On the Complexity of Finding Small Subgradients in Nonsmooth Optimization

Guy Kornowski GUY.KORNOWSKI@WEIZMANN.AC.IL
Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Abstract

We study the oracle complexity of producing (δ, ϵ)-stationary points of Lipschitz functions, in the sense proposed by Zhang et al. [20]. While there exist dimension-free randomized algorithms for producing such points within $\widetilde{O}(1/\delta\epsilon^3)$ first-order oracle calls, we show that no dimension-free rate can be achieved by a deterministic algorithm. On the other hand, we point out that this rate can be derandomized for smooth functions with merely a logarithmic dependence on the smoothness parameter. Moreover, we establish several lower bounds for this task which hold for any randomized algorithm, with or without convexity. Finally, we show how the convergence rate of finding (δ, ϵ)-stationary points can be improved in case the function is convex, a setting which we motivate by proving that in general no finite time algorithm can produce points with small subgradients even for convex functions.

1. Introduction

We consider the problem of optimizing Lipschitz continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ using a first-order algorithm, which utilizes values and derivatives of the function at various points. Though problems of this type are ubiquitous throughout modern machine learning, they are well known to be impossible to solve at a dimension-free rate without further assumptions such as convexity or smoothness. For example, it is generally impossible to obtain local minima or approximate-stationary points of f [14, 20], nor is it even possible to get close to such points within any finite time independent of d [11]. Seeking to design reachable optimization goals for this large function class, Zhang et al. [20] proposed the relaxed notion of a (δ, ϵ)-stationary point. Simply put, these are points for which there exists a convex combination of gradients in a δ-neighborhood whose norm is less than ϵ (see Section 2 for a formal definition). Their main contribution is a randomized algorithm, INGD, which produces (δ, ϵ)-stationary points of any Lipschitz, bounded from below function within $\widetilde{O}(1/\delta\epsilon^3)$ iterations. Although INGD initially required access to a slightly nonstandard oracle, follow-up works modified the algorithm such that it will rely on access to a standard first-order oracle at differentiable points [9, 17]. All of these algorithms are randomized and share the same oracle complexity of $\widetilde{O}(1/\delta\epsilon^3)$.

In this work, we aim towards a better understanding of the oracle complexity of producing (δ, ϵ)-stationary points in various settings. First, we examine whether the algorithms mentioned above can be derandomized. Namely, what rate can be achieved by a deterministic algorithm that produces (δ, ϵ)-stationary points? We solve this question by showing a strong lower bound, proving that deterministic algorithms cannot achieve *any* dimension-free rate. On the other hand, we point

© G. Kornowski & O. Shamir.

out that if the objective function is even slightly smooth, it is possible to provide a deterministic algorithm whose rate is $\widetilde{O}(1/\delta\epsilon^3)$ with only a logarithmic dependence on the smoothness parameter.

Next, we take the first step towards providing lower bounds which hold for any (possibly randomized) first-order method. In terms of ϵ dependence, we provide a lower bound of $\Omega(1/\epsilon^2)$ which holds even for convex functions. In terms of δ dependence, we provide an $\Omega(\log(1/\delta))$ lower bound for nonconvex functions. Though these results are clearly not tight, there are currently no randomized lower bounds whatsoever for this setting in terms of neither ϵ nor δ. Noticeably, Zhang et al. [20] do prove an $\Omega(1/\delta)$ lower bound, though their proof covers only deterministic algorithms which have access to the derivative alone (without function value). As we conjecture the lower bounds are loose, especially with respect to δ, providing tighter and unified lower bounds - in terms of δ *and* ϵ - remains an intriguing open question.

Finally, we also examine the complexity of finding (δ, ϵ)-stationary of *convex* functions. Although the vast majority of convex optimization literature deals with minimizing the function value [16], some applications require the design of algorithms that produce points with small subgradients (see [1, 15] and discussions therein). Moreover, while for *smooth* convex function the picture is relatively well understood in terms of complexity upper and lower bounds [6], for nonsmooth convex optimization much less is known in terms of minimizing subgradient norm. We start by proving that even for convex yet nonsmooth functions, it is generally impossible to produce ϵ-stationary points - namely, with some subgradient of norm at most ϵ (corresponding to $\delta \to 0$). In fact, we prove an explicit δ-dependent lower bound for convex functions, hinting that it might be beneficial to consider the relaxation of (δ, ϵ)-stationarity in this setting as well. Furthermore, we show that the complexity of producing a (δ, ϵ)-stationary point of a convex function improves (when compared to the nonconvex case) to $\widetilde{O}(1/\delta^{2/3}\epsilon^2)$ iterations using a randomized algorithm, or (again) even a deterministic algorithm if the function is slightly smooth. These algorithms have optimal ϵ-dependence as they match our aforementioned lower bound.

Overall, we hope this work will motivate further understanding of complexity guarantees for nonsmooth optimization as a whole, and of producing (δ, ϵ)-stationary points in particular.

This paper is structured as follows. In Section 2 we formally introduce the terminology which we use throughout the paper. In Section 3 we present our results for deterministic algorithms, while in Section 4 we provide our randomized lower bounds. In Section 5 we turn to analyze the case of convex functions. We conclude in Section 6 with a discussion of our results and future directions. The appendices contain a brief account of the notation we use throughout this work, along with the full proofs of our results.

2. Preliminaries

Nonsmooth analysis. We call a function $f : \mathbb{R}^d \to \mathbb{R}$ L-Lipschitz if for any $x, y \in \mathbb{R}^d : |f(x) - f(y)| \leq L \|x - y\|$, and H-smooth if it is differentiable and $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is H-Lipschitz, namely for any $x, y \in \mathbb{R}^d : \|\nabla f(x) - \nabla f(y)\| \leq H \|x - y\|$. By Rademacher's theorem, Lipschitz functions are differentiable almost everywhere (in the sense of Lebesgue). Hence, for any Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and point $x \in \mathbb{R}^d$ the Clarke sub-differential set [7] can be defined as $\partial f(x) := \mathrm{conv}\{g : g = \lim_{n\to\infty} \nabla f(x_n),\ x_n \to x\}$, namely, the convex hull of all limit points of $\nabla f(x_n)$ over all sequences of differentiable points which converge to x. Note that if the function is differentiable at a point or convex, the Clarke sub-differential reduces to the gradient or subgradient in the convex analytic sense, respectively. We say that a point x is an ϵ-stationary point of $f(\cdot)$ if

there exists $g \in \partial f(x) : \|g\| \leq \epsilon$. Furthermore, given $\delta > 0$ the Goldstein δ-subdifferential [10] of f at x is the set $\partial_\delta f(x) := \mathrm{conv}(\bigcup_{\|z-x\|<\delta} \partial f(z))$, namely all convex combinations of gradients at points in a δ-neighborhood of x. We say that a point x is a (δ, ϵ)-stationary of $f(\cdot)$ if there exists $g \in \partial_\delta f(x)$ such that $\|g\| \leq \epsilon$. Note that a point is ϵ-stationary if and only if it is (δ, ϵ)-stationary for all $\delta > 0$ [20, Lemma 7].

Algorithms and complexity. Throughout this work we consider (possibly randomized) iterative first-order algorithms, from an oracle complexity perspective [14]. Such an algorithm first produces x_1 possibly at random and receives $(f(x_1), \partial f(x_1))$.[1] Then, for any $t > 1$ produces x_t possibly at random based on previously observed responses, and receives $(f(x_t), \partial f(x_t))$. We are interested in the minimal number T for which it can guarantee to produce some (δ, ϵ) stationary point. Namely, given some function class $\mathcal{F}$ we look at $\sup_{f \in \mathcal{F}} \min_{T \in \mathbb{N}} \Pr_{\mathcal{A}}[\exists t \in [T] : x_t \text{ is } (\delta, \epsilon)\text{−stationary}] \geq \frac{2}{3}$, where the probability is with respect to the algorithm's internal randomness when acting upon $f \in \mathcal{F}$ (or deterministically, if the algorithm is deterministic).

3. Deterministic algorithms

As discussed in the introduction, [9, 17, 20] all provide randomized algorithms that given any L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and an initial point x_1 that satisfies $f(x_1) - \inf_x f(x) \leq \Delta$, produce a (δ, ϵ)-stationary point of f within $\widetilde{O}(\Delta L^2/\delta\epsilon^3)$ oracle calls to f. We start by showing that this rate, let alone any dimension-free rate whatsoever, cannot be achieved by any deterministic algorithm.

Theorem 1 *For any deterministic first-order algorithm and any iteration budget $T \in \mathbb{N}$, there exists a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, $d = T + 2$ such that $f(x_1) - \inf_x f(x) \leq 1$ yet the T iterates produced by the algorithm when applied to f are not (δ, ϵ)-stationary points for any $\delta \leq \frac{1}{7}$, $\epsilon \leq \frac{1}{252}$.*

We defer the full proof to Section B in the appendix, though the intuition can be described as follows. For any deterministic first-order algorithm, if an oracle can always return the "uninformative" answer $f(x_i) = 0, \nabla f(x_i) = \mathrm{e}_1$ this fixes the algorithm iterates $x_1, \ldots, x_T$. It remains to construct a Lipschitz function that will be consistent with the oracle answers, yet all the queried points are not (δ, ϵ)-stationary. To that end, we construct a function which in a very small neighborhood of each queried point x_t looks like $x \mapsto \mathrm{e}_1^\top(x - x_t)$, yet in most of the space looks like $x \mapsto \max\{v^\top x, -1\}$ which has (δ, ϵ)-stationary points only when x is correlated with $-v$. By letting v be some vector which is orthogonal to all the queried points, we obtain the result.

The construction we just described crucially relies on the function being highly nonsmooth - essentially interpolating between two orthogonal linear functions in an arbitrarily small neighborhood. As it turns out, if the function to be optimized is even slightly smooth, then the theorem above can be bypassed, as manifested in following theorem.

Theorem 2 *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, H-smooth, and $x_1 \in \mathbb{R}^d$ is such that $f(x_1) - \inf_x f(x) \leq \Delta$. Then there is a deterministic first-order algorithm that produces a (δ, ϵ)-stationary point of f within $O(\frac{\Delta L^2 \log(H\delta/\epsilon)}{\delta\epsilon^3})$ oracle calls.*

1. For the purpose of this work it makes no difference whether the algorithm gets to see some subgradient or the whole subdifferential set. That is, the lower bounds to follow hold even if the algorithm has access to the entire subdifferential set, while the upper bounds hold even if the algorithm receives a single subgradient.

The simple idea which proves Theorem 2 is to replace a certain randomized line search in INGD with a deterministic binary search subroutine, which terminates within $O(\log(H\delta/\epsilon))$ steps if the function is H-smooth. Such a procedure is derived by Davis et al. [9] for any H-weakly convex function along differentiable directions, and since any H-smooth function is H-weakly convex and differentiable along any direction, this can be applied as is proving Theorem 2. Although the algorithmic ingredient to this observation appears inside a proof of Davis et al. [9], they use it in a different way in order to produce a randomized algorithm for weakly convex functions in low dimension, with different guarantees suitable for that setting. For completeness we explain in Section B in the appendix how this leads to Theorem 2.

4. Randomized lower bounds

We now turn to establish complexity lower bounds which will hold for any (possibly randomized) algorithm based on a first-order oracle, as common in the optimization literature stemming from the seminal work of Nemirovski and Yudin [14]. Our first lower bound will be for the complexity of producing (δ, ϵ)-stationary points in terms of ϵ-dependence.

Theorem 3 *For any $\epsilon < 1$, $\delta \leq \frac{1}{12\epsilon}$ and any randomized first-order algorithm $\mathcal{A}$, there exists a 1-Lipschitz convex function $f : \mathbb{R}^d \to \mathbb{R}$, $d = \widetilde{O}(1/\epsilon^6)$ such that $f(x_1) - \inf_x f(x) \leq 1$ yet $\Pr_{\mathcal{A}}[\exists t \in [T] : x_t \text{ is } (\delta, \epsilon)\text{−stationary}] < \frac{1}{3}$ unless $T = \Omega(1/\epsilon^2)$.*

An observation we would like to make is that Theorem 3 actually provides an optimal $\Omega(1/\delta\epsilon^3)$ lower bound in a certain parameter regime. Indeed, since it holds even when $\delta = \Omega(1/\epsilon)$, for such a choice of δ we have $\Omega(1/\epsilon^2) = \Omega(1/\delta\epsilon^3)$, resulting in the optimal bound. In standard settings in optimization which involve optimizing with respect to a single complexity parameter (e.g. ϵ-sub-optimality or ϵ-stationarity) providing a lower bound in a certain parameter regime typically extends to any parameter regime through rescaling tricks of the form $f(x) \mapsto C_1 f(C_2 \cdot x)$. Unfortunately, this does not seem to be the case for (δ, ϵ)-stationarity as there aren't enough degrees of freedom to control the scales of both δ and ϵ, while maintaining the Lipschitz and initial sub-optimality constant.

The proof of Theorem 3 appears in Section B in the appendix, and is based on the well established machinery of high-dimensional optimization lower bounds (see for example Carmon et al. [4], Nemirovski and Yudin [14], Nesterov [16], Woodworth and Srebro [18]). To sketch the proof idea, we examine a random orthogonal transformation on top of the so called "Nemirovski function", which is approximately of the form $x \mapsto \max\{|x_i - 1|\}$. Though the Nemirovski function is commonly used to prove lower bounds in terms of sub-optimality (i.e. $f(x_t) - \inf_x f(x)$), we develop an analysis of its (δ, ϵ)-stationary points from which a complexity lower bound can be derived. On the one hand, any oracle response reveals information about a single coordinate which roughly leads to a lower bound of d queries in order to get information about d coordinates. On the other hand, the subgradients at points which are not near $(1, 1, \ldots, 1) \in \mathbb{R}^d$ are convex combinations of $\pm \mathrm{e}_i$ for some standard basis vectors e_i, thus have norm of at least $\Omega(1/\sqrt{d})$ which is smaller than ϵ only whenever $d = \Omega(1/\epsilon^2)$.

As to δ-dependent lower bounds, we provide the following theorem.

Theorem 4 *For any randomized first-order algorithm $\mathcal{A}$ and any $\delta, \epsilon \leq \frac{1}{4}$, there exists a 1-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x_1) - \inf_x f(x) \leq 1$, yet $\Pr_{\mathcal{A}}[\exists t \in [T] : x_t \text{ is } (\delta, \epsilon)\text{−stationary}] < \frac{1}{3}$ unless $T = \Omega(\log(1/\delta))$.*

Underlying the proof of Theorem 4 is a useful observation on the structure of (δ, ϵ)-stationary points in dimension one: Although in general it is not true that a (δ, ϵ)-stationary point is necessarily δ-close to an ϵ-stationary point [11] (yet the opposite implication is trivially true), for univariate functions these two notions are actually equivalent.

Claim 1 *If $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz, then $x \in \mathbb{R}$ is (δ, ϵ)-stationary if and only if it is δ-close to an ϵ-stationary point.*

We provide a proof of Claim 1 in Section B in the appendix. This claim is useful for our purposes since it allows us to convert any one-dimensional lower bound for getting δ-close to ϵ-stationary points, into a lower bound for finding (δ, ϵ)-stationary points. The former can be done using a technique which appears in [11] which itself is based on a technique for the deterministic setting due to Nemirovski [13]. The basic idea is to construct a function which is parameterized by a binary vector, such that optimizing the function essentially reduces to guessing the binary vector. If the function is constructed such that any oracle response reveals a constant number of bits from the vector, this implies that the length of the vector can serve as a complexity lower bound for the optimization problem. Since the result in [11] is asymptotic - for $\delta \to 0$, or equivalently for a binary vector of arbitrary length - providing a specific δ-dependent lower bound boils down to extracting and improving its δ-dependence through a more careful analysis which appears in Section B in the appendix.

5. (δ, ϵ)-stationarity for convex functions

Recalling that for convex functions the Clarke sub-differential coincides with the convex subgradient [7, Proposition 2.2.7], we are interested in the possibility of producing points with small subgradients of convex Lipschitz functions. It turns out that the same discussed proof idea of Theorem 4 also extends to convex functions, albeit with substantially more technical work to ensure that global convexity is maintained. This yields the impossibility of producing ϵ-stationary points within any finite time even for convex functions.

Theorem 5 *For any randomized first-order algorithm $\mathcal{A}$ and any $\delta, \epsilon \leq \frac{1}{4}$, there exists a convex 1-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x_1) - \inf_x f(x) \leq 1$, yet $\Pr_{\mathcal{A}}[\exists t \in [T] : x_t \text{ is } (\delta, \epsilon)\text{−stationary}] < \frac{1}{3}$ unless $T = \Omega(\sqrt{\log(1/\delta)})$. In particular, no finite time algorithm can produce a $\frac{1}{4}$-stationary point of 1-Lipschitz convex functions with constant probability.*

Following the theorem above, we aim towards achievable relaxations of stationarity for convex Lipschitz functions. One natural relaxation is getting δ-close to an ϵ-stationary point. While it is known that without convexity no algorithm can guarantee getting to such points in a dimension-free rate [11], perhaps not surprisingly, this is not the case for convex functions. Indeed, given a Lipschitz convex function and a diameter bound $\mathrm{dist}(x_1, \arg\min_x f(x)) \leq R$, Davis and Drusvyatskiy [8] provide a first-order algorithm that produces a point which is δ-close to an ϵ-stationary point of f within $T = \widetilde{O}(\frac{R^2}{\delta^2 \epsilon^2})$ oracle calls. Furthermore, since any such point is also a (δ, ϵ)-stationary point, then by Theorem 3 this is the best achievable rate for this task in terms of ϵ-dependence.

As in the rest of this work, we consider the coarser relaxation of (δ, ϵ)-stationary points. It is important to emphasize that even for convex functions, the two notions of (δ, ϵ)-stationarity and being δ-close to an ϵ-stationary point are inherently different. Indeed, as we illustrate in Lemma 18

in the appendix, a (δ, ϵ)-stationary point can be arbitrarily far away from any ϵ-stationary point.[2] It turns out that for the sake of producing (δ, ϵ)-stationary points a simple combination of gradient descent and the INGD algorithm of [20] can provide a better dependence in terms of δ than [8] as we claim next.

Theorem 6 *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz and convex, and $x_1 \in \mathbb{R}^d$, $R > 0$ are such that $\mathrm{dist}(x_1, \arg\min_x f(x)) \leq R$. Then there exists a randomized first-order algorithm that produces a (δ, ϵ)-stationary point of f with probability at least $\frac{2}{3}$ within $T = \widetilde{O}\left(\frac{L^2 R^{2/3}}{\delta^{2/3}\epsilon^2}\right)$ oracle calls. Moreover, if f is also H-smooth, then there exists such a deterministic algorithm with $T = \widetilde{O}\left(\frac{L^2 R^{2/3}\log(H)}{\delta^{2/3}\epsilon^2}\right)$.*

The proof is provided in Section B in the appendix. We would like to point out is that the improved rates for convex functions provided by Theorem 6 require the stronger assumption of a domain bound $R := \mathrm{dist}(x_1, \arg\min_x f(x))$, as opposed to a sub-optimality bound $\Delta := f(x_1) - \inf_x f(x)$. This is indeed a stronger assumption since for any L-Lipschitz function we have $\Delta \leq LR$. Furthermore, since the lower bound in Theorem 3 holds for convex functions, and as we explained why it provides an $\Omega(1/\delta\epsilon^3)$ lower bound for a certain parameter regime, we cannot expect such an improvement in general without any further assumption. We note that this is analogous to the smooth convex case for which a gap between Δ-based rate and R-based rate exists as well [6].

6. Discussion

In this paper, we studied the oracle complexity of producing (δ, ϵ)-stationary points in various settings. Following our results, there are several notes and open ends we would like to point out.

In our opinion, the main open question in this realm is narrowing the gap between the $\widetilde{O}(1/\delta\epsilon^3)$ complexity upper bound of the INGD algorithm, and our $\Omega(1/\epsilon^2 + \log(1/\delta))$ complexity lower bound (which follows by combining Theorem 3 and Theorem 4). We observed that the $\Omega(1/\epsilon^2)$ lower bound in Theorem 3 actually matches the upper bound in a certain parameter regime, which we suspect hints that the lower bound can be further improved. This would require in particular a polynomial lower bound in terms of δ^{-1} which we do not currently know how to prove.

We also examined the complexity of producing small subgradients of convex Lipschitz functions, showing that no finite algorithm can obtain such points unless a positive δ-relaxation is introduced, similarly to the nonconvex case. Given an additional assumption of a domain bound as opposed to a sub-optimality bound, we were able to provide improved rates for convex functions. We would like to point out that although the lower bound in Theorem 3 is stated in terms of a sub-optimality bound, the same proof can be easily adapted to the bounded domain case. Indeed, by replacing $|x_i - 1|$ in the proof by $|x_i - R/\sqrt{d}|$, the same result holds for any $\delta = O(R)$.

Another interesting direction for future work is to establish dimension-dependent lower bounds for producing (δ, ϵ)-stationary points. Although several recent works considered this setting [9, 12], no complementing lower bounds currently exist.

2. Similar results appear in [11, 17]. However, the former reference provides such a result for nonconvex functions. The latter reference does provides a result for convex functions, but the construction uses a point which is 2δ-close to an ϵ-stationary one, rather than arbitrarily far. Accordingly, Lemma 18 strengthens both.

References

[1] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

[2] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[3] Yair Carmon. *The Complexity of Optimization Beyond Convexity*. Stanford University, 2020.

[4] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

[5] Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33:19052–19063, 2020.

[6] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.

[7] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[8] Damek Davis and Dmitriy Drusvyatskiy. Complexity of finding near-stationary points of convex functions stochastically. *arXiv preprint arXiv:1802.08556*, 2018.

[9] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *arXiv preprint arXiv:2112.06969*, 2021.

[10] AA Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[11] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[12] Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *arXiv preprint arXiv:2209.05045*, 2022.

[13] Arkadi Nemirovski. *Information-based complexity of convex programming*. Lecture Notes, 1995.

[14] Arkadi Semenovich Nemirovski and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

[15] Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

[16] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[17] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

[18] Blake Woodworth and Nathan Srebro. Lower bound for randomized first order convex optimization. *arXiv preprint arXiv:1709.03594*, 2017.

[19] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 222–227. IEEE Computer Society, 1977.

[20] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

Appendix A. Notation

We denote by $\mathbb{R}^d$ the d-dimensional Euclidean space, any by $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ the standard inner product and its associated Euclidean norm, respectively. Given a point $x \in \mathbb{R}^d$ we denote by x_i its i'th coordinate, and for any subset $A \subset \mathbb{R}^d$ we denote their distance $\mathrm{dist}(x, A) := \inf_{y \in A} \|x - y\|$. We let $\mathrm{e}_i \in \mathbb{R}^d$ be the i'th standard basis vector, $\mathbb{N} = \{1, 2, \dots\}$ be the natural numbers starting from 1, and $[N] = \{1, 2, \dots, N\}$ be the natural numbers up to $N \in \mathbb{N}$. We denote by $B_r(x)$ the open Euclidean ball around $x \in \mathbb{R}^d$ of radius $r > 0$, where d is clear from context. We use standard big-O asymptotic notation: For functions $f, g : \mathbb{R} \to [0, \infty)$ we write $f = O(g)$ if there exists $c > 0$ such that $f(x) \leq c \cdot g(x)$; $f = \Omega(g)$ if $g = O(f)$. We occasionally hide logarithmic factors by writing $f = \widetilde{O}(g)$ if $f = O(g \log(g + 1))$.

Appendix B. Proofs

B.1. Proof of Theorem 1

Fix $T \in \mathbb{N}$. Suppose that for any $i \in [T - 1]$ the first-order oracle response is $f(x_i) = 0, \nabla f(x_i) = \mathrm{e}_1$. Since the algorithm is deterministic this fixes the iterate sequence $x_1, \dots, x_T$. We will show this resisting strategy is indeed consistent with a function which satisfies the conditions in the theorem.

To that end, we denote $r := \min_{1 \leq i \neq j \leq T} \|x_i - x_j\| / 4$ and fix some $v \in (\mathrm{span}\{\mathrm{e}_1, x_1, \dots, x_T\})^\perp$ with $\|v\| = 1$ (which exists since $d = T + 2$). For any $z \in \mathbb{R}^d$ we define

$$g_z(x) := \min\{\|x - z\|^2 / r^2, 1\} v^\top x + (1 - \min\{\|x - z\|^2 / r^2, 1\}) \mathrm{e}_1^\top (x - z) \,,$$

and further define

$$h(x) := \begin{cases} v^\top x \,, & \forall i \in [T] : \|x - x_i\| \geq r \\ g_{x_i}(x) \,, & \exists i \in [T] : \|x - x_i\| < r \end{cases} .$$

Note that h is well defined since by definition of r there cannot be $i \neq j$ such that $\|x - x_i\| < r$ and $\|x - x_j\| < r$.

Lemma 7 *$h : \mathbb{R}^d \to \mathbb{R}$ as defined above is 7-Lipschitz, satisfies for any $i \in [T] : h(x_i) = 0, \nabla h(x_i) = \mathrm{e}_1$ and has no $(\delta, \frac{1}{36})$-stationary points for any $\delta > 0$.*

Proof We start by noting that h is continuous, since for any z and $(y_n)_{n=1}^{\infty} \subset B_r(z)$, $y_n \overset{n\to\infty}{\longrightarrow} y$ such that $\|y-z\| = r$ we have

$$\begin{aligned}
\lim_{n\to\infty} h(y_n) &= \lim_{n\to\infty} g_z(y_n) \\
&= \lim_{n\to\infty} \left(\min\{\|y_n - z\|^2/r^2, 1\} v^\top y_n + (1 - \min\{\|y_n - z\|^2/r^2, 1\}) \mathrm{e}_1^\top (y_n - z)\right) \\
&= \lim_{n\to\infty} \left(\frac{\|y_n - z\|^2}{r^2} \cdot v^\top y_n + \left(1 - \frac{\|y_n - z\|^2}{r^2}\right) \mathrm{e}_1^\top (y_n - z)\right) \\
&= v^\top y \,.
\end{aligned}$$

Having established continuity, since $x \mapsto v^\top x$ is clearly 1-Lipschitz (in particular 7-Lipschitz), in order to prove Lipschitzness of h it is enough to show that $g_{x_i}(x)$ is 7-Lipschitz in $\|x - x_i\| < r$ for any x_i. For any such x, x_i we have

$$\begin{aligned}
\nabla g_{x_i}(x) &= \frac{2v^\top x}{r^2}(x - x_i) + \frac{\|x - x_i\|^2}{r^2} v - \frac{2\mathrm{e}_1^\top (x - x_i)}{r^2}(x - x_i) - \frac{\|x - x_i\|^2}{r^2}\mathrm{e}_1 + \mathrm{e}_1 \\
&\overset{v \perp x_i}{=} \frac{2v^\top (x - x_i)}{r^2}(x - x_i) + \frac{\|x - x_i\|^2}{r^2} v - \frac{2\mathrm{e}_1^\top (x - x_i)}{r^2}(x - x_i) - \frac{\|x - x_i\|^2}{r^2}\mathrm{e}_1 + \mathrm{e}_1 \,,
\end{aligned} \tag{1}$$

hence

$$\begin{aligned}
\|\nabla g_{x_i}(x)\| &= \left\| \frac{2v^\top (x - x_i)}{r^2}(x - x_i) + \frac{\|x - x_i\|^2}{r^2} v - \frac{2\mathrm{e}_1^\top (x - x_i)}{r^2}(x - x_i) - \frac{\|x - x_i\|^2}{r^2}\mathrm{e}_1 + \mathrm{e}_1 \right\| \\
&\leq \frac{2\,\|v\| \cdot \|x - x_i\|^2}{r^2} + \frac{\|x - x_i\|^2}{r^2}\|v\| + \frac{2\,\|\mathrm{e}_1\| \cdot \|x - x_i\|^2}{r^2} + \frac{\|x - x_i\|^2}{r^2}\|\mathrm{e}_1\| + \|\mathrm{e}_1\| \\
&\leq 2 + 1 + 2 + 1 + 1 = 7 \,,
\end{aligned}$$

which proves the desired Lipschitz bound. The fact that for any $i \in [T] : h(x_i) = 0$, $\nabla h(x_i) = \mathbf{e_1}$ is easily verified by construction and by Eq. (1). In order to finish the proof, we need to show that h has no $(\delta, \frac{1}{36})$ stationary-points. By construction we have

$$\partial h(x) = \begin{cases} v \,, & \forall i \in [T] : \|x - x_i\| > r \\ \nabla g_{x_i}(x) \,, & \exists i \in [T] : \|x - x_i\| < r \end{cases} \,,$$

while for $\|x - x_i\| = r$ we would get convex combinations of the two cases.[3] Inspecting the set $\{\nabla g_{x_i}(x) : \|x - x_i\| < r\}$ through Eq. (1), we see that it depends on x, x_i only through $x - x_i$ and that actually

$$\{\nabla g_{x_i}(x) : \|x - x_i\| < r\} = \{\nabla g_{\mathbf{0}}(x) : \|x\| < r\} \,,$$

3. Since we are interested in analyzing the δ-subdifferential set which consists of convex combinations of subgradients, and subgradients are defined as convex combinations of gradients at differentiable points - we do not lose anything by considering convex combinations gradients at differentiable points in the first place.

which is convenient since the latter set does not depend on x_i. Overall, we see that *any* convex combination of gradients of h is in the set

$$\left\{\lambda_1 v + \lambda_2 \left(\frac{2v^\top x}{r^2}x + \frac{\|x\|^2}{r^2}v - \frac{2\mathrm{e}_1^\top x}{r^2}x - \frac{\|x\|^2}{r^2}\mathrm{e}_1 + \mathrm{e}_1\right) : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1, \|x\| \leq r\right\}$$
$$= \left\{\lambda_1 v + \lambda_2 \left(2v^\top x \cdot x + \|x\|^2 v - 2\mathrm{e}_1^\top x \cdot x - \|x\|^2 \mathrm{e}_1 + \mathrm{e}_1\right) : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1, \|x\| \leq 1\right\}$$
$$= \left\{(\lambda_1 + \lambda_2 \|x\|^2)v + 2\lambda_2((v - \mathrm{e}_1)^\top x)x + \lambda_2(1 - \|x\|^2)\mathrm{e}_1 : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1, \|x\| \leq 1\right\} .$$

We aim to show that the set above does not contain any vectors of norm smaller than $\frac{1}{36}$. Let u be an element in the set with corresponding λ_1, λ_2, x as above. If $\|u\| \geq 1$ then there is nothing to show, so we can assume $\|u\| < 1$. We have

$$\begin{aligned}
u^\top v &= \lambda_1 + \lambda_2 \|x\|^2 + 2\lambda_2 (v - e_1)^\top x \cdot x^\top v \\
&= \lambda_1 + \lambda_2 \|x\|^2 + 2\lambda_2 (v^\top x)^2 - 2\lambda_2 e_1^\top x \cdot x^\top v \\
&\geq \lambda_1 + \lambda_2 (v^\top x)^2 + \lambda_2 (e_1^\top x)^2 + 2\lambda_2 (v^\top x)^2 - 2\lambda_2 e_1^\top x \cdot x^\top v \\
&= \lambda_1 + \lambda_2 (v^\top x - e_1^\top x)^2 + 2\lambda_2 (v^\top x)^2 \\
&\geq \lambda_1 + \lambda_2 (v^\top x - e_1^\top x)^2 \\
&\geq \lambda_2 (v^\top x - e_1^\top x)^2 ,
\end{aligned} \tag{2}$$

which gives

$$\begin{aligned}
u^\top (e_1 + v) &= \lambda_1 + \lambda_2 + 2\lambda_2 (v^\top x - e_1^\top x)(e_1^\top x + v^\top x) \\
&\geq 1 - 4\lambda_2 |v^\top x - e_1^\top x| \\
&\overset{(2)}{\geq} 1 - 4\sqrt{\lambda_2 u^\top v} .
\end{aligned}$$

Hence

$$\begin{aligned}
1 &\leq |u^\top (e_1 + v)| + 4\sqrt{\lambda_2 u^\top v} \leq \|u\| \cdot \|e_1 + v\| + 4\sqrt{\|u\|} \\
&\leq \sqrt{2}\,\|u\| + 4\sqrt{\|u\|} \overset{\|u\|<1}{\leq} \sqrt{2\,\|u\|} + 4\sqrt{\|u\|} \\
\implies \|u\| &\geq \frac{1}{(\sqrt{2}+4)^2} > \frac{1}{36} .
\end{aligned}$$

■

Given the previous lemma we can easily finish the proof of the theorem by looking at

$$f(x) := \frac{1}{7}\max\{h(x), -1\} .$$

f is 1-Lipschitz (since h is 7-Lipschitz), and satisfies $f(x_1) - \inf_x f(x) = 0 - (-\frac{1}{7}) < 1$. Furthermore, for any $i \in [T] : h(x_i) = 0 > -1 \implies f(x_i) = \frac{1}{7}h(x_i) = 0$, which by the fact that h is 7-Lipschitz implies that for any $x \in B_{\frac{1}{7}}(x_i) : h(x) > -1 \implies \partial f(x) = \frac{1}{7}\partial h(x)$. In particular, $\partial_{\frac{1}{7}} f(x_i) = \frac{1}{7}\partial_{\frac{1}{7}} h(x_i)$ so we conclude using the lemma shows that no x_i is a $(\frac{1}{7}, \frac{1}{7} \cdot \frac{1}{36}) = (\frac{1}{7}, \frac{1}{252})$ stationary point of f.

B.2. Proof of Theorem 2

The proof is based on a slight modification of the INGD algorithm of Zhang et al. [20]. To give some background, this algorithm is based on a gradient-descent-like procedure, where in each iteration, given current point x_t, it uses a certain element $g_t \in \partial_\delta f(x_t)$, until it finds a g_t such that $\|g_t\| \leq \epsilon$. If g_t does not satisfy this condition, it is ensured that the function value decreases by at least $\delta\epsilon/4$, which can happen at most $4\Delta/\delta\epsilon$ times. g_t itself is produced by repeatedly searching in various directions away from x_t, and maintaining a convex combination of elements in $\partial_\delta f(x)$, until either a sufficient decrease is found or the norm of the convex combination becomes sufficiently small.

In the original INGD algorithm, a given direction is searched by picking a point at random in an interval starting from x_t. If there is no sufficient decrease in that direction, the fundamental theorem of calculus implies that the derivative along that direction is small on average, hence picking a point at random is sufficient. In our case, we de-randomize this part of the algorithm, by replacing it with a deterministic binary search (following Davis et al. [9]).

More formally, we replace the INGD algorithm with Algorithms 1,2,3 below, where the change is in calling the binary search subroutine (Algorithm 3) instead of picking a point at random. Inspecting Algorithm 2, we see that it maintains a convex combination of elements in $\partial_\delta f(x)$, and calls the subroutine $\mathrm{BinarySearch}(x, g_k)$ only whenever $\|g_k\| > \epsilon$ and $f(x) - f(x - \frac{\delta}{\|g_k\|} g_k) \leq \frac{\delta}{4}\|g_k\|$. That being the case, the fundamental theorem of calculus ensures

$$\frac{1}{4}\|g_k\|^2 \geq \frac{\|g_k\|}{\delta}\left(f(x) - f\left(x - \frac{\delta}{\|g_k\|} g_k\right)\right) = \frac{1}{\delta}\int_0^\delta \left\langle \nabla f\left(x - \frac{t}{\|g_k\|} g_k\right), g_k \right\rangle dt \,.$$

Namely, on average along the segment $[x, x - \frac{\delta}{\|g_k\|} g_k]$, points have gradients whose dot product with g_k is small. Accordingly, using the fact the ∇f is H-Lipschitz, $\mathrm{BinarySearch}(x, g_k)$ performs a binary search along this segment and produces a point y_k such that $\langle \nabla f(y_k), g_k \rangle \leq \frac{1}{2}\|g_k\|^2$ within $O(\log(\delta H/\|g_k\|)) = O(\log(\delta H/\epsilon))$ first-order oracle calls [9, Lemma 3.9]. This is the only modification to INGD which is introduced to Algorithm 1, when compared to [9, 20]. Accordingly, $\mathrm{MinNorm}$ finds an approximate minimal norm element of $\partial_\delta f(x_t)$ within $O\left(\frac{L^2 \log(\delta H/\epsilon)}{\epsilon^2}\right)$ first-order oracle calls [9, Corollary 2.5]. Since we claimed that this can happen at most $O(\Delta/\delta\epsilon)$ times, overall this produces a (δ, ϵ)-stationary point of f within $O\left(\frac{\Delta L^2 \log(\delta H/\epsilon)}{\delta\epsilon^3}\right)$ oracle calls.

Algorithm 1 $\mathrm{INGD}(x_1, T)$

Input: Initialization $x_1 \in \mathbb{R}^d$, iteration budget $T \in \mathbb{N}$.
  for $t = 1, \ldots, T-1$ **do**
    $g_t \leftarrow \mathrm{MinNorm}(x_t)$ ▷ Complexity $O(\frac{L^2 \log(\delta H/\epsilon)}{\epsilon^2})$
    if $\|g_t\| \leq \epsilon$ **then**
      return x_t
    else
      $x_{t+1} \leftarrow x_t - \frac{\delta}{\|g_t\|} g_t$
    end if
  end for
  return x_T

Algorithm 2 $\mathrm{MinNorm}(x)$

Input: $x \in \mathbb{R}^d,\ \delta, \epsilon > 0$.
  $k \leftarrow 0$
  $g_0 \leftarrow \nabla f(x)$
  while $\|g_k\| > \epsilon$ and $f(x) - f(x - \frac{\delta}{\|g_k\|} g_k) \leq \frac{\delta}{4} \|g_k\|$ **do**
    $y_k \leftarrow \mathrm{BinarySearch}(x, g_k)$ ▷ Finds $y_k \in [x, x - \frac{\delta}{\|g_k\|} g_k] : \langle \nabla f(y_k), g_k \rangle \leq \frac{1}{2} \|g_k\|^2$
    $g_{k+1} \leftarrow \arg\min_{\lambda \in [0,1]} \|\lambda g_k + (1-\lambda) \nabla f(y_k)\|$
    $k \leftarrow k + 1$
  end while
  return g_k

Algorithm 3 $\mathrm{BinarySearch}(x, g)$

Input: $x, g \in \mathbb{R}^d,\ \delta, H > 0$.
  $(a, b) \leftarrow (0, 1)$
  $\bar{g} \leftarrow \frac{\delta}{\|g\|} g$
  while $b - a > \frac{\|g\|}{8\delta H}$ **do**
    if $f(x - \frac{a+b}{2} \cdot \bar{g}) \geq \frac{1}{2}(f(x - a\bar{g}) + f(x - b\bar{g}))$ **then**
      $(a, b) \leftarrow (a, \frac{a+b}{2})$ ▷ $\mathbb{E}_{t \in [a, \frac{a+b}{2}]} \langle \nabla f(z - t\bar{g}), g \rangle \leq \mathbb{E}_{t \in [\frac{a+b}{2}, b]} \langle \nabla f(z - t\bar{g}), g \rangle$
    else
      $(a, b) \leftarrow (\frac{a+b}{2}, b)$ ▷ $\mathbb{E}_{t \in [a, \frac{a+b}{2}]} \langle \nabla f(z - t\bar{g}), g \rangle > \mathbb{E}_{t \in [\frac{a+b}{2}, b]} \langle \nabla f(z - t\bar{g}), g \rangle$
    end if
  end while
  return $x - a\bar{g}$

B.3. Proof of Theorem 3

We derive the lower bound using the so-called "zero chain" technique [4, 18]. While a full discussion of this technique can be found in [3, 4], we will only state the relevant definitions and propositions we will use in our proof.

Definition 8 *For $x \in \mathbb{R}^d$, we define its α-progress as* $\mathrm{prog}_\alpha(x) := \max\{i \in [d] : |x_i| > \alpha\}$.

Definition 9 *A function $f : \mathbb{R}^d \to \mathbb{R}$ is an α-robust zero-chain if for every $x \in \mathbb{R}^d$:*

$$\mathrm{prog}_\alpha(x) < i \implies f(y) = f(y_1, \ldots, y_i, 0, \ldots, 0) \text{ for all } y \text{ in a neighborhood of } x.$$

Proposition 10 ([3], Proposition 2.3) *Let $R, \alpha > 0$, $T \in \mathbb{N}$ and $d' \geq \lceil T + \frac{2R^2}{\alpha^2} \log(3T^2) \rceil$. Let $f : \mathbb{R}^T \to \mathbb{R}$ be an α-robust zero-chain, let $U \in \mathbb{R}^{d' \times T}$ be a uniformly random orthogonal matrix and denote $f_U(x) = f(U^T x)$. If $\mathcal{A}$ is a (possibly randomized) algorithm interacting with f_U such that its iterates satisfy $\|x^{(t)}\| \leq R$ for all t with probability 1, then with probability at least $\frac{2}{3}$ over the draw of U:*

$$\mathrm{prog}_\alpha(U^T x^{(t)}) < t \text{ for all } t \leq T.$$

Given the proposition, we aim to construct a robust zero chain function which such that all its (δ, ϵ)-stationary points will have large progress. To that end, for any $T \in \mathbb{N}, \alpha > 0$ we define $f : \mathbb{R}^T \to \mathbb{R}$ as follows

$$f_{T,\alpha}(x) := \max_{i \in [T]} \{|x_i - 1| + 3\alpha(T - i)\} \tag{3}$$

Lemma 11 *For any $T \in \mathbb{N}, \alpha > 0$, $f_{T,\alpha}$ is 1-Lipschitz, convex and is an α-robust zero-chain.*

The lemma and its proof are similar to [5, Lemma 23], though we provide a proof for completeness.
Proof Let $T \in \mathbb{N}, \alpha > 0$ and abbreviate $f = f_{T,\alpha}$. The Lipschitz property is easily seen since is a composition of 1-Lipschitz functions. Convexity follows from the fact that the maximum of convex functions is itself convex. To prove that f is an α-robust zero-chain, let $x \in \mathbb{R}^T$ be such that $\mathrm{prog}_\alpha(x) < i$, and $y \in B_\alpha(x)$. Note that $|x_i| \leq \alpha$ and that $\forall j : |x_j - y_j| < \alpha$. So for any $i + 1 \leq j \leq T$ we get

$$\begin{aligned} |y_j - 1| + 3\alpha(T - j) &\leq |x_j - 1| + 3\alpha(T - j) + \alpha \leq 1 + 2\alpha + 3\alpha(T - j) \\ &\leq 1 + 2\alpha + 3\alpha(T - i - 1) = 1 - \alpha + 3\alpha(T - i) \\ &\leq |x_i - 1| - \alpha + 3\alpha(T - i) \leq |y_i - 1| + 3\alpha(T - i) \ . \end{aligned}$$

By Eq. (3) this shows f is an α-robust zero-chain by definition. ■

From now on we fix some $\epsilon < 1$, $\delta \leq \frac{1}{12\epsilon}$ and set $T = \frac{1}{4\epsilon^2}$, $\alpha = \frac{1}{9T}$. The following lemma provides a crucial property of the (δ, ϵ)-stationary point of $f_{T,\alpha}$.

Lemma 12 *Let x be a (δ, ϵ)-stationary point of $f_{T,\alpha}$. Then there exists $I \subset [T], |I| \geq \frac{3T}{4}$ such that $\forall i \in I : x_i \geq \frac{1}{3}$. In particular $\mathrm{prog}_\alpha(x) \geq \frac{3T}{4}$.*

Proof Throughout the proof we abbreviate $f = f_{T,\alpha}$. Denote $f_i(x) := |x_i - 1| + 3\alpha(T - i)$, and note that $f(x) = \max_{i \in [T]} \{f_i(x)\}$. Consequently,

$$\begin{aligned} \partial f(x) &= \mathrm{conv}\{\partial f_i(x) \ : \ f_i(x) = f(x)\} \\ &= \mathrm{conv}\{g_i(x) \ : \ f_i(x) = f(x)\} \ , \ \ g_i(x) := \begin{cases} \{\mathrm{sign}(x_i - 1)\mathrm{e}_i\}, & x_i \neq 1 \\ \{s \cdot \mathrm{e}_i : -1 \leq s \leq 1\}, & x_i = 1 \end{cases} \\ \implies \partial_\delta f(x) &= \mathrm{conv}\{g_i(y) \ : \ f_i(y) = f(y), \ \|x - y\| < \delta\} \ . \end{aligned}$$

From the representation above we can see that if x satisfies that

$$\forall y \in B_\delta(x) \ \forall i \in [T] \textit{ such that } f_i(y) = f(y) \ : \ |y_i - 1| > 0 \tag{4}$$

then any $g \in \partial_\delta f(x)$ is a certain convex combination of vectors of the form $\pm \mathrm{e}_i$ (since in that case the second condition in the definition of g_i is never satisfied). Furthermore, in this case note that either $\mathrm{e}_i \in \partial_\delta f(x)$ or $-\mathrm{e}_i \in \partial_\delta f(x)$ but not both, since if both of them are in $\partial_\delta f(x)$ then by convexity of ∂_δ there exists some $y_i - 1 = 0$ contradicting Eq. (4). We will now show that if that is the case, then x is *not* a (δ, ϵ)-stationary point. Indeed, if x satisfies Eq. (4) then for any $g \in \partial f_\delta(x)$

there exist $S = \{i_1, \dots, i_{|S|}\} \subseteq [T], \lambda_{i_1}, \dots, \lambda_{i_{|S|}} \geq 0 : \sum_{i \in S} \lambda_i = 1$ and $\sigma_{i_1}, \dots, \sigma_{i_{|S|}} \in \{\pm 1\}$ such that $g = \sum_{i \in S} \lambda_i \cdot \sigma_i \mathrm{e}_i$. Hence

$$\|g\| = \sqrt{\sum_{i \in S} \lambda_i^2} \overset{\text{Cauchy Schwarz}}{\geq} \left(\sum_{i \in S} \frac{1}{\sqrt{|S|}} \lambda_i \right) = \frac{1}{\sqrt{|S|}} \geq \frac{1}{\sqrt{T}} = 2\epsilon \ .$$

This establishes that if x satisfies Eq. (4) then it is indeed not a (δ, ϵ)-stationary point. So if x is a (δ, ϵ)-stationary point, we deduce that there exist $y \in B_\delta(x), i \in [T]$ such that $f_i(y) = f(y)$ and also $y_i = 1$. For this y it holds that $\forall j \in [T] : f_j(y) \leq f(y) = f_i(y)$, hence

$$\begin{aligned} |y_j - 1| + 3\alpha(T - j) &\leq |y_i - 1| + 3\alpha(T - i) = 3\alpha(T - i) \\ \implies |y_j - 1| &\leq 3\alpha(j - i) \leq 3\alpha T \leq \frac{1}{3} \ . \end{aligned}$$

We see that for all $j \in [T] : \frac{2}{3} \leq y_j \leq \frac{4}{3}$, and recall that $\|x - y\| < \delta$. Assuming towards contradiction that the lemma is not true, we get $|J| \subset [T]$, $|J| \geq \frac{T}{4}$ such that $\forall j \in J : x_j < \frac{1}{3}$ and in particular $|x_j - y_j| > \frac{1}{3}$. But this means that $\|x - y\| > \frac{1}{3} \cdot \sqrt{\frac{T}{4}} = \frac{\sqrt{T}}{6} = \frac{L}{12\epsilon} \geq \delta$ which is a contradiction. ■

Given the previous lemma, finishing up the proof is done by a standard application of the aforementioned lower bound techniques. All we need to do is to apply a random orthogonal transformation on top of our constructed function, and to handle "large queries" by extending our function such that points of large norm do not depend on the orthogonal transformation (see [4, Section 5]). We do this by defining

$$\widetilde{f}_U(x) := \max\{f_{T,\alpha}(U^T x),\ 2(\|x\| - 2\sqrt{T})\} \ .$$

On the one hand, if $\|x\| < 2\sqrt{T}$ then

$$2(\|x\| - 2\sqrt{T}) < 0 \leq f_{T,\alpha}(U^T x) \implies \widetilde{f}_U(x) = f_{T,\alpha}(U^T x) \ .$$

On the other hand, if $\|x\| > 6\sqrt{T}$ then

$$\begin{aligned} f_{T,\alpha}(U^T x) \leq \|Ux\| + \|\mathbf{1}\| + \frac{1}{3} &< (\|x\| + 2\sqrt{T}) + (\|x\| - 6\sqrt{T}) = 2(\|x\| - 2\sqrt{T}) \\ \implies \widetilde{f}_U(x) &= 2(\|x\| - 2\sqrt{T}) \ , \end{aligned}$$

which does not depend on U. Furthermore, repeating essentially the same calculation as in Lemma 12 shows that the (δ, ϵ)-stationary points of $\widetilde{f}_U(\cdot)$ are exactly the (δ, ϵ)-stationary points of $f(U^T \cdot)$, which in particular shows they satisfy $\mathrm{prog}_\alpha(U^T x) \geq \frac{3T}{4}$. Applying Proposition 10 with $R = 6\sqrt{T}$ finishes the proof.

B.4. Proof of Theorem 5

Let $\delta \leq \frac{1}{4}$, $T < \frac{1}{6}\sqrt{\log(1/\delta)}$. By Yao's lemma [19], we may assume $\mathcal{A}$ is deterministic and provide a distribution over hard functions. Namely, we will construct an index set Σ and a parameterized subset of convex Lipschitz functions $\{f_\sigma : \sigma \in \Sigma\}$ that satisfy for any $\sigma \in \Sigma$:

- f_σ is convex and 1-Lipschitz.
- f_σ attains its minimum with $\arg\min_x f(x) \subset (0,1)$. In particular, by Lipschitzness $f(x_1) - \inf_x f(x) \leq 1$.
- $\forall x \notin \arg\min_x f(x),\ \forall f'(x) \in \partial f(x):\ |f'(x)| \geq \frac{1}{4}$.

Yet, the first T iterates produced by $\mathcal{A}(f_\sigma) : x_1^{f_\sigma}, \ldots, x_T^{f_\sigma}$, satisfy

$$\Pr_{\sigma\sim\mathcal{D}}\left[\exists t \in [T] : \mathrm{dist}(x_t^{f_\sigma}, \arg\min_x f_\sigma(x)) < \delta\right] < \frac{2}{3}\,, \tag{5}$$

for some distribution $\mathcal{D}$ over Σ. In particular, by the third bullet above, with the same probability and Claim 1 they are not $(\delta, \frac{1}{4})$-stationary.

We start by defining a sequence of segments in $[0,1]$ parameterized by binary vectors, which will form the basis for our upcoming function class construction. For any $i \in \mathbb{N}$ we denote

$$I_0^i := \left(\frac{1}{2} - \frac{2}{8^{i+1}}, \frac{1}{2} - \frac{1}{8^{i+1}}\right) \subset (0,1)$$
$$I_1^i := \left(\frac{1}{2} + \frac{1}{8^{i+1}}, \frac{1}{2} + \frac{2}{8^{i+1}}\right) \subset (0,1)\,.$$

We interpret the subscript 0 as "left" and 1 as "right", corresponding to the location of the segment with respect to the center $\frac{1}{2}$. Denote by ϕ_0^i, ϕ_1^i the unique affine functions with positive derivatives which map $(0,1)$ to I_0^i, I_1^i respectively - which we think of as "rescaling" the unit segment to the left or to the right. We can now define for any $k \in \mathbb{N}$, $(\sigma_1, \ldots, \sigma_k) \in \{0,1\}^k$:

$$I_{\sigma_1,\ldots,\sigma_k} := \phi_{\sigma_1}^1 \circ \cdots \circ \phi_{\sigma_k}^k(0,1) \subset (0,1)\,.$$

The following lemma shows that the parameterization of the segments above forms a binary tree structure. This correspondence is given by considering binary vectors as vertices, while its sub-tree consists of vectors for which it serves as a prefix. Later on we will use this intuition to construct functions parameterized by binary vectors, in a way that essentially reduces optimization to traversing the tree. In particular, optimizing the function will require getting to a randomly selected leave which is unknown in advance to the algorithm.

Lemma 13

1. *For any $l < k$ and any $\sigma_1, \ldots, \sigma_l, \ldots, \sigma_k \in \{0,1\}$: $I_{\sigma_1,\ldots,\sigma_l} \supset I_{\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k}$.*
2. *If $(\sigma_1, \ldots, \sigma_{k-1}) \neq (\sigma_1', \ldots, \sigma_{k-1}')$ then for all $\sigma_k, \sigma_k' \in \{0,1\}$: $I_{\sigma_1,\ldots,\sigma_k} \cap I_{\sigma_1',\ldots,\sigma_k'} = \emptyset$.*

Proof

1. $I_{\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k} = \phi_{\sigma_1}^1 \circ \cdots \circ \phi_{\sigma_l}^l(\phi_{\sigma_{l+1}}^{l+1} \circ \cdots \circ \phi_{\sigma_k}^k(0,1)) \subset \phi_{\sigma_1}^1 \circ \cdots \circ \phi_{\sigma_l}^l(0,1) = I_{\sigma_1,\ldots,\sigma_l}$.

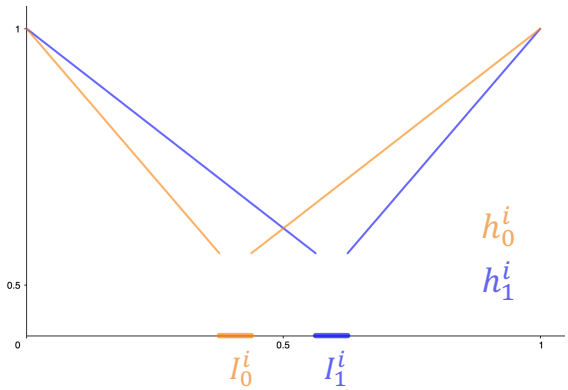

Figure 1: Illustration of I_0, I_1 and h_0, h_1.

2. Let $i \leq k-1$ be the minimal index for which $\sigma_i \neq \sigma_i'$, and assume without loss of generality that $\sigma_i = 0, \sigma_i' = 1$. If $x \in I_{\sigma_1,\ldots,\sigma_k}$, then

$$\begin{aligned}
x &\leq \sup \phi^1_{\sigma_1} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}} \circ \phi^i_0 \circ \cdots \circ \phi^k_{\sigma_k}(0,1) \\
&\leq \sup \phi^1_{\sigma_1} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}} \circ \phi^i_0(0,1) \\
&< \phi^1_{\sigma_1} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}}(\frac{1}{2}) \\
&< \inf \phi^1_{\sigma_1} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}} \circ \phi^i_1(0,1) \\
&= \inf \phi^1_{\sigma_1'} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}'} \circ \phi^i_{\sigma_i'}(0,1) \\
&\leq \inf \phi^1_{\sigma_1'} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}'} \circ \phi^i_{\sigma_i'} \circ \cdots \circ \phi^k_{\sigma_k'}(0,1) \\
&= \inf I_{\sigma_1',\ldots,\sigma_k'} \,,
\end{aligned}$$

hence $x \notin I_{\sigma_1',\ldots,\sigma_k'}$.

■

With the parameterized family of segments in hand, we are ready to define a corresponding family of functions. We define the function $h_1^i : [0,1] \setminus I_1^i \to \mathbb{R}$ as follows

$$h_1^i(x) := \begin{cases} -\left(\frac{8^{i+1}-2}{8^{i+1}+2}\right)x + 1\,, & 0 \leq x \leq \frac{1}{2} + \frac{1}{8^{i+1}} \\ \left(\frac{8^{i+1}-2}{8^{i+1}-4}\right)x - \frac{2}{8^{i+1}-4}\,, & \frac{1}{2} + \frac{2}{8^{i+1}} \leq x \leq 1 \end{cases},$$

which is the piecewise linear function which satisfies $h(0) = h(1) = 1$, $h(\frac{1}{2} + \frac{1}{8^{i+1}}) = h(\frac{1}{2} + \frac{2}{8^{i+1}}) = \frac{1}{2} + \frac{1}{8^{i+1}}$, and also define $h_0^i : [0,1] \setminus I_0^i \to \mathbb{R}$ as $h_0^i(x) := h_1^i(1-x)$ - see Fig. 1 for an illustration. We also denote $\Phi^i(x) := \phi_1^i(x) - \frac{1}{8^{i+1}}$ which is the affine function which satisfies $\Phi^i(0) = \frac{1}{2}$, $\Phi^i(1) = \frac{1}{2} + \frac{1}{8^{i+1}}$, $(\Phi^i)' = (\phi_0^i)' = (\phi_1^i)'$.

Finally, given any $N \geq 2$ and $\sigma = (\sigma_1, \dots, \sigma_N) \in \{0,1\}^N$ we are ready to define $f : \mathbb{R} \to \mathbb{R}$ as

$$f_\sigma^N(x) := \begin{cases} h_{\sigma_1}^1(x) & x \in [0,1] \setminus I_{\sigma_1} \\ \Phi^1 \circ h_{\sigma_2}^2 \circ (\phi_{\sigma_1}^1)^{-1}(x) & x \in I_{\sigma_1} \setminus I_{\sigma_1,\sigma_2} \\ \Phi^1 \circ \Phi^2 \circ h_{\sigma_3}^3 \circ (\phi_{\sigma_1}^1 \circ \phi_{\sigma_2}^2)^{-1}(x) & x \in I_{\sigma_1,\sigma_2} \setminus I_{\sigma_1,\sigma_2,\sigma_3} \\ \vdots & \\ \Phi^1 \circ \dots \circ \Phi^{N-1} \circ h_{\sigma_N}^N \circ (\phi_{\sigma_1}^1 \circ \dots \circ \phi_{\sigma_{N-1}}^{N-1})^{-1}(x) & x \in I_{\sigma_1,\dots,\sigma_{N-1}} \setminus I_{\sigma_1,\dots,\sigma_N} \\ \Phi^1 \circ \dots \circ \Phi^{N-1}(0) & x \in I_{\sigma_1,\dots,\sigma_N} \\ 1 - x & x < 0 \\ x & x > 1 \end{cases} .$$

The following lemma show that the constructed function satisfies our desired requirements.

Lemma 14 *For any $N \geq 2$, $\sigma \in \{0,1\}^N$:*

- *f_σ^N is 2-Lipschitz and convex.*
- *$f_\sigma^N(x)$ attains its minimum, and $\forall x \notin \arg\min_x f_\sigma^N(x)\ \forall g \in \partial f_\sigma^N(x) : |g| \geq \frac{1}{2}$.*
- *$f_\sigma^N(0) - \min_x f_\sigma^N(x) \leq 1$ and $\forall x \in \arg\min_x f_\sigma^N(x) : |x| \leq 1$.*

Proof Let $N \geq 2$, $\sigma \in \{0,1\}^N$. We start by proving that f_σ^N is continuous. It is clear that f_σ^N is piecewise linear since $h_{\sigma_i}^i, \phi_{\sigma_i}^i, \Phi^i$ are, So in order to show continuity we need to show continuity at the endpoints of adjacent linear pieces. Since by Lemma 13 it holds that $[0,1] \supset I_{\sigma_1} \supset I_{\sigma_1,\sigma_2} \supset \dots \supset I_{\sigma_1,\dots,\sigma_N}$, inspecting the definition of f_σ^N we see that we need to verify continuity only at the endpoints of $I_{\sigma_1,\dots,\sigma_i}$ for any $i \in [N]$. For any $i \in [N]$, the left endpoint of $I_{\sigma_1,\dots,\sigma_i}$ satisfies

$$\begin{aligned} &\lim_{x \to (\inf I_{\sigma_1,\dots,\sigma_i})^+} f_\sigma^N(x) \\ &= \lim_{x \to (\inf \phi_{\sigma_1}^1 \circ \dots \circ \phi_{\sigma_i}^i(0,1))^+} \Phi^1 \circ \dots \circ \Phi^i \circ h_{\sigma_{i+1}}^{i+1} \circ (\phi_{\sigma_1}^1 \circ \dots \circ \phi_{\sigma_i}^i)^{-1}(x) \\ &= \lim_{x \to 0^+} \Phi^1 \circ \dots \circ \Phi^i \circ h_{\sigma_{i+1}}^{i+1}(x) \\ &= \Phi^1 \circ \dots \circ \Phi^i(\lim_{x \to 0^+} h_{\sigma_{i+1}}^{i+1}(x)) \\ &= \Phi^1 \circ \dots \circ \Phi^{i-1} \circ \Phi^i(1) \\ &= \Phi^1 \circ \dots \circ \Phi^{i-1} \circ \Phi^i \circ h_{\sigma_i}^i(0) \\ &= \lim_{x \to (\inf \phi_{\sigma_1}^1 \circ \dots \circ \phi_{\sigma_{i-1}}^{i-1}(0,1))^-} \Phi^1 \circ \dots \circ \Phi^{i-1} \circ h_{\sigma_i}^i \circ (\phi_{\sigma_1}^1 \circ \dots \circ \phi_{\sigma_{i-1}}^{i-1})^{-1}(x) \\ &= \lim_{x \to (\inf I_{\sigma_1,\dots,\sigma_i})^-} f_\sigma^N(x) . \end{aligned}$$

The right endpoint follows a similar calculation, resulting in continuity.

Having established that f_σ^N is piecewise linear and continuous, Lipschitzness will follow from the fact the each linear segment has a bounded slope. To see this holds, recall that

$$(\Phi^i)' = (\phi^i)' \implies (\Phi^i)' = \frac{1}{((\phi^i)^{-1})'} , \tag{6}$$

so we get

$$\left|\left(\Phi^1 \circ \cdots \circ \Phi^{i-1} \circ h^i_{\sigma_i} \circ (\phi^1_{\sigma_1} \circ \cdots \circ \phi^{i-1}_{\sigma_{i-1}})^{-1}\right)'\right| = \left|(h^i_{\sigma_i})'\right| \leq \frac{8^{i+1}-2}{8^{i+1}-4} = 1 + \frac{2}{8^{i+1}-4} < 2\,.$$

In order to prove convexity, it is enough to show that each linear segment has slope which is larger than the segment to its left (hence the derivative is increasing which implies convexity). Recalling that by Lemma 13 we have $[0,1] \supset I_{\sigma_1} \supset I_{\sigma_1,\sigma_2} \supset \cdots \supset I_{\sigma_1,\ldots,\sigma_N}$, we see that it is enough to establish monotonicity between the left linear segments of $I_{\sigma_1,\ldots,\sigma_i}$, $I_{\sigma_1,\ldots,\sigma_{i+1}}$ (in that order) and between the right linear segments of $I_{\sigma_1,\ldots,\sigma_{i+1}}$, , $I_{\sigma_1,\ldots,\sigma_i}$ (in that order) for any $i \in [N]$. In order to do that, first note that by definition of h^i_1, h^i_0 we have

$$(h^0_i)'(x) = \begin{cases} -\left(\frac{8^{i+1}-2}{8^{i+1}-4}\right), & 0 \leq x \leq \frac{1}{2} - \frac{2}{8^{i+1}} \\ \left(\frac{8^{i+1}-2}{8^{i+1}+2}\right), & \frac{1}{2} - \frac{1}{8^{i+1}} \leq x \leq 1 \end{cases},$$

$$(h^i_1)'(x) = \begin{cases} -\left(\frac{8^{i+1}-2}{8^{i+1}+2}\right), & 0 \leq x \leq \frac{1}{2} + \frac{1}{8^{i+1}} \\ \left(\frac{8^{i+1}-2}{8^{i+1}-4}\right), & \frac{1}{2} + \frac{2}{8^{i+1}} \leq x \leq 1 \end{cases}. \tag{7}$$

Denoting by $L(I_{\sigma_1,\ldots,\sigma_i})$ the left linear segment of $I_{\sigma_1,\ldots,\sigma_i}$, we see that for any $x \in L(I_{\sigma_1,\ldots,\sigma_i})$ we have by Eq. (6) and Eq. (7):

$$\begin{aligned}(f^N_\sigma)'(x) &= \left(\Phi^1 \circ \cdots \circ \Phi^i \circ h^{i+1}_{\sigma_{i+1}} \circ (\phi^1_{\sigma_1} \circ \cdots \circ \phi^i_{\sigma_i})^{-1}\right)'(x) \\ &= \left(h^{i+1}_{\sigma_{i+1}}\right)'(x \in L([0,1])) \leq -\left(\frac{8^{i+2}-2}{8^{i+2}+2}\right),\end{aligned}$$

while for any $x \in L(I_{\sigma_1,\ldots,\sigma_{i+1}})$ we have

$$\begin{aligned}(f^N_\sigma)'(x) &= \left(\Phi^1 \circ \cdots \circ \Phi^{i+1} \circ h^{i+2}_{\sigma_{i+2}} \circ (\phi^1_{\sigma_1} \circ \cdots \circ \phi^{i+1}_{\sigma_{i+1}})^{-1}\right)'(x) \\ &= \left(h^{i+2}_{\sigma_{i+2}}\right)'(x \in L[0,1]) \geq -\left(\frac{8^{i+3}-2}{8^{i+3}-4}\right).\end{aligned}$$

It is easy to verify for all $i \in \mathbb{N}$ that $-\left(\frac{8^{i+3}-2}{8^{i+3}-4}\right) \leq -\left(\frac{8^{i+2}-2}{8^{i+2}+2}\right)$, which established the desired monotonicity. Checking the right linear segments results in the same calculation, overall proving the first bullet.

For the second bullet, note that f^N_σ attains its minimum at $I_{\sigma_1,\ldots,\sigma_N}$ since till then the function is decreasing, and after that it's increasing (and on that segment it is constant). Elsewhere we already saw that the derivatives are $\pm\left(\frac{8^{i+1}-2}{8^{i+1}+2}\right), \pm\left(\frac{8^{i+1}-2}{8^{i+1}+4}\right)$ which are easily verified to have absolute value larger than $\frac{1}{2}$ for any $i \in \mathbb{N}$.

The third bullet easily follows since $I_{\sigma_1,\ldots,\sigma_N} \subset [0,1]$ and $f^N_\sigma(0) = 1$, $\Phi^1 \circ \cdots \circ \Phi^{N-1}(0) > 0$. ■

The following lemma follows immediately from the definition of f^N_σ and from Lemma 13, though we state it for future reference.

Lemma 15 *For any $N \geq 2$, $\sigma \in \{0,1\}^N$, $1 \leq k < N$, it holds that $f_\sigma^N(x), (f_\sigma^N)'(x)$ do not depend on $\sigma_{k+1}, \ldots, \sigma_N$ for $x \notin I_{\sigma_1,\ldots,\sigma_k}$.*

Given Lemma 13 and Lemma 15, we state the following lemma which whose proof appears in [11, Proof of Proposition 1].

Lemma 16 ([11]) *For any $t \in \mathbb{N}$, $1 \leq l < k \leq N$:*

$$\Pr_\sigma \left[x_{t+1} \in I_{\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k} | x_1, \ldots, x_t \notin I_{\sigma_1,\ldots,\sigma_l}\right] \leq \frac{1}{2^{k-l-1}} \,.$$

We now claim that points that are δ close to $\arg\min_x f_\sigma^N(x) = I_{\sigma_1,\ldots,\sigma_N} \subset I_{\sigma_1,\ldots,\sigma_{N-1}} \subset I_{\sigma_1,\ldots,\sigma_{N-2}} \subset \ldots$ are necessarily inside $I_{\sigma_1,\ldots,\sigma_k}$ with large enough k.

Lemma 17 *Let $k = \lfloor \frac{1}{4}\sqrt{\log(1/\delta)} \rfloor < N \in \mathbb{N}$. Then $\operatorname{dist}(x, I_{\sigma_1,\ldots,\sigma_N}) < \delta$ implies $x \in I_{\sigma_1,\ldots,\sigma_k}$.*

Proof Recall that $I_{\sigma_1,\ldots,\sigma_N} \subset I_{\sigma_1,\ldots,\sigma_k}$, and note that for any $i \in \mathbb{N}$:

$$\begin{aligned}
\inf I_{\sigma_1,\ldots,\sigma_{i+1}} - \inf I_{\sigma_1,\ldots,\sigma_i} &= \phi_{\sigma_1}^1 \circ \cdots \circ \phi_{\sigma_i}^i \circ \phi_{\sigma_{i+1}}^{i+1}(0) - \phi_{\sigma_1}^1 \circ \cdots \circ \phi_{\sigma_i}^i(0) \\
&\overset{(*)}{=} \left[\prod_{j=1}^{i} (\phi_{\sigma_j}^j)'\right] \cdot \left(\phi_{\sigma_{i+1}}^{i+1}(0) - 0\right) \\
&\geq \left[\prod_{j=1}^{i} \frac{2}{8^{j+1}}\right] \cdot \left(\frac{1}{2} - \frac{2}{8^{i+2}}\right) \\
&\geq \frac{2^i}{8^{\frac{i(i+3)}{2}}} \cdot \frac{1}{4} \,.
\end{aligned}$$

where $(*)$ follows from the fact that for any affine mapping $x \mapsto ax + b$ it holds that $(ax_1 + b) - (ax_2 + b) = a(x_1 - x_2)$. Thus

$$\begin{aligned}
\inf I_{\sigma_1,\ldots,\sigma_N} - \inf I_{\sigma_1,\ldots,\sigma_k} &= \sum_{i=k}^{N-1} \left(\inf I_{\sigma_1,\ldots,\sigma_{i+1}} - \inf I_{\sigma_1,\ldots,\sigma_i}\right) \\
&\geq \sum_{i=k}^{N-1} \frac{2^i}{8^{\frac{i(i+3)}{2}}} \cdot \frac{1}{4} \\
&> \frac{2^k}{4 \cdot 8^{\frac{k(k+3)}{2}}} > \delta \,,
\end{aligned}$$

where that last inequality is a straightforward computation for $k = \lfloor \frac{1}{4}\sqrt{\log(1/\delta)} \rfloor$. We conclude that

$$\inf I_{\sigma_1,\ldots,\sigma_k} < \inf I_{\sigma_1,\ldots,\sigma_N} - \delta \,,$$

while the same arguments also show that $\sup I_{\sigma_1,\ldots,\sigma_k} > \sup I_{\sigma_1,\ldots,\sigma_N} + \delta$. Overall, we see that for any x satisfying $\operatorname{dist}(x, I_{\sigma_1,\ldots,\sigma_N}) < \delta$:

$$\inf I_{\sigma_1,\ldots,\sigma_k} < I_{\sigma_1,\ldots,\sigma_N} - \delta < x < \sup I_{\sigma_1,\ldots,\sigma_N} + \delta < \sup I_{\sigma_1,\ldots,\sigma_k} \implies x \in I_{\sigma_1,\ldots,\sigma_k} \,.$$

■

We are now ready to finish the proof. We set $k = \lfloor \frac{1}{4}\sqrt{\log(1/\delta)} \rfloor$, $N = k+1$, $\sigma \sim \mathrm{Unif}\{0,1\}^N$. We consider the algorithm's iterates $x_1, \ldots, x_T$ as random variables which depend on the random choice of the function f_σ fed to the algorithm. Denote the stochastic process

$$Z_0 = 0,\ Z_t := \max\{l \in \mathbb{N} : \exists s \leq t, x_s \in I_{\sigma_1,\ldots,\sigma_l}\}\ ,$$

and note that that $Z_{t+1} - Z_t \geq 0$ with probability 1, yet by Lemma 16: $\Pr[Z_{t+1} - Z_t = m] \leq \frac{1}{2^{m-1}}$. Hence by Lemma 17

$$\begin{aligned}
&\Pr_\sigma\left[\exists t \in [T] : \mathrm{dist}(x_t, \arg\min_x f_\sigma^N(x)) < \delta\right] \leq \Pr_\sigma\left[\exists t \in [T] : x_t \in I_{\sigma_1,\ldots,\sigma_k}\right] = \Pr_\sigma\left[Z_T \geq k\right] \\
&\leq \frac{1}{k}\mathbb{E}[Z_T] = \frac{1}{k}\sum_{j=1}^{T}\mathbb{E}[Z_j - Z_{j-1}] = \frac{1}{k}\sum_{j=1}^{T}\sum_{m=0}^{\infty}\Pr[Z_j - Z_{j-1} = m] \\
&\leq \frac{1}{k}\sum_{j=1}^{T}\sum_{m=0}^{\infty}\frac{1}{2^{m-1}} \leq \frac{1}{k}\sum_{j=1}^{T}\sum_{m=0}^{\infty}\frac{1}{2^{m-1}} \leq \frac{1}{k}\sum_{j=1}^{T}4 = \frac{4T}{k} < \frac{2}{3}\ .
\end{aligned}$$

This proves Eq. (5) for 2-Lipschitz function and $\frac{1}{2}$-stationary points. By rescaling the Lipschitz constant and ϵ by a factor of 2, we have finished the proof.

B.5. Proof of Theorem 4

The proof is identical to the proof of Theorem 5 which appears above, by replacing the constructed family f_σ by the family h_σ constructed in [11, Proof of proposition 1]. The only modification throughout the whole proof is that Lemma 17 holds with $k = \Theta(\log(1/\delta))$ instead of $k = \frac{1}{4}\sqrt{\log(1/\delta)}$, affecting the final bound on T appropriately.

B.6. Proof of Theorem 6

It is well known (see for example [2, Theorem 3.2]) that given a Lipschitz convex function, performing projected gradient descent for $T_1 := \frac{L^2R^{2/3}}{\epsilon^2\delta^{2/3}}$ iterations and setting $x_{\mathrm{GD}} := \frac{1}{T_1}\sum_{j=1}^{T_1} x_j$ satisfies $f(x_{\mathrm{GD}}) - \inf_x f(x) \leq \frac{RL}{\sqrt{T_1}} = R^{2/3}\epsilon\delta^{1/3}$. Furthermore, according to [20, Theorem 8], if we initialize INGD at a point x such that $f(x) - \inf_x f(x) \leq \Delta$ then with probability $\frac{2}{3}$ it produces a (δ, ϵ) stationary point within $T_2 = \widetilde{O}\left(\frac{\Delta L^2}{\delta\epsilon^3}\right)$ oracle calls. In particular, initializing at x_{GD} for which we have $\Delta = R^{2/3}\epsilon\delta^{1/3}$ yields $T_2 = \frac{L^2R^{2/3}}{\epsilon^2\delta^{2/3}}$ and letting $T = T_1 + T_2$ proves the claim. Replacing INGD with its deterministic counterpart from Theorem 2 proves the additional claim.

B.7. Proof of Claim 1

First, if x is δ-close to an ϵ-stationary point it is clear that it is also (δ, ϵ)-stationary. For the nontrivial implication, suppose x is (δ, ϵ)-stationary. Namely, there exists $g \in \mathrm{conv}(\cup_{x-\delta<t<x+\delta}\partial f(t))$ with $|g| \leq \epsilon$. By Caratheodory's theorem (for $d = 1$), this means there exist $t_1, t_2 \in (-\delta, \delta)$, $g_i \in \partial f(t_i)$ and $\lambda_1, \lambda_2 \in [0,1]$ satisfying $\lambda_1 + \lambda_2 = 1$ such that $g = \lambda_1 g_1 + \lambda_2 g_2$. If either g_1 or g_2 satisfy $|g_i| \leq \epsilon$ then we get that x is δ-close the ϵ-stationary point t_i. Otherwise, both g_i satisfy $|g_i| > \epsilon$,

but their convex combination is of magnitude smaller than ϵ - thus they necessarily have different signs. Without loss of generality suppose $g_1 < 0, g_2 > 0$. We claim that in this case there exists $t^* \in [t_1, t_2]$ such that $0 \in \partial f(t^*)$, finishing the proof since x is δ-close to the stationary point t^*. Indeed, let $A := \{t \in [t_1, t_2] : \forall g \in \partial f(t), g < 0\}$ and note that by assumption $t_1 \in A, t_2 \notin A$. Denote $t^* := \sup A \in [t_1, t_2]$ which exists since A is non-empty. On the one hand, by definition of the supremum there exists a non-decreasing sequence $\{y_n\}_{n=1}^{\infty} \subset A$ with $\lim_{n\to\infty} y_n = t^*$. Note that by definition of A all the sub-differentials at all of y_n are negative, thus t^* has some non-positive sub-differential (by Bolzano–Weierstrass one can construct a converging sub-sequence of derivatives). On the other hand, by construction all differentiable points in (t^*, t_2) have only positive differentials, thus $x + t^*$ has some non-negative sub-differential. Finally, recalling that $\partial f(t^*)$ is convex we get $0 \in \partial f(t^*)$ as claimed.

Appendix C. Additional Lemma

Lemma 18 *For any $C > 1$, $\delta < 1$ and any $\epsilon < \frac{1}{8C}$, there exists a 1-Lipschitz convex function $f : \mathbb{R}^d \to \mathbb{R}$ in dimension $d = O(1/\epsilon^2)$ and a point $x \in \mathbb{R}^d$ such that x is a (δ, ϵ)-stationary point of f, yet is at distance at least $C\delta$ away from any ϵ-stationary point of f.*

Proof Fix $\delta < 1 < C,\ \epsilon < \frac{1}{8C}$, and let $d = \lceil \frac{3}{\epsilon^2} \rceil$. We denote by $A := \mathrm{diag}(2\epsilon^2, 1, \ldots, 1) \in \mathbb{R}^{d\times d}$ the diagonal positive-definite matrix whose diagonal is $(2\epsilon^2, 1, \ldots, 1)$, and consider its corresponding Mahalanobis norm $f(x) := \sqrt{x^\top A x}$. For any $x \neq 0$ we have

$$\nabla f(x) = \frac{1}{\sqrt{x^\top A x}} \cdot Ax \quad \Longrightarrow \quad \|\nabla f(x)\|^2 = \frac{x^\top A^2 x}{x^\top A x} \in [\lambda_{\min}(A), \lambda_{\max}(A)] = [2\epsilon^2, 1] \ .$$

In particular, we see that for any $x \neq 0$ the gradient norm is at least $\sqrt{2\epsilon^2} = \sqrt{2} \cdot \epsilon$, thus it is not an ϵ-stationary point. We deduce that f is indeed 1-Lipschitz, and that it has no ϵ-stationary points except for the origin (which is its minimum, hence stationary). Denote $x_0 := \frac{\delta}{8\epsilon}\mathrm{e}_1$ and notice that its distance from the only ϵ-stationary point of f is $\|x_0 - 0\| = \frac{\delta}{8\epsilon} > C\delta$. It remains to show that x_0 is a (δ, ϵ)-stationary point of f. To that end, consider for any $2 \leq j \leq d :\ z_j := x_0 + \frac{\delta}{2} \cdot \mathrm{e}_j = \frac{\delta}{8\epsilon}\mathrm{e}_1 + \frac{\delta}{2} \cdot \mathrm{e}_j$ which are clearly inside a δ-ball around x_0. Furthermore,

$$\nabla f(z_j) = \frac{1}{\sqrt{\frac{\delta^2}{32} + \frac{\delta^2}{4}}} \cdot \left(\frac{\delta\epsilon}{4} \cdot \mathrm{e}_1 + \frac{\delta}{2} \cdot \mathrm{e}_j \right) = \frac{\sqrt{2}}{3} \cdot \epsilon \cdot \mathrm{e}_1 + \frac{2\sqrt{2}}{3} \cdot \mathrm{e}_j \ .$$

Now, define $g := \frac{1}{d-1} \sum_{j=2}^{d} \nabla f(z_j)$ and notice that by definition $g \in \partial f_\delta(x_0)$ since $\|x - z_j\| < \delta$. This establishes that x_0 is indeed a (δ, ϵ)-stationary point since

$$g = \frac{1}{d-1} \sum_{j=2}^{d} \nabla f(z_j) = \frac{\sqrt{2}}{3}\epsilon \cdot \mathrm{e}_1 + \frac{2\sqrt{2}}{3(d-1)} \sum_{j=2}^{d} \mathrm{e}_j$$

$$\Longrightarrow \ \|g\| = \sqrt{\frac{2\epsilon^2}{9} + \frac{8}{9(d-1)^2} \cdot (d-1)} = \sqrt{\frac{2\epsilon^2}{9} + \frac{8}{9(d-1)}} < \epsilon \ ,$$

where the last inequality is easily verified for $d = \lceil \frac{3}{\epsilon^2} \rceil$. ■