1}{2}|\alpha_K - \alpha^*| \leq |\alpha_K^+ - \alpha_K^-| \leq \frac{1}{2}\eta$, as desired.

We will handle the cases when $|\alpha^*|$ is greater than one and less than one separately. First assume that $|\alpha^*| \geq 1$. Using the fact that f is strongly convex (τ), it is straightforward to show that immediately

One Dimensional Pairwise comparison algorithm

Input: $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$, $\eta > 0$

Initialize: $\alpha_0 = 0$, $\alpha_0^+ = \alpha_0 + 1$, $\alpha_0^- = \alpha_0 - 1$, $k = 0$

If $C_f(x, x + d\,\alpha_0^+) > 0$ and $C_f(x, x + d\,\alpha_0^-) < 0$

  $\alpha_0^+ = 0$

end

If $C_f(x, x + d\,\alpha_0^-) > 0$ and $C_f(x, x + d\,\alpha_0^+) < 0$

  $\alpha_0^- = 0$

end

While $C_f(x, x + d\,\alpha_k^+) < 0$

  $\alpha_{k+1}^+ = 2\alpha_k^+$, $k = k + 1$

end

While $C_f(x, x + d\,\alpha_k^-) < 0$

  $\alpha_{k+1}^- = 2\alpha_k^-$, $k = k + 1$

end

$\alpha_k = \frac{1}{2}(\alpha_k^- + \alpha_k^+)$

While $|\alpha_k^+ - \alpha_k^-| \geq \eta/2$

  if $C_f(x + d\,\alpha_k, x + d\,\frac{1}{2}(\alpha_k + \alpha_k^+)) < 0$

    $\alpha_{k+1} = \frac{1}{2}(\alpha_k + \alpha_k^+)$, $\alpha_{k+1}^+ = \alpha_k^+$, $\alpha_{k+1}^- = \alpha_k$

  else if $C_f(x + d\,\alpha_k, x + d\,\frac{1}{2}(\alpha_k + \alpha_k^-)) < 0$

    $\alpha_{k+1} = \frac{1}{2}(\alpha_k + \alpha_k^-)$, $\alpha_{k+1}^+ = \alpha_k$, $\alpha_{k+1}^- = \alpha_k^-$

  else

    $\alpha_{k+1} = \alpha_k$, $\alpha_{k+1}^+ = \frac{1}{2}(\alpha_k + \alpha_k^+)$, $\alpha_{k+1}^- = \frac{1}{2}(\alpha_k + \alpha_k^-)$

  end

end

Output: α_k

Figure 2: Algorithm to minimize a convex function in one dimension.

after exiting the initial while loops, (i) at most $2 + \frac{1}{2}\log_2\left(\frac{8}{\tau}\left(f(x) - f(x + d\,\alpha^*)\right)\right)$ pairwise comparisons were requested, (ii) $\alpha_* \in [\alpha_k^-,\, \alpha_k^+]$, and (iii) $|\alpha_k^+ - \alpha_k^-| \leq \left(\frac{8}{\tau}\left(f(x) - f(x + d\,\alpha^*)\right)\right)^{1/2}$. We also have that $\alpha_* \in [\alpha_{k+1}^-,\, \alpha_{k+1}^+]$ if $\alpha_* \in [\alpha_k^-,\, \alpha_k^+]$ for all k. Thus, it follows that

$$|\alpha_{k+l}^+ - \alpha_{k+l}^-| = 2^{-l}|\alpha_k^+ - \alpha_k^-| \leq 2^{-l}\left(\frac{8}{\tau}\left(f(x) - f(x + d\,\alpha^*)\right)\right)^{1/2}.$$

To make the right-hand-side less than or equal to $\eta/2$, set $l = \log_2\left(\frac{\left(\frac{8}{\tau}(f(x) - f(x + d\,\alpha^*))\right)^{1/2}}{\eta/2}\right)$. This brings the total number of pairwise comparison requests to no more than $2\log_2\left(\frac{32(f(x) - f(x + d\,\alpha^*))}{\tau\eta}\right)$.

Now assume that $|\alpha^*| \leq 1$. A straightforward calculation shows that the while loops will terminate after requesting at most $2 + \frac{1}{2}\log_2\left(\frac{L}{\tau}\right)$ pairwise comparisons. And immediately after exiting the while loops we have $|\alpha_k^+ - \alpha_k^-| \leq 2$. It follows by the same arguments of above that if we want $|\alpha_{k+l}^+ - \alpha_{k+l}^-| \leq \eta/2$ it suffices to set $l = \log_2\left(\frac{4}{\eta}\right)$. This brings the total number of pairwise comparison requests to no more than $2\log_2\left(\frac{8L}{\tau\eta}\right)$. For sufficiently small η both cases are positive and the result follows from adding the two. □

This implies that if the function comparison oracle makes no errors and it is given an iterate x_k and direction d_k then $T_\ell\left(\sqrt{\frac{\epsilon\tau}{4nL^2}}\right) \leq 2\log_2\left(\frac{2048nL^2(f(x_k) - f(x_k + d_k\,\alpha^*))}{\tau^3\epsilon}\right)$ which brings the total number of pairwise comparisons requested to at most $\frac{8nL}{\tau}\log\left(\frac{f(x_0) - f(x^*)}{\epsilon/2}\right)\log_2\left(\frac{2048nL^2\max_k(f(x_k) - f(x_k + d_k\,\alpha^*))}{\tau^3\epsilon}\right)$.

B.3 Proof of Theorem 2

We now introduce a line search algorithm that is robust to a function comparison oracle that makes errors. Essentially, the algorithm consists of nothing more than repeatedly querying the same random pairwise 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]. Consider the subroutine of Figure 3.

Repeated querying subroutine
Input: $x, y \in \mathbb{R}^n$, $\delta > 0$
Initialize: $S = \emptyset$, $l = -1$
do
  $l = l + 1$
  $\Delta_l = \sqrt{\frac{(l+1)\log(2/\delta)}{2^l}}$
  $S = S \cup \{2^l \text{ i.i.d. draws of } C_f(x, y)\}$
while $\left|\frac{1}{2}\sum_{e_i \in S} e_i\right| - \Delta_l < 0$
return sign $\left\{\sum_{e_i \in S} e_i\right\}$.

Figure 3: Subroutine that estimates $\mathbb{E}\left[C_f(x, y)\right]$ by repeatedly querying the random variable.

Lemma 5. *[24] For any $x, y \in \mathbb{R}^n$ with $\mathbb{P}\left(C_f(x, y) = \text{sign}\{f(y) - f(x)\}\right) = p$, then with probability at least $1-\delta$ the algorithm of Figure 3 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 the algorithm of Figure 2. 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 four points per iteration as opposed to just two in the noiseless algorithm to ensure that we can always lower bound $|1/2 - p|$. We will see that the algorithm and analysis naturally adapts to when $\kappa = 1$ or $\kappa > 1$.

Consider the following modification to the algorithm of Figure 2. We discuss the sampling process that takes place in $[\alpha_k,\, \alpha_k^+]$ but it is understood that the same process is repeated symmetrically in $[\alpha_k^-,\, \alpha_k]$. We begin with the first two `while` loops. Instead of repeatedly sampling $C_f(x, x + d\,\alpha_k^+)$ we will have two sampling procedures running in parallel that repeatedly compare α_k to α_k^+ and α_k to $2\alpha_k^+$. As soon as the repeated sampling procedure terminates for one of them we terminate the second sampling strategy and proceed with what the noiseless algorithm would do with α_k^+ assigned to be the sampling location that finished first. Once we're out of the initial `while` loops, instead of comparing α_k to $\frac{1}{2}(\alpha_k + \alpha_k^+)$ repeatedly, we will repeatedly compare α_k to $\frac{1}{3}(\alpha_k + \alpha_k^+)$ and α_k to $\frac{2}{3}(\alpha_k + \alpha_k^+)$. Again, we will treat the location that finishes its sampling first as $\frac{1}{2}(\alpha_k + \alpha_k^+)$ in the noiseless algorithm.

If we perform this procedure every iteration, then at each iteration we are guaranteed to remove at least $1/3$ the search space, as opposed to $1/2$ in the noiseless case, so we realize that the number of iterations of the robust algorithm is within a constant factor of the number of iterations of the noiseless algorithm. However, unlike the noiseless case where at most two pairwise comparisons were requested at each iteration, we must now apply Lemma 2 to determine the number of pairwise comparisons that are requested per iteration.

Intuitively, the repeated sampling procedure requests the most pairwise comparisons when the distance between the two function evaluations being compared smallest. This corresponds to when the distance between probe points is smallest, i.e. when $\eta/2 \leq |\alpha_k - \alpha^*| \leq \eta$. By considering this worst case, we can bound the number of pairwise comparisons that are requested

at any iteration. By strong convexity (τ) we find through a straightforward calculation that $\max\left\{|f(x+d\,\alpha_k)-f(x+d\,\frac{2}{3}(\alpha_k+\alpha_k^+))|, |f(x+d\,\alpha_k)-f(x+d\,\frac{1}{3}(\alpha_k+\alpha_k^+))|\right\} \geq \frac{\tau}{18}\eta^2$ for all k. This implies $|1/2-p| \geq \mu\left(\frac{\tau}{18}\eta^2\right)^{\kappa-1}$ so that on on any given call to the repeated querying subroutine, with probability at least $1-\delta$ the subroutine requests no more than $\widetilde{O}\left(\frac{\log(1/\delta)}{(\tau\eta^2)^{2(\kappa-1)}}\right)$ pairwise comparisons. However, because we want the total number of calls to the subroutine to hold with probability $1-\delta$, not just one, we must union bound over 4 pairwise comparisons per iteration times the number of iterations per line search times the number of line searches. This brings the total number of calls to the repeated query subroutine to no more than $4 \times \frac{3}{2}\log_2\left(\frac{256L\max_k(f(x_k)-f(x_k+d_k\,\alpha_k^*))}{\tau^2\eta^2}\right) \times \frac{4nL}{\tau}\log\left(\frac{f(x_0)-f(x^*)}{\eta^2 2nL^2/\tau}\right) = O\left(n\frac{L}{\tau}\log^2\left(\frac{f(x_0)-f(x^*)}{n\eta^2}\right)\right)$. If we set $\eta = \left(\frac{\epsilon\tau}{4nL^2}\right)^{1/2}$ so that $\mathbb{E}\left[f(x_K)-f(x^*)\right] \leq \epsilon$ by Theorem 7, then the total number of requested pairwise comparisons 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.