OPT2024: 16th Annual Workshop on Optimization for Machine Learning

On the Hardness of Meaningful Local Guarantees in Nonsmooth Nonconvex Optimization

Guy Kornowski GUY.KORNOWSKI@WEIZMANN.AC.IL
Weizmann Institute of Science

Swati Padmanabhan PSWT@MIT.EDU
Massachusetts Institute of Technology

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Abstract

We study the oracle complexity of nonsmooth nonconvex optimization, with the algorithm allowed access only to local function information. It has been shown by Davis, Drusvyatskiy, and Jiang (2023) that for nonsmooth Lipschitz functions satisfying certain regularity and strictness conditions, perturbed gradient descent converges to local minimizers *asymptotically*. Motivated by this and other recent algorithmic advances in nonconvex nonsmooth optimization, we consider the question of obtaining a non-asymptotic rate of convergence to local minima for this problem class.

We provide the following negative answer to this question: Local algorithms acting on regular Lipschitz functions *cannot*, in the worst case, provide meaningful local guarantees in terms of function value in sub-exponential time, even when all near-stationary points are global minima. This sharply contrasts with the smooth setting, for which standard gradient methods are known to do so at a dimension-free rate. Our result complements the rich body of work in the theoretical computer science literature that provide hardness results conditional on conjectures such as $\mathsf{P} \neq \mathsf{NP}$ or cryptographic assumptions, in that ours holds unconditional of any such assumptions.

1. Introduction

Nonconvex optimization problems are ubiquitous throughout the computational and applied sciences. Since globally optimizing nonconvex objectives is infeasible in general, optimization theory has long pursued iterative algorithms that find solutions satisfying some *local* optimality guarantees. For example, given a sufficiently smooth objective $f : \mathbb{R}^d \to \mathbb{R}$, a folklore result asserts that gradient descent, when applied with a suitable step-size, converges to a stationary point at a dimension-independent rate. Similarly, the perturbed gradient descent update

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) + \xi_t, \tag{1.1}$$

where ξ_t is a mean-zero random variable, is known [28] to asymptotically converge only to local minimizers of f for suitable choices of $(\eta_t, \xi_t)_{t\in\mathbb{N}}$. Moreover, under a *"strictness"* assumption [23] stating, roughly, that critical points are either sufficiently negatively curved or local minimizers, this convergence is known to have a favorable polynomial rate nearly independent of the dimension [27].

Even though the state of affairs is relatively well understood for smooth objectives, many modern applications in machine learning, operations research, and statistics (e.g., deep learning with

1. Full version: https://arxiv.org/abs/2409.10323. GK and SP contributed equally.

ReLU activations [34], piecewise affine regression [7]) require solving nonconvex problems that are also inherently *nonsmooth*. This structure presents several important challenges: As a prominent example, it calls into question the correctness of automatic differentiation with PyTorch and TensorFlow [30]. Aiming at this regime, Davis and Drusvyatskiy [12] studied (nonsmooth) weakly-convex functions,[1] showing that strictness enables proximal methods to asymptotically converge to local minimizers. Subsequently, this convergence was shown to have a polynomial rate [15, 26], thus extending to a nonsmooth setting what was previously known only for smooth objectives.

However, many prominent contemporary optimization problems such as neural network training fall outside this class of weakly convex objectives.[2] In this regard, Davis et al. [17] proved a vast generalization of the previously mentioned results, asserting that for nonsmooth Lipschitz functions satisfying mild regularity properties and a strictness assumption, the dynamics dictated by (1.1) asymptotically converge only to local minimizers. Another asymptotic result of similar spirit was also obtained in [4]. These developments motivate the following natural question:

Is it possible to obtain non-asymptotic convergence guarantees to local minima, when optimizing sufficiently regular Lipschitz objectives that satisfy a strictness property?

A priori, recent advances in nonsmooth nonconvex optimization suggest that there is room for optimism for such finite-time guarantees. Following the work of Zhang et al. [50], a surge of recent results showed that it is possible to converge, at a dimension-free polynomial rate, to approximate-stationary points in the sense of Goldstein [24] when optimizing Lipschitz functions [16, 49]. This remains an active area of research, with recent works obtaining finite-time guarantees for optimizing Lipschitz functions in terms of Goldstein-stationarity under a variety of settings such as stochastic [8], constrained [25], and zeroth-order optimization [33, 35].

Nevertheless, as our main result (Theorem 2.1) will soon show, obtaining *any* non-trivial convergence rate in terms of local function decrease is impossible even for strict functions. In particular, we prove that even under suitable regularity assumptions and the non-existence of non-strict saddles, any algorithm whatsoever based on local queries will necessarily get stuck, in the worst case, at points at which there is significant local decrease, unless the number of iterations grows exponentially with the dimension. In fact, this statement holds even under the supposedly easier case in which all approximate-stationary points below some constant (sub)gradient norm are in fact global minima – which trivially precludes the existence of non-strict saddles.

1.1. Related Work

There has been a long line of work on developing efficient algorithms for various classes of nonconvex programs, some of which we discuss in Appendix A. Here, we focus on lower bounds. There has been extensive effort providing hardness results on reaching different solution concepts in nonsmooth nonconvex optimization. The computational intractability of globally minimizing a Lipschitz function up to a small constant tolerance was known since the works of Nemirovski and Yudin [37] and Murty and Kabadi [36]. More recently, Zhang et al [50] showed that local, first-order algorithms acting on nonsmooth nonconvex functions cannot attain either small function error or small gradients: Indeed, approximately-stationary points can be easily "hidden" inside some arbitrarily

1. A function f is called weakly-convex if there exists $\rho > 0$ such that $\mathbf{x} \mapsto f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}\|^2$ is convex.
2. Indeed, this is the case even for a single negated ReLU neuron, namely $z \mapsto -\max\{0, z\}$.

small neighbourhood, which cannot in general be found in a finite number of iterations. Subsequently, Kornowski and Shamir [32] considered if, for general Lipschitz functions, instead of trying to find approximately-stationary points, the more relaxed notion of getting *near* an approximately-stationary point is more tractable. Via the construction of a novel hard function, [32] answer this question negatively (for both deterministic and randomized algorithms). This hardness result was subsequently adapted by Tian and So [47], for the case of deterministic algorithms, endowing it with Clarke regularity by employing a Huber-type smoothing. Next, Jordan et al. [29] provide an improved understanding of achieving Goldstein stationarity by showing that no deterministic algorithm can achieve a dimension-free rate of convergence. The work of Tian and So [48] extend this to deterministic general zero-respecting algorithms for achieving Goldstein stationarity.

Our work adds to this line of hardness results: For oracle-based algorithms seeking to achieve local optimality in Lipschitz functions, we prove a lower bound that is exponential in dimension, even when the algorithm is allowed to employ randomization, the function is Clarke regular, and all near-stationary points of the function are in fact global minima.

In a somewhat different direction, it is interesting to compare our result to a rich body of work in the theoretical computer science literature. Some of the earliest such works include those of Sahni [45, 46], which showed that global optimization of a general quadratic program is NP-hard, and those by Murty and Kabadi [36], Pardalos and Schnitger [40], which showed that it is NP-hard to test whether a given point is a local minimizer for constrained nonconvex quadratic programming. The recent work of Ahmadi and Zhang [1] shows that unless $\mathsf{P} = \mathsf{NP}$, there cannot be a polynomial-time algorithm that finds a point within Euclidean distance c^n (for any constant $c \geq 0$) of a local minimizer of an n-variate quadratic function over a polytope. Additionally, they show that the problem of deciding whether a quadratic function has a local minimizer over an (unbounded) polyhedron, and that of deciding if a quartic polynomial has a local minimizer are NP-hard. In the process, [1] answers a question posed by Pardalos and Vavasis [41]. Another open problem listed by [41] was recently settled by Fearnley et al. [22], who showed that the problem remains hard even if we are searching only for a Karush-Kuhn-Tucker (KKT) point. In particular, they show that the quadratic-KKT problem is CLS-complete (a problem class introduced by [10]), which, by another result of [21], is unlikely to have polynomial-time algorithms.

The most important difference of these results from ours is that while the above lower bounds rely on conditional hardness assumptions from complexity theory (such as $\mathsf{P} \neq \mathsf{NP}$), our framework of oracle complexity, which reduces optimization to information theoretic notions, enables proving lower bounds that are entirely unconditional. Furthermore, many of these results are stated in terms of hardness of verification, which, in general, does not imply hardness of search — namely, finding points of interest, as opposed to verifying that a given point is such. Finally, verification complexity results typically focus on non-Lipschitz polynomials or other function classes, which, as is, do not directly correspond to known complexity upper bounds discussed in our introduction.

1.2. Preliminaries

Nonsmooth Analysis. We say a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is L-Lipschitz if for any $\mathbf{x}, \mathbf{y}$, we have $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$. Recall that by Rademacher's theorem [42], Lipschitz functions are differentiable almost everywhere (in the sense of Lebesgue). Throughout our paper, we will be working with L-Lipschitz functions for some L (we specify our exact set of assumptions in Section 2). We therefore next collect some relevant quantities associated with Lipschitz functions.

Definition 1.1 *For any Lipschitz function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the (ordinary) directional derivative of f at a point $\mathbf{x}$ in the direction $\mathbf{v}$ is defined as* $f'(\mathbf{x};\mathbf{v}) := \lim_{t\to 0^+} \frac{f(\mathbf{x}+t\mathbf{v})-f(\mathbf{x})}{t}$.

Definition 1.2 ([5, 43]) *For any Lipschitz function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the generalized directional derivative of f at a point $\mathbf{x}$ in the direction $\mathbf{v}$ is defined as* $f^\circ(\mathbf{x};\mathbf{v}) := \limsup_{\mathbf{y}\to\mathbf{x},\, t\to 0^+} \frac{f(\mathbf{y}+t\mathbf{v})-f(\mathbf{y})}{t}$.

The generalized directional derivative leads to the following definition of the Clarke subdifferential.

Definition 1.3 (Clarke [5, 6]) *For any Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and point $\mathbf{x} \in \mathbb{R}^d$, the Clarke subdifferential of f at $\mathbf{x}$ is defined as $\partial f(\mathbf{x}) := \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{v}\rangle \le f^\circ(\mathbf{x};\mathbf{v}), \forall \mathbf{v} \in \mathbb{R}^d\}$, with each element $\mathbf{g}$ of this set termed a Clarke subgradient.*

Equivalently, $\partial f(\mathbf{x}) := \mathrm{conv}\{\mathbf{g} : \mathbf{g} = \lim_{n\to\infty} \nabla f(\mathbf{x}_n),\ \mathbf{x}_n \to \mathbf{x}\}$, namely, the Clarke subdifferential is the convex hull of all limit points of $\nabla f(\mathbf{x}_n)$ over all sequences of differentiable points $\mathbf{x}_n$ which converge to $\mathbf{x}$. If the function is continuously differentiable at a point or convex, the Clarke subgradient there reduces to the gradient or subgradient in the convex analytic sense, respectively. Equipped with the notation of the Clarke subdifferential, one may define a Clarke regular function.

Definition 1.4 ([5]; [6, Definition 2.3.4]) *A locally Lipschitz function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is Clarke regular at $\mathbf{x} \in \mathbb{R}^d$ if for every direction $\mathbf{v} \in \mathbb{R}^d$, the ordinary directional derivative $f'(\mathbf{x};\mathbf{v})$ exists and $f'(\mathbf{x};\mathbf{v}) = f^\circ(\mathbf{x};\mathbf{v})$. The function $f : \mathbb{R}^d \to \mathbb{R}$ is regular if it is Clarke regular at all $\mathbf{x} \in \mathbb{R}^d$.*

Finally, $\bar{\partial} f(\mathbf{x}) := \arg\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial f(\mathbf{x})\}$, denotes the minimal norm subgradient at a point $\mathbf{x}$, and we say that $\mathbf{x}$ is an ϵ-stationary point of $f(\cdot)$ if $\left\|\bar{\partial} f(\mathbf{x})\right\| \le \epsilon$.

Local algorithms. We consider iterative algorithms that have access to local information at queried points and proceed based on information gathered along these queries [37]. Formally, we call an oracle *local* if for any point $\mathbf{x}$ and any two functions f, g that are equal over some neighborhood of $\mathbf{x}$, the equation $\mathbb{O}_f(\mathbf{x}) = \mathbb{O}_g(\mathbf{x})$ holds. At every iteration $t \in \mathbb{N}$, a local algorithm which aims to optimize an unknown objective $f : \mathbb{R}^d \to \mathbb{R}$ chooses an iterate $\mathbf{x}_t \in \mathbb{R}^d$, receives the local information $\mathbb{O}_f(\mathbf{x}_t)$, and proceeds to choose the next iterate $\mathbf{x}_{t+1}$ as some (possibly random) mapping of all the oracle outputs seen thus far: $(\mathbb{O}_f(\mathbf{x}_1), \ldots, \mathbb{O}_f(\mathbf{x}_t))$. An important subclass of local algorithms are first-order algorithms, which utilize an oracle of the form $\mathbb{O}_f(\mathbf{x}) = (f(\mathbf{x}), \mathbf{g}_x)$ where $\mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x})$ is some consistent choice of a subgradient.

Organization. Due to the limit on pages, we defer all our proofs to the appendix.

2. Our Main Result

We now present our main theorem. Put simply, it states that local algorithms acting on regular Lipschitz functions cannot, in the worst case, guarantee meaningful local guarantees in terms of function value in sub-exponential time, even when all near-stationary points are global minima.

For comparison, recall that for smooth objectives $f : \mathbb{R}^d \to \mathbb{R}$ and sufficiently small $\delta > 0$, gradient descent gets after T steps to a point $\mathbf{x}_T \in \mathbb{R}^d$ such that $f(\mathbf{x}_T) \le \min_{\mathbf{z}\in B(\mathbf{x}_T,\delta)} f(\mathbf{z}) + O(\frac{\delta}{\sqrt{T}})$, namely, a point locally competitive with nearby function values up to a factor which vanishes in a dimension-free manner. Our main result precludes precisely that when smoothness is absent.

Theorem 2.1 *For any (possibly randomized) local algorithm $\mathcal{A}$ and any $T, d \in \mathbb{N}$, there is a function $f : \mathbb{R}^d \to \mathbb{R}$ so that for some absolute constant $c \geq \frac{1}{100}$, the following properties hold:*

(i) The function f is 1-Lipschitz and Clarke regular, $f(\mathbf{0}) - \inf f \leq 2$, and all c-stationary points of f are global minima.

(ii) With probability at least $1 - 2T\exp(-d/36)$, for all $t \in [T]$ and any $\delta \in (0, 1]$, the following inequality holds:

$$\min_{\mathbf{z} \in B(\mathbf{x}_t, \delta)} f(\mathbf{z}) < f(\mathbf{x}_t) - \delta c\,. \tag{2.1}$$

To contextualize Theorem 2.1, we further note that under the stated 1-Lipschitzness condition, for any $\delta > 0$, the local decrease $f(\mathbf{x}_t) - \min_{\mathbf{z} \in B(\mathbf{x}_t, \delta)} f(\mathbf{z})$ can be *at most* δ. Thus, the theorem shows that unless $T \gtrsim \exp(\Omega(d))$, all iterates $\mathbf{x}_t$ suffer from a nearly-maximal local decrease; as a result, none of these iterates can be regarded as approximate local minima. On the other hand, it is clearly the case that with exponential dimension dependence, a trivial grid search algorithm can guarantee getting anywhere (i.e. over a discretization of a bounded domain), and in particular can achieve approximate local optimality somewhere along the algorithm's trajectory. Hence Theorem 2.1 can be seen as asserting that nothing substantially better than a trivial strategy is possible.

The proof of Theorem 2.1 consists of constructing a variant of the function which was previously used to prove a strong lower bound on the complexity of getting near stationary points of Lipschitz functions [32]. Our analysis further reveals that our constructed function satisfies that all near-stationary points are in fact global minima. The prior construction by [32] does not apply to Clarke regular functions, which is an important consideration for our purposes in two aspects. First, for the sake of interest of the derived result, the upper bounds in the context of local minimality, as discussed throughout the introduction, crucially rely on this property.[3] Second, Clarke regularity implies the so called "Lyapunov property", asserting that the subgradient flow decreases the function value proportionally to the subgradient norm (see [9] for an elaborate discussion on this property and function classes for which it holds). Therefore, having established a Clarke regular function for which the algorithms' iterates are nowhere near a point with sub-constant subgradient norm, the Lyapunov property further ensures that by tracking the subgradient flow, the local decrease in function value is significant, hence implying our desired lower bound in terms of function value.

References

[1] Amir Ali Ahmadi and Jeffrey Zhang. On the complexity of finding a local minimizer of a quadratic function over a polytope. *Mathematical Programming*, 195(1):783–792, 2022.

[2] Keith Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31 (1-58):26, 1997.

[3] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1), 2005.

3. We remark that Tian and So [47] provide a Clarke regular variant of this construction, but their construction implies a lower bound only for *deterministic* algorithms — as a result, this fails to address the aforementioned algorithms which are based on the perturbed gradient descent dynamics (1.1).

[4] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research*, 2023.

[5] Frank H Clarke. Generalized gradients of lipschitz functionals. *Advances in Mathematics*, 40 (1):52–67, 1981.

[6] Frank H Clarke. *Optimization and nonsmooth analysis*. Siam, 1990.

[7] Ying Cui and Jong-Shi Pang. *Modern nonconvex nondifferentiable optimization*. SIAM, 2021.

[8] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670. PMLR, 2023.

[9] Aris Daniilidis and Dmitriy Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020.

[10] Constantinos Daskalakis and Christos Papadimitriou. Continuous local search. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 790–804. SIAM, 2011.

[11] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1), 2019.

[12] Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22(2):561–606, 2022.

[13] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[14] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[15] Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the moreau envelope in nonsmooth optimization. *SIAM Journal on Optimization*, 32(3):1958–1983, 2022.

[16] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

[17] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. *arXiv preprint arXiv:2108.11832*, 2023.

[18] Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

[19] John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4), 2018.

[20] Yu M Ermol'ev and VI Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, 1998.

[21] John Fearnley, Paul Goldberg, Alexandros Hollender, and Rahul Savani. The complexity of gradient descent: Cls= ppad ∩ pls. *Journal of the ACM*, 70(1), 2022.

[22] John Fearnley, Paul W Goldberg, Alexandros Hollender, and Rahul Savani. The complexity of computing kkt solutions of quadratic programs. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 892–903, 2024.

[23] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

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

[25] Benjamin Grimmer and Zhichao Jia. Goldstein stationarity in lipschitz constrained optimization. *arXiv preprint arXiv:2310.03690*, 2023.

[26] Minhui Huang. Escaping saddle points for nonsmooth weakly convex functions via perturbed proximal algorithms. *arXiv preprint arXiv:2102.02837*, 2021.

[27] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[28] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[29] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.

[30] Sham M Kakade and Jason D Lee. Provably correct automatic sub-differentiation for qualified programs. *Advances in neural information processing systems*, 31, 2018.

[31] Siyu Kong and AS Lewis. The cost of nonconvexity in deterministic nonsmooth optimization. *Mathematics of Operations Research*, 2023.

[32] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *The Journal of Machine Learning Research*, 23(1):14161–14204, 2022.

[33] Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122):1–14, 2024.

[34] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[35] Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

[36] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39, 1987.

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

[38] VI Norkin. Stochastic generalized-differentiable functions in the problem of nonconvex nonsmooth stochastic optimization. *Cybernetics*, 22(6), 1986.

[39] EA Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, 1974.

[40] Panos M Pardalos and Georg Schnitger. Checking local optimality in constrained quadratic programming is np-hard. *Operations Research Letters*, 7(1), 1988.

[41] Panos M Pardalos and Stephen A Vavasis. Open questions in complexity theory for numerical optimization. *Mathematical programming*, 57(1):337–339, 1992.

[42] Hans Rademacher. Über partielle und totale differenzierbarkeit von funktionen mehrerer variabeln und über die transformation der doppelintegrale. *Mathematische Annalen*, 79(4):340–359, 1919.

[43] R Tyrrell Rockafellar. Generalized directional derivatives and subgradients of nonconvex functions. *Canadian Journal of Mathematics*, 32(2):257–280, 1980.

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

[45] Sartaj Sahni. Some related problems from network flows, game theory and integer programming. In *13th Annual Symposium on Switching and Automata Theory (swat 1972)*, pages 130–138. IEEE, 1972.

[46] Sartaj Sahni. Computationally related problems. *SIAM Journal on computing*, 3(4):262–279, 1974.

[47] Lai Tian and Anthony Man-Cho So. On the hardness of computing near-approximate stationary points of clarke regular nonsmooth nonconvex problems and certain dc programs. In *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.

[48] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, pages 1–24, 2024.

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

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

Appendix A. Notation and Preliminaries

We let $\mathbb{N} = \{1, 2, \dots, \}$ be the natural numbers starting from one. We let boldfaced letters (e.g., $\mathbf{x}$) denote vectors. We denote the d-dimensional Euclidean space by $\mathbb{R}^d$ and by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ its associated inner product and norm, respectively. We use $\mathbf{0}_d$ (or simply $\mathbf{0}$ when d is clear from context) to denote the zero vector in $\mathbb{R}^d$ and $\mathbf{e}_1, \mathbf{e}_2, \dots$ for the standard basis vectors. Given a vector $\mathbf{x}$, we denote by x_i its i-th coordinate, by $\mathbf{x}_{1:i} := (x_1, \dots, x_i)$ its truncation, and by $\bar{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ the normalized vector (assuming $\mathbf{x} \neq \mathbf{0}$). We use $B(\mathbf{x}, \delta)$ to denote the closed Euclidean ball of radius $\delta > 0$ centered at $\mathbf{x}$.

Fact A.1 ([6, Proposition 2.3.6]) *Let f be Lipschitz near x.*

(i) If f is continuously differentiable at x, then f is regular at x.
(ii) If f is convex, then f is regular at x.
(iii) A finite linear combination by nonnegative scalars of functions regular at x is regular at x.

Finally, we state the following important facts from subdifferential calculus that we use.

Fact A.2 (Proposition 2.3.3 and Theorem 2.3.9 [6]) *We have that $\partial(g_1 + g_2) \subseteq \partial g_1 + \partial g_2$, and if g_1 is univariate, then $\partial(g_1 \circ g_2)(\mathbf{x}) \subseteq \operatorname{conv}\{r_1 \mathbf{r}_2 : r_1 \in \partial g_1(g_2(\mathbf{x})), \mathbf{r}_2 \in \partial g_2(\mathbf{x})\}$.*

Fact A.3 ([44, Theorem 10.29]) *If F is regular, and g is regular at all $F(x)$, then $f(x) = g(F(x))$ is regular at all x.*

Related Work. There has been a long line of work in the optimization literature on studying convergence of first-order methods for various classes of nonconvex programs. Some early works include those of Benaïm et al. [3], Ermol'ev and Norkin [20], Norkin [38], Nurminskii [39]. More recently, Davis et al. [14] showed the first rigorous convergence guarantees for the stochastic subgradient method on Whitney stratifiable functions by building on techniques from Drusvyatskiy et al. [18], Duchi and Ruan [19]. For well-behaved problems with ρ-weakly convex objective functions, Davis and Drusvyatskiy [11], Davis and Grimmer [13] studied convergence to stationary points of the Moreau envelope and provide dimension-free convergence rates for finding these points. As alluded to earlier, the work of Zhang et al [50] provided the first finite-time dimension-free guarantees for converging to a Goldstein stationary point of a given Lipschitz function, a result that was subsequently strengthened to hold under standard first-order oracle access [16, 49]. Finally, by slightly restricting the class of nonsmooth objectives, Kong and Lewis [31] develop a simple de-randomized version of the algorithm of [50] with increased, but still dimension-free, complexity and quantify how the cost in complexity of optimizing nonsmooth objectives grows with their level of nonconvexity.

Appendix B. Proof of Our Main Result (Theorem 2.1)

In this section, we prove our main result (Theorem 2.1). We first state Proposition B.1 (deferring its proof to Appendix C.3), a technical result on one-dimensional functions that we crucially use for the construction of our hard instance for Theorem 2.1.

Proposition B.1 *For any $\gamma > 0$ and $T \in \mathbb{N}$, there exists $\rho > 0$ so that the following holds: For any (possibly randomized) local algorithm $\mathcal{A}$, there exists a function $\bar{h} : \mathbb{R} \to [2, \infty)$ such that*

(i) $\bar{h}$ *is* 1*-Lipschitz, convex, and satisfies* $\bar{h}(0) \le 3$.
(ii) $\bar{h}$ *has a unique minimizer* $x^* \in (0,1)$*, and* $\forall x \neq x^* : \left|\partial\bar{h}(x)\right| \ge \frac{1}{8}$.
(iii) $\Pr_{\mathcal{A}}[\exists t \in [T] : |x_t - x^*| \le \rho] < \gamma$.

The above proposition considers the class of Lipschitz, convex functions bounded from below and with a certain minimum slope at all points that are not the function minimizers. Then, for any randomized local algorithm, there exists a function in this function class, such that, with high probability, the said algorithm cannot reach its local minimum.

By embedding the one-dimensional hard instance of Proposition B.1 into higher dimensions, a simple reduction enables us to extend Proposition B.1 to functions beyond merely one dimension.

Lemma B.2 *For any* $\gamma > 0$, $T \in \mathbb{N}$, *and* $d \ge 2$, *there exists* $\rho > 0$ *(which depends only on* γ, T*) so that the following holds: For any (possibly randomized) local algorithm* $\mathcal{A}$*, there exists a function* $\bar{h} : \mathbb{R} \to [1, \infty)$ *satisfying the following properties.*

(i) $\bar{h}$ *is* $\frac{1}{2}$*-Lipschitz, convex, and satisfies* $\bar{h}(0) \le \frac{3}{2}$.
(ii) $\bar{h}$ *has a unique minimizer* $x^* \in (0,1)$*, and* $\forall x \neq x^* : \left|\partial\bar{h}(x)\right| \ge \frac{1}{16}$.
(iii) When applying $\mathcal{A}$ *to* $h(\mathbf{x}) := \frac{1}{32}\left\|\mathbf{x}_{1:d-1}\right\| + \bar{h}(x_d)$*, we have, for* $\mathbf{x}^* = (\mathbf{0}_{d-1}, x^*)$*, that*

$$\Pr_{\mathcal{A}}[\exists t \in [T] : \|\mathbf{x}_t - \mathbf{x}^*\| \le \rho] < \gamma\,.$$

Proof [Proof of Lemma B.2] Since our goal is to provide a lower bound when applying the local algorithm $\mathcal{A}$, we can assume without loss of generality that $\mathcal{A}$ has access to an even stronger oracle of the form

$$\overline{\mathbb{O}}(\mathbf{x}) = \left(\left\{h(\mathbf{z}_{1:d-1}, x_d) \mid \mathbf{z}_{1:d-1} \in \mathbb{R}^{d-1}\right\}, \mathbb{O}_{h_{1\mathrm{d}}}(x_d)\right),$$

where $h_{1\mathrm{d}}$ is a one-dimensional function we will soon choose. Note that oracle $\overline{\mathbb{O}}$ as defined here provides a full description of the function h over the affine subspace $\{\mathbf{z} \mid z_d = x_d\}$ in addition to the local information with respect to the last coordinate.[4] Moreover, given the algorithm $\mathcal{A}$ with such an oracle, one can simulate a local algorithm $\mathcal{A}'$ which optimizes the one-dimensional function $h_{1\mathrm{d}}$ by restricting to the d'th coordinates $((\mathbf{x}_t)_d)_{t\in\mathbb{N}}$ of the iterates $(\mathbf{x}_t)_{t\in\mathbb{N}}$.

We let $h_{1\mathrm{d}}$ be the one-dimensional function given by Proposition B.1 when applied to $\gamma, T, \mathcal{A}'$. With this choice of $h_{1\mathrm{d}}$, we let $\bar{h} = \frac{1}{2}h_{1\mathrm{d}}$. Then, Lemma B.2(i) and Lemma B.2(ii) are immediate by Proposition B.1(i) and Proposition B.1(ii). We also have $\bar{h} : \mathbb{R} \mapsto [1, \infty)$ from the range $[2, \infty)$ of $h_{1\mathrm{d}}$ from Proposition B.1. Finally, by combining the fact that $\mathcal{A}$ can simulate $\mathcal{A}'$ and using Proposition B.1(iii) for $\mathcal{A}'$, we have the following inequality, which establishes Lemma B.2(iii):

$$\Pr_{\mathcal{A}}[\exists t \in [T] : \|\mathbf{x}_t - \mathbf{x}^*\| \le \rho] \le \Pr_{\mathcal{A}'}[\exists t \in [T] : |(\mathbf{x}_t)_d - x^*| \le \rho] < \gamma\,.$$

■

For our subsequent proofs, we use the setup described next, in Definition B.3.

4. As an example, in the canonical case of a (sub)gradient oracle, $\overline{\mathbb{O}}$ provides the partial derivatives with respect to the first $d-1$ coordinates at all points, while revealing the partial derivative only with respect to the last coordinate at the queried point.

Definition B.3 *We let $\mathcal{A}$ be a local algorithm, $T, d \geq 2$, and set $\gamma := T\exp(-d/36)$. We denote by $h : \mathbb{R}^d \to [1, \infty)$, $\mathbf{x}^* \in \mathbb{R}^d$, $\rho > 0$ the function, minimizer, and positive constant, respectively, given by Lemma B.2(iii) when applied to $\mathcal{A}, T, d$. Given any $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $\mu > 0$, we denote $\bar{\mathbf{w}} := \frac{\mathbf{w}}{\|\mathbf{w}\|}$ and construct*

$$f_{\mathbf{w},\mu}(\mathbf{x}) := \max\left\{h(\mathbf{x}) - \sigma_\mu(q(\mathbf{x} - \mathbf{x}^*)), 0\right\}, \tag{B.1}$$

where $q : \mathbb{R}^d \to \mathbb{R}$ and $\sigma_\mu : \mathbb{R} \to \mathbb{R}$ are defined as follows:

$$q(\mathbf{x}) := \langle \bar{\mathbf{w}}, \mathbf{x} + \mathbf{w}\rangle - \tfrac{1}{2}\|\mathbf{x} + \mathbf{w}\| \text{ and } \sigma_\mu(z) := \begin{cases} 0, & z \leq 0 \\ \frac{z^2}{8\mu}, & z \in (0, \mu] \\ \frac{z}{4} - \frac{\mu}{8}, & z > \mu \end{cases}. \tag{B.2}$$

We need the following technical result about $f_{\mathbf{w},\mu}$, the proof of which we defer to Appendix D.

Lemma B.4 *For the setup in Definition B.3 and $\mathbf{w}$ such that $\mathbf{w} \perp \mathbf{e}_d$ and $\|\mathbf{w}\| = 1000\mu$, the following hold:*

(i) $f_{\mathbf{w},\mu}$ is non-negative, 1-Lipschitz, and Clarke regular.
(ii) For $c = \frac{1}{100}$, any c-stationary point $\mathbf{x}$ of $f_{\mathbf{w},\mu}$ satisfies $f_{\mathbf{w},\mu}(\mathbf{x}) = 0$. In particular, any c-stationary point of $f_{\mathbf{w},\mu}$ is a global minimum.
(iii) There exist $\mathbf{w} \in \mathbb{R}^d$, $\mu > 0$ such that applying $\mathcal{A}$ to $f_{\mathbf{w},\mu}$ satisfies

$$\Pr_{\mathcal{A}}[\exists t \in [T] : f_{\mathbf{w},\mu}(\mathbf{x}_t) < 1] < 2\gamma\,.$$

Remark B.5 *Following Lemma B.4(iii), we set $\mathbf{w}, \mu$ so that $\Pr_{\mathcal{A}}[\exists t \in [T] : f_{\mathbf{w},\mu}(\mathbf{x}_t) < 1] < 2\gamma$ holds. For notational brevity, from hereon, we let $f = f_{\mathbf{w},\mu}$.*

We already see by Lemma B.4(i) that f satisfies the regularity assumptions required by Theorem 2.1. We therefore complete the proof by showing that it satisfies hardness in terms of getting near stationary points.

Lemma B.6 *For the setup in Definition B.3, for any $c = \frac{1}{100}$, we have that*

$$\Pr_{\mathcal{A}}\left[\exists t \in [T] : \mathbf{x}_t \text{ is of distance } < 1 \text{ to a } c\text{-stationary point of } f\right] \leq 2T\exp(-d/36).$$

Proof [Proof of Lemma B.6] By Lemma B.4, with probability at least $1 - 2\gamma : \min_{t\in[T]} f(\mathbf{x}_t) \geq 1$, while for any c-stationary point $\mathbf{x}$: $f(\mathbf{x}) = 0$. Under this probable event, recalling that f is 1-Lipschitz, we get that $(\mathbf{x}_t)_{t=1}^T$ must be of distance of at least 1 from any c-stationary point of f. Finally, we complete the proof by recalling that $\gamma := T\exp(-d/36)$ in Definition B.3. ■

We finalize by showing that if an iterate is far away from any approximate-stationary point, then it cannot be an approximate local minimum.

Lemma B.7 *Let f be a Clarke regular function, and suppose $\mathbf{x} \in \mathbb{R}^d, c_\delta > 0$ are such that $B(\mathbf{x}, c_\delta)$ does not contain any c_ϵ-stationary point. Then for every $\delta \leq c_\delta$:*

$$\min_{\mathbf{z}\in B(\mathbf{x},\delta)} f(\mathbf{z}) < f(\mathbf{x}) - \delta c_\epsilon\,.$$

Proof [Proof of Lemma B.7] Consider the subgradient flow $\mathbf{x}(0) = \mathbf{x}$, $\frac{d\mathbf{x}(t)}{dt} = -\frac{\bar{\partial} f(\mathbf{x}(t))}{\|\bar{\partial} f(\mathbf{x}(t))\|}$. Note that the flow is well defined throughout $t \in [0, c_\delta]$ since it has unit speed, hence $\mathbf{x}(t) \in B(\mathbf{x}, t) \subseteq B(\mathbf{x}, c_\delta)$ and by assumption $B(\mathbf{x}, c_\delta)$ does not contain any point with zero gradient. Defining $\phi(t) := f(\mathbf{x}(t))$, we get by the chain rule

$$\frac{d\phi(t)}{dt} = \bar{\partial} f(\mathbf{x}(t)) \cdot \frac{d\mathbf{x}(t)}{dt} = -\left\|\bar{\partial} f(\mathbf{x}(t))\right\|,$$

thus

$$\min_{\mathbf{z} \in B(\mathbf{x}, \delta)} f(\mathbf{z}) - f(\mathbf{x}) \le f(\mathbf{x}(\delta)) - f(\mathbf{x}) = \phi(\delta) - \phi(0) = -\int_0^\delta \left\|\bar{\partial} f(\mathbf{x}(t))\right\| dt < -\delta c_\epsilon .$$

■

Proof [Proof of Theorem 2.1] To prove our main result of this paper, the function we consider is $f = f_{\mathbf{w},\mu}$ as defined in Definition B.3, with $\mathbf{w}$ and μ chosen so as to have Lemma B.4(iii) be satisfied. We now state where we proved all its claimed properties. We showed in Lemma B.4(i) that this f is 1-Lipschitz and Clarke regular. We also showed in Lemma B.4(i) that f is non-negative, which implies $\inf f \ge 0$. Next, from (B.1), we have that $f(\mathbf{0}) \le h(\mathbf{0}) = \bar{h}(0) \le 3/2$, where we used Lemma B.2(i) in the final step. This shows $f(\mathbf{0}) - \inf f \le 2$ as claimed. Similarly, we showed in Lemma B.4(ii) that any c-stationary point of f is a global minimum for $c = \frac{1}{100}$. To show Theorem 2.1(ii), we combine Lemma B.6 with Lemma B.7 (using $c_\epsilon = c,\ c_\delta = 1$) to complete the proof. ■

Appendix C. Proof of Proposition B.1

The goal of this section is to prove Proposition B.1, which we first recall below.

Proposition B.1 *For any $\gamma > 0$ and $T \in \mathbb{N}$, there exists $\rho > 0$ so that the following holds: For any (possibly randomized) local algorithm $\mathcal{A}$, there exists a function $\bar{h} : \mathbb{R} \to [2, \infty)$ such that*

(i) $\bar{h}$ is 1-Lipschitz, convex, and satisfies $\bar{h}(0) \le 3$.
(ii) $\bar{h}$ has a unique minimizer $x^* \in (0, 1)$, and $\forall x \ne x^* : \left|\bar{\partial}\bar{h}(x)\right| \ge \frac{1}{8}$.
(iii) $\Pr_{\mathcal{A}}[\exists t \in [T] : |x_t - x^*| \le \rho] < \gamma$.

Before proving this result (i.e., constructing the function $\bar{h}$), we describe our high-level idea, followed by definitions of the components that constitute $\bar{h}$; our proof appears in Appendix C.3.

C.1. Intuition for Proposition B.1

Our function $\bar{h}$ is essentially a translated version of $r_\sigma^{(N)}$ defined in (C.10). As can be seen from this defining expression, $r_\sigma^{(N)}$ is a *random* piecewise-affine scalar-valued function defined over the entire real line. The key functions that make up $r_\sigma^{(N)}$ are $\phi_{\sigma_i}^{(i)}$ (Definition C.3), $g_{\sigma_i}^{(i)}$ (Definition C.5), and $\Phi^{(i)}$ (Definition C.7).

Crucially, the randomness in the parameter $\sigma \sim \mathrm{Unif}\{0,1\}^N$ of $r_\sigma^{(N)}$ determines the both the intervals and affine functions that constitute $r_\sigma^{(N)}$. The intervals and their corresponding functions

are carefully chosen so as to ensure continuity of $r_\sigma^{(N)}$: in particular, within each affine function is baked in a term of the form $\left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \dots\right)^{-1}$ that leads to appropriate cancellation at the endpoints of intervals of definition; see, e.g., (C.11) and (C.12). The Lipschitzness of $r_\sigma^{(N)}$ follows from that of each of its pieces, which in turn follows from the chain rule applied to the composition functions making up each of these pieces and Lemma C.6(iv), where $g_{\sigma_i}^{(i)}$ is proven Lipschitz. The same line of reasoning also yields a *lower* bound on the slope of $r_\sigma^{(N)}$. The proof of convexity of $r_\sigma^{(N)}$ is made simple by our parametrization of it in terms of $\theta^{(i)}$: the slope of the constituent affine functions of $r_\sigma^{(N)}$ varies as $\cot(\theta^{(i)})$. The monotonicity of the cotangent function and our chosen range of $\theta^{(i)}$ then imply that the slope of $r_\sigma^{(N)}$ increases as we traverse the real line from left to right, thus immediately giving us the desired convexity. We now proceed to state these technical details with the goal of proving Proposition B.1. The proof is in Appendix C.3.

C.2. Technical Details for Proposition B.1

Definition C.1 *For $i \in \mathbb{N}$, we define a sequence of angles $\{\theta_{\text{base}}^{(i)}\}$ and $\{\theta_{\text{shift}}^{(i)}\}$ that satisfies*

$$\theta_{\text{base}}^{(1)} = \arctan(1), \quad \theta_{\text{shift}}^{(i)} = \frac{\arctan(8) - \theta_{\text{base}}^{(1)}}{2^i}, \quad \theta_{\text{base}}^{(i+1)} = \theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}. \tag{C.1}$$

For $\theta_{\text{base}}^{(i)}$ and $\theta_{\text{shift}}^{(i)}$, now define the quantities $\epsilon^{(i)}$ and $\delta^{(i)}$ as follows:

$$\epsilon^{(i)} = 1 - \frac{3}{2} \cdot \left(\frac{1}{2\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right), \quad \delta^{(i)} = \frac{1}{2} \cdot \left(\frac{\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) - \tan(\theta_{\text{base}}^{(i)})}{2 \cdot \tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right). \tag{C.2}$$

Claim C.2 *The $\delta^{(i)}$ and $\epsilon^{(i)}$ and angles $\theta_{\text{base}}^{(i)}$ from Definition C.1 satisfy, for all $i \in \mathbb{N}$, that*

$$\epsilon^{(i)} > 0, \quad 0 < \delta^{(i)} \le \tfrac{7}{32}, \quad \text{and} \quad \theta_{\text{base}}^{(i)} \in [\arctan(1), \arctan(8)].$$

Proof First, observe that $\theta_{\text{base}}^{(i)}$ is monotonically increasing in i (as seen from (C.1)). Hence, for all $i \in \mathbb{N}$, we have $\theta_{\text{base}}^{(i)} \ge \theta^{(1)} = \arctan(1)$, which finishes one part of the claim. To state a lower bound on $\epsilon^{(i)}$ defined in (C.2), we observe that

$$2\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)}) \ge 3\tan(\theta_{\text{base}}^{(i)}) \ge 3\tan(\theta_{\text{base}}^{(1)}) = 3\tan(\arctan(1)) = 3,$$

where the first step used the monotoniticity of the tangent function, the second step used the fact that $\theta_{\text{base}}^{(i)}$ is monotonically increasing (as seen from (C.1)), and the final step is by evaluation. Therefore, we have that

$$\epsilon^{(i)} \ge 1 - \frac{3}{2 \times 3} > 0,$$

which finishes the proof of the claim of positive $\epsilon^{(i)}$ for all $i \in \mathbb{N}$. To see the bounds on $\delta^{(i)}$, we first note that by monotonicity of the tangent function, $\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) > \tan(\theta_{\text{base}}^{(i)})$, so $\delta^{(i)} > 0$. For the upper bound, we note that

$$\delta^{(i)} = \frac{1}{2} \cdot \left(\frac{\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) - \tan(\theta_{\text{base}}^{(i)})}{2 \cdot \tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right) \le \frac{1}{2} \cdot \frac{\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) - 1}{2\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)})},$$

where the first step is because $\theta_{\text{base}}^{(i)} \geq \arctan(1)$ for all i. Finally, noting that $\theta_{\text{base}}^{(i)} \leq \arctan(8)$ for all i, we can further bound the term above as $\delta^{(i)} \leq \frac{1}{4} - \frac{1}{32} = \frac{7}{32}$, as claimed. ■

Definition C.3 *For $i \in \mathbb{N}$, we use the $\delta^{(i)}$ from Definition C.1 to define the functions $\phi_0^{(i)}$ and $\phi_1^{(i)}$ as follows.*

$$\begin{aligned} \phi_0^{(i)} : [0,1] \to \left[\tfrac{1}{2} - 2\delta^{(i)}, \tfrac{1}{2} - \delta^{(i)}\right] \subseteq (0,1), \text{ where } \phi_0^{(i)}(x) = \delta^{(i)} \cdot x + \tfrac{1}{2} - 2\delta^{(i)}, \\ \phi_1^{(i)} : [0,1] \to \left[\tfrac{1}{2} + \delta^{(i)}, \tfrac{1}{2} + 2\delta^{(i)}\right] \subseteq (0,1), \text{ where } \phi_1^{(i)}(x) = \delta^{(i)} \cdot x + \tfrac{1}{2} + \delta^{(i)}. \end{aligned} \tag{C.3}$$

These are the unique affine maps with positive derivatives mapping $[0,1]$ to their respective ranges. Our assertion that $[\frac{1}{2} - 2\delta^{(i)}, \frac{1}{2} - \delta^{(i)}] \subseteq (0,1)$ and $[\frac{1}{2} + \delta^{(i)}, \frac{1}{2} + 2\delta^{(i)}] \subseteq (0,1)$ is justified by $0 < \delta^{(i)} < \frac{1}{4}$ from Claim C.2. We use the functions in (C.3) to define the interval

$$\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_k} \stackrel{def}{=} \phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_k}^{(k)}(0,1) \subseteq (0,1).$$

We denote the left and right end point of $\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_k}$ as

$$\inf \mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_k} = \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_k}^{(k)}(0) \text{ and } \sup \mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_k} = \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_k}^{(k)}(1). \tag{C.4}$$

Furthermore, we note the following definitions of the specific intervals $\mathbf{I}_0$ and $\mathbf{I}_1$:

$$\mathbf{I}_0 = \phi_0^{(1)}(0,1), \text{ and } \mathbf{I}_1 = \phi_1^{(1)}(0,1).$$

Lemma C.4 *The functions defined in Definition C.3 satisfy the following properties.*

(i) For any $\ell < k$, and any $\sigma_1, \sigma_2, \ldots, \sigma_\ell, \ldots, \sigma_k \in \{0,1\}$, we have that $\mathbf{I}_{\sigma_1\ldots\sigma_\ell} \supseteq \mathbf{I}_{\sigma_1\ldots\sigma_\ell\ldots\sigma_k}$.

(ii) For any $i \geq 1$, the function $\phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_i}^{(i)} : (0,1) \mapsto \mathbf{I}_{\sigma_1\ldots\sigma_i}$ is non-decreasing.

(iii) If $(\sigma_1 \ldots \sigma_{k-1}) \neq (\sigma'_1 \ldots \sigma'_{k-1})$ then for all $\sigma_k, \sigma'_k \in \{0,1\} : \mathbf{I}_{\sigma_1\ldots\sigma_k} \cap \mathbf{I}_{\sigma'_1\ldots\sigma'_k} = \emptyset$.

(iv) Let $k = \frac{1}{4}\sqrt{\log\left(\frac{1}{\rho}\right)}$. Then dist$(x, \mathbf{I}_{\sigma_1\ldots\sigma_N}) \leq \rho$ implies $x \in \mathbf{I}_{\sigma_1\ldots\sigma_k}$.

Proof [Proof of Lemma C.4] We prove each of the parts below.

Proof of Lemma C.4(i). The claim follows by the following observation:

$$\mathbf{I}_{\sigma_1\ldots\sigma_\ell\ldots\sigma_k} = \phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_\ell}^{(\ell)}(\phi_{\sigma_{\ell+1}}^{(\ell+1)} \circ \ldots \circ \phi_{\sigma_k}^{(k)}(0,1)) \subseteq \phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_\ell}^{(\ell)}(0,1) = \mathbf{I}_{\sigma_1\ldots\sigma_\ell}.$$

Proof of Lemma C.4(ii). Observe that from Definition C.3, the function $\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \phi_{\sigma_i}^{(i)}$ is a composition of affine functions and is therefore itself affine; furthermore, by applying the chain rule, one may infer that the derivative of a composition of affine functions equals the product of the derivatives of the individual composing functions, which in our case is the product $\delta^{(1)} \cdot \delta^{(2)} \ldots \delta^{(i)}$, which is positive since each $\delta^{(i)}$ is as well.

Proof of Lemma C.4(iii). Suppose $i \le k-1$ be the minimal index for which $\sigma_i \neq \sigma'_i$ and assume, without loss of generality, that $\sigma_i = 0$ and $\sigma'_i = 1$. Consider a point x satisfying $x \in \mathbf{I}_{\sigma_i \dots \sigma_k}$. In other words, for some $u \in (0,1)$, we have that $x = \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(k)}_{\sigma_k}(0)$. Then, we have:

$$\begin{aligned}
\phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(k)}_{\sigma_k}(u) &\le \sup_{u\in(0,1)} \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}} \circ \phi^{(i)}_{0} \circ \phi^{(i+1)}_{\sigma_{i+1}} \ldots \circ \phi^{(k)}_{\sigma_k}(u) \\
&\le \sup_{u\in(0,1)} \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}} \circ \phi^{(i)}_{0}(u) \\
&< \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}} \left(\tfrac{1}{2}\right) \\
&\le \inf_{u\in(0,1)} \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}} \circ \phi^{(i)}_{1}(u) \\
&= \inf_{u\in(0,1)} \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}} \circ \phi^{(i)}_{\sigma'_i}(u) \\
&\le \inf_{u\in(0,1)} \phi^{(1)}_{\sigma'_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma'_{i-1}} \circ \phi^{(i)}_{\sigma'_i} \circ \phi^{(i+1)}_{\sigma'_{i+1}} \circ \ldots \phi^{(k)}_{\sigma'_k}(u). \qquad \text{(C.5)}
\end{aligned}$$

where the first step is by taking the largest value over all feasible u; the second step is because $\phi^{(i+1)}_{\sigma_{i+1}} \ldots \circ \phi^{(k)}_{\sigma_k}(u) \subseteq (0,1)$ implies we are simply maximimizing over a larger set of arguments of $\phi^{(i)}_0$; the third step uses that by Claim C.2, we have $\phi^{(i)}_0(0) = \frac{1}{2} - 2\delta^{(i)} \in (0, \frac{1}{2})$ and further that $\phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}}$ is non-decreasing; the fourth step again uses monotonicity of $\phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i-1)}_{\sigma_{i-1}}$ and the fact that $\frac{1}{2} < \phi^{(i)}_1(0) = \frac{1}{2} + \delta^{(i)}$; the sixth step uses the fact that we are minimizing over a smaller set as well as replaces, in the first $i-1$ terms of the composition, $\phi^{(j)}_{\sigma_j}$ by $\phi^{(j)}_{\sigma'_j}$. In conclusion, from Inequality (C.5), we have that $x \notin \inf_{u\in(0,1)} \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(k)}_{\sigma'_k}(u)$ and hence $x \notin \mathbf{I}_{\sigma'_1 \dots \sigma'_k}$, which finishes the proof of $\mathbf{I}_{\sigma_1 \dots \sigma_k} \cap \mathbf{I}_{\sigma'_1 \dots \sigma'_k} = \emptyset$.

Proof of Lemma C.4(iv). Recall from Lemma C.4(i) that $\mathbf{I}_{\sigma_1 \dots \sigma_N} \subseteq \mathbf{I}_{\sigma_1 \dots \sigma_k}$, and note that for any $i \in \mathbb{N}$:

$$\inf \mathbf{I}_{\sigma_1 \dots \sigma_{i+1}} - \inf \mathbf{I}_{\sigma_1 \dots \sigma_i} = \phi^{(1)}_{\sigma_1} \circ \ldots \phi^{(i)}_{\sigma_i} \circ \phi^{(i)}_{\sigma_{i+1}}(0) - \phi^{(1)}_{\sigma_1} \circ \ldots \circ \phi^{(i)}_{\sigma_i}(0) = \prod_{j=1}^{i} (\phi^{(j)}_{\sigma_j})' \cdot (\phi^{(i+1)}_{\sigma_{i+1}}(0) - 0).$$

Next, recall from (C.3) that for every $j \in \mathbb{N}$, the slope $(\phi^{(j)}_{\sigma_j})' = \delta^{(j)}$. We get a lower bound on the product $\prod_{j=1}^{i} \delta^{(j)}$ as follows.

$$\delta^{(i)} = \frac{1}{2} \cdot \frac{\tan(\theta^{(i)}_{\text{base}} + \theta^{(i)}_{\text{shift}}) - \tan(\theta^{(i)}_{\text{base}})}{2 \cdot \tan(\theta^{(i)}_{\text{base}} + \theta^{(i)}_{\text{shift}}) + \tan(\theta^{(i)}_{\text{base}})} \ge \frac{\tan(\theta^{(i)}_{\text{shift}})}{12} = \frac{1}{12} \cdot \tan\left(\frac{\arctan(8) - \arctan(1)}{2^i}\right), \qquad \text{(C.6)}$$

where the first inequality is due to the identity $\tan(\theta^{(i)}_{\text{base}} + \theta^{(i)}_{\text{shift}}) = \frac{\tan(\theta^{(i)}_{\text{base}}) + \tan(\theta^{(i)}_{\text{shift}})}{1 - \tan(\theta^{(i)}_{\text{base}}) \tan(\theta^{(i)}_{\text{shift}})}$ and simplifying via the facts that $\tan(\theta^{(i)}_{\text{base}}) \ge 0$, $\tan(\theta^{(i)}_{\text{shift}}) \ge 0$, all angles $\theta^{(i)}_{\text{base}} \in [\arctan(1), \arctan(8)]$ and that since the tan function is monotonically increasing in $(0, \pi/2)$, we have $\tan(\theta^{(i)}_{\text{base}} + \theta^{(i)}_{\text{shift}}) \ge$

$\tan(\theta^{(i)}_{\text{base}})$. We can therefore continue the lower bound on $\prod_{j=1}^{i}(\phi^{(j)}_{\sigma_j})' \cdot (\phi^{(i+1)}_{\sigma_{i+1}}(0) - 0)$ as follows.

$$\begin{aligned}\prod_{j=1}^{i}(\phi^{(j)}_{\sigma_j})' \cdot (\phi^{(i)}_{\sigma_{i+1}}(0) - 0) &\geq \frac{1}{16} \cdot \prod_{j=1}^{i} \tan\left(\frac{\arctan(8) - \arctan(1)}{2^j}\right) \\ &\geq \frac{1}{16} \cdot \prod_{j=1}^{i} \left(\frac{\arctan(8) - \arctan(1)}{2^j}\right) \\ &\geq (\arctan(8) - \arctan(1))^i \cdot \frac{1}{2^{2i^2+4}}, \end{aligned} \tag{C.7}$$

where the first step is by Inequality (C.6) and lower bounding $\phi^{(i+1)}_{\sigma_{i+1}}(0) = \frac{1}{2} - 2\delta^{(i)} \geq \frac{1}{2} - 2 \cdot \frac{7}{32} \geq \frac{1}{16}$ from Definition C.3 and Claim C.2, and the second step is by lower bounding each term $\tan(x)$ of the product via $\tan(x) \geq x$, which in turn follows from the power series expansion of the tangent function. Therefore, we may use Inequality (C.7) as follows:

$$\begin{aligned}\inf \mathbf{I}_{\sigma_1 \ldots \sigma_N} - \inf \mathbf{I}_{\sigma_1 \ldots \sigma_k} &= \sum_{i=k}^{N-1} (\inf \mathbf{I}_{\sigma_1 \ldots \sigma_{i+1}} - \inf \mathbf{I}_{\sigma_1 \ldots \sigma_i}) \\ &\geq \sum_{i=k}^{N} \frac{(\arctan(8) - \arctan(1))^i}{2^{2i^2+4}} \\ &\geq \frac{(\arctan(8) - \arctan(1))^k}{2^{2k^2+4}} \\ &\geq \frac{1}{2^{2k^2+k+4}} > \rho, \end{aligned}$$

where in the second inequality, we dropped all but the first term (valid since all the terms are non-negative), in the penultimate inequality, we plugged in a lower bound for $\arctan(8) - \arctan(1)$, and the final inequality holds for the choice of $k = \frac{1}{4}\sqrt{\log\left(\frac{1}{\rho}\right)}$. We conclude $\inf \mathbf{I}_{\sigma_1 \ldots \sigma_k} < \inf \mathbf{I}_{\sigma_1 \ldots \sigma_N} - \rho$ and analogously $\sup \mathbf{I}_{\sigma_1 \ldots \sigma_k} > \sup \mathbf{I}_{\sigma_1 \ldots \sigma_N} + \rho$. Together, the two bounds imply the claim. ∎

Definition C.5 *Using $\epsilon^{(i)}$ and $\delta^{(i)}$ from Definition C.1 and $\phi^{(i)}_1$ and $\phi^{(i)}_0$ from Definition C.3, we now define, for $i \in \mathbb{N}$, the following family of functions:*

$$g^{(i)}_1 : [0,1] \backslash \phi^{(i)}_1(0,1) \to \mathbb{R}, \text{ where } g^{(i)}_1(x) = \begin{cases} -\frac{1-\epsilon^{(i)}}{\frac{1}{2}+\delta^{(i)}} \cdot x + 1 & \text{if } x \in [0, \frac{1}{2} + \delta^{(i)}] \\ \frac{1-\epsilon^{(i)}}{\frac{1}{2}-2\delta^{(i)}} \cdot x + \frac{-\frac{1}{2}-2\delta^{(i)}+\epsilon^{(i)}}{\frac{1}{2}-2\delta^{(i)}} & \text{if } x \in [\frac{1}{2} + 2\delta^{(i)}, 1], \end{cases}$$

and $g^{(i)}_0(x) = g^{(i)}_1(1-x)$ with the following explicit closed-form expression:

$$g^{(i)}_0 : [0,1] \backslash \phi^{(i)}_0(0,1) \to \mathbb{R}, \text{ where } g^{(i)}_0(x) = \begin{cases} -\frac{1-\epsilon^{(i)}}{\frac{1}{2}-2\delta^{(i)}} \cdot x + 1 & \text{if } x \in [0, \frac{1}{2} - 2\delta^{(i)}] \\ \frac{1-\epsilon^{(i)}}{\frac{1}{2}+\delta^{(i)}} \cdot x + \frac{-\frac{1}{2}+\delta^{(i)}+\epsilon^{(i)}}{\frac{1}{2}+\delta^{(i)}} & \text{if } x \in [\frac{1}{2} - \delta^{(i)}, 1]. \end{cases}$$

Lemma C.6 *The functions $g_{\sigma_i}^{(i)}$ from Definition C.5 satisfy the following properties.*

(i) For both possible choices of σ_i, the end points of the functions $g_{\sigma_i}^{(i)}$ satisfy

$$g_{\sigma_i}^{(i)}(1) = g_{\sigma_i}^{(i)}(0) = 1.$$

(ii) For both possible choices of σ_i, the functions $g_{\sigma_i}^{(i)}$ and $\phi_{\sigma_i}^{(i)}$ (from Definition C.3) satisfy

$$g_{\sigma_i}^{(i)} \circ \phi_{\sigma_i}^{(i)}(0) = \epsilon^{(i)}.$$

(iii) Recall $\theta_{\text{base}}^{(i)}$ and $\theta_{\text{base}}^{(i+1)}$ as defined in Definition C.1. The derivatives of $g_0^{(i)}$ and $g_1^{(i)}$ are given by

$$(g_0^{(i)})'(x) = \begin{cases} -\cot(\theta_{\text{base}}^{(i)}) & \text{if } x \in (0, \frac{1}{2} - 2\delta^{(i)}) \\ \cot(\theta_{\text{base}}^{(i+1)}) & \text{if } x \in (\frac{1}{2} - \delta^{(i)}, 1) \end{cases}$$

and

$$(g_1^{(i)})'(x) = \begin{cases} -\cot(\theta_{\text{base}}^{(i+1)}) & \text{if } x \in (0, \frac{1}{2} + \delta^{(i)}) \\ \cot(\theta_{\text{base}}^{(i)}) & \text{if } x \in (\frac{1}{2} + 2\delta^{(i)}, 1). \end{cases}$$

(iv) The derivative, in absolute value, of $g_{\sigma_i}^{(i)}$ for all $i \in \mathbb{N}$ is at least $\frac{1}{8}$ and at most 1 (in the interior of its domain). In other words, we have:

$$\tfrac{1}{8} \le |(g_0^{(i)})'(x)| \le 1 \text{ for } x \in (0, \tfrac{1}{2} - 2\delta^{(i)}) \cup (\tfrac{1}{2} - \delta^{(i)}, 1)$$

and

$$\tfrac{1}{8} \le |(g_1^{(i)})'(x)| \le 1 \text{ for } x \in (0, \tfrac{1}{2} + \delta^{(i)}) \cup (\tfrac{1}{2} + 2\delta^{(i)}, 1).$$

Proof To check Lemma C.6(i) and Lemma C.6(ii), one may evaluate the functions in question from Definition C.5. To prove Lemma C.6(iii), we observe from Definition C.5 that for $x \in (0, \frac{1}{2} - 2\delta^{(i)})$, the derivative of $g_0^{(i)}$ is given by the expression:

$$(g_0^{(i)})'(x) = -\frac{1-\epsilon^{(i)}}{\frac{1}{2} - 2\delta^{(i)}} = -\frac{\frac{3}{2} \cdot \left(\frac{1}{2\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right)}{\frac{1}{2} - 2 \cdot \frac{1}{2} \cdot \left(\frac{\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) - \tan(\theta_{\text{base}}^{(i)})}{2 \cdot \tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right)} = \frac{-3}{3\tan(\theta_{\text{base}}^{(i)})} = -\cot(\theta_{\text{base}}^{(i)}),$$

and for $x \in (\frac{1}{2} - \delta^{(i)}, 1)$, the derivative of $g_0^{(i)}$ is:

$$(g_0^{(i)})'(x) = \frac{1-\epsilon^{(i)}}{\frac{1}{2} + \delta^{(i)}} = \frac{\frac{3}{2} \cdot \left(\frac{1}{2\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right)}{\frac{1}{2} + \frac{1}{2} \cdot \left(\frac{\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) - \tan(\theta_{\text{base}}^{(i)})}{2 \cdot \tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)}) + \tan(\theta_{\text{base}}^{(i)})} \right)} = \frac{3}{3\tan(\theta_{\text{base}}^{(i)} + \theta_{\text{shift}}^{(i)})} = \cot(\theta_{\text{base}}^{(i+1)}).$$

The derivatives of $g_1^{(i)}$ are computed in a similar fashion. To prove Lemma C.6(iv), we note that the largest angle $\theta_{\text{base}}^{(i)}$ obtained in the sequence described by (C.1) is

$$\theta_{\text{base}}^{(\infty)} = \theta_{\text{base}}^{(1)} + \sum_{i=1}^{\infty} \frac{\arctan(8) - \theta_{\text{base}}^{(1)}}{2^i} \le \theta_{\text{base}}^{(1)} + (\arctan(8) - \theta_{\text{base}}^{(1)}) = \arctan(8). \tag{C.8}$$

Since the cotangent function is decreasing on the positive interval upto $\pi/2$, plugging into Lemma C.6(iii) the upper bound (C.8) implies that the lower bound on the (absolute) value of the derivative of $g_{\sigma_i}^{(i)}$ in the interior of its domain is

$$|(g_{\sigma_i}^{(i)})'(x)| \geq \cot(\arctan(8)) = \tfrac{1}{8}.$$

For the upper bound, we observe that the largest (in absolute value) derivative is attained at the smallest angle:

$$|(g_{\sigma_i}^{(i)})'(x)| \leq \cot(\tan(\theta_{\text{base}}^{(1)})) = \cot(\tan(1)) = 1.$$

■

Definition C.7 *Using the above functions, we define*

$$\Phi^{(i)}(x) = \delta^{(i)} \cdot x + g_{\sigma_i}^{(i)} \circ \phi_{\sigma_i}^{(i)}(0) - \delta^{(i)}.$$

This definition yields the following important consequence:

$$\Phi^{(i)}(1) = g_0^{(i)} \circ \phi_0^{(i)}(0) = g_1^{(i)} \circ \phi_1^{(i)}(0), \tag{C.9}$$

where the second equality is justified in Lemma C.6(ii).

Finally we are ready to provide the following function definition.

Definition C.8 *Given $N \in \mathbb{N}$ and $\sigma \in \{0,1\}^N$, define the function*

$$r_\sigma^{(N)}(x) = \begin{cases}
g_{\sigma_1}^{(1)}(x) & x \in [0, \quad \mathbf{I}_{\sigma_1}[\ell]] \\
\Phi^{(1)} \circ g_{\sigma_2}^{(2)} \circ \left(\phi_{\sigma_1}^{(1)}\right)^{-1}(x) & x \in [\mathbf{I}_{\sigma_1}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2}[\ell]] \\
\Phi^{(1)} \circ \Phi^{(2)} \circ g_{\sigma_3}^{(3)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)}\right)^{-1}(x) & x \in [\mathbf{I}_{\sigma_1\sigma_2}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2\sigma_3}[\ell]], \\
\vdots & \vdots \\
\Phi^{(1)} \circ \ldots \circ \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}\right)^{-1}(x) & x \in \left[\mathbf{I}_{\sigma_1\ldots\sigma_i}[\ell], \quad \mathbf{I}_{\sigma_1\ldots\sigma_{i+1}}[\ell]\right] \\
\vdots & \vdots \\
\cot(\theta_{\text{base}}^{(N+1)}) \cdot \left(\phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(0) - x\right) + \Phi^{(1)} \circ \ldots \circ \Phi^{(N)}(1) & x \in [\mathbf{I}_{\sigma_1\ldots\sigma_N}[\ell], \quad x_{\text{mid}}] \\
\cot(\theta_{\text{base}}^{(N+1)}) \cdot \left(x - \phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(1)\right) + \Phi^{(1)} \circ \ldots \circ \Phi^{(N)}(1) & x \in [x_{\text{mid}}, \quad \mathbf{I}_{\sigma_1\ldots\sigma_N}[r]] \\
\vdots & \\
\Phi^{(1)} \circ \ldots \circ \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}\right)^{-1}(x) & x \in \left[\mathbf{I}_{\sigma_1\ldots\sigma_{i+1}}[r], \quad \mathbf{I}_{\sigma_1\ldots\sigma_i}[r]\right] \\
\vdots & \\
\Phi^{(1)} \circ \Phi^{(2)} \circ g_{\sigma_3}^{(3)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)}\right)^{-1}(x) & x \in [\mathbf{I}_{\sigma_1\sigma_2\sigma_3}[r], \quad \mathbf{I}_{\sigma_1\sigma_2}[r]], \\
\Phi^{(1)} \circ g_{\sigma_2}^{(2)} \circ \left(\phi_{\sigma_1}^{(1)}\right)^{-1}(x) & x \in [\mathbf{I}_{\sigma_1\sigma_2}[r], \quad \mathbf{I}_{\sigma_1}[r]] \\
g_{\sigma_1}^{(1)}(x) & x \in [\mathbf{I}_{\sigma_1}[r], \quad 1] \\
1 - x & x < 0 \\
x & x > 1,
\end{cases} \tag{C.10}$$

where $x_{\text{mid}} = \frac{1}{2} \cdot \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(0) + \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(1)\right)$.

The following lemma follows immediately by combining Lemma C.4(iii) with the construction of $r_\sigma^{(N)}$, and noting that there

Lemma C.9 *For any $k < N \in \mathbb{N}$, $\sigma \in \{0,1\}^N$, any local oracle $\mathbb{O}$ and any $x \notin \mathbf{I}_{\sigma_{1\ldots\sigma_k}}$, it holds that $\mathbb{O}_{r_\sigma^{(N)}}(x)$ does not depend on $\sigma_{k+1}, \ldots, \sigma_N$. Therefore, for any $t \in \mathbb{N}$, $1 \le k < l < N$:*

$$\Pr_{\sigma \sim \text{Unif}\{0,1\}^N}[x_{t+1} \in \mathbf{I}_{\sigma_1 \ldots \sigma_k \ldots \sigma_l} \mid x_1, \ldots, x_t \notin \mathbf{I}_{\sigma_1 \ldots \sigma_k}] \le \frac{1}{2^{l-k-1}} .$$

We now prove some properties of $r_\sigma^{(N)}$ and construct the function $\bar{h}$ referred to in Proposition B.1; our construction of $\bar{h}$ ensures that it inherits all its necessary properties from $r_\sigma^{(N)}$.

C.3. Putting it all together: proof of Proposition B.1

Proof [Proof of Proposition B.1] We first show the properties of continuity, 1-Lipschitzness, and convexity in $r_\sigma^{(N)}$, and its upper bound at zero (i.e., Proposition B.1(i)).

Proof of continuity. We start by proving the continuity of $r_\sigma^{(N)}$. Since $g_{\sigma_i}^{(i)}$, $\phi_{\sigma_i}^{(i)}$,and $\Phi^{(i)}$ are all (piecewise) affine over their domains of definition, their composition and hence, from Definition C.8, $r_\sigma^{(N)}$ is also piecewise affine. To show continuity, we therefore need to show this only at the endpoints of each of the segments in Definition C.8. For some $1 < i < N$, consider the left endpoint of the segment

$$\left[\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_i}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_{i+1}}[\ell]\right] .$$

For continuity at this interval's left endpoint $\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_i}[\ell] = \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}(0)$ ((C.4)), we need to show that at $x = \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}(0)$, the functions $\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_{i-1}}^{(i-1)}\right)^{-1}(x)$ and $\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}\right)^{-1}(x)$ have the same value. To prove this, we simply evaluate the two functions one by one. First, observe that

$$\begin{aligned} &\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}\right)^{-1} \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}(0)\right) \\ &= \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)}(0) = \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i)}(1), \end{aligned} \tag{C.11}$$

where we use Lemma C.6(i) in the final step. Next, observe that

$$\begin{aligned} &\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_{i-1}}^{(i-1)}\right)^{-1} \left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_i}^{(i)}(0)\right) \\ &= \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ \phi_{\sigma_i}^{(i)}(0). \end{aligned} \tag{C.12}$$

From (C.9), the final terms in (C.11) and (C.12) may be concluded to be equal. Next, consider $i = 1$. We need to consider the left endpoint of the interval

$$\left[\mathbf{I}_{\sigma_1}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2}[\ell]\right] .$$

To check continuity at the left endpoint of this interval, we evaluate two functions $g^{(1)}_{\sigma_1}(x)$ and $\Phi^{(1)} \circ g^{(2)}_{\sigma_2} \circ \left(\phi^{(1)}_{\sigma_1}\right)^{-1}(x)$ at $\mathbf{I}_{\sigma_1}[\ell] = \phi^{(1)}_{\sigma_1}(0)$. We have:

$$g^{(1)}_{\sigma_1}(x) = g^{(1)}_{\sigma_1} \circ \phi^{(1)}_{\sigma_1}(0). \tag{C.13}$$

Next, we have

$$\Phi^{(1)} \circ g^{(2)}_{\sigma_2} \circ \left(\phi^{(1)}_{\sigma_1}\right)^{-1}(x) = \Phi^{(1)} \circ g^{(2)}_{\sigma_2} \circ \left(\phi^{(1)}_{\sigma_1}\right)^{-1}(\phi^{(1)}_{\sigma_1}(0)) = \Phi^{(1)} \circ g^{(2)}_{\sigma_2}(0) = \Phi^{(1)}(1). \tag{C.14}$$

Comparing (C.13) and (C.14) using (C.9) completes the proof for $i = 1$. Next, we consider the left endpoint of the "left middle" segment

$$[\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_N}[\ell], \quad x_{\text{mid}}]\,.$$

To evaluate continuity at the left endpoint of this segment, we need to evaluate at $\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_N}[\ell] = \phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0)$ the functions $\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N-1)} \circ g^{(N)}_{\sigma_N} \circ \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N-1)}_{\sigma_{N-1}}\right)^{-1}(x)$ and $\cot(\theta^{(N+1)}_{\text{base}}) \cdot \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0) - x\right) + \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N)}(1)$. To this end, first observe that

$$\begin{aligned}
&\Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N-1)} \circ g^{(N)}_{\sigma_N} \circ \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N-1)}_{\sigma_{N-1}}\right)^{-1}(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0)) \\
&= \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N-1)} \circ g^{(N)}_{\sigma_N} \circ \phi^{(N)}_{\sigma_N}(0).
\end{aligned} \tag{C.15}$$

Next, observe that evaluating the other function at $x = \phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0)$ gives:

$$\begin{aligned}
&\cot(\theta^{(N+1)}_{\text{base}}) \cdot \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0) - \phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0)\right) + \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N)}(1) \\
&= \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \Phi^{(N)}(1) \\
&= \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N-1)} \circ \Phi^{(N)}(1).
\end{aligned} \tag{C.16}$$

We may again use (C.9) to equate (C.15) and (C.16). Finally, we show continuity at the "midpoint"

$$x_{\text{mid}} = \tfrac{1}{2} \cdot \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0) + \phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(1)\right).$$

From Definition C.8, we note that the left of the two middle segments (i.e., $x \in [\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_N}[\ell], \quad x_{\text{mid}}]$) is

$$r^{(N)}_{\sigma}(x) := \cot(\theta^{(N+1)}_{\text{base}}) \cdot \left(\phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(0) - x\right) + \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N)}(1).$$

Similarly, we have that the segment to the right of x_{mid} (i.e., $x \in [x_{\text{mid}}, \quad \mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_N}[r]]$) is described by

$$r^{(N)}_{\sigma}(x) := \cot(\theta^{(N+1)}_{\text{base}}) \cdot \left(x - \phi^{(1)}_{\sigma_1} \circ \phi^{(2)}_{\sigma_2} \circ \ldots \circ \phi^{(N)}_{\sigma_N}(1)\right) + \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N)}(1).$$

From these two definitions, we can check that the values of $r^{(N)}_{\sigma}(x_{\text{mid}})$ from both the definitions coincide. The continuity at the endpoints of segments to the right of x_{mid} may be similarly established.

Proof of 1-Lipschitzness. Since $r_\sigma^{(N)}$ is piecewise linear and continuous, the proof of its Lipschitzness requires only proving Lipschitzness of each of the segments it is composed of. We use the following two observations to establish this fact. First,

$$(\Phi^{(i)})' = (\phi^{(i)})',$$

because $\Phi^{(i)}$ and $\phi^{(i)}$ differ by only a constant additive factor. Second, by application of the chain rule, one may note that the derivative of a composition of affine functions equals the product of the derivatives of the individual composing functions. Using these two facts, we observe that

$$\left|\left(\Phi^{(1)} \circ \dots \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ (\phi_{\sigma_1}^{(1)} \circ \dots \phi_{\sigma_{i-1}}^{(i)})^{-1}\right)'(x)\right| = |(g_{\sigma_i}^{(i)})'(x)| \le 1,$$

where we used the upper bound from Lemma C.6(iv) in the final step. For the two middle segments, the absolute value of the slope is $\cot(\theta_{\text{base}}^{(N+1)})$. From the proof of Lemma C.6(iv) and the fact that the cotangent function is non-increasing in $(0, \pi/2)$, we have $\cot(\theta_{\text{base}}^{(N+1)}) \le \cot(\theta_{\text{base}}^{(1)}) = \cot(\arctan(1)) = 1$. This concludes the proof of $r_\sigma^{(N)}$ being a 1-Lipschitz function.

Proof of convexity. To prove convexity of $r_\sigma^{(N)}$, we show that the derivatives of consecutive segments composing $r_\sigma^{(N)}$ are non-decreasing. Consider the segment $\left[\mathbf{I}_{\sigma_1\sigma_2\dots\sigma_{i-1}}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2\dots\sigma_i}[\ell]\right]$. We have that for some $x \in \left[\mathbf{I}_{\sigma_1\sigma_2\dots\sigma_{i-1}}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2\dots\sigma_i}[\ell]\right]$, the derivative of $r_\sigma^{(N)}$ at this point x is:

$$\left(\Phi^{(1)} \circ \dots \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ (\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \dots \phi_{\sigma_{i-1}}^{(i-1)})^{-1}\right)'(x) = \left(g_{\sigma_i}^{(i)}\right)'(y) \le -\cot(\theta_{\text{base}}^{(i+1)}),$$

where $y \in [0, \phi_{\sigma_i}^{(i)}(0)]$ and for the final bound, we used from Lemma C.6(iii) the largest possible value for the slope of the left segment of $g_{\sigma_i}^{(i)}$ (for $g_{\sigma_i}^{(i)}$, the segment $[0, \phi_{\sigma_i}^{(i)}]$ corresponds to the left segments of $g_{\sigma_i}^{(i)}$). In a similar fashion, the smallest derivative of $r_\sigma^{(N)}$ on some x in the segment $\left[\mathbf{I}_{\sigma_1\sigma_2\dots\sigma_i}, \quad \mathbf{I}_{\sigma_1\sigma_2\dots\sigma_{i+1}}\right]$ may be derived to be at least

$$\left(\Phi^{(1)} \circ \dots \Phi^{(i)} \circ g_{\sigma_{i+1}}^{(i+1)} \circ (\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \dots \phi_{\sigma_i}^{(i)})^{-1}\right)'(x) \ge -\cot(\theta_{\text{base}}^{(i+1)}).$$

Therefore, the derivatives of consecutive segments composing $r_\sigma^{(N)}$ to the left of the middle segments are non-decreasing, going from left to right. The proof analogously extends to the right of the middle segments. For the two middle segments, note that the left segment has slope $-\cot(\theta_{\text{base}}^{(N+1)})$ and the right segment has slope $\cot(\theta_{\text{base}}^{(N+1)})$; since $\theta_{\text{base}}^{(N+1)} \in (\arctan(1), \arctan(8)) \subseteq (0, \pi/2)$, this means the slope is increasing when crossing x_{mid} from left to right. Finally, going from the left of the middle left segment to the middle left segment, the slope can only increase because of our choice of slope of the middle left segment; the analogous argument applies to the right side. Thus, overall, we have shown that the slope increases going from left to right, thus proving the convexity of $r_\sigma^{(N)}$.

Proof of $r_\sigma^{(N)}(0) \le 2$. From Definition C.8, one may check that $r_\sigma^{(N)}(0) = g_{\sigma_1}^{(1)}(0)$, which from Lemma C.6(i) satisfies $g_{\sigma_1}^{(1)}(0) = 1$, thus proving our claim.

We now prove Proposition B.1(ii), i.e., that $r_\sigma^{(N)}$ has a unique minimizer, that the absolute value of its slope is lower bounded by a constant, and obtain the expression for its minimum value.

Proof of lower bound on absolute slope. For any $i \in [2, N]$, consider a point x that lies in the segment $\left(\mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_{i-1}}[\ell], \quad \mathbf{I}_{\sigma_1\sigma_2\ldots\sigma_i}[\ell]\right)$. We have that the derivative of $r_\sigma^{(N)}$ at this point is given by:

$$|\left(\Phi^{(1)} \circ \ldots \Phi^{(i-1)} \circ g_{\sigma_i}^{(i)} \circ (\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \phi_{\sigma_{i-1}}^{(i-1)})^{-1}\right)'(x)| = |\left(g_{\sigma_i}^{(i)}\right)'(y)| \geq \tfrac{1}{8},$$

where $y \in (0, \phi_{\sigma_i}^{(i)}(0))$ and the final lower bound follows by applying Lemma C.6(iv). The same lower bound may be similarly obtained for segments to the right of the middle segment. For the two middle segments, the absolute value of the slope is $\cot(\theta_{\text{base}}^{(N+1)})$. From the non-increasing property of the cotangent function and our initial choice of $\theta_{\text{base}}^{(1)}$, we have $\cot(\theta_{\text{base}}^{(N+1)}) \geq \cot(\arctan(8)) = 1/8$, thus concluding the proof, overall, of the lower bound on the absolute value of the slope.

Proof of unique minimizer of $r_\sigma^{(N)}$. As we just showed, $r_\sigma^{(N)}$ is convex. At the point x_{mid}, the slope of $r_\sigma^{(N)}$ changes from $-\cot(\theta_{\text{base}}^{(N+1)})$ to $\cot(\theta_{\text{base}}^{(N+1)})$, which implies that there exists a zero subgradient of $r_\sigma^{(N)}$ at x_{mid}. Hence, x_{mid} is *a* minimizer of $r_\sigma^{(N)}$. This minimizer is unique because the two segments of the function intersecting at it both have non-zero slopes. This minimum value is obtained by evaluating $r_\sigma^{(N)}$ at x_{mid}:

$$\cot(\theta_{\text{base}}^{(N+1)}) \cdot \tfrac{1}{2}\left(\phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(0) - \phi_{\sigma_1}^{(1)} \circ \phi_{\sigma_2}^{(2)} \circ \ldots \circ \phi_{\sigma_N}^{(N)}(1)\right) + \Phi^{(1)} \circ \Phi^{(2)} \circ \ldots \circ \Phi^{(N)}(1).$$

Construction of $\overline{h}_\sigma^N$. To use $r_\sigma^{(N)}$ to construct $\overline{h}_\sigma^N$ as defined in Proposition B.1, we define

$$\overline{h}_\sigma^N = r_\sigma^{(N)} + 2 - r_\sigma^{(N)}(x_{\text{mid}}).$$

Since we only added a constant term to $r_\sigma^{(N)}$, its properties of continuity, 1-Lipschitzness, convexity, unique minimizer $x^* = x_{\text{mid}}$ in $(0, 1)$, and lower bound of $1/8$ on slope remain preserved. Next, note that since $r_\sigma^{(N)} \geq r_\sigma^{(N)}(x_{\text{mid}})$, it implies $\overline{h}_\sigma^N \geq 2$, which proves the assertion that $\overline{h}_\sigma^N : \mathbb{R} \mapsto [2, \infty)$. Finally, at $x = 0$, we have $\overline{h}_\sigma^N(0) = 2 + r_\sigma^{(N)}(0) - r_\sigma^{(N)}(x_{\text{mid}}) \leq 3$ because $r_\sigma^{(N)}$ is a piecewise affine function over $[0, 1]$ with maximum slope being 1, and so its maximum range of values can be 1. Thus, our constructed $\bar{h}$ satisfies Proposition B.1(i) and Proposition B.1(ii).

Proof of Proposition B.1(iii) Let $k = \frac{1}{4}\sqrt{\log\left(\frac{1}{\rho}\right)}$ for some $\rho > 0$ to be determined, and $N = k + 1$, $\sigma \sim \text{Unif}\{0, 1\}^N$. Consider the iterates of $\mathcal{A}$, $x_1, \ldots, x_T$, as random variables when applied to the (random) function $\overline{h}_\sigma^N$. Since $x^* = x_{\text{mid}} \in \mathbf{I}_{\sigma_1\ldots\sigma_N}$, we have

$$\Pr[\exists t \in [T] : |x_t - x^*| \leq \rho] \leq \Pr[\exists t \in [T] : \text{dist}(x_t, \mathbf{I}_{\sigma_1\ldots\sigma_N}) \leq \rho] \leq \Pr[\exists t \in [T] : x_t \in \mathbf{I}_{\sigma_1\ldots\sigma_k}],$$

where the second inequality follows from Lemma C.4(iv). By denoting the "progress tracking" stochastic process

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

we note that $Z_{t+1} - Z_t \geq 0$ with probability 1, and moreover that $\Pr[Z_{t+1} - Z_t = m] \leq 2^{-(m-1)}$ by Lemma C.9. Hence

$$\begin{aligned}\Pr[\exists t \in [T] : x_t \in \mathbf{I}_{\sigma_1 \dots \sigma_k}] = \Pr[Z_T \geq k] &\leq \frac{1}{k}\mathbb{E}[Z_T] = \frac{1}{k}\sum_{j=1}^{T}\mathbb{E}[Z_j - Z_{j-1}] \\ &= \frac{1}{k}\sum_{j=1}^{T}\sum_{m=0}^{\infty}\Pr[Z_j - Z_{j-1} = m] \leq \frac{1}{k}\sum_{j=1}^{T}\sum_{m=0}^{\infty}2^{-(m-1)} \\ &= \frac{4T}{k} = \frac{16T}{\sqrt{\log(1/\rho)}}\,.\end{aligned}$$

Finally, by setting $\rho := \exp(-256T^2/\gamma^2)$ the quantity above is bounded by γ, completing the proof.

■

Appendix D. Proof of Lemma B.4

The goal of this section is to prove Lemma B.4, which we first recall below.

Lemma B.4 *For the setup in Definition B.3 and* $\mathbf{w}$ *such that* $\mathbf{w} \perp \mathbf{e}_d$ *and* $\|\mathbf{w}\| = 1000\mu$*, the following hold:*

- *(i)* $f_{\mathbf{w},\mu}$ *is non-negative,* 1*-Lipschitz, and Clarke regular.*
- *(ii) For* $c = \frac{1}{100}$*, any* c*-stationary point* $\mathbf{x}$ *of* $f_{\mathbf{w},\mu}$ *satisfies* $f_{\mathbf{w},\mu}(\mathbf{x}) = 0$*. In particular, any* c*-stationary point of* $f_{\mathbf{w},\mu}$ *is a global minimum.*
- *(iii) There exist* $\mathbf{w} \in \mathbb{R}^d$*,* $\mu > 0$ *such that applying* $\mathcal{A}$ *to* $f_{\mathbf{w},\mu}$ *satisfies*

$$\Pr_{\mathcal{A}}[\exists t \in [T] : f_{\mathbf{w},\mu}(\mathbf{x}_t) < 1] < 2\gamma\,.$$

Proof Throughout the proof we omit the subscripts $\mathbf{w}, \mu$. We recall that by Definition B.3

$$f(\mathbf{x}) := \max\{h(\mathbf{x}) - \sigma_\mu(q(\mathbf{x} - \mathbf{x}^*)), 0\}\,.$$

Proof of Lemma B.4(i). It is clear that f is non-negative by design. The function h (in Definition B.3) is $\frac{1}{2}$-Lipschitz (by design in Lemma B.2); the function σ_μ (in Definition B.3) has the following derivative:

$$\sigma'_\mu(z) = \begin{cases} 0, & z \leq 0 \\ \frac{z}{4\mu}, & z \in (0, \mu] \\ \frac{1}{4}, & z > \mu \end{cases} \tag{D.1}$$

and is therefore $\frac{1}{4}$-Lipschitz. Further, $\mathbf{z} \mapsto \|\mathbf{z}\|$, $\mathbf{z} \mapsto \langle \bar{\mathbf{w}}, \mathbf{z}\rangle$, $\mathbf{z} \mapsto \max\{\mathbf{z}, 0\}$ and translations are all 1-Lipschitz. Finally, the summation (respectively, positive scaling) of functions results in a summation (respectively, scaling) of Lipschitz constants; and the composition of functions yields a product of Lipschitz constants. These imply that q (as in (B.2)) is $\frac{3}{2}$-Lipschitz and f is 1-Lipschitz.

To prove Clarke regularity of f, we start by examining the function

$$\phi(\mathbf{x}) := -\sigma_\mu(q(\mathbf{x})), \tag{D.2}$$

and note that ϕ is continuously differentiable over $\mathbb{R}^d \setminus \{-\mathbf{w}\}$ with

$$\nabla\phi(\mathbf{x}) = -\sigma'_\mu(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w}\rangle - \tfrac{1}{2}\|\mathbf{x}+\mathbf{w}\|) \cdot (\bar{\mathbf{w}} - \tfrac{1}{2}(\overline{\mathbf{x}+\mathbf{w}})),\ \forall \mathbf{x} \neq -\mathbf{w}.$$

In particular, by (D.1), we see that $\lim_{\mathbf{x}\to-\mathbf{w}} \nabla\phi(\mathbf{x}) = \mathbf{0}$. Further note that

$$\limsup_{\mathbf{v}\to\mathbf{0}} \frac{|\phi(-\mathbf{w}+\mathbf{v}) - \phi(-\mathbf{w})|}{\|\mathbf{v}\|} = \limsup_{\mathbf{v}\to\mathbf{0}} \frac{|\phi(-\mathbf{w}+\mathbf{v})|}{\|\mathbf{v}\|} = \limsup_{\mathbf{v}\to\mathbf{0}} \frac{|\sigma_\mu(\langle \bar{\mathbf{w}}, \mathbf{v}\rangle - \frac{1}{2}\|\mathbf{v}\|)|}{\|\mathbf{v}\|} = 0,$$

where the last equality follows from $\langle \bar{\mathbf{w}}, \mathbf{v}\rangle - \frac{1}{2}\|\mathbf{v}\| \overset{\mathbf{v}\to\mathbf{0}}{\longrightarrow} \mathbf{0}$ and $\lim_{z\to 0} \sigma'_\mu(z) = 0$. Therefore ϕ is differentiable at $-\mathbf{w}$ with $\nabla\phi(-\mathbf{w}) = \mathbf{0}$, hence everywhere continuously differentiable, which implies by Fact A.1(i) that for ϕ as defined in (D.2), it holds that

$$\phi \text{ is regular.} \tag{D.3}$$

Since ϕ is regular, the shifted function $\phi(\cdot - \mathbf{x}^*)$ is regular as well. Moreover, since h is convex and Lipschitz, it is regular as well (Fact A.1(ii)), hence by Fact A.1(iii) we have that

$$h(\cdot) + \phi(\cdot - \mathbf{x}^*) \text{ is regular.}$$

Finally, since $\max\{\cdot, 0\}$ is convex and Lipschitz, we conclude by Fact A.1(ii) that it is regular; hence, by Fact A.3, the composition $f_{\mathbf{w},\mu}(\mathbf{x}) = \max\{h(\mathbf{x}) - \sigma_\mu(q(\mathbf{x}-\mathbf{x}^*)), 0\}$ is regular.

Proof of Lemma B.4(ii). Our proof for this part of the lemma closely follows that by Tian and So [47] and Kornowski and Shamir [32]. We start by proving that, for $c < \frac{1}{100}$, the function $\varphi(\mathbf{x}) := h(\mathbf{x}+\mathbf{x}^*) - \sigma_\mu\left(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\|\right)$ has no c-stationary points, by an exhaustive case analysis showing a universal positive lower bound (of $\frac{1}{100}$) on the absolute value of the minimum norm element of its Clarke subdifferential in every case.

- $\mathbf{x} = \mathbf{0}$: By applying Fact A.2, the fact that $\|\mathbf{w}\| \gg \mu$, and (D.1), it holds that

  $$\partial\varphi(\mathbf{0}) = \left\{\partial h(\mathbf{x}^*) - \partial\sigma_\mu(\tfrac{1}{2}\|\mathbf{w}\|)\right\} = \left\{\partial h(\mathbf{x}^*) - \tfrac{1}{8}\bar{\mathbf{w}}\right\}.$$

  By further examining the definition of h (from Lemma B.2(iii)), we may simplify this to

  $$\partial\varphi(\mathbf{0}) = \left\{\partial h(\mathbf{x}^*) - \tfrac{1}{8}\bar{\mathbf{w}}\right\} \subseteq \left\{\tfrac{1}{32}\mathbf{u} + \lambda\mathbf{e}_d - \tfrac{1}{8}\bar{\mathbf{w}} \mid \lambda \in \partial\bar{h}(x^*), \|\mathbf{u}\| \le 1, \mathbf{u} \perp \mathbf{e}_d\right\},$$

  where $\bar{h}$ is as defined in Lemma B.2. Since projecting any vector from the above set $\partial\varphi(\mathbf{0})$ onto $\mathrm{span}(\mathbf{e}_d)^\perp$ cannot increase its norm, we may conclude that

  $$\left\|\tfrac{1}{32}\mathbf{u} + \lambda\mathbf{e}_d - \tfrac{1}{8}\bar{\mathbf{w}}\right\| \ge \left\|\tfrac{1}{32}\mathbf{u} - \tfrac{1}{8}\bar{\mathbf{w}}\right\| \ge \tfrac{1}{8} - \tfrac{1}{32} = \tfrac{3}{32}.$$

- $\mathbf{x} = -\mathbf{w}$: Recall in (D.3) we proved the Clarke regularity of $\phi(\mathbf{x}) := -\sigma_\mu(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\|)$. Applying this property at $-\mathbf{w}$ and noting, from (D.1), that $\partial\phi(-\mathbf{w}) = \mathbf{0}$, we get that

  $$\partial\varphi(-\mathbf{w}) = \{\partial h(-\mathbf{w}+\mathbf{x}^*) - \partial\phi(-\mathbf{w})\} \subset \left\{\tfrac{1}{32}\bar{\mathbf{w}} + \lambda\mathbf{e}_d - \mathbf{0} \mid \lambda \in \partial\bar{h}(-w_d + x^*)\right\},$$

  where we used the fact that $\mathbf{x}^* \perp \mathbf{e}_d$ (from Lemma B.2(iii)) and $\mathbf{w} \perp \mathbf{e}_d$. By projecting any such vector onto $\mathrm{span}(\mathbf{e}_d)^\perp$, we see that it clearly has norm of at least $\frac{1}{32}$.

- $x_d \neq 0$: It holds that

$$\partial\varphi(\mathbf{x}) \subseteq \left\{ \tfrac{1}{32}\mathbf{u} + \lambda \mathbf{e}_d - s\left(\bar{\mathbf{w}} - \tfrac{1}{2}\mathbf{v}\right) \mid \lambda \in \partial\bar{h}(x_d + x^*), \mathbf{u} \perp \mathbf{e}_d,\ s \in [0, \tfrac{1}{4}], \|\mathbf{v}\| \leq 1 \right\} ,$$

where we used (D.1) to evaluate $\partial\phi(\mathbf{x})$. Next, since $x_d \neq 0$, it implies that $x_d + x^* \neq x^*$, which implies, by Lemma B.2(ii), that $\forall \lambda \in \partial\bar{h}(x_d + x^*) : |\lambda| \geq \frac{1}{16}$. Then, projecting any vector from the set above onto $\mathbf{e}_d$, we see that

$$\left\| \tfrac{1}{32}\mathbf{u} + \lambda \mathbf{e}_d - s\left(\bar{\mathbf{w}} - \tfrac{1}{2}\mathbf{v}\right) \right\| \geq \left| |\lambda| - \tfrac{1}{4} \cdot \tfrac{1}{2} \right| \geq \tfrac{1}{16} .$$

- $x_d = 0,\ \mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\},\ \langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle < \frac{1}{2}$: Note that for any such $\mathbf{x}$, we have that $\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\| = \|\mathbf{x}+\mathbf{w}\| \cdot (\langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle - \frac{1}{2}) < 0$, which implies that $\sigma_\mu(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\|) = 0$. This in turn implies $\phi(\mathbf{x}) = 0$. Consequently, $\varphi(\mathbf{x})$ is locally identical to $h(\mathbf{x}+\mathbf{x}^*) = \bar{h}(x_d + x^*) + \frac{1}{32}\|\mathbf{x}_{1:d-1}\| = \bar{h}(x^*) + \frac{1}{32}\|\mathbf{x}_{1:d-1}\|$, where we used $\mathbf{x}^*_{1:d-1} = \mathbf{0}_{d-1}$ (from Lemma B.2(iii)) in the first step and $x_d = 0$ (as assumed in this case) in the final step. Since $x_d = 0$ yet $\mathbf{x} \neq \mathbf{0}$, it must be that $\mathbf{x}_{1:d-1} \neq \mathbf{0}_{d-1}$, thus

$$\partial\varphi(\mathbf{x}) = \left\{ \partial\bar{h}(x^*) + \tfrac{1}{32}\partial\|\mathbf{x}_{1:d-1}\| \right\} \subseteq \left\{ \tfrac{1}{32}\overline{\mathbf{x}_{1:d-1}} + \lambda \mathbf{e}_d \mid \lambda \in \partial h(x^*) \right\} .$$

For any $\mathbf{g} \in \partial\varphi(\mathbf{x})$, we have that $\|\mathbf{g}\| \geq \langle \mathbf{g}, \overline{\mathbf{x}_{1:d-1}} \rangle \geq \frac{1}{32}$.

- $x_d = 0,\ \mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\},\ \langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle > \frac{1}{2} + \frac{\mu}{\|\mathbf{x}+\mathbf{w}\|}$: In this case, we have $\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\| = \left(\langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle - \frac{1}{2}\right)\|\mathbf{x}+\mathbf{w}\| \geq \mu$. Then, noting from (B.2) the definition of σ_μ for the appropriate range of the argument yields that for any such $\mathbf{x}$: $\sigma_\mu(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\|) = \frac{1}{4}\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{8}\|\mathbf{x}+\mathbf{w}\| - \frac{\mu}{8}$. Combining this with $x_d = 0$, we have that $\varphi(\mathbf{x})$ is locally identical to

$$\mathbf{x} \mapsto \bar{h}(x^*) + \tfrac{1}{32}\|\mathbf{x}_{1:d-1}\| - \tfrac{1}{4}\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle + \tfrac{1}{8}\|\mathbf{x}+\mathbf{w}\| + \tfrac{\mu}{8} .$$

As in the previous case, since $x_d = 0$ yet $\mathbf{x} \neq \mathbf{0}$, we have that $\mathbf{x}_{1:d-1} \neq \mathbf{0}_{d-1}$, thus

$$\partial\varphi(\mathbf{x}) \subseteq \left\{ \tfrac{1}{32}\overline{\mathbf{x}_{1:d-1}} + \lambda \mathbf{e}_d - \tfrac{1}{4}\bar{\mathbf{w}} + \tfrac{1}{8}(\overline{\mathbf{x}+\mathbf{w}}) \mid \lambda \in \partial h(x^*) \right\} .$$

Hence, for $\mathbf{g} \in \partial\varphi(\mathbf{x})$, it holds that

$$\|\mathbf{g}\| \geq \langle \mathbf{g}, -\bar{\mathbf{w}} \rangle = -\frac{1}{32}\langle \bar{\mathbf{x}}, \bar{\mathbf{w}} \rangle + \frac{1}{4} - \frac{1}{8}\langle \overline{\mathbf{x}+\mathbf{w}}, \bar{\mathbf{w}} \rangle \geq \frac{1}{4} - \frac{5}{32} = \frac{3}{32} .$$

- $x_d = 0,\ \mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\},\ \|\mathbf{x}+\mathbf{w}\| \leq 10\mu,\ \langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle \in [\frac{1}{2}, \frac{1}{2} + \frac{\mu}{\|\mathbf{x}+\mathbf{w}\|}]$: We have that

$$\partial\varphi(\mathbf{x}) = \left\{ \tfrac{1}{32}\overline{\mathbf{x}_{1:d-1}} + \lambda \mathbf{e}_d - v \cdot (\bar{\mathbf{w}} - \tfrac{1}{2}\overline{\mathbf{x}+\mathbf{w}}) \mid \lambda \in \partial h(x^*) \right\} , \tag{D.4}$$

where $0 \leq v := \frac{1}{4\mu}(\langle \bar{\mathbf{w}}, \mathbf{x}+\mathbf{w} \rangle - \frac{1}{2}\|\mathbf{x}+\mathbf{w}\|) \leq 1$. The claimed bounds on v follow from the assumption $\langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}} \rangle \in [\frac{1}{2}, \frac{1}{2} + \frac{\mu}{\|\mathbf{x}+\mathbf{w}\|}]$. The final term in (D.4) comes from plugging in the appropriate argument in (D.1) following the deduced range of v. We now proceed to show a lower bound on $-\bar{\mathbf{w}}^\top \bar{\mathbf{x}}$, which we will use to show the desired lower bound on (D.4). To this end, define $\mathbf{x}' := -\frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{x}\|^2} \cdot \mathbf{x}$ (which is a valid operation because we assume $\mathbf{x} \neq \mathbf{0}$). We can verify that:

$$\langle \mathbf{x}', \mathbf{x}' + \mathbf{w} \rangle = \|\mathbf{x}'\|^2 + \mathbf{w}^\top \mathbf{x}' = 0, \tag{D.5}$$

where the last step follows from plugging in the definition of $\mathbf{x}'$. Next, by starting with the assumption that $10\mu \geq \|\mathbf{x}+\mathbf{w}\|$ and expressing $\mathbf{x}+\mathbf{w}$ as $\mathbf{x}-\mathbf{x}'+\mathbf{x}'+\mathbf{w}$, we have

$$100\mu^2 \geq \|\mathbf{x}+\mathbf{w}\|^2 = \|\mathbf{x}'+\mathbf{w}\|^2 + \|\mathbf{x}-\mathbf{x}'\|^2 + 2\left\langle \mathbf{x}-\mathbf{x}', \mathbf{x}'+\mathbf{w}\right\rangle. \tag{D.6}$$

Now, since $\langle \mathbf{x}', \mathbf{x}'+\mathbf{w}\rangle = 0$ (by (D.5)) and $\mathbf{x}$ is proportional to $\mathbf{x}'$ (by design of $\mathbf{x}'$), we have $\langle \mathbf{x}, \mathbf{x}'+\mathbf{w}\rangle = 0$ as well. Consequently, we have $\langle \mathbf{x}-\mathbf{x}', \mathbf{x}'+\mathbf{w}\rangle = 0$, which implies, in Inequality (D.6), that

$$100\mu^2 \geq \|\mathbf{x}'+\mathbf{w}\|^2 + \|\mathbf{x}-\mathbf{x}'\|^2. \tag{D.7}$$

By recalling that $\|\mathbf{w}\| = 1000\mu \geq \|\mathbf{x}+\mathbf{w}\|$, we have that $\mathbf{w}^\top \mathbf{x} \leq 0$. Combining this with the definition of $\mathbf{x}'$, we obtain $\bar{\mathbf{w}}^\top \bar{\mathbf{x}} = \bar{\mathbf{w}}^\top \bar{\mathbf{x}}'$, from which we conclude that

$$-\bar{\mathbf{w}}^\top \bar{\mathbf{x}} = -\bar{\mathbf{w}}^\top \bar{\mathbf{x}}' = \frac{\|\mathbf{x}'\|}{\|\mathbf{w}\|}. \tag{D.8}$$

Next, by again using $\langle \mathbf{x}', \mathbf{x}'+\mathbf{w}\rangle = 0$, we have $\|\mathbf{w}\|^2 = \|\mathbf{x}'\|^2 + \|\mathbf{x}'+\mathbf{w}\|^2$. Then, we have that $\|\mathbf{w}\|^2 = \|\mathbf{x}'\|^2 + \|\mathbf{x}'+\mathbf{w}\|^2 \leq \|\mathbf{x}'\|^2 + 100\mu^2$, where we used $\|\mathbf{x}'+\mathbf{w}\|^2 \leq 100\mu^2$ from Inequality (D.7). Rearranging and combining with (D.8) yields

$$-\bar{\mathbf{w}}^\top \bar{\mathbf{x}} = \frac{\|\mathbf{x}'\|}{\|\mathbf{w}\|} \geq \sqrt{1 - \frac{100\mu^2}{\|\mathbf{w}\|^2}}. \tag{D.9}$$

Next, we take the inner product of some element in $\partial\varphi(\mathbf{x})$ with $-\bar{\mathbf{w}}$ (as defined in (D.4)) and plug in Inequality (D.9) to obtain:

$$-\tfrac{1}{32}\bar{\mathbf{w}}^\top \bar{\mathbf{x}} + v - \tfrac{v}{2}\cdot \bar{\mathbf{w}}^\top(\overline{\mathbf{x}+\mathbf{w}}) \geq \tfrac{1}{32}\sqrt{1-\tfrac{100\mu^2}{\|\mathbf{w}\|^2}} > \tfrac{1}{33},$$

where the final step follows from our choice of $\mathbf{w}$ such that $\|\mathbf{w}\| = 1000\mu$.

- $x_d = 0,\ \mathbf{x} \notin \{\mathbf{0}, -\mathbf{w}\},\ \|\mathbf{x}+\mathbf{w}\| \geq 10\mu,\ \langle \bar{\mathbf{w}}, \overline{\mathbf{x}+\mathbf{w}}\rangle \in [\frac{1}{2}, \frac{1}{2}+\frac{\mu}{\|\mathbf{x}+\mathbf{w}\|}]$: As in (D.4) in the previous case, we have for $0 \leq v \leq 1$ that

$$\begin{aligned}\partial\varphi(\mathbf{x}) &= \left\{ \tfrac{1}{32}\overline{\mathbf{x}_{1:d-1}} + \lambda \mathbf{e}_d - v\cdot(\bar{\mathbf{w}} - \tfrac{1}{2}\overline{\mathbf{x}+\mathbf{w}}) \mid \lambda \in \partial h(x^*), v \in [0,1]\right\}, \\ &= \left\{ \lambda \mathbf{e}_d + \left(\tfrac{1}{32\|\mathbf{x}\|} + \tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x} + \left(\tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|} - \tfrac{v}{\|\mathbf{w}\|}\right)\mathbf{w} \mid \lambda \in \partial h(x^*), v \in [0,1]\right\}.\end{aligned} \tag{D.10}$$

Denote $\mathbf{x} = \mathbf{x}_| + \mathbf{x}_\perp$, where $\mathbf{x}_\perp = (\mathbf{I} - \bar{\mathbf{w}}\bar{\mathbf{w}}^\top)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathrm{span}(\mathbf{w})^\perp$, and $\mathbf{x}_| \in \mathrm{span}(\mathbf{w})$. Recall also that $x_d = 0$ (by assumption in this case) and $w_d = 0$ (by our choice of $\mathbf{w}$). For any $\mathbf{g} \in \partial\varphi(\mathbf{x})$, we can therefore write, for some scalar α, that

$$\begin{aligned}\|\mathbf{g}\| &\geq \left\| \left(\tfrac{1}{32\|\mathbf{x}\|} + \tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x} + \left(\tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|} - \tfrac{v}{\|\mathbf{w}\|}\right)\mathbf{w}\right\| \\ &= \left\| \left(\tfrac{1}{32\|\mathbf{x}\|} + \tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x}_\perp + \alpha\mathbf{w}\right\| \\ &\geq \left\| \left(\tfrac{1}{32\|\mathbf{x}\|} + \tfrac{v}{2\|\mathbf{x}+\mathbf{w}\|}\right)\mathbf{x}_\perp\right\| \\ &\geq \tfrac{\|\mathbf{x}_\perp\|}{32\|\mathbf{x}\|},\end{aligned} \tag{D.11}$$

where the first step is by projecting onto $\mathrm{span}(\mathbf{e}_d)^\perp$; the second step is by splitting $\mathbf{x}$ into $\mathbf{x}_| + \mathbf{x}_\perp$ and absorbing $\mathbf{x}_|$ into the term written as a multiple of $\mathbf{w}$ (valid since $\mathbf{x}_| \in \mathrm{span}(\mathbf{w})$); the third step is because $\mathbf{x}_\perp \perp \mathbf{w}$, and so the norm of their sum is at least as large as each of them; the fourth step is by $v \geq 0$. Further, since $\mathbf{I} - \bar{\mathbf{w}}\bar{\mathbf{w}}^\top$ is an orthogonal projection, we have

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

Plugging into Inequality (D.11) yields $\|\mathbf{g}\| \geq \frac{1}{32}\sqrt{1 - \langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle^2}$, where the square root operation is valid because $|\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle| \leq 1$. Now, suppose that there exists a $\mathbf{g} \in \partial\varphi(\mathbf{x})$ with $\|\mathbf{g}\| \leq \epsilon \leq \frac{1}{32}$ (note that if $\|\mathbf{g}\| \geq \frac{1}{32}$ for all $\mathbf{g} \in \partial\varphi(\mathbf{x})$, then we are done). It then follows that

$$\tfrac{1}{32}\sqrt{1 - \langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle^2} \leq \epsilon. \tag{D.12}$$

Next, the assumed range implies

$$\tfrac{1}{2} \leq \langle \bar{\mathbf{w}}, \overline{\mathbf{x} + \mathbf{w}}\rangle \leq \tfrac{3}{5}. \tag{D.13}$$

If $\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle \leq 0$, then by Inequality (D.12), it must be that $\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle \leq -\sqrt{1 - 1024\epsilon^2}$. Consider any $\mathbf{u} \in \partial\varphi(\mathbf{x})$; plugging Inequality (D.13) into (D.10), we have

$$\langle \mathbf{u}, \bar{\mathbf{w}}\rangle = \tfrac{1}{32}\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle - v \cdot (1 - \tfrac{1}{2} \cdot \tfrac{3}{5}) \leq -\tfrac{1}{32}\sqrt{1 - 1024\epsilon^2},$$

which implies that $\|\mathbf{u}\| \geq \frac{1}{32}\sqrt{1 - 1024\epsilon^2}$ for any $\mathbf{u} \in \partial\varphi(\mathbf{x})$. Thus, if $\|\mathbf{u}\| \leq \epsilon$, then chaining the two inequalities yields $\epsilon \geq \frac{1}{32}\sqrt{1 - 1024\epsilon^2}$. This implies that $\epsilon \geq \frac{1}{\sqrt{2048}}$. On the other hand, if $\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle \geq 0$, then combining with Inequality (D.12) gives $\langle \bar{\mathbf{w}}, \bar{\mathbf{x}}\rangle \geq \sqrt{1 - 1024\epsilon^2}$. Hence,

$$\langle \bar{\mathbf{w}}, \mathbf{x} + \mathbf{w}\rangle = \langle \bar{\mathbf{w}}, \mathbf{x}\rangle + \|\mathbf{w}\| \geq \|\mathbf{x}\|\sqrt{1 - 1024\epsilon^2} + \|\mathbf{w}\| \geq \sqrt{1 - 1024\epsilon^2}\|\mathbf{x} + \mathbf{w}\|.$$

If $\epsilon < \frac{1}{50}$, then $\frac{3}{5} \geq \langle \bar{\mathbf{w}}, \overline{\mathbf{w} + \mathbf{x}}\rangle > \sqrt{1 - \frac{1024}{2500}}$, which is a contradiction. Thus, combining both the cases, we see that the lower bound must be at least $\min\left(\frac{1}{\sqrt{2048}}, \frac{1}{50}\right)$.

From the above analysis, we conclude that $\varphi(\mathbf{x}) = h(\mathbf{x}+\mathbf{x}^*) - \sigma_\mu\left(\langle \bar{\mathbf{w}}, \mathbf{x} + \mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x} + \mathbf{w}\|\right)$ has no c-stationary points for $c \leq \frac{1}{100}$. We now show that, for $c = \frac{1}{100}$, any c-stationary point of $f(\mathbf{x}) = \max\{\varphi(\mathbf{x} - \mathbf{x}^*), 0\}$ (which matches the definition of f in (B.1) combined with Theorem B.5) satisfies $f(\mathbf{x}) = 0$. Indeed, if there existed $\mathbf{x}$ with $\left\|\bar{\partial} f(\mathbf{x})\right\| \leq c$ and $f(\mathbf{x}) > 0$, then we note by the latter that $f(\cdot) = \varphi(\cdot - \mathbf{x}^*)$ in an open neighborhood of $\mathbf{x}$, thus $\left\|\bar{\partial}\varphi(\mathbf{x} - \mathbf{x}^*)\right\| = \left\|\bar{\partial} f(\mathbf{x})\right\| \leq c$, which is a contradiction by our earlier claim on c-stationarity of φ.

Proof of Lemma B.4(iii). We denote by $(\mathbf{x}_t^h)_{t=1}^T$ the (possibly random) iterates produced by $\mathcal{A}$ when applied to h. We will first show that for some fixed $\mathbf{w} \in \mathbb{R}^d$:

$$\Pr_{\mathcal{A}}\left[\min_{t\in[T]}\left\|\mathbf{x}_t^h - \mathbf{x}^*\right\| \geq \rho \text{ and } \max_{t\in[T]}\left\langle \bar{\mathbf{w}}, \overline{\mathbf{x}_t^h - \mathbf{x}^*}\right\rangle < \frac{1}{3}\right] \geq 1 - 2\gamma\,. \tag{D.14}$$

To see why, recall that by Lemma B.2 we know that $\Pr_{\mathcal{A}}[\min_{t\in[T]} \left\|\mathbf{x}_t^h - \mathbf{x}^*\right\| \geq \rho] \geq 1-\gamma$. Furthermore, letting $\mathbf{w} \in \mathbb{R}^d$ be a random vector that satisfies $\mathbf{w}_{1:d-1} \sim \mathrm{Unif}(\frac{\rho}{99} \cdot \mathbb{S}^{d-2})$, $\Pr[w_d = 0] = 1$, and noting that $(\mathbf{x}_t^h)_{t=1}^T, \mathbf{x}^*$ are independent of $\mathbf{w}$ and that $\bar{\mathbf{w}}_{1:d-1} \sim \mathrm{Unif}(\mathbb{S}^{d-2})$, applying a standard tail bound on the inner product of a uniformly random unit vector (cf. [2, Lemma 2.2]) we get

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

By the union bound, we see that Inequality (D.14) holds with probability at least $1-2\gamma$ over the joint probability of $\mathbf{w}, \mathcal{A}$, thus (via the probabilistic method argument) there exists some fixed $\mathbf{w}$ for which it holds over the randomness of $\mathcal{A}$. We therefore fix $\mathbf{w}$ so that Inequality (D.14) holds, and assume the high probability event indeed occurs. We aim to show that under this event, for all $t \in [T]:\ f(\mathbf{x}_t) \geq 1$, which will then conclude the proof. Indeed, under this event, we see that

$$\max_{t\in[T]} \left\langle \bar{\mathbf{w}}, \overline{\mathbf{x}_t^h - \mathbf{x}^*} \right\rangle < \frac{1}{3} < \frac{1}{2} - \frac{1}{66} \leq \frac{1}{2} - \frac{\rho}{66\left\|\mathbf{x}_t^h - \mathbf{x}^*\right\|}\,.$$

Further note that for any $\mathbf{x} \neq \mathbf{x}^*$, if $\left\langle \bar{\mathbf{w}}, \overline{\mathbf{x}-\mathbf{x}^*} \right\rangle < \frac{1}{2} - \frac{\rho}{66\left\|\mathbf{x}_t^h - \mathbf{x}^*\right\|}$ then since $\|\mathbf{w}\| = \frac{\rho}{99}$ a straightforward calculation yields $\langle \bar{\mathbf{w}}, \mathbf{x}-\mathbf{x}^*+\mathbf{w}\rangle - \frac{1}{2}\|\mathbf{x}-\mathbf{x}^*+\mathbf{w}\| < 0$. As this is an open condition with respect to $\mathbf{x}$, by construction of f this implies that $f(\cdot) = h(\cdot)$ in a neighborhood of $\mathbf{x}$. We therefore get that for all $t \in [T]: h \equiv f$ in a neighborhood of $\mathbf{x}_t^h$, so in particular we see that $\mathbf{x}_t^h = \mathbf{x}_t$, namely applying $\mathcal{A}$ to h results in the same iterate sequence as if the algorithm were applied to f. Thus

$$\min_{t\in[T]} f(\mathbf{x}_t) = \min_{t\in[T]} f(\mathbf{x}_t^h) = \min_{t\in[T]} h(\mathbf{x}_t^h) \geq 1\,.$$

■