Algorithms and matching lower bounds for approximately-convex optimization

Yuanzhi Li
Department of Computer Science
Princeton University
Princeton, NJ, 08450
yuanzhil@cs.princeton.edu

Andrej Risteski
Department of Computer Science
Princeton University
Princeton, NJ, 08450
risteski@cs.princeton.edu

Abstract

In recent years, a rapidly increasing number of applications in practice requires optimizing non-convex objectives, like training neural networks, learning graphical models, maximum likelihood estimation. Though simple heuristics such as gradient descent with very few modifications tend to work well, theoretical understanding is very weak.

We consider possibly the most natural class of non-convex functions where one could hope to obtain provable guarantees: functions that are "approximately convex", i.e. functions $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ for which there exists a *convex function* f such that for all x, $|\tilde{f}(x) - f(x)| \leq \Delta$ for a fixed value Δ. We then want to minimize $\tilde{f}$, i.e. output a point $\tilde{x}$ such that $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$.

It is quite natural to conjecture that for fixed ϵ, the problem gets harder for larger Δ, however, the exact dependency of ϵ and Δ is not known. In this paper, we significantly improve the known lower bound on Δ as a function of ϵ and an algorithm matching this lower bound for a natural class of convex bodies. More precisely, we identify a function $T : \mathbb{R}^+ \to \mathbb{R}^+$ such that when $\Delta = O(T(\epsilon))$, we can give an algorithm that outputs a point $\tilde{x}$ such that $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$ within time $poly\left(d, \frac{1}{\epsilon}\right)$. On the other hand, when $\Delta = \Omega(T(\epsilon))$, we also prove an *information theoretic* lower bound that any algorithm that outputs such a $\tilde{x}$ must use *super polynomial* number of evaluations of $\tilde{f}$.

1 Introduction

Optimization of convex functions over a convex domain is a well studied problem in machine learning, where a variety of algorithms exist to solve the problem efficiently. However, in recent years, practitioners face ever more often non-convex objectives – e.g. training neural networks, learning graphical models, clustering data, maximum likelihood estimation etc. Albeit simple heuristics such as gradient descent with few modifications usually work very well, theoretical understanding in these settings are still largely open.

The most natural class of non-convex functions where one could hope to obtain provable guarantees is functions that are "approximately convex": functions $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ for which there exists a *convex function* f such that for all x, $|\tilde{f}(x) - f(x)| \leq \Delta$ for a fixed value Δ. In this paper, we focus on *zero order optimization* of $\tilde{f}$: an algorithm that outputs a point $\tilde{x}$ such that $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$, where the algorithm in the course of its execution is allowed to pick points $x \in \mathbb{R}^d$ and query the value of $\tilde{f}(x)$.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Trivially, one can solve the problem by constructing a ϵ-net and search through all the net points. However, such an algorithm requires $\Omega\left(\frac{1}{\epsilon}\right)^d$ evaluations of $\tilde{f}$, which is highly inefficient in high dimension. In this paper, we are interested in *efficient algorithms*: algorithms that run in time $poly\left(d, \frac{1}{\epsilon}\right)$ (in particular, this implies the algorithm makes $poly\left(d, \frac{1}{\epsilon}\right)$ evaluations of $\tilde{f}$).

One extreme case of the problem is $\Delta = 0$, which is just standard convex optimization, where algorithms exist to solve it in polynomial time for every $\epsilon > 0$. However, even when Δ is any quantity > 0, none of these algorithms extend without modification. (Indeed, we are not imposing *any* structure on $\tilde{f} - f$ like stochasticity.) Of course, when $\Delta = +\infty$, the problem includes any non-convex optimization, where we cannot hope for an efficient solution for any finite ϵ. Therefore, the crucial quantity to study is the optimal tradeoff of ϵ and Δ: For which ϵ, Δ the problem can be solved in polynomial time, and for which it can not.

In this paper, we study the rate of Δ as a function of ϵ: We identify a function $T : \mathbb{R}^+ \to \mathbb{R}^+$ such that when $\Delta = O(T(\epsilon))$, we can give an algorithm that outputs a point $\tilde{x}$ such that $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$ within time $poly\left(d, \frac{1}{\epsilon}\right)$ over a natural class of *well-conditioned* convex bodies. On the other hand, when $\Delta = \tilde{\Omega}(T(\epsilon))$[1], we also prove an *information theoretic* lower bound that any algorithm outputs such $\tilde{x}$ must use *super polynomial* number of evaluations of $\tilde{f}$. Our result can be summarized as the following two theorems:

Theorem (Algorithmic upper bound, informal)**.** *There exists an algorithm $\mathcal{A}$ that for any function $\tilde{f}$ over a* well-conditioned *convex set in $\mathbb{R}^d$ of diameter 1 which is Δ close to an* 1-Lipschitz convex function [2] *f, and*

$$\Delta = O\left(\max\left(\left\{\frac{\epsilon^2}{\sqrt{d}}, \frac{\epsilon}{d}\right\}\right)\right)$$

$\mathcal{A}$ finds a point $\tilde{x}$ such that $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$ within time $poly\left(d, \frac{1}{\epsilon}\right)$

The notion of well-conditioning will formally be defined in section 3, but intuitively captures the notion that the convex body "curves" in all directions to a good extent.

Theorem (Information theoretic lower bound, informal)**.** *For every algorithm $\mathcal{A}$, every d, Δ, ϵ with*

$$\Delta = \tilde{\Omega}\left(\max\left\{\frac{\epsilon^2}{\sqrt{d}}, \frac{\epsilon}{d}\right\}\right)$$

there exists a function $\tilde{f}$ on a convex set in $\mathbb{R}^d$ of diameter 1, and $\tilde{f}$ is Δ close to an 1-Lipschitz convex function *f, such that $\mathcal{A}$* can not *find a point $\tilde{x}$ with $\tilde{f}(\tilde{x}) \leq \min_x \tilde{f}(x) + \epsilon$ in $poly\left(d, \frac{1}{\epsilon}\right)$ evaluations of $\tilde{f}$.*

2 Prior work

To the best of our knowledge, there are three works on the problem of approximately convex optimization, which we summarize briefly below.

On the algorithmic side, the classical paper by [DKS14] considered optimizing *smooth* convex functions over convex bodies with *smooth* boundaries. More precisely, they assume a bound on both the gradient and the Hessian of F. Furthermore, they assume that for every small ball centered at a point in the body, a large proportion of the volume of the ball lies in the body. Their algorithm is local search: they show that for a sufficiently small r, in a ball of radius r there is with high probability a point which has a smaller value than the current one, as long as the current value is sufficiently larger than the optimum. For *constant-smooth* functions only, their algorithm applies when $\Delta = O(\frac{\epsilon}{\sqrt{d}})$.

Also on the algorithmic side, the work by [BLNR15] considers 1-Lipschitz functions, but their algorithm only applies to the case where $\Delta = O(\frac{\epsilon}{d})$ (so not optimal unless $\epsilon = O(\frac{1}{\sqrt{d}})$). Their methods rely on sampling log-concave distribution via hit and run walks. The crucial idea is to show that for approximately convex functions, one needs to sample from "approximately log-concave"

[1] The $\tilde{\Omega}$ notation hides $polylog(d/\epsilon)$ factors.

[2] The assumptions on the diameter of $\mathcal{K}$ and the Lipschitz condition are for convenience of stating the results. (See Section **??** to extend to arbitrary diameter and Lipschitz constant)

distributions, which they show can be done by a form of rejection sampling together with classical methods for sampling log-concave distributions.

Finally, [SV15] consider information theoretic lower bounds. They show that when $\Delta = 1/d^{1/2-\delta}$ no algorithm can, in polynomial time, achieve achieve $\epsilon = \frac{1}{2} - \delta$, when optimizing a convex function over the hypercube. This translates to a super polynomial information theoretic lower bound when $\Delta = \Omega(\frac{\epsilon}{\sqrt{d}})$. They additionally give lower bounds when the approximately convex function is multiplicatively, rather than additively, close to a convex function. [3]

We also note a related problem is zero-order optimization, where the goal is to minimize a function we only have value oracle access to. The algorithmic motivations here come from various applications where we only have black-box access to the function we are optimizing, and there is a classical line of work on characterizing the oracle complexity of convex optimization.[NY83, NS, DJWW15]. In all of these settings however, the oracles are either noiseless, or the noise is stochastic, usually because the target application is in bandit optimization. [AD10, AFH$^+$11, Sha12]

3 Overview of results

Formally, we will consider the following scenario.

Definition 3.1. *A function $\tilde{f} : \mathcal{K} \to \mathbb{R}^d$ will be called* Δ-approximately convex *if there exists a 1-Lipschitz convex function $f : \mathcal{K} \to \mathbb{R}^d$, s.t. $\forall x \in \mathcal{K}, |\tilde{f}(x) - f(x)| \leq \Delta$.*

For ease of exposition, we also assume that $\mathcal{K}$ has diameter 1[4]. We consider the problem of *optimizing* $\tilde{f}$, more precisely, we are interesting in finding a point $\tilde{x} \in \mathcal{K}$, such that

$$\tilde{f}(\tilde{x}) \leq \min_{x \in \mathcal{K}} \tilde{f}(x) + \epsilon$$

We give the following results:

Theorem 3.1 (Information theoretic lower bound)**.** *For very constant $c \geq 1$, there exists a constant d_c such that for every algorithm $\mathcal{A}$, every $d \geq d_c$, there exists a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ with diameter 1, an Δ-approximate convex function $\tilde{f} : \mathcal{K} \to \mathbb{R}$ and $\epsilon \in [0, 1/64)$ [5] such that*

$$\Delta \geq \max\left\{\frac{\epsilon^2}{\sqrt{d}}, \frac{\epsilon}{d}\right\} \times \left(13c \log \frac{d}{\epsilon}\right)^2$$

Such that $\mathcal{A}$ fails *to output, with probability $\geq 1/2$, a point $\tilde{x} \in \mathcal{K}$ with $\tilde{f}(\tilde{x}) \leq \min_{x \in \mathcal{K}}\{\tilde{f}(x)\} + \epsilon$ in $o((\frac{d}{\epsilon})^c)$ time.*

In order to state the upper bounds, we will need the definition of a well-conditioned body:

Definition 3.2 (μ-well-conditioned)**.** *A convex body $\mathcal{K}$ is said to be μ-well-conditioned for $\mu \geq 1$, if there exists a function $F : \mathbb{R}^d \to \mathbb{R}$ such that $\mathcal{K} = \{x | F(x) \leq 0\}$ and for every $x \in \partial\mathcal{K}$: $\frac{\|\nabla^2 F(x)\|_2}{\|\nabla F(x)\|_2} \leq \mu$.*

This notion of well-conditioning of a convex body to the best of our knowledge has not been defined before, but it intuitively captures the notion that the convex body should "curve" in all directions to a certain extent. In particular, the unit ball has $\mu = 1$.

Theorem 3.2 (Algorithmic upper bound)**.** *Let d be a positive integer, $\delta > 0$ be a positive real number, ϵ, Δ be two positive real number such that*

$$\Delta \leq \max\left\{\frac{\epsilon^2}{\mu\sqrt{d}}, \frac{\epsilon}{d}\right\} \times \frac{1}{16348}$$

Then there exists an algorithm $\mathcal{A}$ such that on given any Δ-approximate convex function $\tilde{f}$ over a μ-rounded convex set $\mathcal{K} \subseteq \mathbb{R}^d$ of diameter 1, $\mathcal{A}$ returns a point $\tilde{x} \in \mathcal{K}$ with probability $1 - \delta$ in time $poly\left(d, \frac{1}{\epsilon}, \log \frac{1}{\delta}\right)$ such that

$$\tilde{f}(\tilde{x}) \leq \min_{x \in \mathcal{K}} \tilde{f}(x) + \epsilon$$

[3] Though these are not too difficult to derive from the additive ones, considering the convex body has diameter bounded by 1.

[4] Generalizing to arbitrary Lipschitz constants and diameters is discussed in Section 6.

[5] Since we normalize f to be 1-Lipschitz and $\mathcal{K}$ to have diameter 1, the problem is only interesting for $\epsilon \leq 1$

For the reader wishing to digest a condition-free version of the above result, the following weaker result is also true (and much easier to prove):

Theorem 3.3 (Algorithmic upper bound (condition-free))**.** *Let d be a positive integer, $\delta > 0$ be a positive real number, ϵ, Δ be two positive real number such that*

$$\Delta \leq \max\left\{\frac{\epsilon^2}{\sqrt{d}}, \frac{\epsilon}{d}\right\} \times \frac{1}{16348}$$

Then there exists an algorithm $\mathcal{A}$ such that on given any Δ-approximate convex function $\tilde{f}$ over a μ-rounded convex set $\mathcal{K} \subseteq \mathbb{R}^d$ of diameter 1, $\mathcal{A}$ returns a point $\tilde{x} \in \mathcal{K}$ with probability $1 - \delta$ in time $poly\left(d, \frac{1}{\epsilon}, \log\frac{1}{\delta}\right)$ such that

$$\tilde{f}(\tilde{x}) \leq \min_{x \in S(\mathcal{K}, -\epsilon)} \tilde{f}(x) + \epsilon$$

Where $S(\mathcal{K}, -\epsilon) = \{x \in \mathcal{K} | \mathbb{B}_\epsilon(x) \subseteq \mathcal{K}\}$

The result merely states that we can output a value that competes with points "well-inside" the convex body – around which a ball of radius of ϵ still lies inside the body.

The assumptions on the diameter of $\mathcal{K}$ and the Lipschitz condition are for convenience of stating the results. It's quite easy to extend both the lower and upper bounds to an arbitrary diameter and Lipschitz constant, as we discuss in Section 6.

3.1 Proof techniques

We briefly outline the proof techniques we use. We proceed with the information theoretic lower bound first. The idea behind the proof is the following. We will construct a function $G(x)$ and a family of convex functions $\{f_w(x)\}$ depending on a direction $w \in \mathcal{S}^d$ ($\mathcal{S}^d$ is the unit sphere in $\mathbb{R}^d$). On one hand, the minimal value of G and f_w are quite different: $\min_x G(x) \geq 0$, and $\min_x f_w(x) \leq -2\epsilon$. On the other hand, the approximately convex function $\tilde{f}_w(x)$ for $f_w(x)$ we consider will be such that $\tilde{f}_w(x) = G(x)$ except in a very small cone around w. Picking w at random, no algorithm with small number of queries will, with high probability, every query a point in this cone. Therefore, the algorithm will proceed as if the function is $G(x)$ and fail to optimize $\tilde{f}_w$.

Proceeding to the algorithmic result, since [BLNR15] already shows the existence of an efficient algorithm when $\Delta = O(\frac{\epsilon}{d})$, we only need to give an algorithm that solves the problem when $\Delta = \Omega(\frac{\epsilon}{d})$ and $\Delta = O(\frac{\epsilon^2}{\sqrt{d}})$ (i.e. when ϵ, Δ are large). There are two main ideas for the algorithm. First, we show that the gradient of a *smoothed* version of $\tilde{f}_w$ (in the spirit of [FKM05]) at any point x will be correlated with $x^* - x$, where $x^* = \operatorname{argmin}_{x \in \mathcal{K}} \tilde{f}_w(x)$. The above strategy will however require averaging the value of $\tilde{f}_w$ along a ball of radius ϵ, which in many cases will not be contained in $\mathcal{K}$ (especially when ϵ is large). Therefore, we come up with a way to *extend* $\tilde{f}_w$ outside of $\mathcal{K}$ in a manner that maintains the correlation with $x^* - x$.

4 Information-theoretic lower bound

In this section, we present the proof of Theorem 3.1.

The idea is to construct a function $G(x)$, a family of convex functions $\{f_w(x)\}$ depending on a direction $w \in \mathcal{S}^d$, such that $\min_x G(x) \geq 0$, $\min_x f_w(x) \leq -2\epsilon$, and an approximately convex $\tilde{f}_w(x)$ for $f_w(x)$ such that $\tilde{f}_w(x) = G(x)$ except in a very small "critical" region depending on w. Picking w at random, we want to argue that the algorithm will with high probability not query the critical region. The convex body $\mathcal{K}$ used in the lower bound will be arguably the simplest convex body imaginable: the unit ball $\mathbb{B}_1(0)$.

We might hope to prove a lower bound for even a linear function f_w for a start, similarly as in [SV15]. A reasonable candidate construction is the following: we set $f_w(x) = -\epsilon\langle w, x\rangle$ for some random chosen unit vector w and define $\tilde{f}(x) = 0$ when $|\langle x, w\rangle| \leq \frac{\log\frac{d}{\epsilon}}{\sqrt{d}}\|x\|_2$ and $\tilde{f}(x) = f_w(x)$ otherwise.[6]

[6]For the proof sketch only, to maintain ease of reading all of the inequalities we state will be only correct up to constants. In the actual proofs we will be completely formal.

Observe, this translates to $\Delta = \frac{\log \frac{d}{\epsilon}}{\sqrt{d}}\epsilon$. It's a standard concentration of measure fact that for "most" of the points x in the unit ball, $|\langle x, w\rangle| \leq \frac{\log \frac{d}{\epsilon}}{\sqrt{d}}\|x\|_2$. This implies that any algorithm that makes a polynomial number of queries to $\tilde{f}$ will with high probability see 0 in all of the queries, but clearly $\min \tilde{f}(x) = -\epsilon$. However, this idea fails to generalize to optimal range as $\Delta = \frac{1}{\sqrt{d}}\epsilon$ is tight for linear, even smooth functions.[7]

In order to obtain the optimal bound, we need to modify the construction to a non-linear, non-smooth function. We will, in a certain sense, "hide" a random linear function inside a non-linear function. For a random unit vector w, we consider two regions inside the unit ball: a core $\mathcal{C} = \mathbb{B}_r(0)$ for $r = \max\{\epsilon, \frac{1}{\sqrt{d}}\}$, and a "critical angle" $\mathcal{A} = \{x \mid |\langle x, w\rangle| \geq \frac{\log \frac{d}{\epsilon}}{\sqrt{d}}\|x\|_2\}$. The *convex* function f will look like $\|x\|_2^{1+\alpha}$ for some $\alpha > 0$ outside $\mathcal{C} \cup \mathcal{A}$ and $-\epsilon\langle w, x\rangle$ for $x \in \mathcal{C} \cup \mathcal{A}$. We construct $\tilde{f}$ as $\tilde{f} = f$ when $f(x)$ is sufficiently large (e.g. $|f(x)| > \frac{\Delta}{2}$) and $\frac{\Delta}{2}$ otherwise. Clearly, such $\tilde{f}$ obtain its minimal at point w, with $\tilde{f}(w) = -\epsilon$. However, since $\tilde{f} = \|x\|_2^{1+\alpha}$ outside $\mathcal{C}$ or $\mathcal{A}$, the algorithm needs either query $\mathcal{A}$ or query $\mathcal{C} \cap \mathcal{A}^c$ to detect w. The former happens with exponentially small probability in high dimensions, and for any $x \in \mathcal{C} \cap \mathcal{A}^c, |f(x)| = \epsilon|\langle w, x\rangle| \leq \frac{\epsilon \log \frac{d}{\epsilon}}{\sqrt{d}}\|x\|_2 \leq \frac{\epsilon \log \frac{d}{\epsilon}}{\sqrt{d}} r \leq \max\{\frac{\epsilon^2}{\sqrt{d}}, \frac{\epsilon}{d}\} \times \log \frac{d}{\epsilon} \leq \frac{\Delta}{2}$, which implies that $\tilde{f}(x) = \frac{\Delta}{2}$. Therefore, the algorithm will fail with high probability.

Now, we move on to the detailed of the constructions. We will consider $\mathcal{K} = \mathbb{B}_{\frac{1}{2}}(0)$: the ball of radius $\frac{1}{2}$ in $\mathbb{R}^d$ centered at 0. [8]

4.1 The family $\{f_w(x)\}$

Before delving into the construction we need the following definition:

Definition 4.1 (Lower Convex Envelope (LCE))**.** *Given a set $\mathcal{S} \subseteq \mathbb{R}^d$, a function $F : \mathcal{S} \to \mathbb{R}$, define the lower convex envelope $F_{\mathsf{LCE}} = \mathsf{LCE}(F)$ as a function $F_{\mathsf{LCE}} : \mathbb{R}^d \to \mathbb{R}$ such that for every $x \in \mathbb{R}^d$,*

$$F_{\mathsf{LCE}}(x) = \max_{y \in \mathcal{S}}\{\langle x - y, \nabla F(y)\rangle + F(y)\}$$

Proposition 4.1. *$LCE(F)$ is convex.*

Proof. LCE(F) is the pointwise maximum of linear functions, so the claim follows. □

Remark : The LCE of a function F is a function defined over the entire $\mathbb{R}^d$, while the input function F is only defined in a set $\mathcal{S}$ (not necessarily convex set). When the input function F is convex, $\mathsf{LCE}(F)$ can be considered as an extension of F to the entire $\mathbb{R}^d$.

To define the family $f_w(x)$, we will need four parameters: a power factor $\alpha > 0$, a shrinking factor β, and a radius factor $\gamma > 0$, and a vector $w \in \mathbb{R}^d$ such that $\|w\|_2 = \frac{1}{2}$, which we specify in a bit.

Construction 4.1. *Given w, α, β, γ, define the* core $\mathcal{C} = \mathbb{B}_\gamma(0)$*, the* critical angle $\mathcal{A} = \{x \mid |\langle x, w\rangle| \geq \beta\|x\|_2\}$ *and let $\mathcal{H} = \mathcal{K} \cap \overline{\mathcal{C}} \cap \overline{\mathcal{A}}$. Let $\tilde{h} : \mathcal{H} \to \mathbb{R}$ be defined as*

$$\tilde{h}(x) = \frac{1}{2}\|x\|_2^{1+\alpha}$$

and define $l_w(x) = -8\epsilon\langle x, w\rangle$. Finally let $f_w : \mathcal{K} \to \mathbb{R}^d$ as

$$f_w(x) = \max\left\{\tilde{h}_{\mathsf{LCE}}(x), l_w(x)\right\}$$

Where $\tilde{h}_{\mathsf{LCE}} = \mathsf{LCE}(\tilde{h})$ as in Definition 4.1.

We then construct the "hard" function $\tilde{f}_w$ as the following:

Construction 4.2. *Consider the function $\tilde{f}_w : \mathcal{K} \to \mathbb{R}$:*

$$\tilde{f}_w(x) = \begin{cases} f_w(x) & \text{if } x \in \mathcal{K} \cap \left(\overline{\mathcal{C}} \cup \mathcal{A}\right); \\ \max\{f_w(x), \frac{1}{2}\Delta\} & \text{otherwise.} \end{cases}$$

[7] This follows from the results in [DKS14]

[8] We pick $\mathbb{B}_{\frac{1}{2}}(0)$ instead of the unit ball in order to ensure the diameter is 1.

Consider the following settings of the parameters β, γ, α (depending on the magnitude of ϵ):

- Case 1, $\frac{1}{\sqrt{d}} \leq \epsilon \leq \frac{1}{(\log d)^2}$: $\beta = \frac{\sqrt{c \log \frac{d}{\epsilon}}}{\sqrt{d}}$, $\gamma = 10c\epsilon(\log \frac{d}{\epsilon})^{1.5}$, $\alpha = \frac{1}{\log(1/\gamma)}$.
- Case 2, $\epsilon \leq \frac{1}{\sqrt{d}}$: $\beta = \frac{\sqrt{c \log d/\epsilon}}{\sqrt{d}}$, $\gamma = \frac{10c}{\sqrt{d}}(\log d/\epsilon)^{3/2}$, $\alpha = \frac{1}{\log(1/\gamma)}$.
- Case 3, $\frac{1}{64} \geq \epsilon \geq \frac{1}{(\log d)^2}$: $\beta = \frac{\sqrt{c \log d}}{\sqrt{d}}$, $\gamma = \frac{1}{2}$, $\alpha = 1$.

For convenience of delivering the ideas, we divide the proof into each of these three cases. The proofs are essentially the same in each case, with minor modifications in the calculation.

4.2 Case 1: $\frac{1}{\sqrt{d}} \leq \epsilon \leq \frac{1}{(\log d)^2}$

Let us set $\beta = \frac{\sqrt{c \log \frac{d}{\epsilon}}}{\sqrt{d}}$, $\gamma = 10c\epsilon(\log \frac{d}{\epsilon})^{1.5}$, $\alpha = \frac{1}{\log(1/\gamma)}$.

Here, we consider sufficiently large d_c such that when $d \geq d_c$, $\gamma < \frac{1}{2}$ so $\alpha < 1$

Following the the proof outline, we first show the minimum of f_w is small, in particular we will show $f_w(w) \leq -2\epsilon$. Note that $f_w(x) = \max\left\{\tilde{h}_{\mathsf{LCE}}(x), l_w(x)\right\}$ and $l_w(w) = -8\epsilon\|w\|_2^2 = -2\epsilon$, therefore, we can just focus on $\tilde{h}_{\mathsf{LCE}}(w)$:

We will need the following proposition:

Proposition 4.2. *Let $\tilde{h}$ be the function defined in Construction 4.1, then we have:*

$$\nabla\tilde{h}(x) = \frac{1+\alpha}{2}\|x\|_2^{\alpha-1}x$$

Lemma 4.1. $\tilde{h}_{\mathit{LCE}}(w) \leq \frac{1}{2}(1+\alpha)\beta\gamma^\alpha - \frac{1}{2}\alpha\gamma^{1+\alpha}$

Proof of Lemma 4.1. By the definition of LCE, we have:

$$\begin{aligned}
&\tilde{h}_{\mathsf{LCE}}(w) = \max_{x\in\mathcal{H}}\{\langle w-x, \nabla\tilde{h}(x)\rangle + \tilde{h}(x)\} \\
= &\max_{x\in\mathcal{H}}\left\{\left\langle w-x, \frac{1+\alpha}{2}\|x\|_2^{\alpha-1}x\right\rangle + \frac{1}{2}\|x\|_2^{1+\alpha}\right\} \\
= &\max_{x\in\mathcal{H}}\left\{\frac{1+\alpha}{2}\|x\|_2^{\alpha-1}\langle w, x\rangle - \frac{\alpha}{2}\|x\|_2^{1+\alpha}\right\} \\
\leq &\max_{x\in\mathcal{H}}\left\{\frac{(1+\alpha)\beta}{2}\|x\|_2^{\alpha} - \frac{\alpha}{2}\|x\|_2^{1+\alpha}\right\}
\end{aligned}$$

where the last inequality is due to the fact that $x \in \mathcal{H}$, so: $|\langle x, w\rangle| \leq \beta\|x\|_2$. Now consider the function $g(y) = \frac{(1+\alpha)\beta}{2}y^\alpha - \frac{\alpha}{2}y^{1+\alpha}$: we know that $g'(y) = \frac{(1+\alpha)\alpha\beta}{2}y^{\alpha-1} - \frac{\alpha(1+\alpha)}{2}y^\alpha$. Notice that $g'(y) = \frac{(1+\alpha)}{2}y^{\alpha-1}(\beta - y)$. For $x \in \mathcal{H}$, since $x \notin \mathcal{C}$, we have

$$\|x\|_2 \geq \gamma = 10c\epsilon\left(\log\frac{d}{\epsilon}\right)^{1.5} \geq \frac{10c\left(\log\frac{d}{\epsilon}\right)^{1.5}}{\sqrt{d}} \geq 10\beta \geq \beta$$

Where the second inequality is due to $\epsilon \geq \frac{1}{\sqrt{d}}$ and the second last inequality is due to $c \geq 1, \epsilon \leq 1$.

Therefore, $g'(\|x\|_2) < 0$, which implies that $g(\|x\|_2)$ in $\mathcal{H}$ increases as $\|x\|_2$ decreases. Hence:

$$\tilde{h}_{\mathsf{LCE}}(w) \leq g(\gamma) = \frac{(1+\alpha)\beta\gamma^\alpha}{2} - \frac{\alpha\gamma^{1+\alpha}}{2}$$

□

As a corollary, we get the following:

Corollary 4.1. $f_w(w) = -2\epsilon$

Proof of Corollary 4.1. Note that $l_w(w) = -2\epsilon$, moreover, since $\gamma^\alpha = \gamma^{\frac{1}{-\log\gamma}} = 2^{\frac{\log\gamma}{-\log\gamma}} = \frac{1}{2}$, have:

$$\begin{aligned}\frac{1}{2}(1+\alpha)\beta\gamma^\alpha - \frac{1}{2}\alpha\gamma^{1+\alpha} &= \frac{1}{4}(1+\alpha)\beta - \frac{1}{4}\alpha\gamma \\ &\leq \frac{1}{4}(1+\alpha)\beta - \frac{1}{4}\alpha\gamma\end{aligned}$$

By $\gamma \geq 10\beta$ we can conclude:

$$\frac{1}{2}(1+\alpha)\beta\gamma^\alpha - \frac{1}{2}\alpha\gamma^{1+\alpha} \leq \frac{\beta}{4} - \frac{9}{40}\alpha\gamma$$

By the definition of α, we have: for every $c \geq 1$, there exists $d_c = 1600c$, for every $d \geq d_c$ and every $1 \geq \epsilon \geq \frac{1}{\sqrt{d}}$

$$\alpha = \frac{1}{\log(1/\gamma)} = \frac{1}{\log\frac{1}{10c\epsilon(\log d/\epsilon)^{1.5}}} \geq \frac{1}{\log d}$$

Therefore,

$$\alpha\gamma \geq \frac{10c\epsilon(\log d/\epsilon)^{1.5}}{\log d} \geq 10c\epsilon\sqrt{\log\frac{d}{\epsilon}}$$

By $\beta = \frac{\sqrt{c\log\frac{d}{\epsilon}}}{\sqrt{d}}$, we can conclude that

$$\frac{1}{2}(1+\alpha)\beta\gamma^\alpha - \frac{1}{2}\alpha\gamma^{1+\alpha} \leq \frac{\beta}{4} - \frac{9}{40}\alpha\gamma \leq \frac{1}{4}\frac{\sqrt{c\log\frac{d}{\epsilon}}}{\sqrt{d}} - \frac{9}{4}c\epsilon\sqrt{\log\frac{d}{\epsilon}} \leq -2\epsilon$$

The last inequality is due to $\epsilon \geq \frac{1}{\sqrt{d}}$, $c \geq 1$.

Which implies $\tilde{h}_{\mathsf{LCE}}(w) \leq -2\epsilon$. This completes the proof. □

4.3 $f_w(x) = \tilde{h}(x)$ in $\mathcal{H}$

We now show that $f_w(x) = \tilde{h}(x)$ in $\mathcal{H}$:

Lemma 4.2. *For every $x \in \mathcal{H}$, $f_w(x) = \tilde{h}(x)$*

Proof of Lemma 4.2. Note that $\tilde{h}$ is a convex function defined on $\mathcal{H}$, therefore, $\tilde{h}_{\mathsf{LCE}} = \tilde{h}$ on $\mathcal{H}$. Now we consider l_w on $\mathcal{H}$: Since $\forall x \in \mathcal{H}$, $|\langle x, w\rangle| \leq \beta\|x\|_2$, $l_w(x) = -8\epsilon\langle x, w\rangle \leq 8\epsilon\beta\|x\|_2 \leq \frac{1}{2}\gamma^\alpha\|x\|_2$ where the last inequality follows by noticing that for every $c \geq 1$, there exists $d_c = 8192c$, such that for every $d \geq d_c$ and $1 \geq \epsilon \geq \frac{1}{\sqrt{d}}$

$$8\beta\epsilon \leq 8\beta \leq 8\frac{\sqrt{c\log\frac{d}{\epsilon}}}{\sqrt{d}} \leq 8\frac{\sqrt{c\log(d^{1.5})}}{\sqrt{d}} \leq \frac{1}{4} = \frac{\gamma^\alpha}{2}$$

Moreover, for $x \in \mathcal{H}$, we know that $\frac{1}{2}\gamma^\alpha\|x\|_2 \leq \tilde{h}(x) = \tilde{h}_{\mathsf{LCE}}(x)$, therefore, $f_w(x) = \tilde{h}(x) = \frac{1}{2}\|x\|_2^{1+\alpha}$.

□

4.4 Approximate convexity and constructing $G(x)$

Finally, we show that $\tilde{f}_w$ is indeed a Δ-approximately convex, by showing $\forall x \in \mathcal{K}, |f_w - \tilde{f}_w| \leq \Delta$ and f_w is 1-Lipschitz and convex. Note that by construction, $\tilde{f}_w$ only differs from f_w on $\mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$, so it's sufficient to focus on the set $\mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$.

We will need the following simple bound on the value of l_w in $\mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$.

Lemma 4.3. *For every $x \in \mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$, we have $|l_w(x)| \leq \frac{1}{2}\Delta$.*

Proof of Lemma 4.3. For $x \in \mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$, $|l_w(x)| = 8\epsilon|\langle x, w\rangle| \leq 8\epsilon\beta\|x\|_2 \leq 8\epsilon\beta\gamma \leq \frac{1}{2}\Delta$.

The last inequality holds since $\Delta \geq 160\frac{\epsilon^2}{\sqrt{d}}\left(c\log\frac{d}{\epsilon}\right)^2$ and $c \geq 1$ □

Proposition 4.3. *$\tilde{f}_w$ is a Δ-approximately convex.*

Proof. First, notice that f_w is convex and 1-Lipschitz.

Since f_w is a point-wise maximum of $\tilde{h}_{\mathsf{LCE}}$ and l_w, it is convex.

Furthermore, we claim both $\tilde{h}_{\mathsf{LCE}}$ and l_w are 1-Lipschitz, which will imply that f_w is 1-Lipschitz. Indeed, l_w is 1-Lipschitz by definition, and the norm of the gradient of $\tilde{h}_{\mathsf{LCE}}$ is upper bounded by $\frac{1}{2}(1+\alpha) \leq 1$ since $\alpha < 1$.

Now we argue $\max_{x\in\mathcal{K}} |f_w(x) - \tilde{f}_w(x)| \leq \Delta$.

By Lemma 4.3, we know that when $x \in \mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$, $l_w(x) \geq -\frac{1}{2}\Delta$. Since $f_w(x) = \max\{l_w(x), \tilde{h}_{\mathsf{LCE}}(x)\}$, we have $f_w(x) \geq -\frac{1}{2}\Delta$. The claim follows. □

Now we construct $G(x)$, which does not depend on w, we want to show that for an algorithm with small number of queries of $\tilde{f}_w$, it can not distinguish f_w from this function.

Construction 4.3. *Let $G : \mathcal{K} \to \mathbb{R}$ be defined as:*

$$G(x) = \begin{cases} \max\left\{\frac{1+\alpha}{4}\|x\|_2 - \frac{\alpha}{4}\gamma, \frac{1}{2}\Delta\right\} & \text{if } x \in \mathcal{K} \cap \mathcal{C}\,; \\ \frac{1}{2}\|x\|_2^{1+\alpha} & \text{otherwise}. \end{cases}$$

Lemma 4.4. *$G(x) \geq 0$ and $\{x \in \mathcal{K} \mid G(x) \neq \tilde{f}_w(x)\} \subseteq \mathcal{A}$*

Proof of Lemma 4.4. By Lemma 4.2, $\tilde{f}_w(x) = f_w(x) = \tilde{h}(x)$ for $x \in \mathcal{H}$. Moreover, by definition, $\tilde{h}(x) = \frac{1}{2}\|x\|_2^{1+\alpha}$. Therefore, $\tilde{f}_w(x) = G(x)$ for $x \in \mathcal{H}$. So we only need consider $\mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$. Note that $|l_w(x)| \leq \frac{1}{2}\Delta$ in $\mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$ by Lemma 4.3. Therefore, for $x \in \mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$, $\tilde{f}(x) = \max\left\{\tilde{h}_{\mathsf{LCE}}(x), \frac{1}{2}\Delta\right\}$.

We conclude the proof by noticing that for every $x \in \mathcal{K} \cap \mathcal{C} \cap \overline{\mathcal{A}}$ (recall $\mathcal{K}$ the ball of radius $1/2$ centered at 0), there exists $y \in \mathcal{H}$ such that $\|y\|_2 = \gamma$ and $\langle x, y\rangle = \|x\|_2\|y\|_2$. Which implies that

$$\begin{aligned} \tilde{h}_{\mathsf{LCE}}(x) &= \max_{x'\in\mathcal{H}}\{\langle x - x', \nabla\tilde{h}(x')\rangle + \tilde{h}(x')\} \\ &= \max_{x'\in\mathcal{H}}\left\{\frac{1+\alpha}{2}\|x'\|_2^{\alpha-1}\langle x, x'\rangle - \frac{\alpha}{2}\|x'\|_2^{1+\alpha}\right\} \\ &= \frac{1+\alpha}{2}\|y\|_2^{\alpha-1}\langle x, y\rangle - \frac{\alpha}{2}\|y\|_2^{1+\alpha} \\ &= \frac{1+\alpha}{4}\|x\|_2 - \frac{\alpha}{4}\gamma \end{aligned}$$

Where the third equality follows from the following observations:

(1). When $\|x'\|_2$ is fixed, the best x' should be aligned with x: $\langle x', x\rangle = \|x'\|_2\|x\|_2$.

(2). defining

$$g(s) = \frac{1+\alpha}{2}\|x\|_2 s^\alpha - \frac{\alpha}{2}s^{1+\alpha}$$

We have:

$$g'(s) = \frac{(1+\alpha)\alpha(\|x\|_2 - s)}{2}s^{\alpha-1} < 0$$

For $\|x\|_2 \leq \gamma \leq s$. □

4.5 Putting everything together

Proof of Theorem 3.1 for $\frac{1}{\sqrt{d}} \leq \epsilon \leq \frac{1}{(\log d)^2}$. With everything prior to this set up, the final claim is somewhat standard. We want to show that no algorithm can, with probability $\geq \frac{1}{2}$, output a point x, s.t. $\tilde{f}_w(x) \leq \min_x \tilde{f}_w(x) + \epsilon$. Since we know that $\tilde{f}_w(x)$ agrees with $G(x)$ everywhere except in $\mathcal{K} \cap \mathcal{A}$, and $G(x)$ satisfies $\min_x G(x) \geq \min_x \tilde{f}_w(x) + \epsilon$, we only need to show that with high probability, any polynomial time algorithm will not query any point in $\mathcal{K} \cap \mathcal{A}$.

Consider a (potentially) randomized algorithm A, making random choices $R_1, R_2, \ldots, R_m$. Conditioned on a particular choice of randomness $r_1, r_2, \ldots, r_m$, for a random choice of w, each r_i lies in $\mathcal{A}$ with probability at most $\exp(-c \log(d/\epsilon))$, by a standard Gaussian tail bound. Union bounding, since $m = o((\frac{d}{\epsilon})^c)$ for an algorithm that runs in time $o((\frac{d}{\epsilon})^c)$, the probability that at least of the queries of A lies in $\mathcal{A}$ is at most $\frac{1}{2}$.

But the claim is true for any choice $r_1, r_2, \ldots, r_m$ of the randomness, by averaging, the claim holds for $r_1, r_2, \ldots, r_m$ being sampled according to the randomness of the algorithm.

□

4.6 Case 2: $\epsilon \leq \frac{1}{\sqrt{d}}$

In the case where $\epsilon \leq \frac{1}{\sqrt{d}}$, the proof proceeds exactly the same as before, but with a different setting of the parameters. In this case we set $\beta = \frac{\sqrt{c \log d/\epsilon}}{\sqrt{d}}$, $\gamma = \frac{10c}{\sqrt{d}}(\log d/\epsilon)^{3/2}$, and $\alpha = \frac{1}{\log(1/\gamma)}$.

We proceed to verify that each Lemma still holds under this parameter setting. We first show that Lemma 4.1 still holds.

Proof of Lemma 4.1. Following the same calculation as in previous section, it is enough to show that $\gamma \geq \beta$ in this setting. We can check that for every $c \geq 1$

$$\frac{\beta}{\gamma} = \frac{\sqrt{c \log d/\epsilon}}{10c(\log d/\epsilon)^{1.5}} = \frac{\sqrt{c}}{10c \log d/\epsilon} < \frac{\sqrt{c}}{10c} \leq \frac{1}{10} < 1$$

□

We then check that Corollary 4.1 still holds.

Proof of Corollary 4.1. Following the same calculation in the previous section, it is sufficient to show that

$$\frac{\beta}{4} - \frac{9}{40}\alpha\gamma \leq -2\varepsilon$$

Putting in the specific numbers, we obtain: since $\epsilon \leq \frac{1}{\sqrt{d}}$

$$\beta = \frac{\sqrt{c \log d/\epsilon}}{\sqrt{d}}$$

On the other hand, we have:

$$\alpha\gamma \geq \frac{10c(\log d/\epsilon)^{1.5}}{\sqrt{d} \log \sqrt{d/\epsilon}} \geq \frac{10c\sqrt{\log d/\epsilon}}{\sqrt{d}}$$

Therefore, for $c \geq 1$

$$\frac{\beta}{4} - \frac{9}{40}\alpha\gamma \leq \frac{\sqrt{c \log d/\epsilon}}{4\sqrt{d}} - \frac{9c\sqrt{\log d/\epsilon}}{4\sqrt{d}} \leq -\frac{2}{\sqrt{d}} \leq -2\epsilon$$

□

We then check that Lemma 4.2 still holds.

Proof of Lemma 4.2. Following the same calculation in the previous section, it is sufficient to show that

$$8\beta\epsilon \leq \frac{1}{4}$$

Putting in the specific bound, we know that for every $c \geq 0$, there exists a $d_c = 32c^2$ such that for every $d \geq d_c$,

$$8\beta\epsilon = 8\frac{\sqrt{c\log d/\epsilon}}{\sqrt{d}}\epsilon \leq \frac{1}{4}$$

□

It remains to check that Lemma 4.3 holds.

Proof of Lemma 4.3. Following the same calculation in the previous section, it is sufficient to show that

$$8\epsilon\beta\gamma \leq \frac{1}{2}\Delta$$

Putting in the specific bound, we know that

$$8\beta\epsilon\gamma = \frac{80c^{1.5}\epsilon(\log d/\epsilon)^2}{d} \leq \frac{\Delta}{2}$$

Where the last inequality follows from $\Delta \geq \frac{\epsilon}{d}(13c\log d/\epsilon)^2$. □

Note that Lemma 4.4 holds regardless of the choice of α, β, γ.

4.7 Case 3: $\frac{1}{64} \geq \epsilon \geq \frac{1}{(\log d)^2}$

In this case, we can choose $\beta = \frac{\sqrt{c\log d}}{\sqrt{d}}$, $\gamma = \frac{1}{2}$ and $\alpha = 1$. Actually, since $\mathcal{C} = \mathcal{K}$ in this case, it reduces to having only linear function l_w.

We can check that for sufficiently large $d_c = 4096c^2$, for every $d \geq d_c$, we have: $\beta \leq \frac{\gamma}{10}$; $\frac{\beta}{4} - \frac{9}{40}\alpha\gamma \leq -\frac{1}{10} \leq -2\epsilon$, $8\beta\epsilon \leq \frac{8\sqrt{c\log d}}{\sqrt{d}} \leq \frac{1}{4}$ and $8\beta\epsilon\gamma = \frac{4\epsilon\sqrt{c\log d}}{\sqrt{d}} \leq \frac{1}{2}\Delta$. So all the Lemma 4.1, 4.1, 4.2, 4.3 holds. Which completes the proof.

5 Algorithmic upper bound

As mentioned before, the algorithm in [BLNR15] covers the case when $\Delta = O(\frac{\epsilon}{d})$, so we only need to give an algorithm when $\Delta = \Omega(\frac{\epsilon}{d})$ and $\Delta = O(\frac{\epsilon^2}{d})$. Our approach will not be making use of simulated annealing, but a more robust version of gradient descent. The intuition comes from [FKM05] who use estimates of the gradient of a convex function derived from Stokes' formula:

$$\mathbb{E}_{w\sim\mathcal{S}^d}\left[\frac{d}{r}f(x+rw)w\right] = \int_{\mathbb{B}}\nabla f(x)dx$$

where $w \sim \mathcal{S}^d$ denotes w being a uniform sample from the sphere $\mathcal{S}^d$. Our observation is the gradient estimation is robust to noise if we instead use $\tilde{f}$ in the left hand side. Crucially, robust is *not* in the sense that it approximates the gradient of f, but it preserves the crucial property of the gradient of f we need: $\langle -\nabla f(x), x^* - x\rangle \geq f(x) - f(x^*)$. In words, this means if we move x at direction $-\nabla f(x)$ for a small step, then x will be closer to x^*, and we will show the property is preserved by $\tilde{f}$ when $\Delta \leq \frac{\epsilon^2}{\sqrt{d}}$. Indeed, we have that:

$$\begin{aligned}&\left\langle -\mathbb{E}_{w\sim\mathcal{S}^d}\left[\frac{d}{r}\tilde{f}(x+rw)w\right], x^* - x\right\rangle \\ &\geq -\mathbb{E}_{w\sim\mathcal{S}^d}\left[\left\langle\frac{d}{r}f(x+rw)w, x^*-x\right\rangle\right] - \frac{d\Delta}{r}\mathbb{E}_{w\sim\mathcal{S}^d}\left[|\langle w, x^*-x\rangle|\right]\end{aligned}$$

The usual [FKM05] calculation shows that

$$\mathbb{E}_{w\sim\mathcal{S}^d)}\left[\left\langle\frac{d}{r}f(x+rw)w, x^*-x\right\rangle\right] = \Omega\left(f(x) - f(x^*) - 2r\right)$$

and $\frac{d}{r}\Delta\mathbb{E}_{w\sim U(S^d)}\left[|\langle w, x^*-x\rangle|\right]$ is bounded by $O(\frac{\Delta\sqrt{d}}{r})$, since $\mathbb{E}_{w\sim U(S^d)}\left[|\langle w, x^*-x\rangle|\right] = O(\frac{1}{\sqrt{d}})$. Therefore, we want $f(x) - f(x^*) - 2r \geq \dfrac{\Delta\sqrt{d}}{r}$ whenever $f(x) - f(x^*) \geq \epsilon$. Choosing the optimal parameter leads to $r = \frac{\epsilon}{4}$ and $\Delta \leq \frac{\epsilon^2}{\sqrt{d}}$.

This intuitive calculation basically proves the simple upper bound guarantee (Theorem 3.3). On the other hand, the argument requires sampling from a ball of radius $\Omega(\epsilon)$ around point x. This is problematic when $\epsilon > \frac{1}{\sqrt{d}}$: many convex bodies (e.g. the simplex, L^1 ball after rescaling to diameter one) will not contain a ball of radius even $\frac{1}{\sqrt{d}}$. The idea is then to make the sampling possible by "extending" $\tilde{f}$ outside of $\mathcal{K}$. Namely, we define a new function $g : \mathbb{R}^d \to \mathbb{R}$ such that ($\Pi_{\mathcal{K}}(x)$ is the projection of x to $\mathcal{K}$)

$$g(x) = \tilde{f}(\Pi_{\mathcal{K}}(x)) + d(x, \mathcal{K})$$

$g(x)$ will not be in general convex, but we instead directly bound $\langle \mathbb{E}_{w\sim}\left[\frac{1}{r}g(x+rw)w\right], x - x^*\rangle$ for $x \in \mathcal{K}$ and show that it behaves like $\langle -\nabla f(x), x^* - x\rangle \geq f(x) - f(x^*)$.

Algorithm 1 Noisy Convex Optimization

1: Input: A convex set $\mathcal{K} \subset \mathbb{R}^d$ with $\mathsf{diam}(\mathcal{K}) = 1$ and $0 \in \mathcal{K}$. A Δ-approximate convex function $\tilde{f}$
2: Define: $g : \mathbb{R} \to \mathbb{R}$ as:

$$\tilde{g}(x) = \tilde{f}(\Pi_{\mathcal{K}}(x)) + d(x, \mathcal{K})$$

where $\Pi_{\mathcal{K}}$ is the projection to $\mathcal{K}$ and $d(x, \mathcal{K})$ is the Euclidean distance from x to $\mathcal{K}$.
3: Initial: $x_1 = 0, r = \frac{\epsilon}{128\mu}, \eta = \frac{\epsilon^3}{4194304d^2}, T = \frac{8388608d^2}{\epsilon^4}$.
4: **for** $t = 1, 2, ..., T$ **do**
5: Let $v_t = \tilde{f}(x_t)$.
6: Estimate up to accuracy $\frac{\epsilon}{4194304}$ in l_2 norm (by uniformly randomly sample w):

$$g_t = \mathbb{E}_{w\sim\mathcal{S}^d}\left[\frac{d}{r}\tilde{g}(x_t + rw)w\right]$$

where $w \sim \mathcal{S}^d$ means w is uniform sample from the unit sphere.
7: Update $x_{t+1} = \Pi_{\mathcal{K}}(x_t - \eta g_t)$
8: **end for**
9: Output $\min_{t\in[T]}\{v_t\}$

The rest of this section will be dedicated to showing the following main lemma for Algorithm 1.

Lemma 5.1 (Main, algorithm)**.** *Suppose $\Delta < \frac{\epsilon^2}{16348\sqrt{d}}$, we have: For every $t \in [T]$, if there exists $x^* \in \mathcal{K}$ such that $\tilde{f}(x^*) < \tilde{f}(x_t) - 2\epsilon$, then*

$$\langle -g_t, x^* - x_t\rangle \geq \frac{\epsilon}{64}$$

Assuming this Lemma, we can prove Theorem 3.2.

Proof of Theorem 3.2. We first focus on the number of iterations:

For every $t \geq 1$, suppose $\tilde{f}(x^*) < \tilde{f}(x_t) - 2\epsilon$, then we have: (since $\|g_t\| \leq 2d/r \leq \frac{256d}{\epsilon}$)

$$\begin{aligned}
\|x^* - x_{t+1}\|_2^2 &\leq \|x^* - (x_t - \eta g_t)\|_2^2 \\
&= \|x^* - x_t\|_2^2 - 2\eta\langle x^* - x_t, g_t\rangle + \eta^2\|g_t\|_2^2 \\
&\leq \|x^* - x_t\|_2^2 - \frac{\eta\epsilon}{64} + \eta^2\frac{65536d^2}{\epsilon^2} \\
&\leq \|x^* - x_t\|_2^2 - \frac{\epsilon^4}{8388608d^2} + \frac{\epsilon^4}{4194304d^2} \\
&= \|x^* - x_t\|_2^2 - \frac{\epsilon^4}{8388608d^2}
\end{aligned}$$

Since originally $\|x^* - x_1\| \leq 1$, the algorithm ends in $poly(d, \frac{1}{\epsilon})$ iterations.

Now we consider the sample complexity. Since we know that

$$\left\| \frac{d}{r} \tilde{g}(x_t + rw)w \right\|_2 \leq \frac{64d}{\epsilon}$$

By standard concentration bound we know that we need $poly(d, \frac{1}{\epsilon})$ samples to estimate the expectation up to error $\frac{\epsilon}{2097152}$ per iteration. □

Proof of Lemma 5.1. Proceeding towards applying the FKM framework, we have, since $|g(x) - h(x)| \leq \Delta$:

$$\mathbb{E}_{w \sim U(S^d)} \left[\left\langle -\frac{d}{r} g(x_t + rw)w, e \right\rangle \right] \geq \mathbb{E}_{w \sim U(S^d)} \left[\left\langle -\frac{d}{r} h(x_t + rw)w, e \right\rangle \right] - \frac{d}{r} \Delta \mathbb{E}_{w \sim U(S^d)} |\langle w, e \rangle|$$

We want to lower bound the quantity on the RHS by $\frac{\epsilon}{64}$. Since the absolute value of the second term is upper bounded by $\frac{d}{r}\Delta(2d^{-1/2}) \leq \frac{\epsilon}{32}$, it is enough to bound the first term by $\frac{\epsilon}{16}$.

Let's proceed to the first term. By applying Stokes' theorem, we have:

$$\begin{aligned} & \mathbb{E}_{w \sim U(S^d)} \left[\left\langle -\frac{d}{r} h(x_t + rw)w, e^* \right\rangle \right] \\ = & -\left\langle \frac{1}{\mathsf{Vol}[\mathbb{B}_1(0)]} \int_{z \in \mathbb{B}_1(0)} \nabla h(x_t + rz) dz, e^* \right\rangle \\ = & -\frac{1}{\mathsf{Vol}[\mathbb{B}_1(0)]} \int_{z \in \mathbb{B}_1(0)} \langle \nabla h(x_t + rz), e^* \rangle \, dz \end{aligned}$$

In order to evaluate gradients of h, we will need to develop machinery for dealing with the projections.

Note the following observation: for any point $y \in \mathbb{R}^d$, $x = \Pi_{\mathcal{K}}(y)$ is given by the solution to the system of equations in x, λ:

$$x + \lambda \nabla F(x) = y, \quad \lambda = \frac{\|y - x\|_2}{\|\nabla F(x)\|_2} \tag{1}$$

For simplicity, we denote x_t to x and let $e = x^* - x$.

Denoting $y_s = y + se, x_s = \Pi_{\mathcal{K}}(y_s)$, we have using (1) and the chain rule:

$$e = \frac{\partial y_s}{\partial s} = \frac{\partial x_s}{\partial s} + \frac{\partial \lambda_s}{\partial s} \nabla F(x) + \lambda_s \nabla^2 F(x) \frac{\partial x_s}{\partial s} \tag{2}$$

We will be evaluating all the partial derivatives at $s = 0$, so as a shorthand, let us denote by $\frac{\partial \lambda}{\partial s}, \frac{\partial x}{\partial s}$ the quantities $\frac{\partial \lambda_s}{\partial s}\Big|_{s=0}, \frac{\partial x_s}{\partial s}\Big|_{s=0}$.

Towards calculating $\frac{\partial h(y_s)}{\partial s}$, we proceed to calculate $\frac{\partial f(x_s)}{\partial s}$ and $\frac{\partial \|y_s - x_s\|_2}{\partial s}$.

For $\frac{\partial f(x_s)}{\partial s}$, by (2) we have:

$$\left. \frac{\partial f(x_s)}{\partial s} \right|_{s=0} = \left\langle \nabla f(x), \frac{\partial x_s}{\partial s} \right\rangle = \langle \nabla f(x), e \rangle - \frac{\partial \lambda}{\partial s} \langle \nabla f(x), \nabla F(x) \rangle - \lambda \nabla f(x)^\top \nabla^2 F(x) \frac{\partial x}{\partial s} \tag{3}$$

On the other hand, we wish to show $\left. \frac{\|y_s - x_s\|_2}{\partial s} \right|_{s=0} = \frac{1}{\|\nabla F(x)\|_2} \langle \nabla F(x), e \rangle$. Using the fact that for any differentiable function $g(x)$, $\frac{d}{dx}\|g(x)\|_2 = \|g(x)\|_2^{-1} \frac{d}{dx} g(x)$ we have

$$\begin{aligned} \left. \frac{\|y_s - x_s\|_2}{\partial s} \right|_{s=0} &= \|y - x\|_2^{-1} \left\langle y - x, \frac{\partial y}{\partial s} - \frac{\partial x}{\partial s} \right\rangle \\ &= \frac{\partial \lambda}{\partial s} \|\nabla F(x)\|_2 + \frac{\lambda}{\|\nabla F(x)\|_2} \nabla F(x)^\top \nabla^2 F(x) \frac{\partial x}{\partial s} \end{aligned} \tag{4}$$

where the second equality follows since by (1), we have $\frac{y - x}{\|y - x\|_2} = \frac{\nabla F(x)}{\|\nabla F(x)\|_2}$.

Since $F(x_s) = 0$ by taking gradients on both sides we have $\langle \nabla F(x), \frac{\partial x}{\partial s} \rangle = 0$ which gives us by (1):

$$\langle \nabla F(x), e \rangle = \frac{\partial \lambda}{\partial s} \|\nabla F(x)\|_2^2 + \lambda \nabla F(x)^\top \nabla^2 F(x) \frac{\partial x}{\partial s} \tag{5}$$

Combine with Equality (4) we get: $\frac{\|y_s - x_s\|_2}{\partial s}\Big|_{s=0} = \frac{1}{\|\nabla F(x)\|_2} \langle \nabla F(x), e \rangle$ as we wanted.

For notational convenience, let's denote $\nabla \tilde{F}(x) = \frac{\nabla F(x)}{\|\nabla F(x)\|_2}$. From the above estimates, putting (3) and (4) together we have:

$$\begin{aligned} \frac{h(y_s)}{\partial s} &= \langle \nabla f(x), e \rangle + \langle \nabla \tilde{F}(x), e \rangle (1 - \langle \nabla f(x), \nabla \tilde{F}(x) \rangle) \\ &\quad + \lambda \left(\nabla \tilde{F}(x) \langle \nabla \tilde{F}(x), f(x) \rangle - \nabla f(x) \right)^\top \nabla^2 F(x) \frac{\partial x}{\partial s} \end{aligned} \tag{6}$$

We will bound each of the terms on the RHS individually. Namely, we show:

$$\langle \nabla f(x), e \rangle \leq -\frac{\epsilon}{2} \tag{7}$$

$$\langle \nabla \tilde{F}(x), e \rangle (1 - \langle \nabla f(x), \nabla \tilde{F}(x) \rangle) \leq 0 \tag{8}$$

$$\lambda \left(\nabla \tilde{F}(x) \langle \nabla \tilde{F}(x), f(x) \rangle - \nabla f(x) \right)^\top \nabla^2 F(x) \frac{\partial x}{\partial s} \leq -\frac{\epsilon}{8} \tag{9}$$

Proceeding to (7), since $f(x^*) \leq f(x) - \epsilon$, $\|x - x^*\|_2 \leq r$ and f is 1-Lipschitz, we know that $f(x^*) \leq f(x) - \frac{\epsilon}{2}$. By convexity of f, we have

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle$$

which by simple rearranging gives $\langle \nabla f(x), e \rangle \leq -\frac{\epsilon}{2}$.

For (8), we know that x^* lies in $\mathcal{K}$, hence using the fact that $\nabla F(x)$ is a normal vector to $\mathcal{K}$ at x, we get $\langle \nabla F(x), e \rangle \leq 0$. Furthermore, since $|\langle \nabla f(x), \nabla \tilde{F}(x) \rangle| \leq 1$, we know that $(1 - \langle \nabla f(x), \nabla F(x) \rangle) \in [0, 2]$, which implies that

$$\langle \nabla \tilde{F}(x), e \rangle \left(1 - \langle \nabla f(x), \nabla \tilde{F}(x) \rangle \right) \leq 0$$

Finally, consider the last term. By (2) and (5) we get:

$$\begin{aligned} (I + \lambda \nabla^2 F(x)) \frac{\partial x_s}{\partial s} &= e - \frac{\partial \lambda}{\partial s} \nabla F(x) \\ &= e - \langle \nabla \tilde{F}(x), e \rangle \nabla \tilde{F}(x) + \langle \nabla \tilde{F}(x), e \rangle \nabla \tilde{F}(x) - \frac{\partial \lambda}{\partial s} \nabla F(x) \\ &= (e - \langle \nabla \tilde{F}(x), e \rangle \nabla \tilde{F}(x)) + \lambda \nabla \tilde{F}(x)^\top \nabla^2 F(x) \frac{\partial x_s}{\partial s} \nabla \tilde{F}(x) \end{aligned}$$

Multiplying on the left and right by $(I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top)$ on both sides of the above equality, and using the fact that $\nabla \tilde{F}(x)^\top \frac{\partial x_s}{\partial s} = 0$ and $(I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top) \nabla \tilde{F}(x) = 0$, we have:

$$(I + \lambda (I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top) \nabla^2 F(x)) \frac{\partial x_s}{\partial s} = (I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top)(e - \langle \nabla \tilde{F}(x), e \rangle \nabla \tilde{F}(x))$$

Denoting $A = (I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top)$, $B = A \nabla^2 F(x)$, we get:

$$A \nabla^2 F(x) \frac{\partial x}{\partial s} = B(I + \lambda B)^{-1} A e$$

We proceed to bound the spectral norm of the RHS (which of course will imply a spectral norm bound on the LHS). Towards that, we first show $\|\lambda B\|_2 \leq \frac{1}{2}$: indeed, by our choice of r, we have $\lambda \|\nabla F(x)\|_2 \leq r \leq \frac{1}{2\mu}$. This implies

$$\|\lambda B\|_2 \leq \lambda \|A\|_2 \|\nabla^2 F(x)\|_2 \leq \|\lambda\|_2 \|\nabla F(x)\|_2 \frac{\|\nabla^2 F(x)\|_2}{\|\nabla F(x)\|_2} \leq r\mu \leq \frac{1}{2} \tag{10}$$

where the first inequality follows by submultiplicativity of the spectral norm, and the second by $\|A\| \leq 1$. Therefore we have:

$$\|B(I + \lambda B)^{-1} A e\|_2 \leq 2\|B\|_2 \leq 2\|A\|_2 \|\nabla^2 F(x)\|_2 \leq 2\mu \|\nabla F(x)\|_2$$

where the first inequality and second inequality follow by (10) and the submultiplicativity of the spectral norm; the third is by well-conditioning of the convex body. Finally, this implies $|\nabla f(x)^\top (I - \nabla \tilde{F}(x) \nabla \tilde{F}(x)^\top) \nabla^2 F(x) \frac{\partial x}{\partial s}| \leq 2\mu \|\nabla F(x)\|_2$

Putting (7), (8), (9) together, we get $\left.\frac{h(y_s)}{\partial s}\right|_{s=0} \leq -\frac{\epsilon}{8}$.

Using the above fact and the 1-Lipschitzness of h, we get

$$\langle \nabla h(y), e^* \rangle = \langle \nabla h(y), e \rangle + \langle \nabla h(y), e^* - e \rangle \leq -\frac{\epsilon}{8} + r \leq -\frac{\epsilon}{16}$$

Which completes the proof.

□

6 Discussion and open problems

6.1 Arbitrary Lipschitz constants and diameter

We assumed throughout the paper that the convex function f is 1-Lipschitz and the convex set $\mathcal{K}$ has diameter 1. Our results can be easily extended to arbitrary functions and convex sets through a simple linear transformation. For f with Lipschitz constant $\|f\|_{\mathsf{Lip}}$ and $\mathcal{K}$ with diameter D, and the corresponding approximately convex $\tilde{f}$, define $\tilde{g} : \frac{\mathcal{K}}{D} \to \mathbb{R}$ as $\tilde{g}(x) = \frac{1}{D\|f\|_{\mathsf{Lip}}} \tilde{f}(rx)$. (Where $\frac{\mathcal{K}}{D}$ is the rescaling of $\mathcal{K}$ by a factor of $\frac{1}{D}$.) This translates to $\|\tilde{g}(x) - g(x)\|_2 \leq \frac{\Delta}{R\|f\|_{\mathsf{Lip}}}$. But $g(x) = \frac{f(Rx)}{R\|f\|_{\mathsf{Lip}}}$ is 1-Lipschitz over a set $\frac{\mathcal{K}}{R}$ of diameter 1. Therefore, for general functions over a general convex sets, our result trivially implies the rate for being able to optimize approximately-convex functions is

$$\frac{\Delta}{R\|f\|_{\mathsf{Lip}}} = \max\left\{ \frac{1}{\sqrt{d}} \left(\frac{\epsilon}{R\|f\|_{\mathsf{Lip}}} \right)^2, \frac{1}{d} \frac{\epsilon}{R\|f\|_{\mathsf{Lip}}} \right\}$$

which simplifies to $\Delta = \max\left\{ \frac{\epsilon^2}{\sqrt{d} R\|f\|_{\mathsf{Lip}}}, \frac{\epsilon}{d} \right\}$.

6.2 Body specific bounds

Our algorithmic result matches the lower bound on well-conditioned bodies. The natural open problem is to resolve the problem for arbitrary bodies. [9]

Also note the lower bound can not hold for any convex body $\mathcal{K}$ in $\mathbb{R}^d$: for example, if $\mathcal{K}$ is just a one dimensional line in $\mathbb{R}^d$, then the threshold should not depend on d at all. But even when the "inherent dimension" of $\mathcal{K}$ is d, the result is still body specific: one can show that for $\tilde{f}$ over the simplex in $\mathbb{R}^d$, when $\epsilon \geq \frac{1}{\sqrt{d}}$, it is possible to optimize $\tilde{f}$ in polynomial time even when Δ is as large as ϵ. [10]

Finally, while our algorithm made use of the well-conditioning – what is the correct property/parameter of the convex body that governs the rate of $T(\epsilon)$ is a tantalizing question to explore in future work.

References

[AD10] Alekh Agarwal and Ofer Dekel. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40. Citeseer, 2010.

[AFH+11] Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.

[9] We do not show it here, but one can prove the upp/lower bound still holds over the hypercube and when one can find a ball of radius ϵ that has most of the mass in the convex body $\mathcal{K}$.

[10] Again, we do not show that here, but essentially one can search through the $d + 1$ lines from the center to the $d + 1$ corners.

[BLNR15] Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Proceedings of The 28th Conference on Learning Theory*, pages 240–265, 2015.

[DJWW15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *Information Theory, IEEE Transactions on*, 61(5):2788–2806, 2015.

[DKS14] Martin Dyer, Ravi Kannan, and Leen Stougie. A simple randomised algorithm for convex optimisation. *Mathematical Programming*, 147(1-2):207–229, 2014.

[FKM05] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[NS] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40.

[NY83] Arkadii Nemirovskii and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York, 1983. A Wiley-Interscience publication.

[Sha12] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. *arXiv preprint arXiv:1209.2388*, 2012.

[SV15] Yaron Singer and Jan Vondrák. Information-theoretic lower bounds for convex optimization with erroneous oracles. In *Advances in Neural Information Processing Systems*, pages 3186–3194, 2015.