An Algorithm with Optimal Dimension-Dependence for Zero-Order Nonsmooth Nonconvex Stochastic Optimization

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

Abstract

We study the complexity of producing (δ, ϵ)-stationary points of Lipschitz objectives which are possibly neither smooth nor convex, using only noisy function evaluations. Recent works proposed several stochastic zero-order algorithms that solve this task, all of which suffer from a dimension-dependence of $\Omega(d^{3/2})$ where d is the dimension of the problem, which was conjectured to be optimal. We refute this conjecture by providing a faster algorithm that has complexity $O(d\delta^{-1}\epsilon^{-3})$, which is optimal (up to numerical constants) with respect to d and also optimal with respect to the accuracy parameters δ, ϵ, thus solving an open question due to Lin et al. [14]. Moreover, the convergence rate achieved by our algorithm is also optimal for smooth objectives, proving that in the nonconvex stochastic zero-order setting, *nonsmooth optimization is as easy as smooth optimization*. We provide algorithms that achieve the aforementioned convergence rate in expectation as well as with high probability. Our analysis is based on a simple yet powerful geometric lemma regarding the Goldstein-subdifferential set, which allows utilizing recent advancements in first-order nonsmooth nonconvex optimization.

1. Introduction

We consider the problem of optimizing a stochastic objective of the form

$$f(\mathbf{x}) = \mathbb{E}_{\xi\sim\Xi}[F(\mathbf{x};\xi)]$$

where the stochastic components $F(\cdot\,;\xi) : \mathbb{R}^d \to \mathbb{R}$ are Lipschitz continuous, yet possibly not smooth nor convex. We consider stochastic zero-order (also known as *gradient-free* or *derivative-free*) algorithms that have access only to noisy function evaluations. At each time step, the algorithm draws $\xi \sim \Xi$ and can observe $F(\mathbf{x}, \xi)$ for points $\mathbf{x} \in \mathbb{R}^d$ of its choice. Problems of this type arise throughout machine learning, control theory and finance, in applications in which gradients are expensive (or even impossible) to evaluate, see for example the book by Spall [17] for an overview. Although in the convex setting the complexity of such algorithms is relatively well understood [1, 8, 15, 16], much less is known about the nonsmooth nonconvex setting, which is of major interest in modern deep learning applications.

Recently, Lin et al. [14] proposed a gradient-free algorithm that produces a (δ, ϵ)-stationary point using $O(d^{3/2}\delta^{-1}\epsilon^{-4})$ function evaluations. Following Zhang et al. [20], recall that a point $\mathbf{x}$ is called a (δ, ϵ)-stationary point if there exists a convex combination of gradients in a δ-neighborhood of $\mathbf{x}$ whose norm is less than ϵ (see Section 2 for a reminder of relevant definitions). Lin et al. [14] posed the question as to whether this super-linear dimension dependence is inevitable or not. The aforementioned complexity was very recently improved to $O(d^{3/2}\delta^{-1}\epsilon^{-3})$ by Chen et al. [4], yet notably, this result still suffers from the same super-linear dimension dependence.

In particular, as pointed out by Lin et al., this dimension dependence is $\Omega(\sqrt{d})$ worse than that of stochastic zero-order *smooth* nonconvex optimization, a setting in which it is possible to find an ϵ-stationary point (i.e. $\mathbf{x}$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$) using $O(d\epsilon^{-4})$ noisy function evaluations [10]. This led the authors to conjecture that stochastic zero-order nonsmooth nonconvex optimization is "likely to be intrinsically harder" than its smooth counterpart.

Our main contribution resolves this open question, showing that this is actually *not* the case. We propose a faster zero-order algorithm for nonsmooth nonconvex optimization, which requires only $O(d\delta^{-1}\epsilon^{-3})$ noisy function evaluations. This complexity has an optimal linear dimension-dependence, while also obtaining the optimal dependence with respect to δ and ϵ (as we will soon argue). Moreover, if $f(\cdot)$ is smooth, the algorithm automatically recovers the best-known $O(d\epsilon^{-4})$ complexity of stochastic gradient-free smooth nonconvex optimization, implying that in the stochastic zero-order setting, *nonsmooth nonconvex optimization is as easy as smooth nonconvex optimization*. Whether this property holds was originally raised as an open question by Zhang et al. [20] in the context of first-order algorithms (that have access to gradient information), and was recently confirmed by Cutkosky et al. [6]. Our result extends the resolution of this question to the case of zero-order algorithms.

As previously mentioned, these ramifications readily show that indeed the dependence on δ, ϵ we obtain is optimal, since the implied ϵ^{-4} factor is known to be inevitable even in the strictly-easier case of stochastic first-order smooth optimization with exact function evaluations [2].[1] Moreover, the linear dimension dependence is well-known to be inevitable for gradient-free algorithms even in the context of smooth or convex optimization [8]. Interestingly, in terms of the dependence on δ and ϵ, the convergence rate we obtain is as fast as the currently best-known *deterministic first-order* algorithms for nonsmooth nonconvex optimization [7, 18, 20]. This is in stark contrast to smooth nonconvex optimization, in which optimal stochastic and deterministic methods have disparate complexities on the order of ϵ^{-4} and ϵ^{-2}, respectively [2, 3]. Finally, we also note that since our derived convergence rate is the same for stochastic and deterministic objectives (i.e. noiseless, when $\Xi = \{\xi\}$), our algorithm is a factor of $\Omega(\sqrt{d})$ faster even than the previously best-known rate for deterministic zero-order nonsmooth nonconvex optimization.

2. Preliminaries.

Notation. We use bold-faced font to denote vectors, e.g. $\mathbf{x} \in \mathbb{R}^d$, and denote by $\|\mathbf{x}\|$ the Euclidean norm. We denote by $[n] := \{1, \ldots, n\}$, $\mathbb{B}(\mathbf{x}, \delta) := \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \leq \delta\}$, and by $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ the unit sphere. We denote by $\mathrm{conv}(\cdot)$ the convex hull operator, and by $\mathrm{Unif}(A)$ the uniform measure over a set A. We use the standard big-O notation, with $O(\cdot)$, $\Theta(\cdot)$ and $\Omega(\cdot)$ hiding absolute constants that do not depend on problem parameters, $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ hiding absolute constants and additional logarithmic factors.

Nonsmooth analysis. We call a function $f : \mathbb{R}^d \to \mathbb{R}$ L-Lipschitz if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d : |f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|$, and H-smooth if it is differentiable and $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is H-Lipschitz, namely for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq H \|\mathbf{x} - \mathbf{y}\|$. By Rademacher's theorem, Lipschitz functions are differentiable almost everywhere (in the sense of Lebesgue). Hence, for any Lipschitz

1. While the lower bound construction in Arjevani et al. [2] is not globally Lipschitz, a slight modification of it which appears in Cutkosky et al. [6, Appendix F] is.

function $f : \mathbb{R}^d \to \mathbb{R}$ and point $\mathbf{x} \in \mathbb{R}^d$ the Clarke subgradient set [5] can be defined as

$$\partial f(\mathbf{x}) := \mathrm{conv}\{\mathbf{g} : \mathbf{g} = \lim_{n\to\infty} \nabla f(\mathbf{x}_n),\ \mathbf{x}_n \to \mathbf{x}\}\ ,$$

namely, the convex hull of all limit points of $\nabla f(\mathbf{x}_n)$ over all sequences of differentiable points which converge to $\mathbf{x}$. Note that if the function is continuously differentiable at a point or convex, the Clarke subdifferential reduces to the gradient or subgradient in the convex analytic sense, respectively. We say that a point $\mathbf{x}$ is an ϵ-stationary point of $f(\cdot)$ if $\min\{\|\mathbf{g}\| : \mathbf{g} \in \partial f(\mathbf{x})\} \le \epsilon$. Furthermore, given $\delta \ge 0$ the Goldstein δ-subdifferential [11] of f at $\mathbf{x}$ is the set

$$\partial_\delta f(\mathbf{x}) := \mathrm{conv}\left(\cup_{\mathbf{y}\in\mathbb{B}(\mathbf{x},\delta)} \partial f(\mathbf{y})\right)\ ,$$

namely all convex combinations of gradients at points in a δ-neighborhood of $\mathbf{x}$.

Definition 1 *Given a Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, a point $\mathbf{x} \in \mathbb{R}^d$ and $\delta \ge 0$, denote $\|\nabla f(\mathbf{x})\|_\delta := \min\{\|\mathbf{g}\| : \mathbf{g} \in \partial_\delta f(\mathbf{x})\}$. A point $\mathbf{x}$ is called a (δ, ϵ)-stationary point of $f(\cdot)$ if*

$$\|\nabla f(\mathbf{x})\|_\delta \le \epsilon\ .$$

Note that a point is ϵ-stationary if and only if it is (δ, ϵ)-stationary for all $\delta \ge 0$ [20, Lemma 7]. Moreover, if f is H-smooth and $\mathbf{x}$ is a $(\frac{\epsilon}{3H}, \frac{\epsilon}{3})$-stationary point of f, then it is also ϵ-stationary [20, Proposition 6].

Randomized smoothing. Given a Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we define its uniform smoothing

$$f_\delta(\mathbf{x}) := \mathbb{E}_{\mathbf{z}\sim\mathrm{Unif}(\mathbb{B}(\mathbf{0},1))}[f(\mathbf{x} + \delta\mathbf{z})]\ .$$

It is well known (cf. 19) that if f is L_0-Lipschitz, then f_δ is L_0-Lipschitz; f_δ is $O(\sqrt{d}L_0\delta^{-1})$-smooth; and $|f(\mathbf{x}) - f_\delta(\mathbf{x})| \le \delta L_0$ for all $\mathbf{x} \in \mathbb{R}^d$.

Setting. We consider optimization objectives of the form $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\Xi}[F(\mathbf{x};\xi)]$, where $\xi \sim \Xi$ is a random variable. We impose the assumption that the stochastic components $F(\,\cdot\,;\xi) : \mathbb{R}^d \to \mathbb{R}$ are Lipschitz continuous, possibly with a varying Lipschitz constant:

Assumption 2 *For any ξ, the function $F(\,\cdot\,;\xi)$ is $L(\xi)$-Lipschitz. Moreover, we assume $L(\xi)$ has a bounded second moment: Namely, there exists $L_0 > 0$ such that*

$$\mathbb{E}_{\xi\sim\Xi}[L(\xi)^2] \le L_0^2\ .$$

We note that Assumption 2 is weaker than assuming $F(\,\cdot\,;\xi)$ is L_0-Lipschitz for all ξ. We also remark that in the deterministic case, namely when Ξ is supported on a single point, the optimization problem reduces to that of a L_0-Lipschitz objective using exact evaluations.

3. Algorithms and Main Results

Before formally presenting our main result, we find it insightful to stress out the key idea, and in particular how our algorithm differs from those of Lin et al. [14], Chen et al. [4]. The main strategy employed by both of these papers is based on the following result.

Proposition 3 (14, Theorem 3.1) *For any $\delta \geq 0 : \nabla f_\delta(\mathbf{x}) \in \partial_\delta f(\mathbf{x})$. Hence, if $\mathbf{x}$ is an ϵ-stationary point of f_δ, then it is a (δ, ϵ)-stationary point of f.*

Following this observation, both papers set out to design algorithms that produce an ϵ-stationary point of f_δ. A well known technique (which we formally recall later on) allows to use two possibly noisy evaluations of f in order to produce a stochastic *first-order* oracle of f_δ whose second moment is bounded by $\sigma^2 = O(d)$. Noting that f_δ is $L_1 = O(\sqrt{d}/\delta)$-smooth, the standard analysis of stochastic gradient descent (SGD) for smooth nonconvex optimization shows that it obtains an ϵ-stationary point of f_δ within $O(\sigma^2 L_1 \epsilon^{-4}) = O(d^{3/2}\delta^{-1}\epsilon^{-4})$ oracle calls, recovering the main result of Lin et al. [14]. The improved ϵ-dependence due to Chen et al. [4] was achieved by employing a variance-reduction method instead of plain SGD, though other than that, their main algorithmic strategy and analysis are the same.

Moreover, the algorithmic strategy we have described seems to reveal a barrier, (mistakenly) suggesting the $d^{3/2}$ dependence is unavoidable. Indeed, it is relatively straightforward to see that any gradient estimator which is based on a constant number of function evaluations must have variance of at least $\sigma^2 = \Omega(d)$, while it is also known that any efficient smoothing technique must suffer from a smoothness parameter of at least $L_1 = \Omega(\sqrt{d})$ [13]. Since the complexity of any stochastic first-order method for smooth nonconvex optimization must scale at least as $\Omega(\sigma^2 L_1)$ [2] which in this case is unavoidably $\Omega(d^{3/2})$, we are stuck with this factor.

The main technical ingredient that allows us to reduce this factor is the following geometric result which examines the Goldstein δ-subdifferential set under randomized smoothing.

Lemma 4 *For any $\delta, \gamma \geq 0 : \partial_\gamma f_\delta(\mathbf{x}) \subseteq \partial_{\delta+\gamma} f(\mathbf{x})$. Hence, if $\mathbf{x}$ is an (γ, ϵ)-stationary point of f_δ, then it is a $(\delta + \gamma, \epsilon)$-stationary point of f. In particular, any $(\frac{\delta}{2}, \epsilon)$-stationary point of $f_{\delta/2}$ is a (δ, ϵ)-stationary point of f.*

Note that the lemma above strictly generalizes Proposition 3 [14, Theorem 3.1] which is readily recovered by plugging $\gamma = 0$. The utility of this result is that it allows to replace the task of finding an ϵ-stationary point of f_δ to that of finding a (δ, ϵ)-stationary point of it (disregarding a constant factor multiplying δ). To see why this is beneficial, recall that while f_δ is $O(\sqrt{d} L_0 \delta^{-1})$-smooth, it is merely L_0-Lipschitz! Thus using a stochastic first-order *nonsmooth* nonconvex algorithm which scales with the Lipschitz parameter (instead of the smoothness parameter), we save a whole $\Omega(\sqrt{d})$ factor, yielding the optimal dimension dependence. It is interesting to note that treating the smoothed objective as if it were nonsmooth leads to such an improvement.

In particular, using the optimal stochastic first-order algorithm of Cutkosky et al. [6] that has complexity $O(\sigma^2 \delta^{-1} \epsilon^{-3})$, as described in Algorithm 1, results in the following convergence guarantee:[2]

Theorem 5 *Let $\delta, \epsilon \in (0, 1)$, and suppose $f(\mathbf{x}_0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leq \Delta$. Under Assumption 2, there exists*

$$T = O\left(d \cdot \left(\frac{\Delta L_0^2}{\delta \epsilon^3} + \frac{L_0^3}{\epsilon^3}\right)\right)$$

2. Notably, using the stochastic algorithm of Zhang et al. [20] (instead of Algorithm 1), when paired with our analysis, yields the desired linear dimension dependence as well – albeit with with a worse convergence rate with respect to ϵ, on the order of $d\delta^{-1}\epsilon^{-4}$.

Algorithm 1 Optimal Stochastic Nonsmooth Nonconvex Optimization Algorithm

Input: Initialization $\mathbf{x}_0 \in \mathbb{R}^d$, smoothing parameter $\delta' > 0$, clipping parameter $D > 0$, step size $\eta > 0$, iteration budget $T \in \mathbb{N}$
Initialize: $\boldsymbol{\Delta}_1 = \mathbf{0}$
for $t = 1, \dots, T$ **do**
  Sample $\xi_t \sim \Xi$
  Sample $s_t \sim \mathrm{Unif}[0, 1]$
  $\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\Delta}_t$
  $\mathbf{z}_t = \mathbf{x}_{t-1} + s_t \boldsymbol{\Delta}_t$
  $\mathbf{g}_t = \textsc{GradEstimator}(\mathbf{z}_t, \delta', \xi_t)$ ▷ `Uses two noisy function evaluations`
  $\boldsymbol{\Delta}_{t+1} = \min\left(1, \frac{D}{\|\boldsymbol{\Delta}_t - \eta \mathbf{g}_t\|}\right) \cdot (\boldsymbol{\Delta}_t - \eta \mathbf{g}_t)$
end for
$M = \lfloor \frac{\delta'}{D} \rfloor,\ K = \lfloor \frac{T}{M} \rfloor$
for $k = 1, \dots, K$ **do**
  $\overline{\mathbf{x}}_k = \frac{1}{M} \sum_{m=1}^{M} \mathbf{z}_{(k-1)M+m}$
end for
Sample $\mathbf{x}^{\mathrm{out}} \sim \mathrm{Unif}\{\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_K\}$
Output: $\mathbf{x}^{\mathrm{out}}$.

Algorithm 2 $\textsc{GradEstimator}(\mathbf{x}, \delta', \xi)$

Input: Point $\mathbf{x} \in \mathbb{R}^d$, smoothing parameter $\delta' > 0$, random seed ξ.
Sample $\mathbf{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
Evaluate $F(\mathbf{x} + \delta'\mathbf{w}; \xi)$ and $F(\mathbf{x} - \delta'\mathbf{w}; \xi)$
$\mathbf{g} = \frac{d}{2\delta'}(F(\mathbf{x} + \delta'\mathbf{w}; \xi) - F(\mathbf{x} - \delta'\mathbf{w}; \xi))\mathbf{w}$ ▷ `Unbiased estimator of` $\nabla f_{\delta'}(\mathbf{x})$
Output: $\mathbf{g}$.

such that setting $\delta' = \frac{\delta}{2},\ \eta = \Theta\left(\frac{\Delta + \delta L_0}{d L_0^2 T}\right),\ D = \Theta\left(\left(\frac{(\Delta + \delta L_0)\sqrt{\delta}}{\sqrt{d} L_0 T}\right)^{2/3}\right)$, *and running Algorithm 1 with Algorithm 2 as a subroutine, outputs a point* $\mathbf{x}^{\mathrm{out}}$ *satisfying* $\mathbb{E}[\left\|\nabla f(\mathbf{x}^{\mathrm{out}})\right\|_\delta\}] \leq \epsilon$ *using* $2T$ *noisy function evaluations.*

Remark 6 (Parallel complexity) *At each iteration, Algorithm 1 determines* $\mathbf{g}_t$ *by calling Algorithm 2 (*GRADESTIMATOR*), which requires* 2 *evaluations of* $F(\cdot\,; \xi_t)$. *More generally,* $\mathbf{g}_t$ *can be set as the average of* k *independent, possibly parallel calls to this subroutine, which would require* $2k$ *function evaluations. An easy generalization of Lemma 8 shows this would result in the second-moment bound on the order of*

$$\mathbb{E}[\|\mathbf{g}_t\|^2 \,|\mathbf{x}_t, s_t, \boldsymbol{\Delta}_t] \lesssim \frac{d L_0^2}{k} + \|\mathbb{E}[\mathbf{g}_t]\|^2 \leq L_0^2 \left(\frac{d}{k} + 1\right) .$$

With the rest of the proof of Theorem 5 as is, this yields an expected number of rounds of

$$T = O\left(\left(\frac{d}{k} + 1\right) \cdot \left(\frac{\Delta L_0^2}{\delta \epsilon^3} + \frac{L_0^3}{\epsilon^3}\right)\right) ,$$

though the total number of queries would be k times larger than above, and equal to

$$O\left((d+k)\cdot\left(\frac{\Delta L_0^2}{\delta\epsilon^3}+\frac{L_0^3}{\epsilon^3}\right)\right).$$

In particular, letting $k=\Theta(d)$ removes the dimension dependence in the parallel complexity altogether, while maintaining the same complexity overall (up to a constant).

High probability guarantee. While Theorem 1 shows that Algorithm 1 yields the desired expected complexity, many practical applications require high probability bounds, namely producing a point $\mathbf{x}$ such that

$$\Pr[\|\nabla f(\mathbf{x})\|_\delta \le \epsilon] \ge 1-\gamma$$

for some small $\gamma>0$. A naive application of Markov's inequality to the expected complexity shows that Algorithm 1 produces such a point within

$$O\left(d\cdot\left(\frac{\Delta L_0^2}{\delta\epsilon^3\gamma^3}+\frac{L_0^3}{\epsilon^3\gamma^3}\right)\right) \tag{1}$$

noisy function evaluations, which is rather crude with respect to the probability parameter γ. Adapting a technique due to Ghadimi and Lan [10] to our setting, we can design an algorithm with a significantly tighter high-probability bound. The original idea of Ghadimi and Lan [10] for the case of smooth stochastic optimization, which was also used by Lin et al. [14], consists of several independent calls to the main algorithm, yielding a list of candidate points. Subsequently, a post-optimization phase estimates the gradient norm of any such point, returning the minimal — which is likely to succeed due to a concentration argument. We note that adapting this technique to our setting is not trivial, since the post-optimization phase should attempt at estimating $\|\nabla f(\cdot)\|_\delta$ rather than $\|\nabla f(\cdot)\|$, which is hard in general. Luckily, using Lemma 4, the former can be bounded by $\left\|\nabla f_{\delta/2}(\cdot)\right\|_{\delta/2}$, which in turn can be bounded (with high probability) using a sequence of evaluations at nearby points. This procedure is described in Algorithm 3, whose convergence rate is presented in the following theorem.

Theorem 7 *Let $\gamma,\delta,\epsilon\in(0,1)$, and suppose $f(\mathbf{x}_0)-\inf_{\mathbf{x}} f(\mathbf{x})\le\Delta$. Under Assumption 2, there exist $T=O\left(d\cdot\left(\frac{\Delta L_0^2}{\delta\epsilon^3}+\frac{L_0^3}{\epsilon^3}\right)\right)$, $R=O(\log(1/\gamma))$, $S=O\left(\frac{\log(1/\gamma)}{\gamma}\right)$ such that setting $\delta'=\frac{\delta}{2}$, $\eta=\Theta\left(\frac{\Delta+\delta L_0}{dL_0^2T}\right)$, $D=\Theta\left(\left(\frac{(\Delta+\delta L_0)\sqrt{\delta}}{\sqrt{d}L_0T}\right)^{2/3}\right)$, and running Algorithm 3 outputs a point $\mathbf{x}^{\mathrm{out}}$ satisfying*

$$\Pr\left[\left\|\nabla f(\mathbf{x}^{\mathrm{out}})\right\|_\delta\} \le \epsilon\right] \ge 1-\gamma$$

using

$$O\left(d\cdot\left(\frac{\Delta L_0^2\log(1/\gamma)}{\delta\epsilon^3}+\frac{L_0^3\log(1/\gamma)}{\epsilon^3}+\frac{L_0^2\log^2(1/\gamma)}{\gamma\epsilon^2}\right)\right)$$

noisy function evaluations.

Notably, the number of function evaluations guaranteed by the theorem above is significantly smaller than in Eq. (1). We also remark that even in the (easier) case in which the function evaluations are noiseless, the lack of smoothness or convexity of the objective function provably necessitates the use of randomization in the optimization algorithm, resorting to high probability guarantees [12].

Algorithm 3 Algorithm with Post-Optimization Validation

Input: Initialization $\mathbf{x}_0 \in \mathbb{R}^d$, smoothing parameter $\delta' > 0$, clipping parameter $D > 0$, step size $\eta > 0$, iteration budget per round $T \in \mathbb{N}$, number of rounds $R \in \mathbb{N}$, validation sample size $S \in \mathbb{N}$.
Initialize: $M = \lfloor \frac{\delta'}{D} \rfloor$
for $r = 1, \dots, R$ **do**
  Call Algorithm 1 with $\mathbf{x}_0, \frac{\delta}{2}, D, \eta, T$ and obtain $\mathbf{z}_1^r, \dots, \mathbf{z}_M^r$, $\mathbf{x}_{\text{out}}^r = \frac{1}{M} \sum_{m \in [M]} \mathbf{z}_m^r$
  for $s = 1, \dots, S$ **do**
    for $m = 1, \dots, M$ **do**
      Sample $\xi_{m,s} \sim \Xi$
      $\mathbf{g}_{m,s}^r = \textsc{GradEstimator}(\mathbf{z}_m^r, \delta', \xi_{m,s})$ ▷ `Unbiased estimator of` $\nabla f_{\delta'}(\mathbf{z}_m^r)$
    end for
    $\hat{\mathbf{g}}_s^r = \frac{1}{M} \sum_{m=1}^M \mathbf{g}_{m,s}^r$
  end for
  $\hat{\mathbf{g}}^r = \frac{1}{S} \sum_{s=1}^S \mathbf{g}_s^r$
end for
$r^* = \arg\min_{r \in [R]} \|\hat{\mathbf{g}}^r\|$
Output: $\mathbf{x}_{\text{out}}^{r^*}$.

References

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

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[3] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

[4] Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2301.06428*, 2023.

[5] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[6] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

[7] D. Davis, D. Drusvyatskiy, Y. T. Lee, S. Padmanabhan, and G. Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *NeurIPS*, 2022.

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

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

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

[11] A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[12] 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.

[13] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

[14] 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.

[15] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[16] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

[17] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.

[18] L. Tian, K. Zhou, and A. M-C. So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *ICML*, pages 21360–21379. PMLR, 2022.

[19] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

[20] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *ICML*, pages 11173–11182. PMLR, 2020.

Appendix A. Proof of Theorem 5

As previously discussed, the key to obtaining the improved rate is Lemma 4. We start by proving it, followed by two additional propositions, after which we combine the ingredients in order to conclude the proof.

Proof [Proof of Lemma 4] Let $\mathbf{g} \in \partial_\gamma f_\delta(\mathbf{x})$. Then, by definition, there exist $\mathbf{y}_1, \ldots, \mathbf{y}_k \in \mathbb{B}(\mathbf{x}, \gamma)$ (for some $k \in \mathbb{N}$) such that $\mathbf{g} = \sum_{i\in[k]} \lambda_i \nabla f_\delta(\mathbf{y}_i)$, where $\lambda_1, \ldots, \lambda_k \geq 0$ with $\sum_{i\in[k]} \lambda_i = 1$. By Proposition 3 we have for all $i \in [k]$:

$$\nabla f_\delta(\mathbf{y}_i) \in \partial_\delta f(\mathbf{y}_i) \,. \tag{2}$$

Further note that since $\|\mathbf{y}_i - \mathbf{x}\| \leq \gamma$, then by definition

$$\partial_\delta f(\mathbf{y}_i) \subseteq \partial_{\delta+\gamma} f(\mathbf{x}) \,. \tag{3}$$

By combining Eq. (2) and Eq. (3) we get that for all $i \in [k]$: $\nabla f_\delta(\mathbf{y}_i) \in \partial_{\delta+\gamma} f(\mathbf{x})$. Since $\partial_{\delta+\gamma} f(\mathbf{x})$ is a convex set, we get that

$$\mathbf{g} = \sum_{i\in[k]} \lambda_i \nabla f_\delta(\mathbf{y}_i) \in \partial_{\delta+\gamma} f(\mathbf{x}) \,,$$

which finishes the proof. ■

The following lemma is essentially due to Shamir [16], showing that it is possible to construct a gradient estimator whose second moment scales linearly with respect to the dimension d.

Lemma 8 *Let*

$$\mathbf{g}_t = \frac{d}{2\delta'} \left(F(\mathbf{x}_t + s_t \boldsymbol{\Delta}_t + \delta' \mathbf{w}_t; \xi_t) - F(\mathbf{x}_t + s_t \boldsymbol{\Delta}_t - \delta' \mathbf{w}_t; \xi_t) \right) \mathbf{w}_t \,,$$

as generated by GRADESTIMATOR *(Algorithm 2) when called at iteration t of Algorithm 1. Then*

$$\mathbb{E}_{\xi_t, \mathbf{w}_t}[\mathbf{g}_t | \mathbf{x}_{t-1}, s_t, \boldsymbol{\Delta}_t] = \nabla f_{\delta'}(\mathbf{x}_{t-1} + s_t \boldsymbol{\Delta}_t) = \nabla f_{\delta'}(\mathbf{z}_t)$$

and

$$\mathbb{E}_{\xi_t, \mathbf{w}_t}[\|\mathbf{g}_t\|^2 | \mathbf{x}_{t-1}, s_t, \boldsymbol{\Delta}_t] \leq 16\sqrt{2\pi} d L_0^2 \,.$$

Proof For the sake of notational simplicity, we omit the subscript t throughout the proof. For the first claim, since $-\mathbf{w} \sim \mathbf{w}$ we have

$$\begin{aligned}
\mathbb{E}_{\xi,\mathbf{w}}[\mathbf{g} | \mathbf{x}, s, \boldsymbol{\Delta}] &= \mathbb{E}_{\xi,\mathbf{w}} \left[\frac{d}{2\delta'} \left(F(\mathbf{x} + s\boldsymbol{\Delta} + \delta' \mathbf{w}; \xi) - F(\mathbf{x} + s\boldsymbol{\Delta} - \delta' \mathbf{w}; \xi) \right) \mathbf{w} \,|\, \mathbf{x}, s, \boldsymbol{\Delta} \right] \\
&= \frac{1}{2} \Big(\mathbb{E}_{\xi,\mathbf{w}} \left[\tfrac{d}{\delta'} F(\mathbf{x} + s\boldsymbol{\Delta} + \delta' \mathbf{w}; \xi) \mathbf{w} \,|\, \mathbf{x}, s, \boldsymbol{\Delta} \right] \\
&\qquad + \mathbb{E}_{\xi,\mathbf{w}} \left[\tfrac{d}{\delta'} F(\mathbf{x} + s\boldsymbol{\Delta} + \delta'(-\mathbf{w}); \xi)(-\mathbf{w}) \,|\, \mathbf{x}, s, \boldsymbol{\Delta} \right] \Big) \\
&= \mathbb{E}_{\xi,\mathbf{w}} \left[\tfrac{d}{\delta'} F(\mathbf{x} + s\boldsymbol{\Delta} + \delta' \mathbf{w}; \xi) \mathbf{w} \,|\, \mathbf{x}, s, \boldsymbol{\Delta} \right] \,.
\end{aligned}$$

Using the law of total expectation, we get

$$\begin{aligned}\mathbb{E}_{\xi,\mathbf{w}}[\mathbf{g}|\mathbf{x},s,\boldsymbol{\Delta}] &= \mathbb{E}_{\mathbf{w}}\left[\tfrac{d}{\delta'}\mathbb{E}_{\xi}\left[F(\mathbf{x}+s\boldsymbol{\Delta}+\delta'\mathbf{w};\xi)\mathbf{w}\,|\,\mathbf{w},\mathbf{x},s,\boldsymbol{\Delta}\right]\,|\,\mathbf{x},s,\boldsymbol{\Delta}\right]\\ &= \mathbb{E}_{\mathbf{w}}\left[\tfrac{d}{\delta'}f(\mathbf{x}+s\boldsymbol{\Delta}+\delta'\mathbf{w})\mathbf{w}\,|\,\mathbf{x},s,\boldsymbol{\Delta}\right]\\ &= \nabla f_{\delta'}(\mathbf{x}+s\boldsymbol{\Delta})\,,\end{aligned}$$

where the last equality is due to Flaxman et al. [9, Lemma 2.1]. The second moment bound follows from Shamir [16, Lemma 10], with the explicit constant pointed out by Lin et al. [14, Lemma E.1]. ■

The following result of Cutkosky et al. [6] provides a stochastic first-order nonsmooth nonconvex optimization method, whose convergence scales linearly with the second-moment of the gradient estimator.

Theorem 9 (6) *Let $\delta',\epsilon\in(0,1)$, and let $h:\mathbb{R}^d\to\mathbb{R}$ be an L-Lipschitz function such that $h(\mathbf{x}_0)-\inf_{\mathbf{x}}h(\mathbf{x})\le\Delta_h$. Suppose* GRADESTIMATOR$(\mathbf{x},\xi)$ *returns an unbiased gradient estimator of $\nabla h(\mathbf{x})$ whose second moment is bounded by σ^2. Then there exists $T=O\left(\frac{\sigma^2\Delta_h}{\delta'\epsilon^3}\right)$ such that setting $\eta=\frac{\Delta_h}{\sigma^2 T}$, $D=\left(\frac{(\delta')^{1/2}\Delta_h}{\sigma T}\right)^{2/3}$ and running Algorithm 1 uses T calls to* GRADESTIMATOR *and satisfies*

- $\mathbf{z}_{(k-1)M+m}\in\mathbb{B}(\overline{\mathbf{x}}_k,\delta')$ *for all $m\in[M],k\in[K]$ (where $M,K,(\mathbf{z}_t)_{t=1}^T$ are defined in the algorithm).*
- $\mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_T}\left[\frac{1}{K}\sum_{k=1}^K\left\|\frac{1}{M}\sum_{m=1}^M\nabla h(\mathbf{z}_{(k-1)M+m})\right\|\right]\le\epsilon\,.$

In particular, its output $\mathbf{x}^{\mathrm{out}}\sim\mathrm{Unif}\{\overline{\mathbf{x}}_1,\ldots,\overline{\mathbf{x}}_K\}$ satisfies $\mathbb{E}[\|\nabla f(\mathbf{x}^{\mathrm{out}})\|_{\delta'}\}]\le\epsilon$.

We are now ready to complete the proof of Theorem 5. By Lemma 8, GRADESTIMATOR (Algorithm 2) returns an unbiased estimator of $\nabla f_{\delta/2}$ whose second moment is bounded by $\sigma^2=O(dL_0^2)$, using two evaluations of $F(\,\cdot\,;\xi)$. Thus applying Theorem 9 with $h=f_{\delta/2}$, $\delta'=\frac{\delta}{2}$ ensures that Algorithm 1 returns a $(\frac{\delta}{2},\epsilon)$-stationary point of $f_{\delta/2}$, which by Lemma 4 is a (δ,ϵ)-stationary point of f. Recall that $\|f-f_{\delta/2}\|_\infty\le\frac{\delta L_0}{2}$ and $f(\mathbf{x}_0)-\inf_{\mathbf{x}}f(\mathbf{x})\le\Delta$, thus $f_{\delta/2}(\mathbf{x}_0)-\inf_{\mathbf{x}}f_{\delta/2}(\mathbf{x})\le\Delta+\frac{\delta L_0}{2}=:\Delta_h$. We finish the proof by noting that overall we have obtained

$$T=O\left(\frac{\sigma^2\Delta_h}{\delta'\epsilon^3}\right)=O\left(\frac{dL_0^2(\Delta+\delta L_0)}{\delta\epsilon^3}\right)=O\left(d\cdot\left(\frac{\Delta L_0^2}{\delta\epsilon^3}+\frac{L_0^3}{\epsilon^3}\right)\right)\,.$$

Appendix B. Proof of Theorem 7

Recall that we saw in the proof of Theorem 5 that $\partial_{\delta'}f_{\delta'}(\cdot)\subseteq\partial_{2\delta'}f(\cdot)$ according to Lemma 4, and that for all $m\in[M]:\mathbf{z}_m^{r^*}\in\mathbb{B}(\mathbf{x}_{\mathrm{out}}^{r^*},\delta')$ by according to Theorem 9. Thus

$$\left\|\nabla f(\mathbf{x}_{\mathrm{out}}^{r^*})\right\|_{\delta}=\left\|\nabla f(\mathbf{x}_{\mathrm{out}}^{r^*})\right\|_{2\delta'}\le\left\|\nabla f_{\delta'}(\mathbf{x}_{\mathrm{out}}^{r^*})\right\|_{\delta'}\le\left\|\frac{1}{M}\sum_{m\in[M]}\nabla f_{\delta'}(\mathbf{z}_m^{r^*})\right\|\,.$$

By denoting $\mathbf{g}^r := \frac{1}{M}\sum_{m\in[M]} \nabla f_{\delta'}(\mathbf{z}_m^r),\ r \in [R]$ we get that it suffices to show

$$\Pr\left[\left\|\mathbf{g}^{r^*}\right\|^2 \le \epsilon^2\right] = \Pr\left[\left\|\mathbf{g}^{r^*}\right\| \le \epsilon\right] \ge 1-\gamma\,. \tag{4}$$

By definition of r^* we have

$$\left\|\hat{\mathbf{g}}^{r^*}\right\|^2 = \min_{r\in[R]} \|\hat{\mathbf{g}}^r\|^2 \le \min_{r\in[R]} \left(2\|\mathbf{g}^r\|^2 + 2\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2\right) \le 2\left(\min_{r\in[R]}\|\mathbf{g}^r\|^2 + \max_{r\in[R]}\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2\right)\,,$$

thus

$$\begin{aligned}\left\|\mathbf{g}^{r^*}\right\|^2 &\le 2\left\|\hat{\mathbf{g}}^{r^*}\right\|^2 + 2\left\|\hat{\mathbf{g}}^{r^*} - \mathbf{g}^{r^*}\right\|^2 \\ &\le 4\left(\min_{r\in[R]}\|\mathbf{g}^r\|^2 + \max_{r\in[R]}\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2\right) + 2\left\|\hat{\mathbf{g}}^{r^*} - \mathbf{g}^{r^*}\right\|^2 \\ &\le 4\cdot\min_{r\in[R]}\|\mathbf{g}^r\|^2 + 6\cdot\max_{r\in[R]}\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2\,.\end{aligned} \tag{5}$$

We now turn to bound each of the summand above with high probability. First, by Theorem 5 we can set $T = O\left(d\cdot\left(\frac{\Delta L_0^2}{\delta\epsilon^3} + \frac{L_0^3}{\epsilon^3}\right)\right)$ so that $\mathbb{E}[\|\mathbf{g}^r\|] \le \frac{\epsilon}{8}$ for any $r \in [R]$, hence by Markov's inequality

$$\Pr\left[4\cdot\min_{r\in R}\|\mathbf{g}^r\|^2 > \frac{\epsilon^2}{4}\right] = \Pr\left[\min_{r\in R}\|\mathbf{g}^r\| > \frac{\epsilon}{4}\right] \le \prod_{r\in[R]}\Pr\left[\|\mathbf{g}^r\| > \frac{\epsilon}{4}\right] \le 2^{-R}\,.$$

By setting $R \ge \lceil\log_2(2/\gamma)\rceil \implies 2^{-R} \le \frac{\gamma}{2}$ we conclude that

$$\Pr\left[4\cdot\min_{r\in R}\|\mathbf{g}^r\|^2 > \frac{\epsilon^2}{4}\right] \le \frac{\gamma}{2}\,. \tag{6}$$

For the second summand in Eq. (5), note that for all $r \in [R]$:

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{g}}^r\right] &= \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^S \mathbf{g}_s^r\right] = \frac{1}{S}\sum_{s=1}^S\left(\frac{1}{M}\sum_{m=1}^M \mathbb{E}\left[\mathbf{g}_{m,s}^r\right]\right) \\ &= \frac{1}{M}\sum_{m=1}^M\left(\frac{1}{S}\sum_{s=1}^S \mathbb{E}\left[\mathbf{g}_{m,s}^r\right]\right) \overset{\text{Lemma 8}}{=} \frac{1}{M}\sum_{m=1}^M \nabla f_{\delta'}(\mathbf{z}_m^r) = \mathbf{g}^r\,,\end{aligned}$$

thus $\mathbb{E}[\hat{\mathbf{g}}^r - \mathbf{g}^r] = 0$, and that it follows from the second claim in Lemma 8 that for any $r \in [R], s \in [S]$:

$$\mathbb{E}\left[\|\hat{\mathbf{g}}_s^r - \mathbf{g}^r\|^2\right] \le \frac{16\sqrt{2\pi}dL_0^2}{M}\,.$$

Noting that $(\mathbf{g}_1^r - \mathbf{g}^r), \ldots, (\mathbf{g}_S^r - \mathbf{g}^r)$ are independent as they are functions of the independent samples $\xi_{m,1}, \ldots, \xi_{m,S},\ m \in [M]$, we apply a simple concentration bound (Lemma 10) to get for

any $\lambda > 0$:

$$\Pr\left[\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2 \geq \lambda \frac{16\sqrt{2\pi}dL_0^2}{MS}\right] = \Pr\left[\left\|\frac{1}{S}\sum_{s=1}^{S}(\mathbf{g}_s^r - \mathbf{g}^r)\right\|^2 \geq \lambda \frac{16\sqrt{2\pi}dL_0^2}{MS}\right]$$
$$= \Pr\left[\left\|\sum_{s=1}^{S}(\mathbf{g}_s^r - \mathbf{g}^r)\right\|^2 \geq \lambda S \cdot \frac{16\sqrt{2\pi}dL_0^2}{M}\right] \leq \frac{1}{\lambda}\,,$$

hence by the union bound

$$\Pr\left[\max_{r\in[R]}\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2 \geq \lambda \frac{16\sqrt{2\pi}dL_0^2}{MS}\right] \leq \frac{R}{\lambda}\,.$$

Setting $\lambda := \lceil \frac{2R}{\gamma} \rceil \implies \frac{R}{\lambda} \leq \frac{\gamma}{2}$, we see that $S \gtrsim \frac{dL_0^2 \log(1/\gamma)}{M\epsilon^2\gamma}$ suffices for having $\lambda \frac{16\sqrt{2\pi}dL_0^2}{MS} \leq \frac{\epsilon^2}{4}$, under which the inequality above shows that

$$\Pr\left[\max_{r\in[R]}\|\hat{\mathbf{g}}^r - \mathbf{g}^r\|^2 \geq \frac{\epsilon^2}{4}\right] \leq \frac{\gamma}{2}\,. \tag{7}$$

By combining Eq. (6) and Eq. (7) and applying the union bound to get Eq. (5), we have proved Eq. (4) as required. Finally, recalling that GRADESTIMATOR (Algorithm 2) requires 2 noisy function evaluations, it is clear that the total number of evaluations performed by Algorithm 3 is bounded by

$$2R \cdot (T + MS) = O\left(d \cdot \left(\frac{\Delta L_0^2 \log(1/\gamma)}{\delta\epsilon^3} + \frac{L_0^3 \log(1/\gamma)}{\epsilon^3} + \frac{L_0^2 \log^2(1/\gamma)}{\gamma\epsilon^2}\right)\right)\,.$$

Appendix C. Concentration Lemma

Lemma 10 *Let $X_1, \ldots, X_N \in \mathbb{R}^d$ be independent random vectors such that for all $i \in [N]$: $\mathbb{E}[X_i] = \mathbf{0}$, $\mathbb{E}[\|X_i\|^2] \leq \sigma_i^2$. Then $\mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|^2\right] \leq \sum_{i=1}^N \sigma_i^2$. In particular,*

$$\Pr\left[\left\|\sum_{i=1}^{N} X_i\right\|^2 \geq \lambda \cdot \sum_{i=1}^{N}\sigma_i^2\right] \leq \lambda^{-1}$$

for any $\lambda > 0$.

Proof By linearity of expectation we have

$$\mathbb{E}\left[\left\|\sum_{i\in[N]} X_i\right\|^2\right] = \sum_{i\in[N]}\mathbb{E}\left[\|X_i\|^2\right] + \sum_{i\neq j\in[N]}\mathbb{E}[\langle X_i, X_j\rangle] = \sum_{i\in[N]}\mathbb{E}\left[\|X_i\|^2\right] \leq \sum_{i\in[N]}\sigma_i^2\,,$$

where we used the assumption that for any $i \neq j$: X_i, X_j are independent, thus $\mathbb{E}[\langle X_i, X_j\rangle] = 0$. The second claim follows from Markov's inequality. ■