Journal of Machine Learning Research 23 (2022) 1-43 Submitted 12/21; Revised 10/22; Published 10/22

Oracle Complexity in Nonsmooth Nonconvex Optimization

Guy Kornowski GUY.KORNOWSKI@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel

Editor: Tong Zhang

Abstract

It is well-known that given a smooth, bounded-from-below, and possibly nonconvex function, standard gradient-based methods can find ϵ-stationary points (with gradient norm less than ϵ) in $\mathcal{O}(1/\epsilon^2)$ iterations. However, many important nonconvex optimization problems, such as those associated with training modern neural networks, are inherently not smooth, making these results inapplicable. In this paper, we study nonsmooth nonconvex optimization from an oracle complexity viewpoint, where the algorithm is assumed to be given access only to local information about the function at various points. We provide two main results: First, we consider the problem of getting *near* ϵ-stationary points. This is perhaps the most natural relaxation of *finding* ϵ-stationary points, which is impossible in the nonsmooth nonconvex case. We prove that this relaxed goal cannot be achieved efficiently, for any distance and ϵ smaller than some constants. Our second result deals with the possibility of tackling nonsmooth nonconvex optimization by reduction to smooth optimization: Namely, applying smooth optimization methods on a smooth approximation of the objective function. For this approach, we prove under a mild assumption an inherent trade-off between oracle complexity and smoothness: On the one hand, smoothing a nonsmooth nonconvex function can be done very efficiently (e.g., by randomized smoothing), but with dimension-dependent factors in the smoothness parameter, which can strongly affect iteration complexity when plugging into standard smooth optimization methods. On the other hand, these dimension factors can be eliminated with suitable smoothing methods, but only by making the oracle complexity of the smoothing process exponentially large.

Keywords: nonconvex nonsmooth optimization, optimization theory, oracle complexity, smoothing

1. Introduction

We consider optimization problems associated with functions $f : \mathbb{R}^d \to \mathbb{R}$, where $f(\cdot)$ is Lipschitz continuous and bounded from below, but otherwise satisfies no special structure, such as convexity. Clearly, in high dimensions, it is generally impossible to efficiently find a global minimum of a nonconvex function. However, if we relax our goal to finding (approximate) stationary points, then the nonconvexity is no longer an issue. In particular,

it is known that if $f(\cdot)$ is *smooth* – namely, differentiable and with a Lipschitz continuous gradient – then for any $\epsilon > 0$, simple gradient-based algorithms can find $\mathbf{x}$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$, using only $\mathcal{O}(1/\epsilon^2)$ gradient computations, independent of the dimension (see for example Nesterov 2012; Jin et al. 2017; Carmon et al. 2019).

Unfortunately, many optimization problems of interest are inherently *not* smooth. For example, when training modern neural networks, involving max operations and rectified linear units, the associated optimization problem is virtually always nonconvex as well as nonsmooth. Thus, the positive results above, which crucially rely on smoothness, are inapplicable. Although there are positive results even for nonconvex nonsmooth functions, they tend to be either purely asymptotic in nature (e.g., Benaïm et al. 2005; Kiwiel 2007; Zhang and Chen 2009; Davis et al. 2018; Majewski et al. 2018), depend on the internal structure and representation of the objective function, or require additional assumptions which many problems of interest do not satisfy, such as weak convexity or some separation between nonconvex and nonsmooth components[1] (e.g., Cartis et al. 2011; Chen 2012; Duchi and Ruan 2018; Bolte et al. 2018; Davis and Drusvyatskiy 2019; Drusvyatskiy and Paquette 2019; Beck and Hallak 2020). This leads to the interesting question of developing black-box algorithms with non-asymptotic guarantees for nonsmooth nonconvex functions, without assuming any special structure.

In this paper, we study this question via the well-known framework of oracle complexity (Nemirovski and Yudin, 1983): Given a class of functions $\mathcal{F}$, we associate with each $f \in \mathcal{F}$ an *oracle*, which for any $\mathbf{x}$ in the domain of $f(\cdot)$, returns local information about $f(\cdot)$ at $\mathbf{x}$ (such as its value and gradient).[2] We consider iterative algorithms which can be described via an interaction with such an oracle: At every iteration t, the algorithm chooses an iterate $\mathbf{x}_t$, and receives from the oracle local information about $f(\cdot)$ at $\mathbf{x}_t$, which is then used to choose the next iterate $\mathbf{x}_{t+1}$. This framework captures essentially all iterative algorithms for black-box optimization. In this framework, we fix some iteration budget T, and ask what properties can be guaranteed for the iterates $\mathbf{x}_1, \ldots, \mathbf{x}_T$, as a function of T and uniformly over all functions in $\mathcal{F}$ (for example, how close to optimal they are, whether they contain an approximately-stationary point, etc.). Unfortunately, as recently pointed out in Zhang et al. (2020), neither small optimization error (in terms of the function value) nor small gradients can be obtained for nonsmooth nonconvex functions with such local-information algorithms: Indeed, approximately-stationary points can be easily "hidden" inside some arbitrarily small neighborhood, which cannot be found in a bounded number of iterations.

Instead, we consider here two alternative approaches to tackle nonsmooth nonconvex optimization, and provide new oracle complexity results for each. We note that Zhang et al. (2020) recently proposed another promising approach, by defining a certain relaxation of approximate stationarity (so-called (δ, ϵ)-stationarity), and remarkably, prove that points satisfying this relaxed goal can be found via simple iterative algorithms with provable guarantees. However, there exist cases where their definition does not resemble a stationary point in any intuitive case, and thus it remains to be seen whether it is the most appropriate one. We further discuss the pros and cons of their approach in Appendix A.

1. A trivial example arising in deep learning, which does not satisfy most such structural assumptions, is the negative of the ReLU function, $x \mapsto -\max\{0, x\}$.
2. For non-differentiable functions, we use a standard generalization of gradients following Clarke (1990), see Sec. 2 for details.

In our first contribution, we consider relaxing the goal of finding approximately-stationary points, to that of finding *near*-approximately-stationary points: Namely, getting δ-close to a point $\mathbf{x}$ with a (generalized) gradient of norm at most ϵ. This is arguably the most natural way to relax the goal of finding ϵ-stationary points, while hopefully still getting meaningful algorithmic guarantees. Moreover, approaching stationary points is feasible in an asymptotic sense (see for instance Drusvyatskiy and Paquette 2019). Unfortunately, we formally prove that this relaxation already sets the bar too high: For any possibly randomized algorithm interacting with a local oracle, it is impossible to find near-approximately-stationary point with efficient worst-case guarantees, for small enough constant δ, ϵ.

In our second contribution, we consider tackling nonsmooth nonconvex optimization by reduction to smooth optimization: Given the target function $f(\cdot)$, we first find a smooth function $\tilde{f}(\cdot)$ (with Lipschitz gradients) which uniformly approximates it up to some arbitrarily small parameter ϵ, and then apply a smooth optimization method on $\tilde{f}(\cdot)$. Such reductions are common in convex optimization (e.g., Nesterov 2005; Beck and Teboulle 2012; Allen-Zhu and Hazan 2016), and intuitively, should usually lead to points with meaningful properties with respect to the original function $f(\cdot)$, at least when ϵ is small enough. For example, it is known that stationary points of $f(\cdot)$ are the limit of approximately-stationary points of appropriate smoothings of $f(\cdot)$, as $\epsilon \to 0$ (Rockafellar and Wets, 2009, Theorem 9.67).

This naturally leads to the question of how we can find a smooth approximation of a Lipschitz function $f(\cdot)$. Inspecting existing approaches for smoothing nonconvex functions, we notice an interesting trade-off between computational efficiency and the smoothness of the approximating function: On the one hand, there exist optimization-based methods from the functional analysis literature (in particular, Lasry-Lions regularization (Lasry and Lions, 1986)) which yield essentially optimal gradient Lipschitz parameters, but are not computationally efficient. On the other hand, there exist simple, computationally tractable methods (such as randomized smoothing Duchi et al., 2012), which unfortunately lead to much worse gradient Lipschitz parameters, with strong scaling in the input dimension. This in turn leads to larger iteration complexity, when plugging into standard smooth optimization methods. Is this kind of trade-off between computational efficiency and smoothness necessary?

Considering this question from an oracle complexity viewpoint, we prove that this trade-off is indeed necessary under mild assumptions: If we want a smoothing method whose oracle complexity is polynomial in the problem parameters, we must accept that the gradient Lipschitz parameter may be no better than that obtained with randomized smoothing (up to logarithmic factors). Thus, in a sense, randomized smoothing is an optimal nonconvex smoothing method among computationally efficient ones.

It is important to stress that although we formalize our results for general Lipschitz functions, all of our results readily apply to more restricted classes of Lipschitz functions which are often studied in the optimization literature such as Hadamard semi-differentiable, Whitney-stratifiable and semi-algebraic functions. We further discuss this in Remark 2.

Overall, we hope that our work motivates additional research into black-box algorithms for nonconvex, nonsmooth optimization problems, with rigorous finite-time guarantees.

Our paper is structured as follows. In Sec. 2, we formally introduce the notations and terminology that we use. In Sec. 3, we present our results for getting near approximately

stationary points. In Sec. 4, we present our results for smoothing nonsmooth nonconvex functions. In Sec. 5, we provide full proofs. We conclude in Sec. 6 with a discussion of open questions. Our paper also contains a few appendices, which beside technical proofs and lemmas, include further discussion of the notion of (δ, ϵ)-stationarity from Zhang et al. (2020) (Appendix A), and a proof that the dimension-dependency arising from randomized smoothing provably affects the iteration complexity of vanilla gradient descent (Appendix B).

2. Preliminaries

Notation. We let bold-face letters (e.g., $\mathbf{x}$) denote vectors. $\mathbf{0}$ is the zero vector in $\mathbb{R}^d$ (where d is clear from context), and $\mathbf{e}_1, \mathbf{e}_2, \ldots$ are the standard basis vectors. Given a vector $\mathbf{x}$, x_i denotes its i-th coordinate, and $\bar{\mathbf{x}}$ denotes the normalized vector $\mathbf{x}/\|\mathbf{x}\|$ (assuming $\mathbf{x} \neq \mathbf{0}$). $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ denote the standard Euclidean dot product and its induced norm over $\mathbb{R}^d$, respectively. For any real number x, we denote $[x]_+ := \max\{x, 0\}$. Given two functions $f(\cdot), g(\cdot)$ on the same domain $\mathcal{X}$, we define $\|f - g\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - g(\mathbf{x})|$. We denote by $\mathbb{S}^{d-1} := \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \|\mathbf{x}\| = 1 \right\}$ the unit sphere. For any natural N, we abbreviate $[N] := \{1, \ldots, N\}$. We occasionally use standard big-O asymptotic notation: For functions $f, g : \mathcal{X} \to [0, \infty)$ we write $f = \mathcal{O}(g)$ if there exists $c > 0$ such that $f(x) \leq c \cdot g(x)$; $f = \Omega(g)$ if $g = \mathcal{O}(f)$; $f = \Theta(g)$ if $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$. We occasionally hide logarithmic factors by writing $f = \tilde{\mathcal{O}}(g)$ if $f = \mathcal{O}(g \log(g + 1))$, or $f = \tilde{\Omega}(g)$ if $g = \tilde{\mathcal{O}}(f)$, and also denote by $\mathrm{poly}(\cdot)$ polynomial factors.

Gradients and generalized gradients. If a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at $\mathbf{x}$, we denote its gradient by $\nabla f(\mathbf{x})$. For possibly non-differentiable functions, we let $\partial f(\mathbf{x})$ denote the set of *generalized* gradients (following Clarke 1990), arguably the most standard extension of gradients to certain classes of nonsmooth nonconvex functions. For Lipschitz functions (which are almost everywhere differentiable by Rademacher's theorem), one simple way to define it is

$$\partial f(\mathbf{x}) \;:=\; \mathrm{conv}\{\mathbf{u} : \mathbf{u} = \lim_{k \to \infty} \nabla f(\mathbf{x}_k), \mathbf{x}_k \to \mathbf{x}\}$$

(namely, the convex hull of all limit points of $\nabla f(\mathbf{x}_k)$, over all sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of differentiable points of $f(\cdot)$ which converge to $\mathbf{x}$). With this definition, a *stationary point* with respect to $f(\cdot)$ is a point $\mathbf{x}$ satisfying $\mathbf{0} \in \partial f(\mathbf{x})$. Also, given some $\epsilon \geq 0$, we say that $\mathbf{x}$ is an *ϵ-stationary point* with respect to $f(\cdot)$, if there is some $\mathbf{u} \in \partial f(\mathbf{x})$ such that $\|\mathbf{u}\| \leq \epsilon$. Furthermore, when some small δ, ϵ will be clear from context, we will say that $\mathbf{x}$ is a near-approximately-stationary point if $\|\mathbf{x} - \mathbf{x}'\| < \delta$ for some ϵ-stationary point $\mathbf{x}'$.

Local oracles. We consider oracles that given a function $f(\cdot)$ and a point $\mathbf{x}$, return some quantity $\mathbb{O}_f(\mathbf{x})$ which conveys local information about the function near that point. Formally (following Braun et al. 2017), we call an oracle *local* if for any $\mathbf{x}$ and any two functions f, g that are equal over some neighborhood of $\mathbf{x}$, it holds that $\mathbb{O}_f(\mathbf{x}) = \mathbb{O}_g(\mathbf{x})$. An important example is the first order oracle $\mathbb{O}_f(\mathbf{x}) = (f(\mathbf{x}), \partial f(\mathbf{x}))$, but one can consider more sophisticated oracles such as those which return all high order derivative information, wherever they exist. We choose to work in this generality since our results hold for any local oracle whatsoever, although we note that in the nonconvex-nonsmooth setting, it remains

to be seen whether in the general Lipschitz setting discussed in this paper there is any use for any local information which is not first order.

3. Hardness of Getting Near Approximately-Stationary Points

In this section, we present our first main result, which establishes the hardness of getting near approximately-stationary points.

To avoid trivialities, and following Zhang et al. (2020), we will focus on functions $f(\cdot)$ which are Lipschitz and bounded from below. In particular, we will assume that $f(\mathbf{0}) - \inf_{\mathbf{x}} f(\mathbf{x})$ is upper bounded by a constant. We note that this is without loss of generality, as $\mathbf{0}$ can be replaced by any other fixed reference point. We consider possibly randomized algorithms which interact with some local oracle. Such an algorithm first produces $\mathbf{x}_1$ possibly at random, receives $\mathbb{O}_f(\mathbf{x}_1)$ for some local oracle $\mathbb{O}_f$, and for every $t > 1$ produces $\mathbf{x}_t$ possibly at random based on previously observed responses. Our main result in this section is the following:

Theorem 1 *There exist absolute constants $c, C > 0$ such that for any algorithm $\mathcal{A}$ interacting with a local oracle, and any $T \in \mathbb{N}$, $d \geq 2$, there is a function $f(\cdot)$ on $\mathbb{R}^d$ such that*

- *$f(\cdot)$ is C-Lipschitz, $f(\mathbf{0}) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq C$, $\inf\{\|\mathbf{x}\| \mid \partial f(\mathbf{x}) = \{\mathbf{0}\}\} \leq C$.*
- *With probability at least $1 - CT\exp(-cd)$ over the algorithm's randomness, the iterates $\mathbf{x}_1, \ldots, \mathbf{x}_T$ produced by the algorithm satisfy*

$$\inf_{\mathbf{x} \in \Sigma} \min_{t \in [T]} \|\mathbf{x}_t - \mathbf{x}\| \geq c \ ,$$

where Σ is the set of c-stationary points of $f(\cdot)$.

The theorem implies that it is impossible to obtain worst-case guarantees for finding near-approximately-stationary points of Lipschitz, bounded-from-below functions unless T has exponential dependence on d. Note that if an exponential dimension dependence is allowed, then under the theorem's assumptions, one can trivially produce points close to a stationary point (or to any point, for that matter) using an exhaustive grid search. The theorem implies that no oracle-based algorithm will be significantly more efficient than this trivial strategy. Further notice that since every deterministic algorithm can be viewed as a randomized algorithm with a trivial distribution over its decisions, the theorem above readily holds for deterministic algorithms as well. That being the case, the lower bound described in the second bullet holds deterministically.

Remark 2 (More assumptions on $f(\cdot)$) *The Lipschitz functions $f(\cdot)$ used to prove the theorem are based on a composition of affine functions, orthogonal projections, the Euclidean norm function $\mathbf{x} \mapsto \|\mathbf{x}\|$, and the max function. Thus, the result also holds for more specific families of functions considered in the literature, which satisfy additional regularity properties, as long as they contain any Lipschitz functions composed as above. These include, for example, Hadamard semi-differentiable functions (Zhang et al., 2020), Whitney-stratifiable functions (Bolte et al., 2007; Davis et al., 2018) and semi-algebraic functions.*

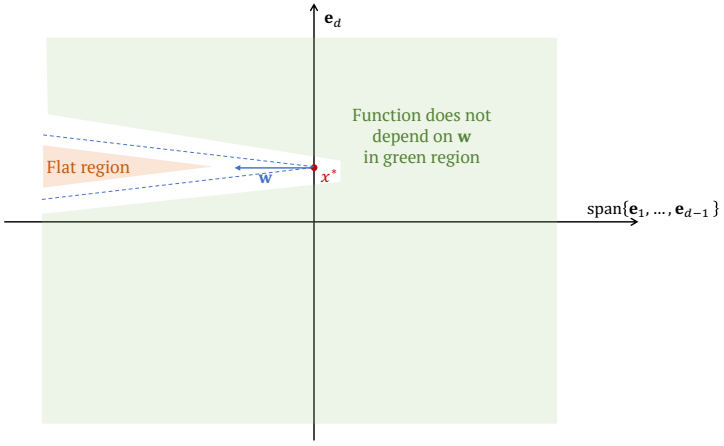

Figure 1: Illustration of the function used in the proof of Thm. 1.

The formal proof of the theorem appears in Sec. 5, but can be informally described as follows: First, we construct an algorithm-dependent one dimensional "hard" Lipschitz function $h : \mathbb{R} \to \mathbb{R}$ that has large derivatives everywhere except for a single point x^*. By "hard", we mean that after any finite number of steps of the algorithm, we can provide some small neighborhood of x^* which the algorithm is likely not to enter. Based on h, we construct a Lipschitz function on $\mathbb{R}^d$, specified by a small vector $\mathbf{w} \in \text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{d-1}\}$, which resembles the function $\mathbf{x} \mapsto \|(x_1, \ldots, x_{d-1})\|_2 + h(x_d)$ in "most" of $\mathbb{R}^d$, but with a "channel" leading away from a small neighborhood of $x^* \cdot \mathbf{e}_d$ in the direction of $\mathbf{w}$, and reaching a completely flat region (see Fig. 1). In high dimensions, the channel and the flat region contain a vanishingly small portion of $\mathbb{R}^d$. This function has the property of having ϵ-stationary points only in the flat region which is in the direction of $x^* \cdot \mathbf{e}_d + \mathbf{w}$, even though the function appears in most places like a function which does not depend on $\mathbf{w}$. As a result, any oracle-based algorithm that doesn't get to close to x^* in the d'th coordinate and doesn't know $\mathbf{w}$, is unlikely to hit the vanishingly small region where the function differs from $\mathbf{x} \mapsto \|(x_1, \ldots, x_{d-1})\|_2 + h(x_d)$, receiving no information about $\mathbf{w}$, and thus cannot determine where the ϵ-stationary points lie. As a result, such an algorithm cannot return near-approximately-stationary points.

4. Smoothing Nonsmooth Nonconvex Functions

In this section, we turn to our second main contribution, examining the possibility of reducing nonsmooth nonconvex optimization to smooth nonconvex optimization, by running a smooth optimization method on a smooth approximation of the objective function. This longstanding approach for nonsmooth nonconvex optimization has been examined by many works, initially driven by practical applications (Blake and Zisserman, 1987; Wu, 1996) complemented by further theoretical analyses (Mobahi and Fisher III, 2015; Hazan et al., 2016). In what follows, we focus our discussion on 1-Lipschitz functions: This is without loss of generality, since if our objective is L-Lipschitz, we can simply rescale it by L (and a

Lipschitz assumption is always necessary if we wish to obtain a Lipschitz-gradient smooth approximation). Also, we focus on smoothing functions over all of $\mathbb{R}^d$ for simplicity, but our results and proofs easily extend to the case where we are only interested in smoothing over some bounded domain on which the function is Lipschitz.

For a nonsmooth *convex* function $f(\cdot)$, a well-known smoothing approach is proximal smoothing (also known as the Moreau envelope or Moreau-Yosida regularization, see Moreau, 1965; Bauschke et al., 2011) defined as $P_\delta(f)$ where

$$P_\delta(f)(\mathbf{x}) := \min_{\mathbf{y}} \left(f(\mathbf{y}) + \frac{1}{2\delta}\|\mathbf{y}-\mathbf{x}\|^2 \right) . \tag{1}$$

By picking δ appropriately, $P_\delta(f)$ is an arbitrarily good smooth approximation of f; more formally, if f is 1-Lipschitz, then for any $\epsilon > 0$, there exists a choice of $\delta = \Theta(\epsilon)$ such that $\|P_\delta(f) - f\|_\infty \leq \epsilon$, with the gradients of $P_\delta(f)$ being $\frac{1}{\epsilon}$-Lipschitz (Beck and Teboulle, 2012, Section 4.2). This is essentially optimal, as no ϵ-approximation can attain a gradient Lipschitz parameter better than $\Omega(1/\epsilon)$ (see Lemma 30 in Appendix D for a formal proof). Finally, computing gradients of $P_\delta(f)$ (which can then be fed into a gradient-based smooth optimization method) is feasible, given a solution to Eq. (1), which is a convex optimization problem and hence efficiently solvable in general.

Unfortunately, for nonconvex functions, proximal smoothing (or other smoothing methods from convex optimization) generally fails in producing smooth approximations. However, it turns out that similar guarantees can be obtained with a slightly more complicated procedure, known as *Lasry-Lions* regularization in the functional analysis literature (Lasry and Lions, 1986; Attouch and Aze, 1993), which is essentially a double application of proximal smoothing combined with function flipping. One way to define it is as follows:

$$P_{\delta,\nu}(f)(\mathbf{x}) \;:=\; -P_\delta(-P_\nu(f))(\mathbf{x}) \;=\; \max_{\mathbf{y}} \min_{\mathbf{z}} \left(f(\mathbf{z}) + \frac{1}{2\nu}\|\mathbf{z}-\mathbf{y}\|^2 - \frac{1}{2\delta}\|\mathbf{y}-\mathbf{x}\|^2 \right) .$$

Once more, if f is 1-Lipschitz, then choosing $\delta, \nu = \Theta(\epsilon)$ appropriately, we get an ϵ-accurate approximation of f, with gradients which are $\frac{c}{\epsilon}$-Lipschitz for some absolute constant c. However, unlike the convex case, implementing this smoothing involves solving a non-convex optimization problem, which may not be computationally tractable.

Alternatively, a very simple smoothing approach, which works equally well on convex and non-convex problems, is randomized smoothing, or equivalently, convolving the objective function with a smoothness-inducing density function. Formally, given the objective function f and a distribution P, we define $\tilde{f}(\mathbf{x}) := \mathbb{E}_{\mathbf{y}\sim P}[f(\mathbf{x}+\mathbf{y})]$. In particular, letting P be a uniform distribution on an origin-centered ball of radius ϵ, the resulting function is an ϵ-approximation of $f(\cdot)$, and its gradient Lipschitz parameter is $\frac{c\sqrt{d}}{\epsilon}$, where c is an absolute constant and d is the input dimension (Duchi et al., 2012, Lemma 8). Moreover, given access to values and gradients of $f(\cdot)$, computing unbiased stochastic estimates of the values or gradients of $\tilde{f}(\cdot)$ is computationally very easy: We just sample a single $\mathbf{y} \sim P$, and return $f(\mathbf{x}+\mathbf{y})$ or $\nabla f(\mathbf{x}+\mathbf{y})$.[3] These stochastic estimates can then be plugged into stochastic methods for smooth optimization (see Duchi et al. 2012; Ghadimi and Lan 2013).

3. Recall that by Rademacher's theorem, f is differentiable almost everywhere hence $\nabla f(\mathbf{x}+\mathbf{y})$ exists almost surely.

Comparing these two approaches, we see an interesting potential trade-off between the smoothness obtained and computational efficiency, summarized in the following table:

	$\nabla\tilde{f}$ Lipschitz param.	Computationally Efficient?
Randomized Smoothing	$c \cdot \sqrt{d}/\epsilon$	✓
Lasry-Lions Regularization	c/ϵ	×

In words, randomized smoothing is computationally efficient (unlike Lasry-Lions regularization), but at the cost of a much larger gradient Lipschitz parameter. Since the iteration complexity of smooth optimization methods strongly depend on this Lipschitz parameter, it follows that in high-dimensional problems, we pay a high price for computational tractability in reducing nonsmooth to smooth problems. As we demonstrate in Appendix B, this is a real phenomenon and not just an artifact of iteration complexity analysis, at least for gradient descent. Roughly speaking, we prove in Proposition 24 that when gradient descent is combined with randomized smoothing, it is impossible to guarantee getting to ϵ-stationary points of the smoothed function within $\Omega(\sqrt{d}/\epsilon)$ iterations.

This discussion leads to a natural question: Is this trade-off necessary, or perhaps there exist computationally efficient methods which can improve on randomized smoothing, in terms of the gradient Lipschitz parameter? Using an oracle complexity framework, we prove that this trade-off is indeed necessary (under mild assumptions), and that randomized smoothing is essentially an optimal method under the constraint of black-box access to the objective $f(\cdot)$, and a reasonable oracle complexity. We note that Duchi et al. (2012) proved that the Lipschitz constant cannot be improved by simple randomized smoothing schemes, but here we consider a much larger class of possible methods.

4.1 Smoothing Algorithms

Before presenting our main result for this section, we need to carefully formalize what we mean by an efficient smoothing method, since "returning" a smooth approximating function over $\mathbb{R}^d$ is not algorithmically well-defined. Recalling the motivation to our problem, we want a method that given a nonsmooth objective function $f(\cdot)$, allows us to estimate values and gradients of a smooth approximation $\tilde{f}(\cdot)$ at arbitrary points, which can then be fed into standard black-box methods for smooth optimization (hence, we need a uniform approximation property). Moreover, for black-box optimization, it is desirable that this smoothing method operates in an oracle complexity framework, where it only requires local information about $f(\cdot)$ at various points. Finally, we are interested in controlling various parameters of the smoothing procedure, such as the degree of approximation, the smoothness of the resulting function, and the complexity of the procedure. In light of these considerations, a natural way to formalize smoothing methods is the following:

Definition 3 *An algorithm $\mathcal{S}$ is an (L, ϵ, T, M, r)-smoother if for any 1-Lipschitz function f on $\mathbb{R}^d$, there exists a differentiable function $\tilde{f}$ on $\mathbb{R}^d$ with the following properties:*

1. *$\|f - \tilde{f}\|_\infty \leq \epsilon$, and $\nabla\tilde{f}$ is L-Lipschitz.*
2. *Given any $\mathbf{x} \in \mathbb{R}^d$, the algorithm produces a (possibly randomized) query sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \subset \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq r\}$, of the form $\mathbf{x}_{i+1} = \mathcal{S}^{(i)}(\xi, \mathbf{x}, \mathbb{O}_f(\mathbf{x}_1), \ldots, \mathbb{O}_f(\mathbf{x}_i))$,*

where $\mathcal{S}^{(i)}$ maps all the previous information gathered by the queries of some local oracle $\mathbb{O}_f$ to a new query, possibly based on a draw of some random variable ξ.[4] *Finally, the algorithm produces a vector*

$$\mathcal{S}(f, \mathbf{x}) := \mathcal{S}^{(out)}\left(\xi, \mathbf{x}, \mathbb{O}_f(\mathbf{x}_1), \ldots, \mathbb{O}_f(\mathbf{x}_T)\right) \ ,$$

where $\mathcal{S}^{(out)}$ is some mapping to $\mathbb{R}^d$, such that

$$\left\| \mathbb{E}_\xi \left[\mathcal{S}(f, \mathbf{x})\right] - \nabla \tilde{f}(\mathbf{x}) \right\| \leq \epsilon \quad \text{and} \quad \Pr_\xi \left[\|\mathcal{S}(f, \mathbf{x})\| \leq M\right] = 1 \ . \tag{2}$$

Some comments about this definition are in order. First, the definition is only with respect to the ability of the algorithm to approximate gradients of $\tilde{f}(\cdot)$: It is quite possible that the algorithm also has additional output (such as an approximation of the value of $\tilde{f}(\cdot)$), but this is not required for our results. Second, we do not require the algorithm to return $\nabla \tilde{f}(\mathbf{x})$: It is enough that the expectation of the vector output is close to it (up to ϵ). This formulation captures both deterministic optimization-type methods (such as Lasry-Lions regularization, where in general we can only hope to solve the auxiliary optimization problem up to some finite precision) as well as stochastic methods (such as randomized smoothing, which returns $\nabla \tilde{f}(\mathbf{x})$ in expectation). Third, we assume that the queries returned by the algorithm lie at a bounded distance r from the input point $\mathbf{x}$. In the context of randomized smoothing, this corresponds (for example) to using a uniform distribution over a ball of radius r centered on $\mathbf{x}$. As we discuss later on, some assumption on the magnitude of the queries is necessary for our proof technique. However, requiring almost-sure boundedness is merely for simplicity, and it can be replaced by a high-probability bound with appropriate tail assumptions (e.g., if we are performing randomized smoothing with a Gaussian distribution), at the cost of making the analysis a bit more complicated.

Remark 4 *We are mostly interested (though do not limit our results) to the following parameter regimes:*

- *$T = poly\left(d, L, \epsilon^{-1}\right)$, essentially meaning that a single call to $\mathcal{S}$ is computable in a reasonable amount of time.*
- *$M = poly\left(L\right)$. As we formally prove in Lemma 31 in Appendix D, if we require $\tilde{f}$ to approximate f and also have L-Lipschitz gradients, we must have $\|\nabla \tilde{f}(x)\| = \mathcal{O}(L)$. In particular, whenever M is sufficiently larger than L, Eq. (2) is interchangeable with the seemingly more natural condition $\Pr_\xi \left[\|\mathcal{S}(f, \mathbf{x}) - \nabla \tilde{f}(\mathbf{x})\| \leq M\right] = 1$.*
- *$r = \mathcal{O}(\epsilon)$. If we are interested in smoothing a 1-Lipschitz, nonconvex function up to an accuracy ϵ around a given point $\mathbf{x}$, we generally expect that only its $\mathcal{O}(\epsilon)$-neighborhood will convey useful information for the smoothing process. We note that this regime is indeed satisfied by randomized smoothing (with a uniform distribution around a radius-ϵ ball, or with high probability if we use a Gaussian distribution), as well as Lasry-Lions regularization (in the sense that the smooth approximation at $\mathbf{x}$ does not change if we alter the function arbitrarily outside an $\mathcal{O}(\epsilon)$-neighborhood of $\mathbf{x}$).*

4. We assume nothing about ξ, allowing the algorithm to utilize an arbitrary amount of randomness.

We consider all three of the above to be quite permissive. In particular, notice that randomized smoothing over a ball satisfies the much stronger $T = 1,\ M = 1,\ r = \epsilon$.[5]

Our result will require the assumption that the smoothing algorithm $\mathcal{S}$ is *translation invariant with respect to constant functions*, in the sense that it treats all constant functions and regions of the input space in the same manner. We formalize our desired translation-invariance property as follows:

Definition 5 *A smoothing algorithm* $\mathcal{S}$ *satisfies* TICF *(translation invariance w.r.t. constant functions) if for any two constant functions* f, g*, any* $\mathbf{x} \in \mathbb{R}^d$ *and* $i \in [T]$*, and any realization of* ξ*,*

$$\mathcal{S}^{(i)}\left(\xi, \mathbf{x}, \mathbb{O}_f\left(\mathbf{x}_1\right), \ldots, \mathbb{O}_f\left(\mathbf{x}_i\right)\right) = \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_g\left(\mathbf{x}_1 - \mathbf{x}\right), \ldots, \mathbb{O}_g\left(\mathbf{x}_i - \mathbf{x}\right)\right) + \mathbf{x}\ , \tag{3}$$

and

$$\mathcal{S}^{(out)}\left(\xi, \mathbf{x}, \mathbb{O}_f\left(\mathbf{x}_1\right), \ldots, \mathbb{O}_f\left(\mathbf{x}_T\right)\right) = \mathcal{S}^{(out)}\left(\xi, \mathbf{0}, \mathbb{O}_g\left(\mathbf{x}_1 - \mathbf{x}\right), \ldots, \mathbb{O}_g\left(\mathbf{x}_1 - \mathbf{x}\right)\right)\ .$$

In other words, if instead of a constant function f and an input point $\mathbf{x}$, we pick some other constant function g and the origin, the distribution of the algorithm's sequence of queries remain the same (up to a shift by $-\mathbf{x}$), and the gradient estimate returned by the algorithm remains the same. We consider this to be a mild and natural assumption, which is clearly satisfied by standard smoothing techniques.

4.2 Main result

With these definitions in hand, we are finally ready to present our main result for this section, which is the following:

Theorem 6 *Let* $\mathcal{S}$ *be an* (L, ϵ, T, M, r)*-smoother which satisfies* TICF*. Then*

$$L\sqrt{\log\left(\left(M+1\right)\left(T+1\right)\right)} \geq c_1 \cdot \frac{\sqrt{d}}{r}\left(c_2 - \epsilon\right) \tag{4}$$

for some absolute constants $c_1, c_2 > 0$*.*

This theorem holds for general values of the parameters L, ϵ, T, M, r. Concretely, for parameter regimes of interest (see Remark 4) we have the following corollary:

Corollary 7 *Suppose that the accuracy parameter satisfies* $\epsilon \leq c_2/2$*. Then any smoothing algorithm which makes at most* $T = poly(d, L, \epsilon^{-1})$ *queries at a distance at most* $r = \mathcal{O}(\epsilon)$ *from the input point, and returns vectors of norm at most* $M = poly(d, L, \epsilon^{-1})$*, must correspond to a smooth approximation* $\tilde{f}(\cdot)$ *with Lipschitz gradient parameter* at least $L = \tilde{\Omega}(\sqrt{d}/\epsilon)$*.*

5. Indeed, $\mathcal{S}(f, \mathbf{x}) := \nabla f(\mathbf{x}+\mathbf{z})$ where $\mathbf{z} \sim \mathrm{Unif}(\epsilon\mathbb{S}^{d-1})$ requires only a single query (namely $T = 1$) at $\mathbf{x}+\mathbf{z}$ which is of distance at most ϵ away from $\mathbf{x}$ (hence $r = \epsilon$) and satisfies $\|\mathcal{S}(f, \mathbf{x})\| \leq 1$ due to Lipschitzness (hence $M = 1$).

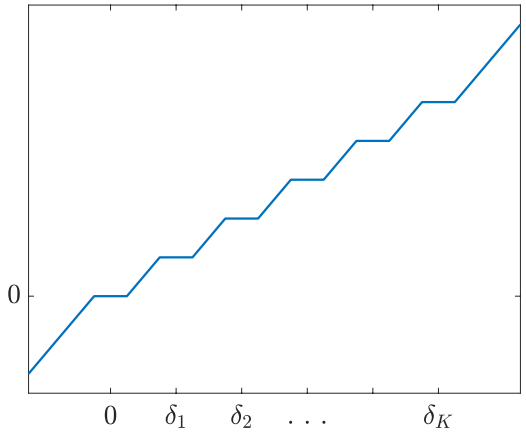

Figure 2: Illustration of $g(x)$, where $\Delta = \{0, \delta_1, \ldots, \delta_K\} \subset [0, 1]$.

We note that the lower bound on L in this corollary matches (up to logarithmic factors) the upper bound attained by randomized smoothing. This implies that at least under our framework and assumptions, randomized smoothing is an essentially optimal efficient smoothing method.

Another implication of Thm. 6 is that even if we relax our assumption that $r = \mathcal{O}(\epsilon)$, then as long as r does not scale with the dimension d, the gradient Lipschitz parameter of any efficient smoothing algorithm must scale with the dimension (even though there exist dimension-free smooth approximations, as evidenced by Lasry-Lions regularization):

Corollary 8 *Fix any accuracy parameter $\epsilon \leq c_2/2$, and any $r > 0$. Then as long as the number of queries is $T = poly(d)$ and the output is of size $M = poly(d)$, we must have $L \geq \tilde{\Omega}(\sqrt{d})$.*

A third corollary of our theorem is that (perhaps unsurprisingly), there is no way to implement Lasry-Lions regularization efficiently in an oracle complexity framework:

Corollary 9 *If $\epsilon \leq c_2/2$ and $M = poly(d)$, then any smoothing algorithm for which $\tilde{f}(\cdot)$ corresponds to the Lasry-Lions regularization (which satisfies $L = \mathcal{O}(1)$) must use a number of queries $T = \exp(\tilde{\Omega}(d))$.*

Remark 10 (Dependence on M) *Our definition of a smoothing algorithm focuses on the expectation of the algorithm's output. This leads to a logarithmic dependence on M (an upper bound on the algorithm's output) in Thm. 6, since in the proof we need to bound the influence of exponentially-small-probability events on the expectation. It is plausible that the dependence on M can be eliminated altogether, by changing the definition of a smoothing algorithm to focus on the expectation of its output, conditioned on all but highly-improbably events. However, that would complicate the definition and our proof.*

Before presenting the formal proof of Thm. 6 in the next section, we outline the main proof idea. Consider a one dimensional monotonically increasing function g, which is locally constant at a $\Omega(1/\sqrt{d})$ neighborhood of a grid $\Delta = \{0, \delta_1, \ldots, \delta_K\}$ of points in $[0, 1]$, with K

roughly of order $\sqrt{d}$ (see Fig. 2). We define $f(\mathbf{x}) = g(\mathbf{w}^\top \mathbf{x})$, where $\mathbf{w} \in \mathbb{S}^{d-1}$ is a uniformly random unit vector. We note that f is a simple function, easily implemented by (say) a one-hidden layer ReLU neural network, and that it belongs to all of the restricted classes discussed in Remark 2.

We now proceed to analyze what happens when a smoothing algorithm is given points of the form $\delta_i \mathbf{w}$, for $\delta_i \in \Delta$. Since $\mathbf{w}$ is random, and the algorithm is assumed to be translation invariant, it can be shown (via a concentration of measure argument) that the algorithm is overwhelmingly likely to produce queries in directions which all have $\tilde{\mathcal{O}}(1/\sqrt{d})$ correlation with $\mathbf{w}$, as long as the number of queries is polynomial. Consequently, with high probability, the queries all lie in a region where the function $f(\cdot)$ is flat, and the algorithm cannot distinguish between it and a constant function. By the translation-invariance property, this implies that the gradient estimates $\nabla \tilde{f}(\delta_i \mathbf{w})$ must be of small norm, uniformly for all $\delta_i \mathbf{w}$. Combining the observation that $\nabla \tilde{f}(\cdot)$ is small along order-of-$\sqrt{d}$-many points between $\mathbf{0}$ and the unit vector $\mathbf{w}$, together with the fact that $\nabla \tilde{f}(\cdot)$ is L-Lipschitz, we can derive an upper bound on how much $\tilde{f}(\cdot)$ can increase along the line segment between $\mathbf{0}$ and $\mathbf{w}$, roughly on the order of $L/\sqrt{d}$. On the other hand, $\tilde{f}(\cdot)$ is an approximation of f, which has a constant increase between $\mathbf{0}$ and $\mathbf{w}$. Overall this allows us to deduce a lower bound on L scaling as $\sqrt{d}$, which results in the theorem.

5. Proofs

5.1 Proof of Thm. 1

We start by claiming that when optimizing a one dimensional Lipschitz function that has large derivatives everywhere apart from it's global minimum, no algorithm can guarantee this minimum can be *exactly* found with some positive probability within any finite time. The following proposition formalizes this claim.

Proposition 11 *For any algorithm $\mathcal{A}$ interacting with a local oracle, any $T \in \mathbb{N}$ and any $\delta > 0$, there exists a function $h : \mathbb{R} \to [0, \infty)$ and $\rho(T, \delta) > 0$ such that*

- *h is 2-Lipschitz, with a single global minimum $x^* \in (0, 1)$, and $h(0) = 1$.*
- *$\forall x \neq x^*,\ \forall g \in \partial h(x) :\ |g| \geq 1$.*
- *The first T iterates produced by $\mathcal{A}(h)$: $x_1^{\mathcal{A}(h)}, \ldots, x_T^{\mathcal{A}(h)}$, satisfy $\Pr_{\mathcal{A}}[\min_{t \in [T]} |x_t - x^*| < \rho] < \delta$.*

Remark 12 *The quantity $\rho(T, \delta)$ given by Proposition 11 depends only on T, δ. In particular, it is worth noticing that it is uniform over any algorithm $\mathcal{A}$.*

The proof of Proposition 11 is inspired by a classic lower bound construction for convex optimization due to Nemirovski (Nemirovski, 1995, Lemma 1.1.1). The basic idea is to construct two functions that are identical outside some segment, in which their distinct minima lie at distance 2ρ one from another. Any algorithm that queries only outside that segment throughout it's first $T - 1$ iterates cannot distinguish between the two functions, thus has probability at least $\frac{1}{2}$ to produce it's next iterate at least ρ away from the minimum of the

function it is actually optimizing (see Fig. 5 in Appendix C for an illustration). By looking at smaller and smaller segments, this idea can be generalized to any number of queries. Unfortunately, the gradient size of Nemirovski's original construction shrinks exponentially with T, therefore cannot be applied to our setting as it does not satisfy the second bullet of Proposition 11, which is crucial to the rest of our proof. Nonetheless, we were able to provide a nonconvex construction similar in spirit which satisfies our desired qualities. Due to the substantial length and technicality of the proof, we defer it to Appendix C. We continue by showing that given Proposition 11, this claim generalizes to any dimension.

Lemma 13 *For any algorithm $\mathcal{A}$ interacting with a local oracle, any $T \in \mathbb{N}$, $d \geq 2$ and any $\delta > 0$, there exists a function $h : \mathbb{R} \to [0, \infty)$ and $\rho(T, \delta) > 0$ such that*

- *h is 2-Lipschitz, with a single global minimum $x^* \in (0, 1)$, and $h(0) = 1$.*
- *$\forall x \neq x^*,\ \forall g \in \partial h(x) :\ |g| \geq 1$.*
- *If we define $\bar{f}_h(x_1, \ldots, x_d) := h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}$, then the first T iterates produced by $\mathcal{A}(\bar{f}_h)$: $\mathbf{x}_1^{\mathcal{A}(\bar{f}_h)}, \ldots, \mathbf{x}_T^{\mathcal{A}(\bar{f}_h)}$, satisfy $\Pr_{\mathcal{A}}\left[\min_{t\in[T]} \|\mathbf{x}_t^{\mathcal{A}(\bar{f}_h)} - \mathbf{x}^*\| < \rho\right] < \delta$, where $\mathbf{x}^* := (0, \ldots, 0, x^*)$.*

Proof

Since our goal is to provide a lower bound ρ for optimizing $\bar{f}_h$ with the algorithm $\mathcal{A}$, we can assume without loss of generality that $\mathcal{A}$ interacts with an even stronger oracle which provides more than just local information about $\bar{f}_h$. Specifically, suppose $\mathcal{A}$ has access to an oracle of the form

$$\overline{\mathbb{O}}_{\bar{f}_h}(\mathbf{x}) = \left(\left\{((z_1, \ldots, z_{d-1}, x_d), \bar{f}_h(z_1, \ldots, z_{d-1}, x_d)) \middle| (z_1, \ldots, z_{d-1}) \in \mathbb{R}^{d-1}\right\}, \mathbb{O}_h(x_d)\right)$$

for some local oracle $\mathbb{O}$. Namely, a full global description of the function $\bar{f}_h$ over the affine subspace $\{\mathbf{z} |\ z_d = x_d\}$, coupled with local information with respect to the last coordinate. Note that by definition of $\bar{f}_h(\mathbf{x})$, changing x_d only affects $h(x_d)$, thus all the information about x_d is indeed conveyed through $h(x_d)$.

Assuming $\mathcal{A}$ interacts with the described $\overline{\mathbb{O}}$, we turn to describe *another* algorithm $\mathcal{A}'$, which given local oracle access to a one dimensional function h works as follows:

- $\mathcal{A}'$ simulates $\mathcal{A}$ and receives it's first iterate $\mathbf{x}_1$. It then produces the first iterate $(\mathbf{x}_1)_d$.
- Given $\mathcal{A}$'s current iterate $\mathbf{x}_t$, $\mathcal{A}'$ queries $\mathbb{O}_h((\mathbf{x}_t)_d)$. Then, $\mathcal{A}'$ feeds into $\mathcal{A}$

  $$\left(\left\{((z_1, \ldots, z_{d-1}, (\mathbf{x}_t)_d), \bar{f}_h(z_1, \ldots, z_{d-1}, (\mathbf{x}_t)_d) \middle| (z_1, \ldots, z_{d-1}) \in \mathbb{R}^{d-1}\right\}, \mathbb{O}_h((\mathbf{x}_t)_d)\right),$$

  and receives from $\mathcal{A}$ it's next iterate $\mathbf{x}_{t+1}$. $\mathcal{A}'$ then produces it's next iterate $(\mathbf{x}_{t+1})_d$.

It is clear that $\mathcal{A}'$ is indeed a well defined algorithm which interacts with a local oracle. Hence, let $h(\cdot), \rho(T, \delta)$ be the function and the positive parameter given by Proposition 11 for $\mathcal{A}', T, \delta$. We will show these h, ρ satisfy all three bullets in the lemma. The first two

bullets are immediate, thus it only remains to prove the third. To that end, note that by construction of $\mathcal{A}'$, $\mathcal{A}'$'s iterates when applied to h are exactly the d'th coordinates of $\mathcal{A}$ when applied to $\bar{f}$. That is, $(\mathbf{x}_t^{\mathcal{A}(\bar{f})})_d = x_t^{\mathcal{A}'(h)}$. Consequently,

$$\begin{aligned}\Pr_{\mathcal{A}}\left[\min_{t\in[T]} \|\mathbf{x}_t^{\mathcal{A}(\bar{f})} - \mathbf{x}^*\| < \rho\right] &\leq \Pr_{\mathcal{A}}\left[\min_{t\in[T]} |(\mathbf{x}_t^{\mathcal{A}(\bar{f})})_d - (\mathbf{x}^*)_d| < \rho\right] \\ &= \Pr_{\mathcal{A}'}\left[\min_{t\in[T]} |x_t^{\mathcal{A}'(h)} - x^*| < \rho\right] < \delta \ ,\end{aligned}$$

where the last inequality follows by definition of h, ρ. ■

From now on, we fix some algorithm $\mathcal{A}$ interacting with a local oracle, $T \in \mathbb{N}$, $d \geq 2$ and set $\delta = T\exp(-d/36)$. We denote by $h(\cdot),\ \rho > 0,\ \mathbf{x}^* \in \mathbb{R}^d$ their associated function, positive parameter and point given by Lemma 13. Given any nonzero vector $\mathbf{w} \in \mathbb{R}^d$ such that $w_d = 0$, we define

$$f_{\mathbf{w}}(x_1, \ldots, x_d) := h(x_d + x^*) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2} - \left[\langle \overline{\mathbf{w}}, \mathbf{x} + \mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x} + \mathbf{w}\|\right]_+ \ ,$$

where $\overline{\mathbf{w}} := \mathbf{w}/\|\mathbf{w}\|$. This function looks like the "hard" function given by Lemma 13 (up to a shift along the d'th axis) as long as $\mathbf{x}$ is not highly correlated with $\mathbf{w}$, which makes the ReLU term vanish. However, unlike the hard function which has a stationary point, the ReLU term adds at that point a large gradient component in the $\mathbf{w}$ direction, preventing it from being even ϵ-stationary. Moreover, the following lemma shows that this function has no ϵ-stationary points for small ϵ.

Lemma 14 *For any nonzero $\mathbf{w} \in \mathbb{R}^d$ such that $w_d = 0$, $f_{\mathbf{w}}$ is $\frac{15}{4}$-Lipschitz and has no ϵ-stationary points for any $\epsilon < \frac{1}{4\sqrt{2}}$.*

Proof Throughout the proof we omit the $\mathbf{w}$ subscript and refer to $f_{\mathbf{w}}(\cdot)$ as $f(\cdot)$.

The functions $\mathbf{x} \mapsto h(x_d + x^*)$, $\mathbf{x} \mapsto \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}$, $\mathbf{x} \mapsto \langle \overline{\mathbf{w}}, \mathbf{w} + \mathbf{x}\rangle$, $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{x} + \mathbf{w}\|$ and $x \mapsto [x]_+$ are 2-Lipschitz, $\frac{1}{4}$-Lipschitz, 1-Lipschitz, $\frac{1}{2}$-Lipschitz and 1-Lipschitz respectively, from which it follows that f is $2 + \frac{1}{4} + 1 + \frac{1}{2} = \frac{15}{4}$-Lipschitz.

In order to prove that no point $\mathbf{x}$ is ϵ-stationary for any $\epsilon < \frac{1}{4\sqrt{2}}$, we will use the facts that $\partial(g_1 + g_2) \subseteq \partial g_1 + \partial g_2$, and that if g_1 is univariate, $\partial(g_1 \circ g_2)(\mathbf{x}) \subseteq \mathrm{conv}\{r_1\mathbf{r}_2 : r_1 \in \partial g_1(g_2(\mathbf{x})), \mathbf{r}_2 \in \partial g_2(\mathbf{x})\}$ (see Clarke, 1990, Proposition 2.3.3 and Theorem 2.3.9). We examine six exhaustive cases:

- $x_d \neq 0$. In this case we have

$$\partial f(\mathbf{x}) \subseteq \left\{ t \cdot \mathbf{e}_d + \frac{1}{4}\mathbf{u} - s\left(\overline{\mathbf{w}} - \frac{1}{2}\mathbf{v}\right) \,\middle|\, |t| \geq 1,\ u_d = 0,\ s \in [0,1],\ \|\mathbf{v}\| \leq 1 \right\} \ .$$

  For any $\mathbf{g} \in \partial f(\mathbf{x})$ corresponding to some $t, \mathbf{u}, s, \mathbf{v}$, using the fact that $u_d = w_d = 0$ we get

$$\|\mathbf{g}\| \geq \langle \mathbf{g}, \mathrm{sign}(t) \cdot \mathbf{e}_d\rangle \geq |t| - \frac{1}{2} \geq \frac{1}{2} > \frac{1}{4\sqrt{2}} \ .$$

- $\mathbf{x} = \mathbf{0}$. In this case we have

$$\partial f(\mathbf{x}) \subseteq \left\{ t \cdot \mathbf{e}_d + \frac{1}{4}\mathbf{u} - \overline{\mathbf{w}} + \frac{1}{2}\overline{\mathbf{w}} \,\middle|\, t \in [-2,2],\ \sum_{i=1}^{d-1} u_i^2 \le 1,\ u_d = 0 \right\} .$$

  For any vector in $\partial f(\mathbf{x})$ corresponding to some $t, \mathbf{u}$, we use the fact that projecting any vector onto $\mathrm{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{d-1}\}$ cannot increase it's norm, in order to get

$$\left\| t \cdot \mathbf{e}_d + \frac{1}{4}\mathbf{u} - \frac{1}{2}\overline{\mathbf{w}} \right\| \ge \left\| \frac{1}{4}\mathbf{u} - \frac{1}{2}\overline{\mathbf{w}} \right\| = \left\| \frac{1}{2}\overline{\mathbf{w}} - \frac{1}{4}\mathbf{u} \right\| \ge \frac{1}{2} - \frac{1}{4} > \frac{1}{4\sqrt{2}} .$$

- $\mathbf{x} = -\mathbf{w}$. In this case we have

$$\partial f(\mathbf{x}) \subseteq \left\{ t \cdot \mathbf{e}_d - \frac{1}{4}\overline{\mathbf{w}} - s\left(\overline{\mathbf{w}} - \frac{1}{2}\mathbf{u}\right) \,\middle|\, t \in [-2,2],\ s \in [0,1],\ \|\mathbf{u}\| \le 1 \right\} .$$

  For any $\mathbf{g} \in \partial f(\mathbf{x})$ corresponding to some $t, s, \mathbf{u}$, using the fact that $w_d = 0$ we get

$$\begin{aligned} \|\mathbf{g}\| \ge \langle \mathbf{g}, -\overline{\mathbf{w}} \rangle &= \left\langle -\frac{1}{4}\overline{\mathbf{w}}, -\overline{\mathbf{w}} \right\rangle + \langle -s\overline{\mathbf{w}}, -\overline{\mathbf{w}} \rangle + \left\langle \frac{s}{2}\mathbf{u}, -\overline{\mathbf{w}} \right\rangle \\ &= \frac{1}{4} + s - \frac{s}{2}\langle \mathbf{u}, \mathbf{w} \rangle \ge \frac{1}{4} + s - \frac{s}{2} \ge \frac{1}{4} > \frac{1}{4\sqrt{2}} . \end{aligned}$$

- $x_d = 0$, $\mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\}$, $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle < \frac{1}{2}$. Note that

$$\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle < \frac{1}{2} \implies \langle \overline{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\| < 0 ,$$

  and that the set $\left\{ \mathbf{x} \middle|\ \langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle < \frac{1}{2} \right\} \setminus \{\mathbf{0}, -\mathbf{w}\}$ is an open set in $\mathbb{R}^d$. Thus for every such point, the function $\mathbf{x} \mapsto f(\mathbf{x})$ is locally identical to $\mathbf{x} \mapsto h(x_d + x^*) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}$, which in particular implies that their gradient sets are identical. Furthermore, combining the assumptions $x_d = 0, \mathbf{x} \neq \mathbf{0}$ reveals that $x_1, \ldots, x_{d-1}$ are not all zeroes. Consequently, we get

$$\partial f(\mathbf{x}) \subseteq \left\{ t \cdot \mathbf{e}_d + \frac{1}{4}\overline{\mathbf{x}} \,\middle|\, t \in [-2,2] \right\} .$$

  For any $\mathbf{g} \in \partial f(\mathbf{x})$ corresponding to some t, we get

$$\|\mathbf{g}\| \ge \langle \mathbf{g}, \overline{\mathbf{x}} \rangle = \frac{1}{4} > \frac{1}{4\sqrt{2}} .$$

- $x_d = 0$, $\mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\}$, $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle > \frac{1}{2}$. Note that

$$\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle > \frac{1}{2} \implies \langle \overline{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\| > 0 ,$$

and that the set $\left\{\mathbf{x}\middle|\ \langle\overline{\mathbf{w}},\overline{\mathbf{x}+\mathbf{w}}\rangle > \frac{1}{2}\right\}\setminus\{\mathbf{0},-\mathbf{w}\}$ is an open set in $\mathbb{R}^d$. Thus for every such point, the function $\mathbf{x}\mapsto f(\mathbf{x})$ is locally identical to

$$\mathbf{x}\mapsto h(x_d+x^*)+\frac{1}{4}\sqrt{\sum_{i=1}^{d-1}x_i^2}-\langle\overline{\mathbf{w}},\mathbf{x}+\mathbf{w}\rangle+\frac{1}{2}\|\mathbf{x}+\mathbf{w}\|\ ,$$

which in particular implies that their gradient sets are identical. Furthermore, combining the assumptions $x_d=0,\mathbf{x}\neq\mathbf{0}$ reveals that $x_1,\ldots,x_{d-1}$ are not all zeroes. Consequently, we get

$$\partial f(\mathbf{x})\subseteq\left\{t\cdot\mathbf{e}_d+\frac{1}{4}\overline{\mathbf{x}}-\overline{\mathbf{w}}+\frac{1}{2}\left(\overline{\mathbf{x}+\mathbf{w}}\right)\middle|\ t\in[-2,2]\right\}\ .$$

For any $\mathbf{g}\in\partial f(\mathbf{x})$ corresponding to some t, using the fact that $w_d=0$ we get

$$\begin{aligned}\|\mathbf{g}\|&\geq\langle\mathbf{g},-\overline{\mathbf{w}}\rangle=\frac{1}{4}\langle\overline{\mathbf{x}},-\overline{\mathbf{w}}\rangle+\langle-\overline{\mathbf{w}},-\overline{\mathbf{w}}\rangle+\frac{1}{2}\langle\overline{\mathbf{x}+\mathbf{w}},-\overline{\mathbf{w}}\rangle\\&\geq-\frac{1}{4}+1-\frac{1}{2}>\frac{1}{4\sqrt{2}}\ .\end{aligned}$$

- $x_d=0$, $\mathbf{x}\notin\{\mathbf{0},-\mathbf{w}\}$, $\langle\overline{\mathbf{w}},\overline{\mathbf{x}+\mathbf{w}}\rangle=\frac{1}{2}$. In this case we have

$$\begin{aligned}&\partial f(\mathbf{x})\subseteq\left\{t\cdot\mathbf{e}_d+\frac{1}{4}\overline{\mathbf{x}}-s\left(\overline{\mathbf{w}}-\frac{1}{2}(\overline{\mathbf{x}+\mathbf{w}})\right)\middle|\ t\in[-2,2],\ s\in[0,1]\right\}\qquad(5)\\&=\left\{\left(\frac{1}{4\|\mathbf{x}\|}+\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x}+\left(\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}-\frac{s}{\|\mathbf{w}\|}\right)\mathbf{w}+t\cdot\mathbf{e}_d\middle|\ t\in[-2,2],\ s\in[0,1]\right\}\ .\end{aligned}$$

Denote $\mathbf{x}=\mathbf{x}_{|}+\mathbf{x}_{\perp}$ where $\mathbf{x}_{\perp}=(I-\bar{\mathbf{w}}\bar{\mathbf{w}}^T)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathrm{span}(\mathbf{w})^{\perp}$, and $\mathbf{x}_{|}\in\mathrm{span}(\mathbf{w})$. For any $\mathbf{g}\in\partial f(\mathbf{x})$ corresponding to some t,s, using the fact that $x_d=w_d=0$ we get for some scalar α:

$$\begin{aligned}\|\mathbf{g}\|&\geq\left\|\left(\frac{1}{4\|\mathbf{x}\|}+\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x}+\left(\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}-\frac{s}{\|\mathbf{w}\|}\right)\mathbf{w}\right\|\\&=\left\|\left(\frac{1}{4\|\mathbf{x}\|}+\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x}_{\perp}+\alpha\cdot\mathbf{w}\right\|\\&\geq\left(\frac{1}{4\|\mathbf{x}\|}+\frac{s}{2\|\mathbf{x}+\mathbf{w}\|}\right)\|\mathbf{x}_{\perp}\|\\&\geq\frac{1}{4\|\mathbf{x}\|}\cdot\|\mathbf{x}_{\perp}\|\ .\end{aligned}$$

Since $(I-\bar{\mathbf{w}}\bar{\mathbf{w}}^T)$ is an orthogonal projection, in particular symmetric, we also have

$$\|\mathbf{x}_{\perp}\|^2=\langle\mathbf{x},(I-\bar{\mathbf{w}}\bar{\mathbf{w}}^T)^2\mathbf{x}\rangle=\langle\mathbf{x},(I-\bar{\mathbf{w}}\bar{\mathbf{w}}^T)\mathbf{x}\rangle=\|\mathbf{x}\|^2-\langle\overline{\mathbf{w}},\mathbf{x}\rangle^2=\|\mathbf{x}\|^2(1-\langle\overline{\mathbf{w}},\overline{\mathbf{x}}\rangle^2)\ .$$

Plugging into the above, it follows that $\|\mathbf{g}\|$ is at least $\frac{1}{4}\sqrt{1-\langle\overline{\mathbf{w}},\overline{\mathbf{x}}\rangle^2}$. Assuming that there exists such $\mathbf{g}\in\partial f(\mathbf{x})$ with norm at most ϵ, it follows that

$$\frac{1}{4}\sqrt{1-\langle\overline{\mathbf{w}},\overline{\mathbf{x}}\rangle^2}\leq\epsilon\ .\qquad(6)$$

However, we will show that for any $\epsilon < \frac{1}{4\sqrt{2}}$, we must arrive at a contradiction. To that end, let us consider two cases:

- If $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}} \rangle > 0$, then by rearranging Eq. (6), we have $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}} \rangle \geq \sqrt{1-16\epsilon^2}$. Hence,

  $$\langle \overline{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle \geq \|\mathbf{x}\|\sqrt{1-16\epsilon^2}+\|\mathbf{w}\| \geq (\|\mathbf{x}\|+\|\mathbf{w}\|)\sqrt{1-16\epsilon^2} \geq \|\mathbf{x}+\mathbf{w}\|\sqrt{1-16\epsilon^2}\,.$$

  However, dividing both sides by $\|\mathbf{x}+\mathbf{w}\|$, we get that $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle \geq \sqrt{1-16\epsilon^2}$. If $\epsilon < \frac{1}{4\sqrt{2}}$, we get that $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle > \frac{1}{2}$, contradicting our assumption on $\mathbf{x}$.
- If $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}} \rangle \leq 0$, then by Eq. (6), we must have $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}} \rangle \leq -\sqrt{1-16\epsilon^2}$. But then, by recalling that $w_d = 0$, we use Eq. (5) and our assumption that $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle = \frac{1}{2}$ in order to obtain

  $$\begin{aligned} -\|\mathbf{g}\| \leq \langle \overline{\mathbf{w}}, \mathbf{g} \rangle &= \frac{1}{4}\langle \overline{\mathbf{w}}, \overline{\mathbf{x}} \rangle - s\left(1 - \frac{1}{2}\cdot\frac{1}{2}\right) \leq -\frac{1}{4}\sqrt{1-16\epsilon^2} - \frac{3}{4}s \\ &\leq -\frac{1}{4}\sqrt{1-16\epsilon^2}\,. \end{aligned}$$

  This implies that $\frac{1}{4}\sqrt{1-16\epsilon^2} \leq \|\mathbf{g}\| \leq \epsilon$, which does not hold for any $\epsilon < \frac{1}{4\sqrt{2}}$.

■

Finally, given some nonzero $\mathbf{w} \in \mathbb{R}^d$ such that $w_d = 0$, we are ready to consider the function

$$\begin{aligned} F_{\mathbf{w}}(x_1, \ldots, x_d) &:= \max\{-1, f_{\mathbf{w}}(\mathbf{x}-\mathbf{x}^*)\} \\ &= \max\left\{-1, h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2} - \left[\langle \overline{\mathbf{w}}, \mathbf{x}-\mathbf{x}^*+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}-\mathbf{x}^*+\mathbf{w}\|\right]_+\right\}. \end{aligned}$$

Lemma 15 *The following hold:*

- $F_{\mathbf{w}}(\cdot)$ *is* $\frac{15}{4}$*-Lipschitz,* $F_{\mathbf{w}}(\mathbf{0}) - \inf_{\mathbf{x}} F_{\mathbf{w}}(\mathbf{x}) \leq 2$ *and* $\inf\{\|\mathbf{x}\| \mid \partial F_{\mathbf{w}}(\mathbf{x}) = \{\mathbf{0}\}\} \leq 13$.
- *Any* ϵ*-stationary point* $\mathbf{x}$ *for* $\epsilon < \frac{1}{4\sqrt{2}}$ *satisfies* $F_{\mathbf{w}}(\mathbf{x}) = -1$.
- *There exists a choice of* $\mathbf{w}$*, such that if we run* $\mathcal{A}$ *on* $F_{\mathbf{w}}(\cdot)$*, then with probability at least* $1-2T\exp(-d/36)$ *the algorithm's iterates* $\mathbf{x}_1^{F_{\mathbf{w}}}, \ldots, \mathbf{x}_T^{F_{\mathbf{w}}}$ *satisfy* $\min_{t\in[T]} F_{\mathbf{w}}(\mathbf{x}_t^{F_{\mathbf{w}}}) > 0$.

Proof First, recall that $f_{\mathbf{w}}(\cdot)$ is $\frac{15}{4}$-Lipschitz by Lemma 14. Combining this with the fact that $\mathbf{x} \mapsto \mathbf{x}-\mathbf{x}^*$, $z \mapsto \max\{-1, z\}$ are both 1-Lipschitz yields the desired Lipschitz bound. Moreover, we see that $\inf_{\mathbf{x}} F_{\mathbf{w}}(\mathbf{x}) \geq -1$, and by definition of $F_{\mathbf{w}}(\cdot)$ and Lemma 13: $F_{\mathbf{w}}(\mathbf{0}) \leq h(0) = 1$. Combining the two observations gives

$$F_{\mathbf{w}}(\mathbf{0}) - \inf_{\mathbf{x}} F_{\mathbf{w}}(\mathbf{x}) \leq 1 + 1 = 2\,.$$

For the remaining claim in the first bullet, consider $\mathbf{v} = 12\overline{\mathbf{w}} + \mathbf{x}^*$. By Lemma 13 we have $\|\mathbf{v}\| \leq 12+1 = 13$, thus it is enough to show that $\mathbf{v}$ is a stationary point of $F_{\mathbf{w}}$. In particular,

it is enough to show that $f_{\mathbf{w}}(\mathbf{v}-\mathbf{x}^*) < -1$, since by the continuity of $f_{\mathbf{w}}$ this will imply that $F_{\mathbf{w}} \equiv -1$ in a neighborhood of $\mathbf{v}$. Indeed, using the facts that $\mathbf{v}-\mathbf{x}^* = 12\overline{\mathbf{w}},\ w_d = 0$ we get

$$\begin{aligned} f_{\mathbf{w}}(\mathbf{v}-\mathbf{x}^*) &= h(x^*) + \frac{1}{4}\cdot 12\|\overline{\mathbf{w}}\| - \left[\langle \overline{\mathbf{w}}, 12\overline{\mathbf{w}}+\mathbf{w}\rangle - \frac{1}{2}\|12\overline{\mathbf{w}}+\mathbf{w}\|\right]_+ \\ &< h(0) + 3 - \frac{1}{2}\|12\overline{\mathbf{w}}+\mathbf{w}\| < 1+3-\frac{1}{2}\cdot 12 < -1\ . \end{aligned}$$

As to the second bullet, suppose $\mathbf{x}$ is an ϵ-stationary point for some $\epsilon < \frac{1}{4\sqrt{2}}$. Namely, there exists $\mathbf{g} \in \partial F_{\mathbf{w}}(\mathbf{x})$ such that $\|\mathbf{g}\| \le \epsilon$. Assume by contradiction that $F_{\mathbf{w}}(\mathbf{x}) > -1$. Since the set $\{\mathbf{y}|\ F_{\mathbf{w}}(\mathbf{y}) > -1\}$ is an open set, it follows that for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$: $F_{\mathbf{w}}(\mathbf{y}) > -1$. Hence, for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$: $F_{\mathbf{w}}(\mathbf{y}) = f_{\mathbf{w}}(\mathbf{y}-\mathbf{x}^*)$, which in particular implies that $\partial F_{\mathbf{w}}(\mathbf{x}) = \partial f_{\mathbf{w}}(\mathbf{x}-\mathbf{x}^*)$. We conclude that $\mathbf{g} \in \partial f_{\mathbf{w}}(\mathbf{x}-\mathbf{x}^*)$ and satisfies $\|\mathbf{g}\| < \frac{1}{4\sqrt{2}}$. Thus $(\mathbf{x}-\mathbf{x}^*)$ is an ϵ-stationary point of $f_{\mathbf{w}}(\cdot)$ for some $\epsilon < \frac{1}{4\sqrt{2}}$, which is a contradiction to Lemma 14.

In order to prove the third bullet, we start by noticing that

$$\mathbf{x} \in \left\{\mathbf{x} : \langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*+\mathbf{w}}\rangle \le \frac{1}{2}\right\} \cup \{\mathbf{x}^*-\mathbf{w}\} \implies \langle \overline{\mathbf{w}}, \mathbf{x}-\mathbf{x}^*+\mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x}-\mathbf{x}^*+\mathbf{w}\| \le 0\ ,$$

from which it follows that

$$\forall \mathbf{x} \in \left\{\mathbf{x} : \langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*+\mathbf{w}}\rangle \le \frac{1}{2}\right\} \cup \{\mathbf{x}^*-\mathbf{w}\}\ :\ F_{\mathbf{w}}(\mathbf{x}) = \bar{f}(\mathbf{x}) := h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}\ . \quad (7)$$

Indeed, for any such $\mathbf{x}$ the ReLU term in the definition of $F_{\mathbf{w}}(\cdot)$ vanishes, and the remaining function (which is non-negative) is greater than -1. We continue by showing that Eq. (7) holds over a set of more convenient form. In order to do that, fix some $\mathbf{x}$ such that $\langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*+\mathbf{w}}\rangle > \frac{1}{2}$ (i.e. the *opposite* condition). Multiplying by $\|\mathbf{x}-\mathbf{x}^*+\mathbf{w}\|$ gives

$$\langle \overline{\mathbf{w}}, \mathbf{x}-\mathbf{x}^*\rangle + \|\mathbf{w}\| = \langle \overline{\mathbf{w}}, \mathbf{x}-\mathbf{x}^*+\mathbf{w}\rangle > \frac{1}{2}\|\mathbf{x}-\mathbf{x}^*+\mathbf{w}\| \ge \frac{1}{2}\left(\|\mathbf{x}-\mathbf{x}^*\| - \|\mathbf{w}\|\right)\ .$$

For $\mathbf{x} = \mathbf{x}^*$ the inequality above is trivially satisfied. For $\mathbf{x} \ne \mathbf{x}^*$, dividing by $\|\mathbf{x}-\mathbf{x}^*\|$ and rearranging yields

$$\langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*}\rangle > \frac{1}{2} - \frac{3\|\mathbf{w}\|}{2\|\mathbf{x}-\mathbf{x}^*\|}\ .$$

The contrapositive allows us to deduce that any $\mathbf{x}$ which does *not* satisfy the condition above belongs to $\{\mathbf{x} : \langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*+\mathbf{w}}\rangle \le \frac{1}{2}\} \cup \{\mathbf{x}^*-\mathbf{w}\}$, so we get by Eq. (7):

$$\forall \mathbf{x} \ne \mathbf{x}^* : \langle \overline{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*}\rangle \le \frac{1}{2} - \frac{3\|\mathbf{w}\|}{2\|\mathbf{x}-\mathbf{x}^*\|} \implies F_{\mathbf{w}}(\mathbf{x}) = \bar{f}(\mathbf{x}) := h(x_d) + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2}\ . \quad (8)$$

With this equation in hand, we turn to describe how $\mathbf{w}$ should be set in order to establish the third bullet. Consider a random vector $\mathbf{w} \in \mathbb{R}^d$ which is distributed as follows:

$$(w_1, \ldots, w_{d-1}) \sim \text{Unif}\left(\frac{\rho}{99}\cdot \mathbb{S}^{d-2}\right),\ \Pr[w_d = 0] = 1\ , \quad (9)$$

where $\frac{\rho}{99}\cdot\mathbb{S}^{d-2} := \{(y_1,\ldots,y_{d-1})|\sum_{i=1}^{d-1} y_i^2 = \frac{\rho}{99}\}$ is the $(d-2)$-dimensional sphere of radius $\frac{\rho}{99}$. Note that $\|\mathbf{w}\| = \frac{\rho}{99}$, which by plugging into Eq. (8) gives

$$\forall \mathbf{x}\neq\mathbf{x}^*:\langle\overline{\mathbf{w}},\overline{\mathbf{x}-\mathbf{x}^*}\rangle \leq \frac{1}{2}-\frac{\rho}{66\|\mathbf{x}-\mathbf{x}^*\|} \implies F_{\mathbf{w}}(\mathbf{x}) = \bar{f}(\mathbf{x}) := h(x_d)+\frac{1}{4}\sqrt{\sum_{i=1}^{d-1}x_i^2}\ . \quad (10)$$

Let $\mathbf{x}_1^{\bar{f}},\ldots,\mathbf{x}_T^{\bar{f}}$ be the (possibly random) iterates produced by $\mathcal{A}$ when ran on $\bar{f}(\cdot)$. Note that if

$$\left(\min_{t\in[T]}\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|\geq\rho>0\right)\wedge\left(\max_{t\in[T]}\langle\overline{\mathbf{w}},\overline{(\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)}\rangle<\frac{1}{3}\right) \quad (11)$$

then for all $t\in[T]$:

$$\langle\overline{\mathbf{w}},\overline{(\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)}\rangle<\frac{1}{3}<\frac{1}{2}-\frac{\rho}{66\rho}\leq\frac{1}{2}-\frac{\rho}{66\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|}\ ,$$

as well as $\mathbf{x}_t^{\bar{f}}\neq\mathbf{x}^*$. Thus, by Eq. (10), this means that Eq. (11) implies that $F_{\mathbf{w}}(\mathbf{x}_t^{\bar{f}}) = \bar{f}(\mathbf{x}_t^{\bar{f}})$ for all $t\in[T]$. Moreover, using the fact that $\mathbf{x}_t^{\bar{f}}$ is bounded away from $\mathbf{x}^*$, it is easily verified that the condition in Eq. (10) also holds for all $\mathbf{x}$ in some neighborhood of $\mathbf{x}_t^{\bar{f}}$, so actually $F_{\mathbf{w}}(\cdot)$ is identical to $\bar{f}(\cdot)$ on these neighborhoods, implying the same local oracle response. Hence, assuming the event in Eq. (11) occurs, if we run the algorithm on $F_{\mathbf{w}}(\cdot)$ rather than $\bar{f}(\cdot)$, then the produced iterates $\mathbf{x}_1^{F_{\mathbf{w}}},\ldots,\mathbf{x}_T^{F_{\mathbf{w}}}$ are identical to $\mathbf{x}_1^{\bar{f}},\ldots,\mathbf{x}_T^{\bar{f}}$. That being the case, we would get

$$\min_{t\in[T]}F_{\mathbf{w}}(\mathbf{x}_t^{F_{\mathbf{w}}}) = \min_{t\in[T]}F_{\mathbf{w}}(\mathbf{x}_t^{\bar{f}}) = \min_{t\in[T]}\bar{f}(\mathbf{x}_t^{\bar{f}})>0\ ,$$

where the last inequality utilizes the fact that $\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|>0$. Overall we see that

$$\left(\min_{t\in[T]}\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|\geq\rho\right)\wedge\left(\max_{t\in[T]}\langle\overline{\mathbf{w}},\overline{(\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)}\rangle<\frac{1}{3}\right)\implies\min_{t\in[T]}F_{\mathbf{w}}(\mathbf{x}_t^{F_{\mathbf{w}}})>0\ .$$

Thus, in order to finish the proof, it is enough to show that there exists $\mathbf{w}$, such that

$$\Pr_{\mathcal{A}}\left[\left(\min_{t\in[T]}\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|\geq\rho\right)\wedge\left(\max_{t\in[T]}\langle\overline{\mathbf{w}},\overline{(\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)}\rangle<\frac{1}{3}\right)\right]\geq 1-2T\exp(-d/36)\ . \quad (12)$$

In order to prove this claim, we observe that:

1. By Lemma 13 we know that $\Pr_{\mathcal{A}}[\min_{t\in[T]}\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|\geq\rho]\geq 1-\delta = 1-T\exp(-d/36)$.
2. If we fix some vectors $\mathbf{u}_1,\ldots,\mathbf{u}_T$ in $\mathbb{R}^{d-1}$ such that $\forall t:\|\mathbf{u}_t\|\leq 1$, and pick a unit vector $\mathbf{u}\in\mathbb{R}^{d-1}$ uniformly at random, then by a union bound and a standard concentration of measure on the sphere argument (e.g., Ball et al., 1997, Lemma 2.2), $\Pr(\max_t\langle\mathbf{u},\mathbf{u}_t\rangle\geq\alpha)\leq T\cdot\Pr(\langle\mathbf{u},\mathbf{u}_1\rangle\geq\alpha)\leq T\exp(-(d-1)\alpha^2/2)$. For any realization of $\mathcal{A}$'s randomness such that such that for all $t\in[T]:\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|\geq\rho$, by setting

$$\alpha = 1/3,\quad \mathbf{u} = \overline{(w_1,\ldots,w_{d-1})},\quad \mathbf{u}_t = \frac{1}{\|\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*\|}((\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)_1,\ldots,(\mathbf{x}_t^{\bar{f}}-\mathbf{x}^*)_{d-1})\ ,$$

while noticing that $\langle \mathbf{u}, \mathbf{u}_t \rangle = \langle \overline{\mathbf{w}}, \overline{(\mathbf{x}_t^{\bar{f}} - \mathbf{x}^*)} \rangle$ since $w_d = 0$, we get

$$\Pr_{\mathbf{w}} \left[\max_{t \in [T]} \langle \overline{\mathbf{w}}, \overline{(\mathbf{x}_t^{\bar{f}} - \mathbf{x}^*)} \rangle \geq 1/3 \right] \leq T \exp(-d/36) \ .$$

Combining the two observations in a formal manner results in

$$\Pr_{\mathcal{A}} \left[\Pr_{\mathbf{w}} \left[E_{\mathcal{A},\mathbf{w}} | \mathcal{A} \right] \geq 1 - T \exp(-d/36) \right] \geq 1 - T \exp(-d/36) \ , \tag{13}$$

where

$$E_{\mathcal{A},\mathbf{w}} := \left(\min_{t \in [T]} \| \mathbf{x}_t^{\bar{f}} - \mathbf{x}^* \| \geq \rho \right) \wedge \left(\max_{t \in [T]} \langle \overline{\mathbf{w}}, \overline{(\mathbf{x}_t^{\bar{f}} - \mathbf{x}^*)} \rangle < \frac{1}{3} \right) \ .$$

Finally, using the law of total expectation and Eq. (13) we get

$$\begin{aligned}
&\Pr_{\mathcal{A},\mathbf{w}} [E_{\mathcal{A},\mathbf{w}}] = \mathbb{E}_{\mathcal{A}} [\Pr_{\mathbf{w}} [E_{\mathcal{A},\mathbf{w}} | \mathcal{A}]] \\
&\geq \mathbb{E}_{\mathcal{A}} \left[\Pr_{\mathbf{w}} \left[E_{\mathcal{A},\mathbf{w}} | \mathcal{A} : \Pr_{\mathbf{w}} [E_{\mathcal{A},\mathbf{w}} | \mathcal{A}] \geq 1 - T \exp(-d/36) \right] \cdot \Pr_{\mathcal{A}} \left[\Pr_{\mathbf{w}} [E_{\mathcal{A},\mathbf{w}} | \mathcal{A}] \geq 1 - T \exp(-d/36) \right] \right] \\
&\geq \mathbb{E}_{\mathcal{A}} \left[(1 - T \exp(-d/36) \cdot (1 - T \exp(-d/36)) \right] \\
&\geq (1 - T \exp(-d/36))^2 \\
&\geq 1 - 2T \exp(-d/36) \ .
\end{aligned}$$

Consequently, by the probabilistic method, there exists some fixed choice of $\mathbf{w}$ such that

$$\Pr_{\mathcal{A}} [E_{\mathcal{A},\mathbf{w}}] \geq 1 - 2T \exp(-d/36) \ ,$$

which is exactly Eq. (12), finishing the proof. ■

The theorem is an immediate corollary of the previous lemma: With the specified high probability, $\min_t F_{\mathbf{w}}(\mathbf{x}_t) > 0$, even though all ϵ-stationary points (for any $\epsilon < \frac{1}{4\sqrt{2}}$) have a value of -1. Since $F_{\mathbf{w}}$ is also $\frac{15}{4}$-Lipschitz, we get that the distance of any $\mathbf{x}_t$ from an ϵ-stationary point must be at least $\frac{0-(-1)}{\frac{15}{4}} = \frac{4}{15}$. Simplifying the numerical terms by choosing a large enough constant C and a small enough constant c, and relabeling $F_{\mathbf{w}}$ as f, the theorem follows.

5.2 Proof of Thm. 6

We start the proof by showing that without loss of generality we can impose certain assumptions on the parameters of interest. First, if $\epsilon \geq 1$ then the right hand side of Eq. (4) is negative for any $c_2 < 1$, which makes the theorem trivial. Consequently, we can assume $\epsilon < 1$. Using Lemma 30 in Appendix D, this also implies that $L \geq \frac{1}{8}$ since otherwise an L-smooth ϵ-approximation does not exist in the first place in case of 1-Lipschitz function $\mathbf{x} \mapsto |x_1|$ (in particular, no such smoother exists). Therefore, if $\sqrt{\log\left((M+1)\left(T+1\right)\right)} \geq \frac{\sqrt{d}}{32r}$ then

$$L \sqrt{\log\left((M+1)\left(T+1\right)\right)} \geq \frac{1}{8} \cdot \frac{\sqrt{d}}{32r} > \frac{1}{256} \cdot \frac{\sqrt{d}}{r} \left(1 - \epsilon\right) \ ,$$

which proves the theorem. Thus we can assume throughout the proof that

$$\sqrt{\log\left((M+1)\left(T+1\right)\right)} < \frac{\sqrt{d}}{32r} \implies \frac{1}{16r}\sqrt{\frac{d}{\log\left((M+1)\left(T+1\right)\right)}} > 2\ . \tag{14}$$

Our strategy is to define a distribution over a family of "hard" 1-Lipschitz functions over $\mathbb{R}^d$, for which we will show that Eq. (4) must hold for some function supported by this distribution. Before we turn to do so, we will show that when a smoother which satisfies TICF acts on a constant function, it returns a gradient estimate of small norm. This will be crucial later on, since we will construct a function which looks "locally constant" at many points of interest, thus deceiving the smoother.

Lemma 16 *If $\mathcal{S}$ is an (L,ϵ,T,M,r)-smoother satisfying* TICF*, then for any constant function f and any $\mathbf{x}\in\mathbb{R}^d:\ \|\mathbb{E}\left[\mathcal{S}\left(f,\mathbf{x}\right)\right]\|\leq\epsilon$.*

Proof Denote $\mathbf{v}:=\mathbb{E}\left[\mathcal{S}\left(f,\mathbf{x}\right)\right]$, and note that by the TICF property $\mathbf{v}$ does not depend on $\mathbf{x}$. Let $\tilde{f}$ be the ϵ-approximation of f implicitly computed by $\mathcal{S}$, then by the definition of a smoothing algorithm, we have for all $\mathbf{x}\in\mathbb{R}^d$:

$$\begin{gathered}\left\|\mathbf{v}-\nabla\tilde{f}\left(\mathbf{x}\right)\right\|\leq\epsilon\\ \implies\|\mathbf{v}\|^2-\left\langle\nabla\tilde{f}\left(\mathbf{x}\right),\mathbf{v}\right\rangle=\left\langle\mathbf{v}-\nabla\tilde{f}\left(\mathbf{x}\right),\mathbf{v}\right\rangle\leq\left\|\mathbf{v}-\nabla\tilde{f}\left(\mathbf{x}\right)\right\|\cdot\|\mathbf{v}\|\leq\epsilon\|\mathbf{v}\|\\ \implies\left\langle\nabla\tilde{f}\left(\mathbf{x}\right),\mathbf{v}\right\rangle\geq\|\mathbf{v}\|^2-\epsilon\|\mathbf{v}\|\ .\end{gathered}$$

Define the one dimensional projected function $\tilde{f}_{\mathbf{v}}(t):=\tilde{f}(t\cdot\mathbf{v})$. Then for all $t\geq 0$,

$$\begin{aligned}\tilde{f}_{\mathbf{v}}\left(t\right)-\tilde{f}_{\mathbf{v}}\left(0\right)&=\int_0^t\tilde{f}'_{\mathbf{v}}(z)dz=\int_0^t\left\langle\nabla\tilde{f}\left(z\cdot\mathbf{v}\right),\mathbf{v}\right\rangle dz\\ &\geq\int_0^t\left(\|\mathbf{v}\|^2-\epsilon\|\mathbf{v}\|\right)dz=t\left(\|\mathbf{v}\|^2-\epsilon\|\mathbf{v}\|\right)=t\|\mathbf{v}\|\left(\|\mathbf{v}\|-\epsilon\right)\ .\end{aligned}\tag{15}$$

On the other hand, $\tilde{f}_{\mathbf{v}}(t),\tilde{f}_{\mathbf{v}}(0)$ are both ϵ-approximations of the same constant, since f is a constant function. Thus, $|\tilde{f}_{\mathbf{v}}(t)-\tilde{f}_{\mathbf{v}}(0)|\leq 2\epsilon$. Combining this with Eq. (15) yields for all $t\geq 0:\ 2\epsilon\ \geq\ t\|\mathbf{v}\|\left(\|\mathbf{v}\|-\epsilon\right)$. This can hold for all $t\geq 0$ only if $\left(\|\mathbf{v}\|-\epsilon\right)\leq 0$, implying the lemma. ■

We are now ready to define a family of functions, with the rest of the proof devoted to analyze how a smoother acts on them. Relying on Eq. (14) we can define the set

$$\Delta:=\left\{16r\sqrt{\frac{\log\left((M+1)\left(T+1\right)\right)}{d}}\cdot k\ \middle|\ k=0,1,\ldots,\left\lfloor\frac{1}{16r}\sqrt{\frac{d}{\log\left((M+1)\left(T+1\right)\right)}}\right\rfloor\right\}\ .$$

That is, a grid on $[0,1]$ which consists of points of distance $16r\sqrt{\frac{\log((M+1)(T+1))}{d}}$ one from another. We further define the "inflation" of Δ by $4r\sqrt{\frac{\log((M+1)(T+1))}{d}}$ around every point:[6]

$$\overline{\Delta} := \left\{ x \in \mathbb{R} \;\middle|\; \exists p \in \Delta : |p - x| \leq 4r\sqrt{\frac{\log\left((M+1)\left(T+1\right)\right)}{d}} \right\} .$$

Now we define the function $g : \mathbb{R} \to \mathbb{R}$ as the unique continuous function which satisfies (see Fig. 2 for an illustration)

$$g(0) = 0$$
$$g'\left(x\right) = \begin{cases} 1 \ , & x \notin \overline{\Delta} \\ 0 \ , & x \in \overline{\Delta} \end{cases} .$$

Finally, we are ready to consider

$$f_{\mathbf{w}}\left(\mathbf{x}\right) = g\left(\left\langle \mathbf{x}, \mathbf{w} \right\rangle\right) \ ,$$

where $\mathbf{w} \in \mathbb{S}^{d-1}$ is drawn uniformly from the unit sphere. The distribution over $\mathbf{w}$ specifies a distribution over the functions $f_{\mathbf{w}}$. We start by claiming that these functions are indeed in our function class of interest:

Lemma 17 *For all $\mathbf{w} \in \mathbb{S}^{d-1}$, $f_{\mathbf{w}}(\cdot)$ is 1-Lipschitz.*

Proof It is clear by construction that g is 1-Lipschitz. Thus

$$\left|f\left(\mathbf{x}\right) - f\left(\mathbf{y}\right)\right| = \left|g\left(\left\langle \mathbf{x}, \mathbf{w} \right\rangle\right) - g\left(\left\langle \mathbf{y}, \mathbf{w} \right\rangle\right)\right| \leq \left|\left\langle \mathbf{x}, \mathbf{w} \right\rangle - \left\langle \mathbf{y}, \mathbf{w} \right\rangle\right| = \left|\left\langle \mathbf{x} - \mathbf{y}, \mathbf{w} \right\rangle\right| \leq \left\|\mathbf{x} - \mathbf{y}\right\| \ .$$

■

The following lemma is the key lemma of the proof. It will show that there exists a function supported by the distribution we defined, such that many points mislead the smoother by appearing as if the function is constant - hence, the the smoother returns gradient estimates of small norm. Formally:

Lemma 18 *There exists $\mathbf{w} \in \mathbb{S}^{d-1}$ such that for all $\delta \in \Delta$: $\mathbb{E}_{\xi}\left[\left\|\mathcal{S}\left(f_{\mathbf{w}}, \delta\mathbf{w}\right)\right\|\right] \leq \epsilon + \frac{1}{32}$.*

Proof Let $\mathbf{x}_1^{(\mathbf{w})}, \ldots, \mathbf{x}_T^{(\mathbf{w})}$ be the (possibly randomized) queries produced by $\mathcal{S}\left(f_{\mathbf{w}}, \mathbf{0}\right)$. Consider the event $E_{\mathbf{w}}$, in which for all $i \in [T]$: $\left|\left\langle \mathbf{x}_i^{(\mathbf{w})}, \mathbf{w} \right\rangle\right| < 4r\sqrt{\frac{\log((M+1)(T+1))}{d}}$. Note that if $E_{\mathbf{w}}$ occurs then for all $\delta \in \Delta, i \in [T], \mathbf{v} \in \mathbb{R}^d$:

$$f_{\mathbf{w}}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w} + \mathbf{v}\right) = g\left(\left\langle \mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w} + \mathbf{v}, \mathbf{w} \right\rangle\right) = g\left(\delta + \left\langle \mathbf{x}_i^{(\mathbf{w})}, \mathbf{w} \right\rangle + \left\langle \mathbf{v}, \mathbf{w} \right\rangle\right) \ . \quad (16)$$

6. Note we use the quantities $T+1, M+1$ instead of the seemingly more natural T, M, since otherwise the logarithmic term in Eq. (4) can vanish, resulting in an invalid theorem. This would have occurred for randomized smoothing, where $T = M = 1$.

In particular, as long as $\|\mathbf{v}\| < 4r\sqrt{\frac{\log((M+1)(T+1))}{d}} - \left|\left\langle \mathbf{x}_i^{(\mathbf{w})}, \mathbf{w} \right\rangle\right|$, which by Cauchy-Schwarz implies

$$\left|\left\langle \mathbf{x}_i^{(\mathbf{w})}, \mathbf{w} \right\rangle + \langle \mathbf{v}, \mathbf{w} \rangle\right| < 4r\sqrt{\frac{\log\left((M+1)\left(T+1\right)\right)}{d}} \ ,$$

we get by construction of g and Eq. (16) that

$$f_{\mathbf{w}}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w} + \mathbf{v}\right) = g\left(\delta\right) \ .$$

In other words, if $E_{\mathbf{w}}$ occurs then inside some neighborhood of $\mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w}$, the function $f_{\mathbf{w}}$ is identical to the constant function $g\left(\delta\right)$. Therefore, if $E_{\mathbf{w}}$ occurs the local oracle $\mathbb{O}$ satisfies for any $\delta \in \Delta, i \in [T]$:

$$\mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w}\right) = \mathbb{O}_{\mathbf{x}\mapsto g(\delta)}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta\mathbf{w}\right) \ . \tag{17}$$

Fix some $\delta_0 \in \Delta$, and let $\tilde{\mathbf{x}}_1^{(\mathbf{w})}, \ldots, \tilde{\mathbf{x}}_T^{(\mathbf{w})}$ be the (possibly randomized) queries produced by $\mathcal{S}\left(f_{\mathbf{w}}, \delta_0\mathbf{w}\right)$. We will now show that conditioned on $E_{\mathbf{w}}$, for all $i \in [T]$:

$$\tilde{\mathbf{x}}_i^{(\mathbf{w})} = \mathbf{x}_i^{(\mathbf{w})} + \delta_0\mathbf{w} \ , \tag{18}$$

in the sense that for every realization of $\mathcal{S}'$'s randomness ξ they are equal. We show this by induction on i. For $i = 1$, using TICF:

$$\tilde{\mathbf{x}}_1^{(\mathbf{w})} = \mathcal{S}^{(1)}\left(\xi, \delta_0\mathbf{w}\right) = \mathcal{S}^{(1)}\left(\xi, \mathbf{0}\right) + \delta_0\mathbf{w} = \mathbf{x}_1^{(\mathbf{w})} + \delta_0\mathbf{w} \ .$$

Assuming this is true up until i, then by the induction hypothesis, Eq. (17) and TICF:

$$\begin{aligned}
\tilde{\mathbf{x}}_{i+1}^{(\mathbf{w})} &= \mathcal{S}^{(i)}\left(\xi, \delta_0\mathbf{w}, \mathbb{O}_{f_{\mathbf{w}}}\left(\tilde{\mathbf{x}}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\tilde{\mathbf{x}}_i^{(\mathbf{w})}\right)\right) \\
&= \mathcal{S}^{(i)}\left(\xi, \delta_0\mathbf{w}, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_1^{(\mathbf{w})} + \delta_0\mathbf{w}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta_0\mathbf{w}\right)\right) \\
&= \mathcal{S}^{(i)}\left(\xi, \delta_0\mathbf{w}, \mathbb{O}_{\mathbf{x}\mapsto g(\delta_0)}\left(\mathbf{x}_1^{(\mathbf{w})} + \delta_0\mathbf{w}\right), \ldots, \mathbb{O}_{\mathbf{x}\mapsto g(\delta_0)}\left(\mathbf{x}_i^{(\mathbf{w})} + \delta_0\mathbf{w}\right)\right) \\
&= \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_{\mathbf{x}\mapsto g(0)}\left(\mathbf{x}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{\mathbf{x}\mapsto g(0)}\left(\mathbf{x}_1^{(\mathbf{w})}\right)\right) + \delta_0\mathbf{w} \\
&= \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_i^{(\mathbf{w})}\right)\right) + \delta_0\mathbf{w} \\
&= \mathbf{x}_{i+1}^{(\mathbf{w})} + \delta_0\mathbf{w} \ .
\end{aligned}$$

Having established Eq. (18) for any $\delta \in \Delta$, we turn to show that for all $\delta \in \Delta$:

$$\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}}, \delta\mathbf{w}\right)|E_{\mathbf{w}}\right] = \mathbb{E}_{\xi}\left[\mathcal{S}\left(\mathbf{x}\mapsto 0, \mathbf{0}\right)|E_{\mathbf{w}}\right] \ . \tag{19}$$

Indeed, by Eq. (18), Eq. (17) and TICF:

$$\begin{aligned}
\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}}, \delta\mathbf{w}\right)|E_{\mathbf{w}}\right] &= \mathbb{E}_{\xi}\left[\mathcal{S}^{(out)}\left(\xi, \delta\mathbf{w}, \mathbb{O}_{f_{\mathbf{w}}}\left(\tilde{\mathbf{x}}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\tilde{\mathbf{x}}_T^{(\mathbf{w})}\right)\right)\middle|E_{\mathbf{w}}\right] \\
&= \mathbb{E}_{\xi}\left[\mathcal{S}^{(out)}\left(\xi, \delta\mathbf{w}, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_1^{(\mathbf{w})} + \delta\mathbf{w}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_T^{(\mathbf{w})} + \delta\mathbf{w}\right)\right)\middle|E_{\mathbf{w}}\right] \\
&= \mathbb{E}_{\xi}\left[\mathcal{S}^{(out)}\left(\xi, \delta\mathbf{w}, \mathbb{O}_{\mathbf{x}\mapsto g(\delta)}\left(\mathbf{x}_1^{(\mathbf{w})} + \delta\mathbf{w}\right), \ldots, \mathbb{O}_{\mathbf{x}\mapsto g(\delta)}\left(\mathbf{x}_T^{(\mathbf{w})} + \delta\mathbf{w}\right)\right)\middle|E_{\mathbf{w}}\right] \\
&= \mathbb{E}_{\xi}\left[\mathcal{S}^{(out)}\left(\xi, \mathbf{0}, \mathbb{O}_{\mathbf{x}\mapsto 0}\left(\mathbf{x}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{\mathbf{x}\mapsto 0}\left(\mathbf{x}_T^{(\mathbf{w})}\right)\right)\middle|E_{\mathbf{w}}\right] \\
&= \mathbb{E}_{\xi}\left[\mathcal{S}\left(\mathbf{x}\mapsto 0, \mathbf{0}\right)|E_{\mathbf{w}}\right] \ .
\end{aligned}$$

We now turn to show that $E_{\mathbf{w}}$ is likely to occur. Fix some realization of $\mathcal{S}$'s randomness ξ, and let $\mathbf{q}_1^\xi, \ldots, \mathbf{q}_T^\xi$ be the (deterministic) queries produced by $\mathcal{S}(\mathbf{y} \mapsto 0, \mathbf{0})$. We claim that if for all $i \in [T]$: $\left|\langle \mathbf{q}_i^\xi, \mathbf{w}\rangle\right| < 4r\sqrt{\frac{\log((M+1)(T+1))}{d}}$ then $\left(\mathbf{q}_1^\xi, \ldots, \mathbf{q}_T^\xi\right) = \left(\mathbf{x}_1^{(\mathbf{w})}, \ldots, \mathbf{x}_T^{(\mathbf{w})}\right)$ independently of $\mathbf{w}$. We show this by induction on i. For $i = 1$:

$$\mathbf{q}_1^\xi = \mathcal{S}^{(1)}(\xi, \mathbf{0}) = \mathbf{x}_1^{(\mathbf{w})} \ .$$

Assuming true up until i, then

$$\begin{aligned}
\mathbf{q}_{i+1}^\xi &= \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_{\mathbf{x}\mapsto 0}\left(\mathbf{q}_1^\xi\right), \ldots, \mathbb{O}_{\mathbf{x}\mapsto 0}\left(\mathbf{q}_i^\xi\right)\right) \\
&= \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{q}_1^\xi\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{q}_i^\xi\right)\right) \\
&= \mathcal{S}^{(i)}\left(\xi, \mathbf{0}, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_1^{(\mathbf{w})}\right), \ldots, \mathbb{O}_{f_{\mathbf{w}}}\left(\mathbf{x}_i^{(\mathbf{w})}\right)\right) \\
&= \mathbf{x}_{i+1}^{(\mathbf{w})} \ ,
\end{aligned}$$

where we used the assumption on $\mathbf{q}_i^\xi$ and the induction hypothesis. Recall that by assumption on the algorithm $\left\|\mathbf{q}_i^\xi\right\| \le r$ for all $i \in [T]$. Using the union bound and concentration of measure on the sphere (e.g., Ball et al., 1997, Lemma 2.2) we can bound the probability of the complementary event

$$\begin{aligned}
\Pr_{\mathbf{w}}\left[E_{\mathbf{w}}^c \mid \xi\right] =& \Pr_{\mathbf{w}}\left[\exists i \in [T] : \ \left|\langle \mathbf{q}_i^\xi, \mathbf{w}\rangle\right| \ge 4r\sqrt{\frac{\log\left((M+1)(T+1)\right)}{d}}\right] \\
=& \Pr_{\mathbf{w}}\left[\exists i \in [T] : \ \left|\left\langle \frac{1}{r}\mathbf{q}_i^\xi, \mathbf{w}\right\rangle\right| \ge 4\sqrt{\frac{\log\left((M+1)(T+1)\right)}{d}}\right] \\
\le& T \cdot 2\exp\left(-\frac{d \cdot \left(4\sqrt{\frac{\log((M+1)(T+1))}{d}}\right)^2}{2}\right) \\
=& \frac{2T}{(M+1)^8 (T+1)^8} \le \frac{2}{(M+1)^8 (T+1)^7} \ .
\end{aligned}$$

This inequality holds for any realization of $\mathcal{S}$'s randomness ξ, hence by the law of total probability

$$\Pr_{\xi, \mathbf{w}}\left[E_{\mathbf{w}}^c\right] \le \frac{2}{(M+1)^8 (T+1)^7} \ .$$

In particular, since $\Pr_{\xi,\mathbf{w}}\left[E_{\mathbf{w}}^c\right] = \mathbb{E}_{\mathbf{w}}\left[\Pr_\xi\left[E_{\mathbf{w}}^c | \mathbf{w}\right]\right]$, there exists $\mathbf{w} \in \mathbb{S}^{d-1}$ such that

$$\Pr_{\xi}\left[E_{\mathbf{w}}^c\right] \le \frac{2}{(M+1)^8 (T+1)^7} \ . \tag{20}$$

For this fixed $\mathbf{w}$, we have for all $\delta \in \Delta$ by the law of total expectation and the triangle inequality:

$$\left\|\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}},\delta\mathbf{w}\right)\right]\right\| \leq \left\|\underbrace{\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}},\delta\mathbf{w}\right)|E_{\mathbf{w}}\right]\cdot\Pr_{\xi}\left[E_{\mathbf{w}}\right]}_{(*)}\right\| + \left\|\underbrace{\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}},\delta\mathbf{w}\right)|E_{\mathbf{w}}^{c}\right]\cdot\Pr_{\xi}\left[E_{\mathbf{w}}^{c}\right]}_{(**)}\right\| \ . \quad (21)$$

On one hand, by Eq. (19):

$$(*) = \mathbb{E}_{\xi}\left[\mathcal{S}\left(\mathbf{x}\mapsto 0,\mathbf{0}\right)|E_{\mathbf{w}}\right]\cdot\Pr_{\xi}\left[E_{\mathbf{w}}\right] = \mathbb{E}_{\xi}\left[\mathcal{S}\left(\mathbf{x}\mapsto 0,\mathbf{0}\right)\right] - \mathbb{E}_{\xi}\left[\mathcal{S}\left(\mathbf{x}\mapsto 0,\mathbf{0}\right)|E_{\mathbf{w}}^{c}\right]\cdot\Pr_{\xi}\left[E_{\mathbf{w}}^{c}\right] \ .$$

Using Lemma 16, and by incorporating the definition of M in Eq. (2) and Eq. (20) we get

$$\|(*)\| \leq \epsilon + M\cdot\frac{2}{\left(M+1\right)^{8}\left(T+1\right)^{7}} \leq \epsilon + \frac{2}{\left(M+1\right)^{7}\left(T+1\right)^{7}} \ . \quad (22)$$

On the other hand, by Eq. (2) and Eq. (20) again we have

$$\|(**)\| \leq \left\|\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}},\delta\mathbf{w}\right)|E_{\mathbf{w}}^{c}\right]\right\|\cdot\Pr_{\xi}\left[E_{\mathbf{w}}^{c}\right] \leq M\cdot\frac{2}{\left(M+1\right)^{8}\left(T+1\right)^{7}} \leq \frac{2}{\left(M+1\right)^{7}\left(T+1\right)^{7}} \ . \quad (23)$$

Overall, plugging Eq. (22) and Eq. (23) into Eq. (21), gives

$$\left\|\mathbb{E}_{\xi}\left[\mathcal{S}\left(f_{\mathbf{w}},\delta\mathbf{w}\right)\right]\right\| \leq \epsilon + \frac{4}{\left(M+1\right)^{7}\left(T+1\right)^{7}} \leq \epsilon + \frac{1}{32} \ ,$$

where the last inequality simply follows from the fact that $M > 0,\ T \geq 1$. ■

From now on, we fix $\mathbf{w} \in \mathbb{S}^{d-1}$ which is given by the previous lemma and denote $f = f_{\mathbf{w}}$. Denote by $\tilde{f}$ the ϵ-approximation of f with L-Lipschitz gradients implicitly computed by $\mathcal{S}$. We turn our focus to the directional projection:

$$\begin{aligned}\varphi &: [0,1] \to \mathbb{R}\\ \varphi(t) &= \tilde{f}\left(t\cdot\mathbf{w}\right) \ .\end{aligned}$$

Note that by assumption on $\tilde{f}$, φ is differentiable, and φ' is L-Lipschitz. Lemma 18 ensures us that φ' is relatively close to zero on the grid Δ, as showed in the following lemma.

Lemma 19 $\forall\delta\in\Delta:\ \left|\varphi'\left(\delta\right)\right| \leq 2\epsilon + \frac{1}{32}$

Proof By Cauchy-Schwarz, Lemma 18 and the definition of a smoother, we get that for all $\delta \in \Delta$:

$$\begin{aligned}\left|\varphi'\left(\delta\right)\right| &= \left|\left\langle\nabla\tilde{f}\left(\delta\mathbf{w}\right),\mathbf{w}\right\rangle\right| \leq \left\|\nabla\tilde{f}\left(\delta\mathbf{w}\right)\right\|\cdot\|\mathbf{w}\| = \left\|\nabla\tilde{f}\left(\delta\mathbf{w}\right)\right\|\\ &\leq \left\|\mathbb{E}\left[\mathcal{S}\left(f,\delta\mathbf{w}\right)\right] - \nabla\tilde{f}\left(\delta\mathbf{w}\right)\right\| + \left\|\mathbb{E}\left[\mathcal{S}\left(f,\delta\mathbf{w}\right)\right]\right\| \leq \epsilon + \epsilon + \frac{1}{32} \ .\end{aligned}$$

■

By combining the fact that φ' has small values along the grid Δ, with the fact that φ' is L-Lipschitz, we can bound the oscillation of φ along the unit interval.

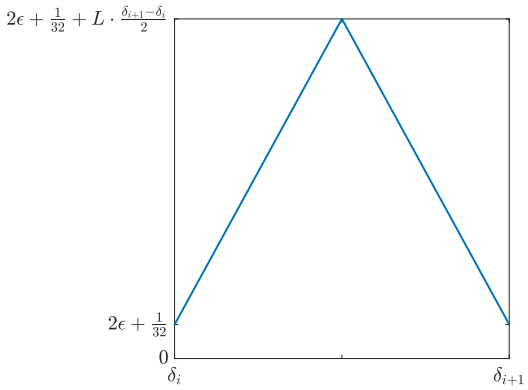

Figure 3: Illustration of $l(t)$

Lemma 20 $|\varphi(1) - \varphi(0)| \leq 2\epsilon + \frac{1}{32} + \frac{4Lr\sqrt{\log((M+1)(T+1))}}{\sqrt{d}}$.

Proof Denote $\delta_i = 16r\sqrt{\frac{\log((M+1)(T+1))}{d}} \cdot i$, and note that for all $i \in \left[\left\lfloor \frac{1}{16r}\sqrt{\frac{d}{\log((M+1)(T+1))}} \right\rfloor\right]$: $\delta_i \in \Delta$. Then

$$|\varphi(1) - \varphi(0)| = \left|\int_0^1 \varphi'(t)\, dt\right| \leq \int_0^1 \left|\varphi'(t)\right| dt = \sum_{i=0}^{\left\lfloor \frac{1}{16r}\sqrt{\frac{d}{\log((M+1)(T+1))}} \right\rfloor - 1} \int_{\delta_i}^{\delta_{i+1}} \left|\varphi'(t)\right| dt$$
$$\leq \left(\frac{1}{16r}\sqrt{\frac{d}{\log((M+1)(T+1))}}\right) \cdot \max_i \int_{\delta_i}^{\delta_{i+1}} \left|\varphi'(t)\right| dt \ . \tag{24}$$

By Lemma 19 we have $\left|\varphi'(\delta_i)\right|, \left|\varphi'(\delta_{i+1})\right| \leq 2\epsilon + \frac{1}{32}$. Recall that φ' is L-Lipschitz, so $|\varphi'(t)|$ is majorized on the interval $[\delta_i, \delta_{i+1}]$ by the piecewise linear function (see Fig. 3)

$$l(t) = \begin{cases} 2\epsilon + \frac{1}{32} + L(t - \delta_i) & \delta_i \leq t \leq \frac{\delta_i + \delta_{i+1}}{2} \\ 2\epsilon + \frac{1}{32} + L(\delta_{i+1} - t) & \frac{\delta_i + \delta_{i+1}}{2} < t \leq \delta_{i+1} \end{cases} .$$

Consequently,

$$\int_{\delta_i}^{\delta_{i+1}} \left|\varphi'(t)\right| dt \leq \int_{\delta_i}^{\delta_{i+1}} l(t)\, dt$$
$$= \left(2\epsilon + \frac{1}{32}\right) \cdot 16r\sqrt{\frac{\log((M+1)(T+1))}{d}} + L\left(8r\sqrt{\frac{\log((M+1)(T+1))}{d}}\right)^2 , \tag{25}$$

where the last equality is a direct calculation. Plugging Eq. (25) into Eq. (24), we get that

$$|\varphi(1) - \varphi(0)| \leq 2\epsilon + \frac{1}{32} + \frac{4Lr\sqrt{\log((M+1)(T+1))}}{\sqrt{d}} .$$

■

We are now ready to finish the proof. Notice that $\varphi(0) = \tilde{f}(0)$, $\varphi(1) = \tilde{f}(\mathbf{w})$. Additionally,

a direct calculation shows that $f(\mathbf{0}) = 0,\ f(\mathbf{w}) \geq \frac{1}{2}$. Using the fact that $\|\tilde{f} - f\|_\infty \leq \epsilon$, Lemma 20 reveals

$$\begin{aligned}\frac{1}{2} &\leq |f(\mathbf{w}) - f(0)| \leq \left|\tilde{f}(\mathbf{w}) - \tilde{f}(0)\right| + 2\epsilon = |\varphi(1) - \varphi(0)| + 2\epsilon \\ &\leq 4\epsilon + \frac{1}{32} + \frac{4Lr\sqrt{\log\left((M+1)(T+1)\right)}}{\sqrt{d}} \\ &\implies L\sqrt{\log\left((M+1)(T+1)\right)} \geq \frac{\sqrt{d}}{r}\left(\frac{15}{128} - \epsilon\right)\ .\end{aligned}$$

6. Discussion

In this paper, we studied the problem of nonconvex, nonsmooth optimization from an oracle complexity perspective, and provided two main results: One (in Sec. 3) is an impossibility result for efficiently getting near approximately-stationary points, and the second (in Sec. 4) proving an inherent trade-off between oracle complexity and the smoothness parameter when smoothing nonsmooth functions. The second result also establishes the optimality of randomized smoothing as an efficient smoothing method, under mild assumptions.

Our work leaves open several questions. First, at a more technical level, there is the question of whether some or all of our assumptions in Sec. 4 can be relaxed. The result currently requires the algorithm to be translation invariant w.r.t. constant functions, as well as querying at some bounded distance from the input point $\mathbf{x}$. We conjecture that the translation invariance assumption can be relaxed, possibly by a suitable reduction that shows that any smoothing algorithm can be converted to a translation invariant one. However, how to formally perform this remains unclear at the moment. As to the bounded distance of the queries, it is currently an essential assumption for our proof technique, which relies on a function which looks "locally" constant at many different points, but is globally nonconstant, and this can generally be determined by querying far enough away from the input point (even along some random direction). Thus, relaxing this assumption may necessitate a different proof technique.

Another open question is whether randomized smoothing can be "derandomized": Our results indicate that the gradient Lipschitz parameter of the smooth approximation cannot be improved, but leave open the possibility of an efficient method returning the actual gradients of some smooth approximation (up to machine precision), in contrast to randomized smoothing which only provides noisy stochastic estimates of the gradient. These can then be plugged into smooth optimization methods which assume access to the exact gradients (rather than noisy stochastic estimates), generally improving the resulting iteration complexity. We note that naively computing the exact gradient of $\tilde{f}(\cdot)$ arising from randomized smoothing is infeasible in general, as it involves a high-dimensional integral.

At a more general level, our work leaves open the question of what is the "right" metric to study for nonsmooth-nonconvex optimization, where neither minimizing optimization error nor finding approximately-stationary points is feasible. In this paper, we show that the goal of getting near approximately stationary points is not feasible, at least in the worst case, whereas smoothing can be done efficiently, but not in a dimension-free manner. Can we find other compelling goals to consider? One very appealing notion is the (δ, ϵ)-stationarity of Zhang et al. (2020) that we mentioned in the introduction, which comes with clean,

finite-time and dimension-free guarantees. Our negative result in Thm. 1 provides further motivation to consider it, by showing that a natural variation of this notion will not work. However, as we discuss in Appendix A, we need to accept that this stationarity notion can have unexpected behavior, and there exist cases where it will not resemble a stationary point in any intuitive sense. In any case, using an oracle complexity framework to study this and other potential metrics for nonsmooth nonconvex optimization, which combine computational efficiency and finite-time guarantees, remains an interesting direction for future research.

Acknowledgements

This research is supported by the European Research Council (ERC) grant 754705.

Appendix A. (δ, ϵ)-Stationarity (Zhang et al., 2020)

In the recent work by Zhang, Lin, Jegelka, Sra and Jadbabaie (Zhang et al., 2020), the authors prove that for nonconvex nonsmooth functions, finding ϵ-approximately stationary points is infeasible in general. Instead, they study the following relaxation (based on the notion of δ-differential introduced by Goldstein 1977): Letting $\partial f(\mathbf{x})$ denote the generalized gradient set (as defined in Sec. 2) of $f(\cdot)$ at $\mathbf{x}$, we say that a point $\mathbf{x}$ is a (δ, ϵ)-*stationary point*, if

$$\min\{\|\mathbf{u}\| : \mathbf{u} \in \mathrm{conv}\{\cup_{\mathbf{y}:\|\mathbf{y}-\mathbf{x}\|\leq\delta} \partial f(\mathbf{y})\}\} \;\leq\; \epsilon \;, \tag{26}$$

where $\mathrm{conv}\{\cdot\}$ is the convex hull. In words, there exists a convex combination of gradients at a δ-neighborhood of $\mathbf{x}$, whose norm is at most ϵ. Remarkably, the authors then proceed to provide a dimension-free, gradient-based algorithm for finding (δ, ϵ)-stationary points, using $\mathcal{O}(1/\delta\epsilon^3)$ gradient and value evaluations, as well as study related settings. Subsequently, several works have continued the study of optimization in terms of (δ, ϵ)-stationary points (Davis et al., 2022; Tian et al., 2022).

Although this constitutes a very useful algorithmic contribution to nonsmooth optimization, it is important to note that a (δ, ϵ)-stationary point $\mathbf{x}$ (as defined above) *does not* imply that $\mathbf{x}$ is δ-close to an ϵ-stationary point of $f(\cdot)$, nor that $\mathbf{x}$ necessarily resembles a stationary point. Intuitively, this is because the convex hull of the gradients might contain a small vector, without any of the gradients being particular small. This is formally demonstrated in the following proposition:

Proposition 21 *For any $\delta > 0$, there exists a differentiable function $f(\cdot)$ on $\mathbb{R}^2$ which is 2π-Lipschitz on a ball of radius 2δ around the origin, and the origin is a $(\delta, 0)$-stationary point, yet $\min_{\mathbf{x}:\|\mathbf{x}\|\leq\delta} \|\nabla f(\mathbf{x})\| \geq 1$.*

Proof Fixing some $\delta > 0$, consider the function

$$f(u, v) \;:=\; (2\delta + u)\sin\left(\frac{\pi}{2\delta}v\right)$$

(see Fig. A for an illustration). This function is differentiable, and its gradient satisfies

$$\nabla f(u, v) \;=\; \left(\sin\left(\frac{\pi}{2\delta}v\right) \;,\; \frac{\pi}{2\delta}(2\delta + u)\cos\left(\frac{\pi}{2\delta}v\right)\right) \;.$$

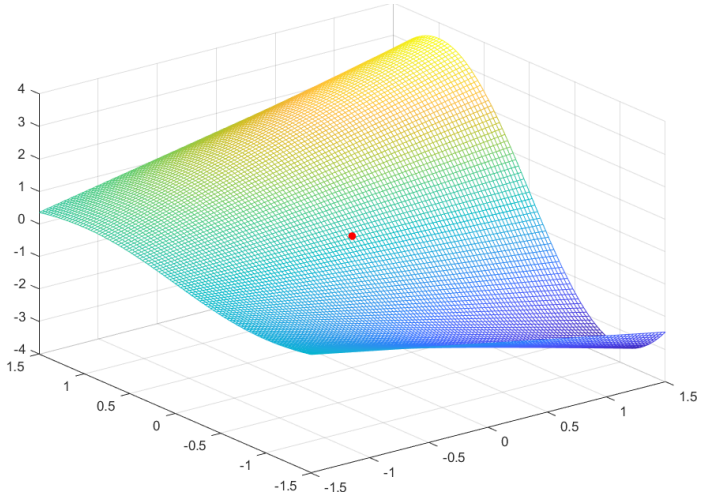

Figure 4: The function used in the proof of Proposition 21, for $\delta = 1$. The origin (which fulfills the definition of a $(1,0)$-stationary point) is marked with a red dot. Best viewed in color.

First, we note that

$$\frac{1}{2}\left(\nabla f(0,\delta) + \frac{1}{2}\nabla f(0,-\delta)\right) \;=\; \frac{1}{2}\left((1,0)+(-1,0)\right) \;=\; (0,0),$$

which implies that $(0,0)$ is in the convex hull of the gradients at a distance at most δ from the origin, hence the origin is a $(\delta,0)$-stationary point. Second, we have that

$$\|\nabla f(u,v)\|^2 = \sin^2\left(\frac{\pi}{2\delta}v\right) + \left(\frac{\pi}{2\delta}\right)^2 (2\delta+u)^2 \cos^2\left(\frac{\pi}{2\delta}v\right) \;. \tag{27}$$

For any (u,v) of norm at most 2δ, we must have $|u| \le 2\delta$, and therefore the above is at most

$$\sin^2\left(\frac{\pi}{2\delta}v\right) + \left(\frac{\pi}{2\delta}\right)^2 (2\delta+2\delta)^2 \cos^2\left(\frac{\pi}{2\delta}v\right) \;\le\; 4\pi^2\left(\sin^2\left(\frac{\pi}{2\delta}v\right) + \cos^2\left(\frac{\pi}{2\delta}v\right)\right) \;=\; 4\pi^2 \;,$$

which implies that the function is 2π-Lipschitz on a ball of radius 2δ around the origin. Finally, for any (u,v) of norm at most δ, we have $|u| \le \delta$, so Eq. (27) is at least

$$\sin^2\left(\frac{\pi}{2\delta}v\right) + \left(\frac{\pi}{2\delta}\right)^2 (2\delta-\delta)^2 \cos^2\left(\frac{\pi}{2\delta}v\right) \;\ge\; \sin^2\left(\frac{\pi}{2\delta}v\right) + \cos^2\left(\frac{\pi}{2\delta}v\right) \;=\; 1 \;.$$

■

Remark 22 (Extension to globally Lipschitz functions) *Although the function $f(\cdot)$ in the proof has a constant Lipschitz parameter only close to the origin, it can be easily*

modified to be globally Lipschitz and bounded, for example by considering the function

$$\tilde{f}(\mathbf{x}) \;=\; \begin{cases} f(\mathbf{x}) & \|\mathbf{x}\| \leq 2\delta \\ \max\left\{0, 2 - \frac{\|\mathbf{x}\|}{2\delta}\right\} \cdot f\left(\frac{2\delta}{\|\mathbf{x}\|}\mathbf{x}\right) & \|\mathbf{x}\| > 2\delta \end{cases} ,$$

which is identical to $f(\cdot)$ in a ball of radius 2δ around the origin, but decays to 0 for larger $\mathbf{x}$, and can be verified to be globally bounded and Lipschitz independent of δ.

Remark 23 (Extension to constant distances) *The proof of Thm. 1 uses a (more complicated) construction that actually strengthens Proposition 21: It implies that for any δ, ϵ smaller than some constants, there is a Lipschitz, bounded-from-below function on $\mathbb{R}^d$, such that the origin is $(\delta, 0)$-stationary, yet there are no ϵ-stationary points even at a constant distance from the origin. In more details, consider the function*

$$\hat{g}_{\mathbf{w}}(\mathbf{x}) := \max\{g_{\mathbf{w}}(\mathbf{0}) - 1 \;,\; g_{\mathbf{w}}(\mathbf{x})\} \;,$$

$$g(\mathbf{x})_{\mathbf{w}} = |x_d| + \frac{1}{4}\sqrt{\sum_{i=1}^{d-1} x_i^2} - \left[\langle \overline{\mathbf{w}}, \mathbf{x} + \mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x} + \mathbf{w}\|\right]_+ .$$

Using exactly the same proof as for Lemma 14, one can show that $g_{\mathbf{w}}(\cdot)$ is $\frac{15}{4}$-Lipschitz and has no ϵ-stationary points for $\epsilon < 1/4\sqrt{2}$. Therefore, it is easily verified that for any $\mathbf{w}$, $\hat{g}_{\mathbf{w}}(\cdot)$ is $\frac{15}{4}$-Lipschitz, bounded from below, and any ϵ-stationary point is at a distance of at least $4/15$ from the origin.[7] *However, we also claim that the origin is a $(\delta, 0)$-stationary point for any $\delta \in (0, 4/15)$. To see this, note first that for such δ, by the Lipschitz property of $g_{\mathbf{w}}(\cdot)$, we have $\hat{g}_{\mathbf{w}}(\mathbf{x}) = g_{\mathbf{w}}(\mathbf{x})$ in a δ-neighborhood of the origin. Fix any $\mathbf{w}$ such that $\|\mathbf{w}\| = \frac{\delta}{2}$, $w_d = 0$, and let $\mathbf{v} \in \{-\delta \cdot \mathbf{e}_d, \delta \cdot \mathbf{e}_d\}$. It is easily verified that $\langle \bar{\mathbf{w}}, \mathbf{v} + \mathbf{w}\rangle - \frac{1}{2}\|\mathbf{v} + \mathbf{w}\| < 0$, in which case*

$$\mathrm{sign}(v_d) \cdot \mathbf{e}_d \in \partial g_{\mathbf{w}}(\mathbf{v}) \;=\; \partial \hat{g}_{\mathbf{w}}(\mathbf{v}) \;,$$

and therefore $\frac{1}{2}\left(\nabla \hat{g}_{\mathbf{w}}(\mathbf{v}) + \nabla \hat{g}_{\mathbf{w}}(-\mathbf{v})\right) = \mathbf{0}$, where $\nabla \hat{g}_{\mathbf{w}}(\mathbf{v})$ denotes the subgradient defined above.

We end by noting that if we drop the the $\mathrm{conv}\{\cdot\}$ operator from the definition of (δ, ϵ)-stationarity in Eq. (26), the goal becomes equivalent to finding points which are δ-close to ϵ-approximately stationary points – which is exactly the goal we study in Sec. 3, and for which we show a strong impossibility result. This impossibility result implies that a natural strengthening of the notion of (δ, ϵ)-stationarity is already too strong to be feasible in general.

Appendix B. Runtime of smoothed GD suffers from a dimension dependency

In this appendix, we formally prove that randomized smoothing can indeed lead to strong dimension dependencies in the iteration complexity of simple gradient methods – in particular, vanilla gradient descent with constant step size – even for simple convex functions.

7. The last point follows from the fact that if $\mathbf{y}$ is an ϵ-stationary point of $\hat{g}_{\mathbf{w}}(\cdot)$, then we can find a point $\mathbf{x}$ arbitrarily close to $\mathbf{y}$ such that $\hat{g}_{\mathbf{w}}(\mathbf{x}) \neq g_{\mathbf{w}}(\mathbf{x})$, hence $g_{\mathbf{w}}(\mathbf{x}) < g_{\mathbf{w}}(\mathbf{0}) - 1$, and as a result $g_{\mathbf{w}}(\mathbf{0}) - f_{\mathbf{w}}(\mathbf{x}) > 1$. But $g_{\mathbf{w}}(\cdot)$ is $\frac{15}{4}$-Lipschitz, hence $\|\mathbf{x}\| > 4/15$, and therefore $\|\mathbf{y}\| \geq 4/15$.

Thus, the dimension dependency arising from applying gradient descent on a randomly-smoothed function is real and not merely an artifact of the analysis (where the standard upper bound on the number of iterations scales with the gradient Lipschitz parameter). We note that we focus on constant step-size gradient descent for simplicity, and a similar analysis can be performed for other gradient-based methods, such as variable step-size gradient descent or stochastic gradient descent.

Given a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, denote the smooth approximation $\tilde{f}(\mathbf{x}) = \mathbb{E}_{\|\mathbf{v}\|\leq 1}[f(\mathbf{x} + \epsilon\mathbf{v})]$ where $\mathbf{v}$ is distributed uniformly over the unit ball. Let $\mathbf{x}_0$ be a point which is of distance at most 1 to an ϵ-stationary point of $\tilde{f}$, and consider vanilla gradient descent with a constant step size $\eta > 0$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \cdot \nabla \tilde{f}\left(\mathbf{x}_t\right) \ .$$

The following proposition shows that for any step size, applying gradient descent to find an approximately-stationary point of $\tilde{f}$ will necessitate a number of iterations scaling strongly with the dimension:

Proposition 24 *There exists a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ such that the following holds: For any $\epsilon < \frac{1}{2}$, $\eta > 0$, there exists $\mathbf{x}_0$ as above such that $\min\{t : \|\nabla \tilde{f}(\mathbf{x}_t)\| \leq \epsilon\} = \Omega\left(\frac{\sqrt{d}}{\epsilon}\right)$.*

Proof We will show the claim holds for $f\left(\mathbf{x}\right) := |x_1|$. In a nutshell, the proof is based on the observation that $\nabla \tilde{f}(\mathbf{x})$ is close to zero only when $|x_1| = \mathcal{O}(1/\sqrt{d})$. Thus, gradient descent must hit an interval of size $\mathcal{O}(1/\sqrt{d})$. But in order to guarantee this, and with an arbitrary bounded starting point, the step size must be small, and hence the number of iterations required will be large.

Proceeding with the formal proof, note that $\tilde{f}\left(\mathbf{x}\right) = \mathbb{E}_{\|\mathbf{v}\|\leq 1}\left[|x_1 + \epsilon v_1|\right]$, hence

$$\begin{aligned}
\nabla \tilde{f}\left(\mathbf{x}\right) &= \mathbb{E}_{\|\mathbf{v}\|\leq 1}\left[\operatorname{sign}\left(x_1 + \epsilon v_1\right)\right] \cdot \mathbf{e}_1 \\
&= \left(\Pr_{\|\mathbf{v}\|\leq 1}\left[x_1 + \epsilon v_1 > 0\right] - \Pr_{\|\mathbf{v}\|\leq 1}\left[x_1 + \epsilon v_1 < 0\right]\right) \cdot \mathbf{e}_1 \\
&= \left(1 - 2 \cdot \Pr_{\|\mathbf{v}\|\leq 1}\left[x_1 + \epsilon v_1 < 0\right]\right) \cdot \mathbf{e}_1 \\
&= \left(1 - 2 \cdot \Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right]\right) \cdot \mathbf{e}_1 \ . \qquad (28)
\end{aligned}$$

We draw several consequences from Eq. (28). First, if $x_1 = 0$ then $\Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right] = \frac{1}{2}$ due to symmetry around the origin, so in particular

$$\nabla \tilde{f}\left(\mathbf{0}\right) = \mathbf{0} \ . \qquad (29)$$

Second, if $x_1 \geq \epsilon$ then $\Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right] = 0$, and if $x_1 \leq -\epsilon$ then $\Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right] = 1$. Overall

$$|x_1| \geq \epsilon \implies \nabla \tilde{f}\left(\mathbf{x}\right) = \operatorname{sign}(x_1) \cdot \mathbf{e}_1 \ . \qquad (30)$$

Third, since probabilities are bounded between zero and one, we obtain the global upper estimate

$$\left\|\nabla \tilde{f}\left(\mathbf{x}\right)\right\| \leq 1 \ . \qquad (31)$$

Lastly, $\Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right]$ equals to the volume of the intersection of the halfspace $\{\mathbf{v} \in \mathbb{R}^d \mid v_1 < -\frac{x_1}{\epsilon}\}$ with the unit ball, normalized by the unit ball volume. In particular, since this intersection is a subset of the spherical sector associated with the spherical cap $\left\{\mathbf{v} \in \mathbb{S}^{d-1} \;\middle|\; v_1 < -\frac{x_1}{\epsilon}\right\}$, its normalized volume is less then the surface area of the cap. By well known estimates of spherical cap (for example Ball et al., 1997, Lemma 2.2):

$$\Pr_{\|\mathbf{v}\|\leq 1}\left[v_1 < -\frac{x_1}{\epsilon}\right] \leq \Pr_{\|\mathbf{v}\|=1}\left[v_1 < -\frac{x_1}{\epsilon}\right] \leq \exp\left(-\frac{dx_1^2}{2\epsilon^2}\right) \ . \tag{32}$$

By combining Eq. (28) and Eq. (32) we get

$$\left\|\nabla \tilde{f}\left(\mathbf{x}\right)\right\| \geq 1 - 2\exp\left(-\frac{dx_1^2}{2\epsilon^2}\right) \ .$$

In particular,

$$|x_1| \geq \frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}} \implies \left\|\nabla \tilde{f}\left(\mathbf{x}\right)\right\| \geq \frac{4}{5} \ . \tag{33}$$

We are now ready to describe the choice of $\mathbf{x}_0$ which will prove the claim, depending on the value of η.

Case I: $\eta \leq \frac{5\sqrt{2\log(10)}\epsilon}{2\sqrt{d}}$

We set $\mathbf{x}_0 = \mathbf{e}_1$. First, $\mathbf{x}_0$ is indeed at distance 1 from $\mathbf{0}$, which by Eq. (29) is a stationary point. Furthermore, by the definition of gradient descent, Eq. (28) and Eq. (31), for all $t \leq \frac{2\sqrt{d}}{5\sqrt{2\log(10)}\epsilon} - \frac{2}{5}$:

$$\begin{aligned}
\left(\mathbf{x}_{t+1}\right)_1 &= \left(\mathbf{x}_0 - \eta\left(\sum_{i=1}^{t}\nabla \tilde{f}\left(\mathbf{x}_i\right)\right)\right)_1 \\
&\geq 1 - \frac{5\sqrt{2\log(10)}\epsilon}{2\sqrt{d}} \cdot t \cdot 1 \\
&\geq \frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}} \ .
\end{aligned}$$

So by Eq. (33), for every $t \leq \frac{2\sqrt{d}}{5\sqrt{2\log(10)}\epsilon} - \frac{2}{5}$: $\left\|\nabla \tilde{f}\left(\mathbf{x}_t\right)\right\| \geq \frac{4}{5}$. Consequently, the minimal t for which the gradient norm is less than ϵ satisfies $t > \frac{2\sqrt{d}}{5\sqrt{2\log(10)}\epsilon} - \frac{2}{5} = \Omega(\frac{\sqrt{d}}{\epsilon})$.

Case II: $\frac{5\sqrt{2\log(10)}\epsilon}{2\sqrt{d}} < \eta \leq 2$

In this case, we define the real function

$$\phi\left(s\right) := 2s - \eta\left(\nabla \tilde{f}\left(s \cdot \mathbf{e}_1\right)\right)_1 \ .$$

On on hand, by assumption on η and Eq. (33):

$$\begin{aligned}\phi\left(\frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}}\right) &= \frac{2\sqrt{2\log(10)}\epsilon}{\sqrt{d}} - \eta\left(\nabla\tilde{f}\left(s\cdot\mathbf{e}_1\right)\right)_1 \\ &\leq \frac{2\sqrt{2\log(10)}\epsilon}{\sqrt{d}} - \frac{5\sqrt{2\log(10)}\epsilon}{2\sqrt{d}}\cdot\frac{4}{5} \\ &= 0\ .\end{aligned}$$

On the other hand, $\frac{\eta}{2} > \frac{5\sqrt{2\log(10)}\epsilon}{4\sqrt{d}} > \frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}}$ and by Eq. (31):

$$\begin{aligned}\phi\left(\frac{\eta}{2}\right) &= \eta - \eta\left(\nabla\tilde{f}\left(\frac{\eta}{2}\cdot\mathbf{e}_1\right)\right)_1 \\ &\geq \eta - \eta\cdot 1 \\ &= 0\ .\end{aligned}$$

Notice that ϕ is continuous since $\tilde{f}$ is smooth, so by the intermediate value theorem there exists $s^* \in \left[\frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}}, \frac{\eta}{2}\right]$ such that $\phi\left(s^*\right) = 0$. Equivalently,

$$s^* - \eta\left(\nabla\tilde{f}\left(s\cdot\mathbf{e}_1\right)\right)_1 = -s^*\ . \tag{34}$$

We set $\mathbf{x}_0 = s^*\mathbf{e}_1$. First, $\mathbf{x}_0$ is of distance at most $\frac{\eta}{2} \leq 1$ from $\mathbf{0}$, which by Eq. (29) is a stationary point. Furthermore, by the definition of gradient descent and Eq. (34) we get

$$\mathbf{x}_1 = s^*\mathbf{e}_1 - \eta\nabla\tilde{f}\left(s^*\mathbf{e}_1\right) = -s^*\mathbf{e}_1 = -\mathbf{x}_0\ .$$

Inductively, due to the symmetry of $\tilde{f}$ with respect to the origin, we obtain $\mathbf{x}_t = (-1)^t\mathbf{x}_0$. In particular, since $s^* \geq \frac{\sqrt{2\log(10)}\epsilon}{\sqrt{d}}$ Eq. (33) ensures that for all $t\in\mathbb{N}:\ \left\|\nabla\tilde{f}\left(\mathbf{x}_t\right)\right\| \geq \frac{4}{5} > \epsilon$.

Case III: $\eta > 2$

Set $\mathbf{x}_0 = \mathbf{e}_1$, which satisfies the distance assumption as explained in case I. By the definition of gradient descent and Eq. (30):

$$\mathbf{x}_1 = \mathbf{e}_1 - \eta\nabla\tilde{f}\left(\mathbf{e}_1\right) = \left(1-\eta\right)\mathbf{e}_1\ .$$

Notice that $(1-\eta) < -1$, so by invoking Eq. (30) we get

$$\mathbf{x}_2 = \mathbf{x}_1 - \eta\nabla\tilde{f}\left(\mathbf{x}_1\right) = \left(1-\eta\right)\mathbf{e}_1 + \eta\mathbf{e}_1 = \mathbf{x}_0\ .$$

We deduce that for all $t\in\mathbb{N}:\ \mathbf{x}_{t+2} = \mathbf{x}_t$, and in particular by Eq. (30): $\left\|\nabla\tilde{f}\left(\mathbf{x}_t\right)\right\| = 1 > \epsilon$.
■

Appendix C. Proof of Proposition 11

Denote by $\mathcal{H}$ the set of non-negative 2-Lipschitz functions h such that $h(0) = 1$, $x^* := \arg\min_{x\in\mathbb{R}} h(x) \in (0,1)$ is unique, and $\forall x \neq x^* \ \forall g \in \partial h(x) : |g| \geq 1$. We start by showing that if the proposition does not hold, then there exists an algorithm that finds the minimum of any function in $\mathcal{H}$ within some finite time, with high probability. We will use this implication in order to obtain a contradiction.

Assume by contradiction that Proposition 11 does not hold. That is, that there exist $\mathcal{A}, T, \delta$ such that for any $h \in \mathcal{H}, \rho > 0$,

$$\Pr_{\mathcal{A}}\left[\min_{t\in[T]} |x_t^h - x^*| < \rho\right] \geq \delta \ . \tag{35}$$

Let $(\rho_n)_{n=1}^{\infty} > 0$ be a decreasing sequence such that $\lim_{n\to\infty} \rho_n = 0$. By continuity of probability measures with respect to decreasing events, assuming Eq. (35) for any $\rho > 0$ implies

$$\begin{aligned}\Pr_{\mathcal{A}}\left[\exists t \in [T] : x_t^h = x^*\right] &= \Pr_{\mathcal{A}}\left[\min_{t\in[T]} |x_t^h - x^*| = 0\right] = \Pr_{\mathcal{A}}\left[\bigcap_{n=1}^{\infty} \min_{t\in[T]} |x_t^h - x^*| < \rho_n\right] \\ &= \lim_{n\to\infty} \Pr_{\mathcal{A}}\left[\min_{t\in[T]} |x_t^h - x^*| < \rho_n\right] \geq \delta \ . \end{aligned} \tag{36}$$

Namely, there exists some algorithm $\mathcal{A}$ that gets to the exact minimum of any function in $\mathcal{H}$ within some finite time T, with some positive probability δ. But note that a classic confidence boosting argument shows that if Eq. (36) holds for *some* $0 < \delta < 1$, then it actually holds for *any* $0 < \delta < 1$ (with an appropriate blow up in running time). Indeed, assuming it is true for some $\delta_0, \mathcal{A}_0, T_0$, we define an algorithm $\mathcal{A}$ which simulates N independent copies of $\mathcal{A}_0$, and returns the point with the smallest function value over all seen iterates along all the independent copies. By standard Chernoff-Hoeffding bounds one easily gets that for N being some function of δ_0 and δ, $\mathcal{A}$ satisfies Eq. (36) for $T = N \cdot T_0$. Hence,

$$\forall \delta < 1 \ \exists \mathcal{A}_\delta, T_\delta \ \forall h \in \mathcal{H} \ : \ \Pr_{\mathcal{A}_\delta}\left[\exists t \in [T] : x_t^h = x^*\right] \geq \delta \ . \tag{37}$$

We fix $\delta = \frac{1}{2}$, and let $\mathcal{A}, T_0$ be it's associated algorithm and iteration number. By Yao's lemma (Yao, 1977), we can assume $\mathcal{A}$ is deterministic and provide a distribution over hard functions. Namely, in order to arrive at a contradiction it is enough to show that

$$\Pr_{\sigma}\left[\exists t \in [T_0] : x_t^{h_\sigma} = x^*\right] < \frac{1}{2} \ , \tag{38}$$

for some distribution over σ, such that $\forall \sigma : h_\sigma \in \mathcal{H}$.

Before delving into the technical details, we turn to explain the intuition behind the construction. We consider functions h_σ^N indexed by $\sigma \in \{0,1\}^N, N \in \mathbb{N}$. The function $h_{\sigma_1}^1$ "tilts to the left" if $\sigma_1 = 0$, and "tilts to the right" if $\sigma_1 = 1$ (see Fig. 5). Given these two functions, we define the functions for $N = 2$, such that σ_1 determines an "outer" tilt, while σ_2 determines an "inner" tilt which behaves like $h_{\sigma_2}^1$ (once again, see Fig. 5). These functions are such that for any point outside the outer tilted segment, it's value does not

depend on the inner tilt - that is, on σ_2. We continue this process recursively for any $N \in \mathbb{N}$, such that the larger i is, σ_i determines finer tilts in smaller segments. The construction has the property that for all points outside the tilt determined by σ_i, the function's values do not depend on $\sigma_{i+1}, \ldots, \sigma_N$. Thus, by setting N large enough relatively to T, we will be able to ensure that $x_1, \ldots, x_T$ are not likely to depend on σ_N, therefore missing the minimum which does depend on it. To that end, we define the following functions:

$$h_0^1(x) = \begin{cases} 1-x & x \in (-\infty, 0) \\ 1-2x & x \in \left[0, \frac{3}{8}\right] \\ \frac{6}{5}x - \frac{1}{5} & x \in \left[\frac{3}{8}, 1\right] \\ x & x \in (1, \infty) \end{cases} \quad , \quad h_1^1(x) = \begin{cases} 1-x & x \in (-\infty, 0) \\ -\frac{6}{5}x + 1 & x \in \left[0, \frac{5}{8}\right] \\ 2x-1 & x \in \left[\frac{5}{8}, 1\right] \\ x & x \in (1, \infty) \end{cases} .$$

Next, in a recursive manner, for any $\hat{\sigma} := (\sigma_2, \ldots, \sigma_N) \in \{0,1\}^{N-1}$ we define (see Fig. 5):

$$h_{0,\hat{\sigma}}^N(x) = \begin{cases} 1-x & x \in (-\infty, 0) \\ 1-2x & x \in \left[0, \frac{1}{4}\right] \\ \frac{1}{4}h_{\hat{\sigma}}^{(N-1)}(4x-1) + \frac{1}{4} & x \in \left[\frac{1}{4}, \frac{1}{2}\right] \\ x & x \in \left[\frac{1}{2}, 1\right] \\ x & x \in (1, \infty) \end{cases} ,$$

$$h_{1,\hat{\sigma}}^N(x) = \begin{cases} 1-x & x \in (-\infty, 0) \\ 1-x & x \in \left[0, \frac{1}{2}\right] \\ \frac{1}{4}h_{\hat{\sigma}}^{(N-1)}(4x-2) + \frac{1}{4} & x \in \left[\frac{1}{2}, \frac{3}{4}\right] \\ 2x-1 & x \in \left[\frac{3}{4}, 1\right] \\ x & x \in (1, \infty) \end{cases} .$$

Lemma 25 *For any $N \in \mathbb{N}$, it holds that for all $\sigma \in \{0,1\}^N$: $h_\sigma^N \in \mathcal{H}$.*

Proof The proof is by induction on N. For $N = 1$ it is easy to verify that $h_0^1, h_1^1 \in \mathcal{H}$, and $h_0^1(1) = h_1^1(1) = 1$, as well as the fact that h_0^1, h_1^1 are both piecewise linear. Assume the claim is true for some $N-1$, and that for all $\hat{\sigma} \in \{0,1\}^{N-1} : h_{\hat{\sigma}}^{N-1}(1) = 1$ as well as that they are all piecewise linear. Consider $h_{0,\hat{\sigma}}^N$ for some $\hat{\sigma} \in \{0,1\}^{N-1}$. First, $h_{0,\hat{\sigma}}^N(0) = 1 - 2 \cdot 0 = 1$ as required. Moreover, it is clear by definition that $h_{0,\hat{\sigma}}^N(x) \geq 0$ for all $x \notin [\frac{1}{4}, \frac{1}{2}]$. For $x \in [\frac{1}{4}, \frac{1}{2}]$, we have by the induction hypothesis

$$h_{0,\hat{\sigma}}^N(x) = \frac{1}{4} \cdot \underbrace{h_{\hat{\sigma}}^{N-1}(4x-1)}_{\geq 0} + \frac{1}{4} \geq 0 ,$$

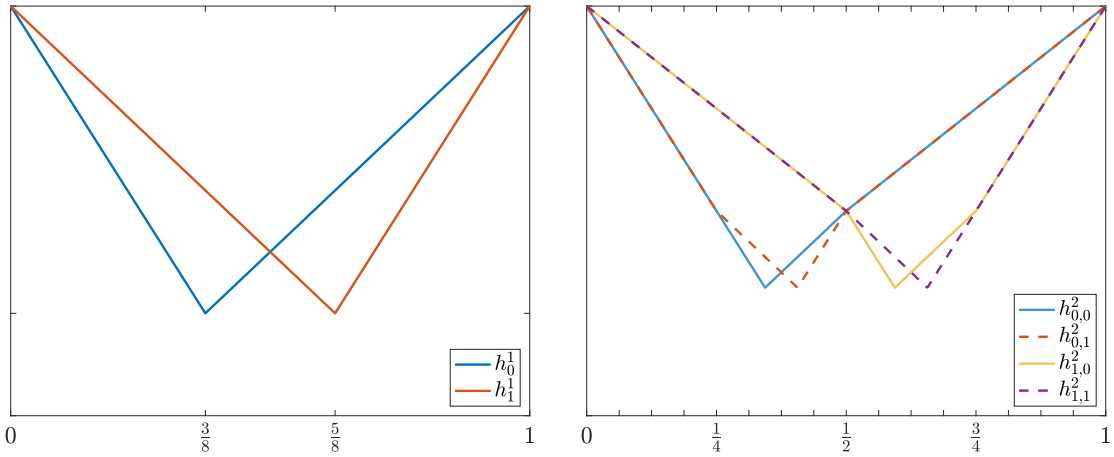

Figure 5: Plots of the functions $h_\sigma^N, \sigma \in \{0,1\}^N$, $N = 1$ (on the left) and $N = 2$ (on the right).

which establishes the required non-negativity property. Note that $h_{0,\hat{\sigma}}^N$ is continuous. Indeed, it is easy to verify continuity for any $x \notin \{\frac{1}{4}, \frac{1}{2}\}$, and for those two points we have

$$h_{0,\hat{\sigma}}^N\left(\frac{1}{4}\right) = \frac{1}{4}h_{\hat{\sigma}}^{N-1}(0) + \frac{1}{4} = \frac{1}{4}\cdot 1 + \frac{1}{4} = \frac{1}{2} = \lim_{x\to\frac{1}{4}}(1-2x)$$
$$h_{0,\hat{\sigma}}^N\left(\frac{1}{2}\right) = \frac{1}{4}h_{\hat{\sigma}}^{N-1}(1) + \frac{1}{4} = \frac{1}{4}\cdot 1 + \frac{1}{4} = \frac{1}{2} = \lim_{x\to\frac{1}{2}}(x)\ .$$

Using our induction hypothesis we see that $h_{0,\hat{\sigma}}^N$ is a piecewise linear continuous function. Thus, in order to prove the remaining properties it is enough to inspect $\partial h_{0,\hat{\sigma}}^N$. It is easy to verify that for all $x \notin [\frac{1}{4}, \frac{1}{2}]:\ \partial h_{0,\hat{\sigma}}^N(x) \subseteq \{-1, -2, 1\}$. For $x \in [\frac{1}{4}, \frac{1}{2}]$, we have

$$\begin{aligned}\partial h_{0,\hat{\sigma}}^N(x) &= \partial\left(\left(z \mapsto \frac{1}{4}z + \frac{1}{4}\right) \circ \left(y \mapsto h_{\hat{\sigma}}^{N-1}(y)\right) \circ (x \mapsto 4x-1)\right)(x) \\ &= \left\{\frac{1}{4}\cdot g \cdot 4 \,\middle|\, g \in \partial h_{\hat{\sigma}}^{N-1}(4x-1)\right\} \\ &= \left\{g \,\middle|\, g \in \partial h_{\hat{\sigma}}^{N-1}(4x-1)\right\}\ . \end{aligned} \tag{39}$$

By the induction hypothesis, all elements of the set above are of magnitude at most 2, which establishes the desired Lipschitz property. Furthermore, note that if $x \in [\frac{1}{4}, \frac{1}{2}]$ then $4x - 1 \in [0, 1]$. Using the induction hypothesis, let $x^* \in [\frac{1}{4}, \frac{1}{2}]$ be the unique number such that $4x^* - 1 = \arg\min h_{\hat{\sigma}}^{N-1}$. Then by the induction hypothesis and Eq. (39), for all $x \in [\frac{1}{4}, \frac{1}{2}] \setminus \{x^*\}\ \forall g \in \partial h_{0,\hat{\sigma}}^N(x) : |g| \geq 1$ which is exactly the desired property, assuming we show that $x^* = \arg\min h_{0,\hat{\sigma}}^N$. Finally, noticing that $h_{0,\hat{\sigma}}^N$ is decreasing for all $x < x^*$ and increasing for all $x > x^*$ (which inside the interval $[\frac{1}{4}, \frac{1}{2}]$ follows once again from the

induction hypothesis and our choice of x^*) finishes the proof for $h^N_{0,\hat{\sigma}}$. The proof for $h^N_{1,\hat{\sigma}}$ is essentially the same. ■

We define

$$I_{\sigma_1,\ldots,\sigma_k} := \left[\sum_{i=1}^{k} \frac{\sigma_i+1}{4^i},\ \sum_{i=1}^{k} \frac{\sigma_i+1}{4^i} + \frac{1}{4^k}\right] \subset \mathbb{R} \ .$$

Lemma 26 *For any $l < k$ and any $\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k \in \{0,1\}$: $I_{\sigma_1,\ldots,\sigma_l} \supset I_{\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k}$.*

Proof On one hand,

$$\sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} < \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \sum_{i=l+1}^{k} \frac{\sigma_i+1}{4^i} = \sum_{i=1}^{k} \frac{\sigma_i+1}{4^i} \ .$$

On the other hand,

$$\begin{aligned}
\sum_{i=1}^{k} \frac{\sigma_i+1}{4^i} + \frac{1}{4^k} &= \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \sum_{i=l+1}^{k} \frac{\sigma_i+1}{4^i} + \frac{1}{4^k} \le \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \sum_{i=l+1}^{k} \frac{2}{4^i} + \frac{1}{4^k} \\
&= \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \frac{2}{3}\left(\frac{1}{4^l} - \frac{1}{4^k}\right) + \frac{1}{4^k} = \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \frac{2}{3}\cdot\frac{1}{4^l} + \frac{1}{3}\cdot\frac{1}{4^k} \\
&< \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \frac{2}{3}\cdot\frac{1}{4^l} + \frac{1}{3}\cdot\frac{1}{4^l} = \sum_{i=1}^{l} \frac{\sigma_i+1}{4^i} + \frac{1}{4^l} \ .
\end{aligned}$$

■

Lemma 27 *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 Let $i_0 \le k-1$ the the minimal index i for which $\sigma_i \neq \sigma'_i$. Assume without loss of generality that $\sigma_{i_0} = 0, \sigma'_{i_0} = 1$. If $x \in I_{\sigma_1,\ldots,\sigma_k}$ then by definition

$$\begin{aligned}
x &\le \sum_{i=1}^{k} \frac{\sigma_i+1}{4^i} + \frac{1}{4^k} \\
&= \sum_{i=1}^{i_0-1} \frac{\sigma_i+1}{4^i} + \frac{\sigma_{i_0}+1}{4^{i_0}} + \sum_{i=i_0+1}^{k} \frac{\sigma_i+1}{4^i} + \frac{1}{4^k} \\
&\le \sum_{i=1}^{i_0-1} \frac{\sigma'_i+1}{4^i} + \frac{1}{4^{i_0}} + \sum_{i=i_0+1}^{k} \frac{2}{4^i} + \frac{1}{4^k} \\
&= \sum_{i=1}^{i_0-1} \frac{\sigma'_i+1}{4^i} + \frac{1}{4^{i_0}} + \frac{2}{3}\left(\frac{1}{4^{i_0}} - \frac{1}{4^k}\right) + \frac{1}{4^k} \\
&= \sum_{i=1}^{i_0-1} \frac{\sigma'_i+1}{4^i} + \frac{2}{4^{i_0}} + \frac{1}{3}\cdot\frac{1}{4^k} - \frac{1}{3}\cdot\frac{1}{4^{i_0}} \ . \qquad (40)
\end{aligned}$$

Using the fact that $i_0 < k$ we also get

$$\frac{1}{3}\cdot\frac{1}{4^k}-\frac{1}{3}\cdot\frac{1}{4^{i_0}}<\frac{1}{3}\cdot\frac{1}{4^{i_0}}-\frac{1}{3}\cdot\frac{1}{4^k}=\sum_{i=i_0+1}^{k}\frac{1}{4^i}\leq\sum_{i=i_0+1}^{k}\frac{\sigma'_i+1}{4^i}\ . \tag{41}$$

Plugging Eq. (41) into Eq. (40) reveals than for any $x \in I_{\sigma_1,\ldots,\sigma_k}$:

$$x<\sum_{i=1}^{i_0-1}\frac{\sigma'_i+1}{4^i}+\frac{2}{4^{i_0}}+\sum_{i=i_0+1}^{k}\frac{\sigma'_i+1}{4^i}=\sum_{i=1}^{k}\frac{\sigma'_i+1}{4^i}\ .$$

Hence $x \notin I_{\sigma'_1,\ldots,\sigma'_k}$, which finishes the proof. ■

Lemma 28 *For any $N \geq 2$, any $1 \leq k < N$ and any local oracle $\mathbb{O}$, it holds that $\mathbb{O}_{h^N_{\sigma_1,\ldots,\sigma_N}}(x)$ does not depend on $\sigma_{k+1},\ldots,\sigma_N$ for $x \notin I_{\sigma_1,\ldots,\sigma_k}$.*

Proof First, notice that since $\mathbb{R} \setminus I_{\sigma_1,\ldots,\sigma_k}$ is an open set it is enough to prove that the function value $h^N_{\sigma_1,\ldots,\sigma_N}(x)$ does not depend on $\sigma_{k+1},\ldots,\sigma_N$ for $x \notin I_{\sigma_1,\ldots,\sigma_k}$, which implies the desired claim about $\mathbb{O}_{h^N_{\sigma_1,\ldots,\sigma_N}}(x)$ by definition of a local oracle. We will prove the claim for all natural pairs (N,k) such that $k < N$, using the following inductive argument:

- For any $N \geq 2$, the claim holds for $(N,1)$.
- If the claim holds for $(N-1,k-1)$ then it holds for (N,k).

Combining both bullets proves the claim for any pair (N,k), through the chain of implications

$$(N-k+1,1)\implies(N-k+2,2)\implies\cdots\implies(N,k)\ .$$

For the first bullet, fix any $N \geq 2$. We need to prove that $h^N_{\sigma_1,\ldots,\sigma_N}(x)$ does not depend on $\sigma_2,\ldots,\sigma_N$ for $x \notin \left[\frac{\sigma_1+1}{4},\frac{\sigma_1+1}{4}+\frac{1}{4}\right]$. Indeed, if $\sigma_1 = 0$ then by construction $h^N_{0,\ldots,\sigma_N}(x)$ does not depend on $\sigma_2,\ldots,\sigma_N$ for $x \notin \left[\frac{1}{4},\frac{1}{2}\right]$. Similarly, if $\sigma_1 = 1$ then by construction $h^N_{1,\ldots,\sigma_N}(x)$ does not depend on $\sigma_2,\ldots,\sigma_N$ for $x \notin \left[\frac{1}{2},\frac{3}{4}\right]$.

For the second bullet, assume the claim is true for some pair $(N-1,k-1)$. By renaming the variables $\sigma_i \leftarrow \sigma_{i+1}$, the induction hypothesis states that $h^{N-1}_{\sigma_2,\ldots,\sigma_N}(x)$ does not depend on $\sigma_{k+1},\ldots,\sigma_N$ for $x \notin [\sum_{i=1}^{k-1}\frac{\sigma_{i+1}+1}{4^i},\ \sum_{i=1}^{k-1}\frac{\sigma_{i+1}+1}{4^i}+\frac{1}{4^{k-1}}]$. Thus, $h^{N-1}_{\sigma_2,\ldots,\sigma_N}(4x-1-\sigma_1)$ does not depend on $\sigma_{k+1},\ldots,\sigma_N$ for

$$\begin{aligned}(4x-1-\sigma_1)&\notin\left[\sum_{i=1}^{k-1}\frac{\sigma_{i+1}+1}{4^i},\sum_{i=1}^{k-1}\frac{\sigma_{i+1}+1}{4^i}+\frac{1}{4^{k-1}}\right]\\&\iff x\notin\left[\sum_{i=1}^{k}\frac{\sigma_i+1}{4^i},\sum_{i=1}^{k}\frac{\sigma_i+1}{4^i}+\frac{1}{4^k}\right]\ .\end{aligned}$$

Noticing that by definition, $h^N_{\sigma_1,\sigma_2,\ldots,\sigma_N}(x)$ depends on $\sigma_2,\ldots,\sigma_N$ only when $(4x-1-\sigma_i)$ is fed to $h^{N-1}_{\sigma_2,\ldots,\sigma_N}$ finishes the proof. ■

From now on we fix some N we will specify later, and abbreviate $h_\sigma = h^N_\sigma$. Let $\sigma \sim \mathrm{Unif}(\{0,1\}^N)$, and consider the random sequence $x_1, x_2, \ldots$ produced by $\mathcal{A}$ when applied to h_σ (where the randomness comes from the random choice of σ).

Lemma 29 *For any $t \in \mathbb{N}$, $1 \le l < k \le N$: $\Pr_\sigma\left[x_{t+1} \in I_{\sigma_1,\ldots,\sigma_l,\ldots,\sigma_k} | x_1, \ldots, x_t \notin I_{\sigma_1,\ldots,\sigma_l}\right] \le \frac{1}{2^{k-l-1}}$.*

Proof If $x_1, \ldots, x_t \notin I_{\sigma_1,\ldots,\sigma_l}$, then Lemma 28 tells us that $\mathbb{O}_{h_\sigma}(x_1), \ldots, \mathbb{O}_{h_\sigma}(x_t)$ do not depend on $\sigma_{l+1}, \ldots, \sigma_N$. Since x_{t+1} is some deterministic function of $\mathbb{O}_{h_\sigma}(x_1), \ldots, \mathbb{O}_{h_\sigma}(x_t)$, which is the information that $\mathcal{A}$ obtains along it's first t iterations, we can deduce that in this case

1. Conditioning on past information, $\sigma_{l+1}, \ldots, \sigma_N \sim \mathrm{Unif}(\{0,1\})$. In particular we get $(\sigma_{l+1}, \ldots, \sigma_{k-1}) \sim \mathrm{Unif}(\{0,1\}^{k-l-1})$.
2. x_{t+1} is chosen independently of $\sigma_{l+1}, \ldots, \sigma_{k-1}$.

Note that by Lemma 27, $I_{\sigma_1,\ldots,\sigma_l,\sigma_{l+1},\ldots,\sigma_{k-1},\sigma_k}, I_{\sigma_1,\ldots,\sigma_l,\sigma'_{l+1},\ldots,\sigma'_{k-1},\sigma'_k}$ are disjoint for any $(\sigma_{l+1}, \ldots, \sigma_{k-1}) \ne (\sigma'_{l+1}, \ldots, \sigma'_{k-1})$, thus the events $x \in I_{\sigma_1,\ldots,\sigma_l,\sigma_{l+1},\ldots,\sigma_{k-1},\sigma_k}$ and $x \in I_{\sigma_1,\ldots,\sigma_l,\sigma'_{l+1},\ldots,\sigma'_{k-1},\sigma'_k}$ are disjoint. Since there are $2^{(k-1)-(l+1)+1} = 2^{k-l-1}$ such binary vectors, we obtain 2^{k-l-1} disjoint events with equal probabilities. Thus, their probability is at most $\frac{1}{2^{k-l-1}}$, which proves the claim. ■

Recall that $I_{\sigma_1} \supset I_{\sigma_1,\sigma_2} \supset \cdots \supset I_{\sigma_1,\ldots,\sigma_N}$ by Lemma 26. Combined with the previous lemma, we get

$$
\begin{aligned}
&\Pr_\sigma\left[\exists t \in [T_0] : x_t = x^*\right] \le \Pr_\sigma\left[\exists t \in [T_0] : x_t \in I_{\sigma_1,\ldots,\sigma_N}\right] \\
&\le \sum_{t=1}^{T_0} \Pr_\sigma\left[x_t \in I_{\sigma_1,\ldots,\sigma_N}\right] \\
&\le \sum_{t=1}^{T_0} \Pr_\sigma\left[\exists s \in [t], l \in [N] : x_1, \ldots, x_s \notin I_{\sigma_1,\ldots,\sigma_l} \wedge x_{s+1} \in I_{\sigma_1,\ldots,\sigma_{l+\frac{N}{t}}}\right] \\
&\le \sum_{t=1}^{T_0} \sum_{s=1}^{t} \sum_{l=1}^{N} \Pr_\sigma\left[x_1, \ldots, x_s \notin I_{\sigma_1,\ldots,\sigma_l} \wedge x_{s+1} \in I_{\sigma_1,\ldots,\sigma_{l+\frac{N}{t}}}\right] \\
&= \sum_{t=1}^{T_0} \sum_{s=1}^{t} \sum_{l=1}^{N} \Pr_\sigma\left[x_{s+1} \in I_{\sigma_1,\ldots,\sigma_{l+\frac{N}{t}}} \middle| x_1, \ldots, x_s \notin I_{\sigma_1,\ldots,\sigma_l}\right] \cdot \Pr_\sigma\left[x_1, \ldots, x_s \notin I_{\sigma_1,\ldots,\sigma_l}\right] \\
&\le \sum_{t=1}^{T_0} \sum_{s=1}^{t} \sum_{l=1}^{N} \frac{1}{2^{\frac{N}{t}-1}} \cdot 1 = \sum_{t=1}^{T_0} \frac{Nt}{2^{\frac{N}{t}-1}} \le T_0 \cdot \frac{NT_0}{2^{\frac{N}{T_0}-1}} \quad .
\end{aligned}
$$

Since $\frac{N(T_0)^2}{2^{\frac{N}{T_0}-1}} \overset{N\to\infty}{\longrightarrow} 0$, there exists some finite N (which depends on T_0) such that $\frac{NT_0^2}{2^{\frac{N}{T_0}-1}} \le \frac{1}{4}$. With this N, we get

$$\Pr_\sigma\left[\exists t \in [T_0] : x_t = x^*\right] \le \frac{1}{4} < \frac{1}{2} \ ,$$

proving Eq. (38) and finishing the proof.

Appendix D. Technical lemmas

Lemma 30 *Denote by $f(\cdot)$ the L_0-Lipschitz function $\mathbf{x} \mapsto L_0|x_1|$. Assume $\tilde{f}(\cdot)$ has L-Lipschitz gradients, and satisfies $\left\|f - \tilde{f}\right\|_\infty \leq \epsilon$. Then $L \geq \frac{L_0}{8\epsilon}$.*

Proof Due to rescaling we can assume without loss of generality that $L_0 = 1$. Denoting by $\mathbf{e}_1$ the first standard basis vector, we have

$$\begin{aligned}\tilde{f}\left(-4\epsilon \cdot \mathbf{e}_1\right) &\geq f\left(-4\epsilon \cdot \mathbf{e}_1\right) - \epsilon = 4\epsilon - \epsilon = 3\epsilon \ , \\ \tilde{f}\left(4\epsilon \cdot \mathbf{e}_1\right) &\geq f\left(4\epsilon \cdot \mathbf{e}_1\right) - \epsilon = 4\epsilon - \epsilon = 3\epsilon \ , \\ \tilde{f}\left(\mathbf{0}\right) &\leq f\left(\mathbf{0}\right) + \epsilon = \epsilon \ .\end{aligned}$$

By the mean value theorem, there exist $-4\epsilon < t_0 < 0, \ 0 < t_1 < 4\epsilon$ such that

$$\begin{aligned}\frac{\partial}{\partial x_1}\tilde{f}\left(t_0\right) &= \frac{\tilde{f}\left(\mathbf{0}\right) - \tilde{f}\left(-4\epsilon \cdot \mathbf{e}_1\right)}{4\epsilon} \leq \frac{\epsilon - 3\epsilon}{4\epsilon} = -\frac{1}{2} \ , \\ \frac{\partial}{\partial x_1}\tilde{f}\left(t_1\right) &= \frac{\tilde{f}\left(4\epsilon \cdot \mathbf{e}_1\right) - \tilde{f}\left(\mathbf{0}\right)}{4\epsilon} \geq \frac{3\epsilon - \epsilon}{4\epsilon} = \frac{1}{2} \ .\end{aligned}$$

So by Cauchy-Schwarz and L-smoothness of $\tilde{f}$:

$$\begin{aligned}1 &= \left|\frac{\partial}{\partial x_1}\tilde{f}\left(t_1\right) - \frac{\partial}{\partial x_1}\tilde{f}\left(t_0\right)\right| = \left|\left\langle \nabla\tilde{f}\left(t_1 \cdot \mathbf{e}_1\right) - \nabla\tilde{f}\left(t_0 \cdot \mathbf{e}_1\right), \mathbf{e}_1 \right\rangle\right| \\ &\leq \left\|\nabla\tilde{f}\left(t_1 \cdot \mathbf{e}_1\right) - \nabla\tilde{f}\left(t_0 \cdot \mathbf{e}_1\right)\right\| \leq L\left|t_1 - t_0\right| \leq L \cdot 8\epsilon \ .\end{aligned}$$

■

Lemma 31 *If $\tilde{f}(\cdot)$ has L-Lipschitz gradients and satisfies $\|f - \tilde{f}\|_\infty \leq \epsilon$ for some 1-Lipschitz function $f(\cdot)$, then for all $\mathbf{x} \in \mathbb{R}^d$: $\|\nabla\tilde{f}(\mathbf{x})\| \leq 1 + 2\epsilon + \frac{L}{2}$.*

Proof Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Denote $\gamma(t) := (1-t)\cdot\mathbf{x} + t\cdot\mathbf{y}$, and notice that

$$\begin{aligned}\tilde{f}\left(\mathbf{y}\right) - \tilde{f}\left(\mathbf{x}\right) &= \tilde{f}\left(\gamma\left(1\right)\right) - \tilde{f}\left(\gamma\left(0\right)\right) = \int_0^1 \left(\tilde{f}\circ\gamma\right)'\left(t\right)dt = \int_0^1 \left\langle \nabla\tilde{f}\left(\gamma\left(t\right)\right), \gamma'\left(t\right)\right\rangle dt \\ &= \int_0^1 \left\langle \nabla\tilde{f}\left(\gamma\left(t\right)\right), \mathbf{y} - \mathbf{x}\right\rangle dt \ . \end{aligned} \tag{42}$$

Combining Cauchy-Schwarz with the fact that $\nabla\tilde{f}$ is L-Lipschitz, we get

$$\begin{aligned}&\left\langle \nabla\tilde{f}\left(\mathbf{x}\right) - \nabla\tilde{f}\left(\gamma\left(t\right)\right), \mathbf{y} - \mathbf{x}\right\rangle \leq \left\|\nabla\tilde{f}\left(\mathbf{x}\right) - \nabla\tilde{f}\left(\gamma\left(t\right)\right)\right\| \cdot \left\|\mathbf{y} - \mathbf{x}\right\| \leq L\left\|\mathbf{x} - \gamma\left(t\right)\right\| \cdot \left\|\mathbf{y} - \mathbf{x}\right\| \\ &\implies \left\langle \nabla\tilde{f}\left(\gamma\left(t\right)\right), \mathbf{y} - \mathbf{x}\right\rangle \geq \left\langle \nabla\tilde{f}\left(\mathbf{x}\right), \mathbf{y} - \mathbf{x}\right\rangle - L\left\|\gamma\left(t\right) - \mathbf{x}\right\| \cdot \left\|\mathbf{y} - \mathbf{x}\right\| \ .\end{aligned}$$

Plugging this into Eq. (42) gives

$$\begin{aligned}
\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) &\geq \int_0^1 \left(\left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle - L \|\gamma(t) - \mathbf{x}\| \cdot \|\mathbf{y} - \mathbf{x}\| \right) dt \\
&= \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle - L \|\mathbf{y} - \mathbf{x}\| \cdot \int_0^1 \|\gamma(t) - \mathbf{x}\| \, dt \\
&= \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle - L \|\mathbf{y} - \mathbf{x}\| \cdot \left[\frac{1}{2} \|\gamma(1) - \mathbf{x}\|^2 - \frac{1}{2} \|\gamma(0) - \mathbf{x}\|^2 \right] \\
&= \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle - \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^3 \\
\implies \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle &\leq \tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^3 \ .
\end{aligned}$$

We assume $\|\nabla \tilde{f}(\mathbf{x})\| \neq 0$ since otherwise the desired claim is trivial. In particular, if $\mathbf{y} = \mathbf{x} + \frac{\nabla \tilde{f}(\mathbf{x})}{\|\nabla \tilde{f}(\mathbf{x})\|}$ then $\|\mathbf{y} - \mathbf{x}\| = 1$ and inequality above reveals

$$\left\| \nabla \tilde{f}(\mathbf{x}) \right\| \leq \tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{x}) + \frac{L}{2} \leq f(\mathbf{y}) - f(\mathbf{x}) + 2\epsilon + \frac{L}{2} \leq \|\mathbf{y} - \mathbf{x}\| + 2\epsilon + \frac{L}{2} = 1 + 2\epsilon + \frac{L}{2} \ ,$$

where we used the fact that $\|\tilde{f} - f\|_\infty \leq \epsilon$, and that f is 1-Lipschitz. ■

References

Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. *Advances in Neural Information Processing Systems*, 29, 2016.

Hédy Attouch and Dominique Aze. Approximation and regularization of arbitrary functions in hilbert spaces by the lasry-lions method. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 10(3):289–312, 1993.

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Amir Beck and Nadav Hallak. On the convergence to stationary points of deterministic and randomized feasible descent directions methods. *SIAM Journal on Optimization*, 30(1): 56–79, 2020.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Andrew Blake and Andrew Zisserman. *Visual reconstruction*. MIT press, 1987.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2019.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

Xiaojun Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming*, 134(1):71–99, 2012.

Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, pages 1–36, 2018.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for general lipschitz functions in high and low dimensions. *arXiv preprint arXiv:2112.06969*, 2022.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International conference on machine learning*, pages 1833–1841. PMLR, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

Krzysztof C Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

Jean-Michel Lasry and Pierre-Louis Lions. A remark on regularization in hilbert spaces. *Israel Journal of Mathematics*, 55(3):257–266, 1986.

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Hossein Mobahi and John W Fisher III. A theoretical analysis of optimization by gaussian continuation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

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

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

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

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

R. Tyrrell Rockafellar and Roger J.B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

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 *Proceedings of the 39th International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

Zhijun Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3):748–768, 1996.

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.

Chao Zhang and Xiaojun Chen. Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIAM Journal on Optimization*, 20(2): 627–649, 2009.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonsmooth nonconvex functions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11173–11182, 2020.