Mathematical Programming
https://doi.org/10.1007/s10107-022-01822-7

FULL LENGTH PAPER

Series A

Lower bounds for non-convex stochastic optimization

Yossi Arjevani[1] · Yair Carmon[2] · John C. Duchi[3] · Dylan J. Foster[4] · Nathan Srebro[5] · Blake Woodworth[6]

Received: 26 May 2020 / Accepted: 14 April 2022

Abstract
We lower bound the complexity of finding ϵ-stationary points (with gradient norm at most ϵ) using stochastic first-order methods. In a well-studied model where algorithms access smooth, potentially non-convex functions through queries to an unbiased stochastic gradient oracle with bounded variance, we prove that (in the worst case) any algorithm requires at least ϵ^{-4} queries to find an ϵ-stationary point. The lower bound is tight, and establishes that stochastic gradient descent is minimax optimal in this model. In a more restrictive model where the noisy gradient estimates satisfy a mean-squared smoothness property, we prove a lower bound of ϵ^{-3} queries, establishing the optimality of recently proposed variance reduction techniques.

Mathematics Subject Classification 90C06 · 90C15 · 90C26 · 90C30 · 90C60 · 68Q25

✉ Yair Carmon
ycarmon@tauex.tau.ac.il

Yossi Arjevani
yossi.arjevani@gmail.com

John C. Duchi
jduchi@stanford.edu

Dylan J. Foster
dylanfoster@microsoft.com

Nathan Srebro
nati@ttic.edu

Blake Woodworth
blake.woodworth@inria.fr

1 The Hebrew University, Jerusalem, Israel

2 Tel Aviv University, Tel Aviv, Israel

3 Stanford University, Stanford, USA

4 Microsoft Research New England, Cambridge, MA, USA

5 TTIC, Chicago, USA

6 Inria, Paris, France

Published online: 09 June 2022

1 Introduction

Stochastic gradient methods—especially variants of stochastic gradient descent (SGD)—are the workhorse of modern machine learning and data-driven optimization [10, 11] more broadly. Much of the success of these methods stems from their broad applicability: any problem that admits an unbiased gradient estimator is fair game. Consequently, there is considerable interest in understanding the fundamental performance limits of methods using stochastic gradients across broad problem classes. For *convex* problems, a long line of work [1, 36, 37, 50] sheds lights on these limits, and they are by now well-understood. However, many problems of interest (e.g., neural network training) are not convex. This has led to intense development of improved methods for non-convex stochastic optimization, but little is known about the optimality of these methods. In this paper, we establish new fundamental limits for stochastic first-order methods in the non-convex setting.

In general non-convex optimization, it is intractable to find approximate global minima [36] or even to test if a point is a local minimum or a high-order saddle point [34]. As an alternative measure of optimization convergence, we consider ϵ-approximate stationarity. That is, given differentiable $F : \mathbb{R}^d \to \mathbb{R}$, our goal is to find a point $x \in \mathbb{R}^d$ with

$$\|\nabla F(x)\| \leq \epsilon. \tag{1}$$

The use of stationarity as a convergence criterion dates back to the early days of nonlinear optimization [cf. 40, 48]. Recent years have seen rapid development of a body of work that studies non-convex optimization through the lens of non-asymptotic convergence rates to ϵ-stationary points [14, 25, 26, 30, 32, 38, 56]. Another growing body of work motivates this study by identifying sub-classes of non-convex problems for which all stationary (or second-order stationary) points are globally optimal [28, 29, 33, 45].

We prove our lower bounds in an oracle model [36, 46], where algorithms access the function F through a *stochastic first-order oracle* consisting of a gradient estimator $g : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ and distribution P_z on $\mathcal{Z}$ satisfying

$$\mathbb{E}_z[g(x, z)] = \nabla F(x), \quad \text{and} \quad \mathbb{E}_z\|g(x, z) - \nabla F(x)\|^2 \leq \sigma^2. \tag{2}$$

At the tth optimization step, the algorithm queries at a point $x^{(t)}$, the oracle draws $z^{(t)} \sim P_z$, and the algorithm observes the noisy gradient estimate $g(x^{(t)}, z^{(t)})$. We make the standard assumption that the objective F has bounded initial suboptimality and Lipschitz gradient:

$$F(x^{(0)}) - \inf_x F(x) \leq \Delta \quad \text{and} \quad \|\nabla F(x) - \nabla F(y)\| \leq L \cdot \|x - y\| \ \ \forall x, y \in \mathbb{R}^d. \tag{3}$$

Following common practice, we refer to functions F with L-Lipschitz gradients as "L-smooth."

For problem instances (F, g) satisfying (2) and (3), given a tolerance ϵ, SGD finds a point x such that $\mathbb{E}\|\nabla F(x)\| \le \epsilon$ using $O(\Delta L \epsilon^{-2}(1+\sigma^2\epsilon^{-2}))$ oracle queries [30], which reduces to $O(\Delta L \sigma^2 \epsilon^{-4})$ in the typical regime where $\epsilon \le \sigma$. The literature on variance reduction for finding stationary points [25, 32, 56] considers the following additional assumptions:

1. The stochastic gradient g satisfies a *mean-squared smoothness* property

$$\mathbb{E}_z \|g(x,z) - g(y,z)\|^2 \le \bar{L}^2 \cdot \|x-y\|^2 \quad \forall x, y \in \mathbb{R}^d. \tag{4}$$

2. The algorithm is allowed K *simultaneous queries*: at step t, the algorithm queries $x^{(t,1)}, \ldots, x^{(t,K)}$ and observes $g(x^{(t,1)}, z^{(t)}), \ldots, g(x^{(t,K)}, z^{(t)})$, where the random seed $z^{(t)} \sim P_z$ is shared.

Under the mean-squared smoothness assumption and using $K = 2$ simultaneous queries the SPIDER [25] and SNVRG [57] algorithms find a point x such that $\mathbb{E}\|\nabla F(x)\| \le \epsilon$ using $O(\Delta \bar{L} \sigma \epsilon^{-3} + \sigma^2 \epsilon^{-2})$ oracle queries. This improvement over the ϵ^{-4} rate of SGD raises natural questions. Can we improve this rate further? Alternatively, can we improve the rate of SGD without the additional assumption (4)? We settle both questions in the negative.

1.1 Contributions

We prove lower bounds for finding stationary points in the stochastic first-order oracle model. Our main result is Theorem 3, which states:

1. There exists a distribution over instances (F, g) satisfying assumptions (2) and (3) under which every randomized[1] algorithm requires at least $c \cdot (\Delta L \sigma^2 \epsilon^{-4} + \Delta L \epsilon^{-2})$ oracle queries to find x satisfying $\mathbb{E}\|\nabla F(x)\| \le \epsilon$, where $c > 0$ is a universal constant and where the expectation is taken over the randomness in both the oracle and the algorithm.
2. When g also satisfies the mean-squared smoothness property (4), every randomized algorithm requires $c \cdot \left(\Delta \bar{L} \sigma \epsilon^{-3} + \Delta \bar{L} \epsilon^{-2} + \sigma^2 \epsilon^{-2}\right)$ oracle queries.

Both lower bounds hold for any number K of simultaneous queries, with the dimension d of the hard instance depending polynomially on ϵ^{-1} and at most logarithmically on K (see expressions for d in Sect. 1.2 below).

Our lower bounds continue to hold when the oracle is subject to more stringent assumptions. In particular, we show that gradient estimators of the form $g(x,z) = \nabla_x f(x,z)$ give rise to the same lower bounds; these gradient estimators arise in statistical learning problems such as empirical risk minimization. Furthermore, our results extend to *active* oracles where the algorithm may choose the seed z. This setting includes the special case of finite sum minimization, where $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, each oracle query consists of point x and index i, and the oracle response is $\nabla f_i(x)$.

The main implications of our results are as follows.

[1] As is common in the optimization literature, we describe algorithms which use random coins in their execution as "randomized," as opposed to "deterministic" algorithms which do not. Likewise, we distinguish between "noiseless" and "stochastic" first-order oracles, which provide exact and noisy gradient information, respectively.

- *Optimality of SGD and recent variance-reduction schemes* Our ϵ^{-4} lower bound matches (up to a numerical constant) the rate of convergence of SGD [30] under assumptions (2) and (3), thereby characterizing the optimal complexity and proving that SGD attains it. Similarly, under the additional assumption (4) our ϵ^{-3} lower bound matches the rates of [25] and [56], thereby proving their optimality.
- *Separation between smoothness assumptions* Our results highlight that the mean-squared smoothness assumption (4) is critical for variance reduction: we prove that in its absence, any scheme will require a number of queries that scales as ϵ^{-4} at least. These results are salient, as this assumption appears in numerous recent works on non-convex optimization [25, 26, 55, 56].
- *Separation between convex and non-convex stochastic optimization* [27] show that for *convex* functions satisfying assumptions (2) and (3), the optimal rate for finding ϵ-stationary points is $\widetilde{\Theta}(\sqrt{\Delta L\epsilon^{-2}} + \sigma^2\epsilon^{-2})$. Our $\Omega(\Delta L\sigma^2\epsilon^{-4})$ lower bound thus implies a gap between the convex and non-convex setting that scales as ϵ^{-2}. Conceptually, both rates admit a simple interpretation. The convex complexity is the sum of the noiseless convex optimization complexity $\sqrt{\Delta L\epsilon^{-2}}$ [16] and the estimation complexity $\sigma^2\epsilon^{-2}$. In contrast, in the non-convex case the noiseless complexity $\Delta L\epsilon^{-2}$ [15] and the estimation complexity $\sigma^2\epsilon^{-2}$ *multiply* rather than add. This observation underpins our proofs.

1.2 Our approach

We build on the noiseless lower bound construction of [15], itself inspired by Nesterov's notion of a chain-like function [37]. The key technique is to construct a function such that any noiseless oracle query reveals the index of at most a single "relevant" coordinate; the lower bound follows from the fact that any ϵ-stationary point is non-zero in $\Omega(L\Delta\epsilon^{-2})$ relevant coordinates. We amplify this lower bound by designing a noisy oracle that reveals a relevant coordinate only with low probability $p = \Theta(\epsilon^2/\sigma^2)$. This increases the number of required queries by a factor proportional to $1/p = \Theta(\sigma^2\epsilon^{-2})$, giving our ϵ^{-4} lower bound. The main challenge lies in making sure that the oracle is not too noisy, in the sense that the variance requirement (2) is met. To do so, we focus all of the noise on the single new coordinate i_x that the query x would discover next via the noiseless gradient. More specifically, we let $z \sim \text{Bernoulli}(p)$, and set $g_{i_x}(x, 0) = 0$ and $g_{i_x}(x, 1)$ to be such that g is unbiased. By careful analysis of the noiseless construction of [15] we show that the variance bound holds and we obtain our lower bound.

Proving the ϵ^{-3} lower bound requires additional nuance, as the "incoming coordinate" index i_x is not continuous in x, and so the gradient estimator above does not satisfy the mean-square smoothness requirement (4). Leveraging the special structure of the noiseless construction once more, we design a continuous surrogate for i_x, and arrive at a mean-square smooth construction for which $g_{i_x}(x, z)$ is again non-zero only with probability p. Scaling this construction such that $\bar{L} = \Theta(\bar{L}\epsilon/\sigma)$ yields the ϵ^{-3} lower bound.

For ease of exposition, we first carry out our proof strategy for the sub-class of "zero-respecting" algorithms, whose queries are non-zero only in coordinates where

previous oracle responses were not zero. We then lift our results to the class of all randomized algorithms using the method of random rotations [15, 51]. On a high level, we argue that in a random coordinate system, any algorithm operating on our constructions is essentially zero-respecting.

Our lower bound constructions are high-dimensional. For zero-respecting algorithms, the dimension we require is exactly the number of relevant coordinates: $d_{\text{zr}} = \Theta(\Delta L \epsilon^{-2})$ for the bounded variance case and $d_{\text{zr}} = \Theta(\Delta \bar{L} \sigma^{-1} \epsilon^{-1})$ for the mean-square smooth case. To handle general, potentially randomized algorithms that allow K simultaneous oracle queries for every random realization $z \sim P_z$, we require a dimension that is larger by a modest logarithmic factor. Lower bound constructions with dimension that scales polynomially in ϵ^{-1} are common [27, 36, 37, 50], and natural for algorithms that (nominally) work in arbitrary Hilbert spaces. In the noiseless setting, obtaining tight and algorithm-independent lower bounds on dimension-independent convergence rates necessitates high-dimensional constructions; see [15, Section 1.2] for additional discussion. Since the noiseless setting is a special case of our noisy setting, it seems likely that here too high-dimensional constructions are to some extent unavoidable. While it is possible that faster rates could be achieved in lower dimensions, our results answer the question of what can be guaranteed without explicitly taking advantage of the dimension being "sufficiently small" in some sense. Finally, when σ^2 is of the order of ΔL, our lower bounds in Theorem 3 apply for dimensions roughly above the square root of the iteration complexity. In many practical applications of non-convex optimization, e.g., training large machine learning models, the dimension is often in the millions or billions which is much larger than the square root of the number of iterations used. Therefore, we argue that our theorem's dimension requirement is not too stringent.

1.3 Related work

Lower bounds for first-order convex optimization in the noiseless setting are well-studied [36, 37]. For L-smooth functions in the high-dimensional regime, it is well-known that $\Theta\left(\sqrt{D^2 L \epsilon^{-1}}\right)$ gradient evaluations are necessary and sufficient to find an ϵ-suboptimal point given $x^{(0)}$ with $\|x^{(0)} - x^\star\| \leq D$; Nesterov's accelerated gradient method [39] achieves this rate.

For smooth high-dimensional non-convex optimization in the noiseless setting, [15] establish that $\Theta(\Delta L \epsilon^{-2})$ gradient evaluations are necessary and sufficient for finding ϵ-stationary points; this rate is achieved by gradient descent. An earlier line of work develops lower bounds for finding stationary points of non-convex functions in the low-dimensional regime where d is constant, but they obtain either weaker lower bounds [48] or tight bounds that hold only for specific algorithm classes [17–20].

A long line of work on lower bounds for stochastic convex optimization traces back to Nemirovski and Yudin's seminal information-based complexity [36]. Extensions since then have allowed sharp dimension-dependent bounds via reductions to statistical estimation problems [1, 41], as well as extension to structured problems common in machine learning, such as finite sums, by restrictions on the form of the update rules [7] and high-dimensional constructions [27, 50]. Our technique for proving stochastic

lower bounds differs qualitatively from these methods in that we preserve the sequential hardness of the noiseless non-convex lower bound construction of [15], and use the noise in the stochastic setting to amplify the hardness of this construction.

For non-convex stochastic optimization, few lower bounds are known. [24] recently showed that SGD itself cannot obtain a rate better than ϵ^{-4} for finding ϵ-stationary points, even for convex functions. This is an algorithm-specific result, whereas we show that *no algorithm* can improve over this rate. For finite sum problems where $F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, [25] show that $\Omega\left(\Delta\bar{L}\epsilon^{-2}\sqrt{n}\right)$ stochastic gradient queries are required to find a ϵ-stationary point; SPIDER and SNVRG [25, 56] have matching upper bounds. This lower bound is incomparable to ours: the stochastic gradient construction in the paper [25] has unbounded variance, so it cannot imply results along the lines of Theorem 3. Indeed, [25] leave obtaining the ϵ^{-3} lower bound we provide in Theorem 3 as an open problem.

We now turn to upper bounds for finding stationary points in the stochastic setting. In the *convex* setting (where achieving approximate global optimality is possible and hence usually the goal) [2] proposes algorithms with rates for finding stationary points improving over SGD, and [27] give improvements on these bounds and establish their optimality. For the non-convex setting, [30] establish an $O(\Delta L\sigma^2\epsilon^{-4})$ upper bound for SGD, and a large body of recent work attempts to improve this rate. These attempts roughly divide into two categories: *variance reduction* and *high-order information*.

Works in the variance reduction category make either the mean-squared smoothness assumption (4) or a stronger variant wherein every $g(\cdot, z)$ is $\bar{L}$-Lipschitz. The earliest results consider only the finite sum setting, and establish improved dependence on the number of summands [4, 42]. Under the bounded variance assumption (2), [32] obtain a rate of $\epsilon^{-10/3}$, demonstrating that in the non-convex setting variance reduction provides benefits beyond finite sum optimization. Subsequent algorithms by [25] and [56] obtain an improved rate of ϵ^{-3}, which we prove is optimal. Recent work [21, 49] offers further refinements of these algorithms that also obtain the ϵ^{-3} rate.

Smoothness in higher derivatives, such as Lipschitz continuity of the Hessian, allows additional possibilities [3, 5, 25, 52]. [47] provide a sub-sampled cubic regularization method that uses stochastic Hessian-vector products and attains a rate of $\epsilon^{-3.5}$ without relying on mean-squared smoothness (4) or simultaneous gradient queries. [26] show that it is possible to obtain the rate $\epsilon^{-3.5}$ using SGD with perturbed gradients and restarts without the need for Hessian-vector products. Most works that assume Lipschitz Hessian also provide guarantees for finding second-order stationary points.

1.4 Organization

Section 2 introduces the formal oracle model in which we prove our lower bounds. In Sect. 3, we develop the ideas behind our main result by proving lower bounds for the subclass of zero-respecting algorithms. In Sect. 4 we apply random rotations to generalize the results from Sect. 3 into lower bounds for all randomized algorithms, leading to our main result. Section 5 describes the extensions of our results to statistical

learning and active oracles, and Sect. 6 concludes with discussion of some remaining open problems.

Notation For a vector $x \in \mathbb{R}^d$, we let $\text{support}(x) := \{i | x_i \neq 0\}$ and $x_{\leq i} := (x_1, \ldots, x_i, 0, \ldots, 0) \in \mathbb{R}^d$ while $x_{\geq i} := (0, \ldots, 0, x_i, \ldots, x_d) \in \mathbb{R}^d$. For $\alpha \in [0, 1)$ we define the "progress" of x as $\text{prog}_\alpha(x) := \max\{i \geq 0 | |x_i| > \alpha\}$, where we assume $x_0 \equiv 1$. For a differentiable function f, we adopt the convention $[\nabla f(x)]_i = \nabla_i f(x) = \frac{\partial}{\partial x_i} f(x)$. When f is twice-differentiable, we likewise define $[\nabla^2 f(x)]_{ij} = \nabla^2_{ij} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$. Throughout, $\|x\|$ denotes the Euclidean norm of x and $\|x\|_\infty$ denotes its ℓ_∞ norm. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, $\|A\|_{\text{op}}$ denotes the operator norm. Given functions $f, g : \mathcal{X} \to [0, \infty)$ where $\mathcal{X}$ is any set, we use non-asymptotic big-O notation: $f = O(g)$ if there exists a numerical constant $c < \infty$ such that $f(x) \leq c \cdot g(x)$ for all $x \in \mathcal{X}$ and $f = \Omega(g)$ if there is a numerical constant $c > 0$ such that $f(x) \geq c \cdot g(x)$. The $\widetilde{O}(\cdot)$ notation hides constants and factors logarithmic in the problem parameters. $f = \widetilde{O}(g)$ to hide logarithmic . Finally, we write the indicator of conditions cond as $\mathbb{1}\{\mathsf{cond}\}$, i.e., $\mathbb{1}\{\mathsf{cond}\} = 1$ if the cond holds, and $\mathbb{1}\{\mathsf{cond}\} = 0$ otherwise.

2 Setup

We study the stochastic optimization problem of finding an ϵ-stationary point through the well-known framework of oracle complexity [36], which we set up formally in this section.

Function class We develop lower bounds for algorithms that find stationary points of functions in the set

$$\mathcal{F}(\Delta, L) := \Big\{ F : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } F(0) - \inf_x F(x) \leq \Delta, \\ \|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\| \text{ for all } x, y \Big\}.$$

We state explicitly the value of the dimension d required for each lower bound construction; the reader may otherwise regard d as a free parameter.

Optimization protocol We consider algorithms that access an unknown function $F \in \mathcal{F}(\Delta, L)$ through a *stochastic first-order oracle* O. Each oracle O consists of a distribution P_z on a measurable space $\mathcal{Z}$ and an unbiased mapping $\mathsf{O}_F(x, z) = g(x, z)$ such that for each $F \in \mathcal{F}(\Delta, L)$ and x, if $z \sim P_z$ then $\mathbb{E}[g(x, z)] = \nabla F(x)$. We consider a protocol in which algorithms interact with the oracle through multiple rounds of batch queries. At each round i, the algorithm queries a batch

$$x^{(i)} := \left(x^{(i,1)}, \ldots, x^{(i,K)}\right), \text{ where } x^{(i,k)} \in \mathbb{R}^d \text{ and } k \in [K] \tag{5}$$

of size K, and for each batch query $x^{(i)}$, the oracle O performs an independent draw $z^{(i)} \sim P_z$ and responds with

$$\mathsf{O}_F(x^{(i)}, z^{(i)}) := \left(\mathsf{O}_F(x^{(i,1)}, z^{(i)}), \ldots, \mathsf{O}_F(x^{(i,K)}, z^{(i)})\right).$$

When $K = 1$ this is the classical first-order stochastic optimization framework. By considering larger batches we can subsume variance-reduction methods such as SPIDER and SNVRG [25, 56], both of which query each stochastic gradient at $K = 2$ points.[2]

Optimization algorithms An algorithm A consists of a distribution Pr over a measurable set $\mathcal{R}$ and a sequence of measurable mappings $\{\mathsf{A}^{(i)}\}_{i\in\mathbb{N}}$ such that $\mathsf{A}^{(i)}$ takes in the first $i-1$ oracle responses and the random seed $r \in \mathcal{R}$ to produce the ith query. We let $\{x^{(i)}_{\mathsf{A}[\mathsf{O}_F]}\}_{i\in\mathbb{N}}$ denote the (random) sequence of queries resulting from applying algorithm A with O, defined recursively as

$$x^{(i)}_{\mathsf{A}[\mathsf{O}_F]} = \mathsf{A}^{(i)}\Big(r, \mathsf{O}_F\big(x^{(1)}_{\mathsf{A}[\mathsf{O}_F]}, z^{(1)}\big), \ldots, \mathsf{O}_F\big(x^{(i-1)}_{\mathsf{A}[\mathsf{O}_F]}, z^{(i-1)}\big)\Big), \tag{6}$$

where $r \sim \Pr$ is drawn a single time at the beginning of the protocol (this is no loss of generality [36]). We define $\mathcal{A}_{\mathsf{rand}}(K)$ to be the class of all algorithms that follow the protocol (6) with K batch queries per round.

Oracle classes We consider two natural classes of oracles. For the *bounded variance* class, denoted $\mathcal{O}(K, \sigma^2)$, we require that the stochastic gradient be unbiased and have the bounded variance property (2), but otherwise allow arbitrary $g(x, z)$. This well-studied setting subsumes the standard analysis of stochastic gradient descent for finding approximate stationary points [30].

The bounded variance setting places few restrictions on the stochastic gradient function $g(x, z)$, but there are many applications in which the stochastic gradients may have additional structure. In the *mean-squared smooth* setting, we require that in addition to the bounded-variance property (2), the stochastic gradient satisfies the mean-squared smoothness property (4). We use $\mathcal{O}(K, \sigma^2, \bar{L})$ to denote the class of all such oracles. By Jensen's inequality, any function that admits an $\bar{L}$-mean-squared smooth oracle must itself be $\bar{L}$-smooth.

Our results also extend to more structured oracles appearing in the statistical learning and/or finite-sum settings, as well as to oracles that provide zeroth-order information on F. We defer the details to Sect. 5.

Complexity measures Our main results are tight lower bounds on the *distributional complexity* [12, 36, 53] of finding ϵ-stationary points. Let $\mathcal{P}[\mathcal{F}(\Delta, L)]$ be set of all distributions over $\mathcal{F}(\Delta, L)$; the distributional complexity in the bounded variance setting is

$$\mathfrak{m}^{\mathsf{rand}}_{\epsilon}(K, \Delta, L, \sigma^2) := \sup_{\mathsf{O}\in\mathcal{O}(K,\sigma^2)} \sup_{P_F\in\mathcal{P}[\mathcal{F}(\Delta,L)]} \inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{rand}}(K)} \inf\left\{T \in \mathbb{N} \;\middle|\; \mathbb{E}\left\|\nabla F\big(x^{(T,1)}_{\mathsf{A}[\mathsf{O}_F]}\big)\right\| \le \epsilon\right\}, \tag{7}$$

where the expectation is over the sampling of F from P_F, the randomness in the oracle O, and the randomness in the algorithm A, though randomization in A does not affect distributional complexity [36, 53]. The distributional complexity for the mean-squared smooth setting is

[2] See also the K-parallel model of [35].

$$\bar{\mathfrak{m}}_{\epsilon}^{\mathsf{rand}}(K, \Delta, \bar{L}, \sigma^2) := \sup_{\mathsf{O} \in \mathcal{O}(K, \sigma^2, \bar{L})} \sup_{P_F \in \mathcal{P}[\mathcal{F}(\Delta, \bar{L})]} \inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}(K)} \inf \left\{ T \in \mathbb{N} \;\middle|\; \mathbb{E} \left\| \nabla F\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(T,1)}\right) \right\| \leq \epsilon \right\}. \tag{8}$$

Lower bounds on distributional complexity imply lower bounds on minimax complexity [cf. 12, 36]. That is, $\mathfrak{m}_{\epsilon}^{\mathsf{rand}}(K, \Delta, L, \sigma^2) > T$ implies that there exists $\mathsf{O} \in \mathcal{O}(K, \sigma^2)$ such that for every $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}(K)$ there exists a function $F \in \mathcal{F}(\Delta, L)$ for which $\mathbb{E} \left\| \nabla F\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(T,1)}\right) \right\| > \epsilon$, where here the expectation is over randomness in A and O.

3 Lower bounds for zero-respecting algorithms

Before presenting our results in full generality, we first develop the key components of our technique by proving lower bounds for a restricted class of *zero-respecting algorithms* [15]. The class of zero-respecting algorithms generalizes the well-known linear span-assumption [see 37, Section 2.1.2], and encompasses many standard optimization algorithms. More importantly, the lower bound instances we introduce in this section form the core of our lower bounds for general algorithms via a reduction in the next section.

An algorithm A is zero-respecting if its queries at each round have support in the supports of all previous oracle responses:

Definition 1 A stochastic first-order algorithm A is *zero-respecting* if for any oracle O and any realization of $z^{(1)}, z^{(2)}, \ldots$, for all $t \geq 1$ and $k \in [K]$,

$$\text{support}\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(t,k)}\right) \subseteq \bigcup_{i<t, k' \in [K]} \text{support}\left(g^{(i,k')}\right), \tag{9}$$

where $\left(f^{(t,1)}, g^{(t,1)}\right), \ldots, \left(f^{(t,K)}, g^{(t,K)}\right) = \mathsf{O}_F\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(t)}, z^{(t)}\right)$ denote the oracle responses for round t.

We let $\mathcal{A}_{\mathsf{zr}}(K)$ denote the class of all zero-respecting algorithms. Our main result for this section is to establish tight lower bounds on the minimax oracle complexity for zero-respecting algorithms, which we denote by $\mathfrak{m}_{\epsilon}^{\mathsf{zr}}(K, \Delta, L, \sigma^2)$ for the bounded variance setting and $\bar{\mathfrak{m}}_{\epsilon}^{\mathsf{zr}}(K, \Delta, \bar{L}, \sigma^2)$ for the mean-squared smooth setting; these complexities are as in (7) and (8), with $\mathcal{A}_{\mathsf{zr}}(K)$ replacing $\mathcal{A}_{\mathsf{rand}}(K)$. The zero-respecting structure allows us to attain tight lower bounds using P_F supported on a single hard function.

3.1 Probabilistic zero-chains

At the core of our development is an embedding of the task of finding a stationary point into that of finding a point x with high coordinate *progress*, which we define as

$$\mathrm{prog}_\alpha(x) := \max\{i \geq 0 | |x_i| > \alpha\} \ \text{(where } x_0 \equiv 1\text{)}, \tag{10}$$

i.e., $\mathrm{prog}_\alpha(x)$ is the highest index whose entry is α-far from zero, for some threshold $\alpha \in [0, 1)$. The starting point for our lower bounds is the notion of a *first-order zero-chain* [15], which is a function F that satisfies $\mathrm{prog}_0(\nabla F(x)) \leq \mathrm{prog}_0(x) + 1$ for all x, generalizing Nesterov's concept of a "chain-like" function [37]. In the noiseless case ($g^{(i)} = \nabla F(x^{(i)})$), zero-chains control the rate of progress of zero-respecting algorithms: every query can "discover" at most one coordinate, and therefore $\mathrm{prog}_0(x^{(i)}) < T$ for all $i \leq T$.

Our key insight is that in the stochastic setting noise can amplify progress control: we construct stochastic gradient functions for which any zero-respecting algorithm requires *many* queries in order to activate one coordinate. We call such functions *probabilistic zero-chains*. For the formal definition recall the truncation notation $[x_{\leq j}]_i = x_i \mathbb{1}\{i \leq j\}$.

Definition 2 A stochastic gradient function $g(x, z)$ is a *probability-p zero-chain* if

$$\mathbb{P}\Big(\forall x : \mathrm{prog}_0(g(x,z)) \leq \mathrm{prog}_{\frac{1}{4}}(x) \text{ and } g(x,z) = g(x_{\leq \mathrm{prog}_{\frac{1}{4}}(x)}, z)\Big) \geq 1 - p, \tag{11}$$

and

$$\mathbb{P}\Big(\forall x : \mathrm{prog}_0(g(x,z)) \leq 1 + \mathrm{prog}_{\frac{1}{4}}(x) \text{ and } g(x,z) = g(x_{\leq 1+\mathrm{prog}_{\frac{1}{4}}(x)}, z)\Big) = 1. \tag{12}$$

The constant $1/4$ in (11) is only used in our lower bound for general algorithms, and any non-zero constant would suffice in its place. Even the constant zero is sufficient for the constructions in this section; we keep $\mathrm{prog}_{\frac{1}{4}}(x)$ in the definition only for notational consistency. Moreover, since $x = x_{\leq \mathrm{prog}_0(x)}$ for all x, the requirements on invariance of $g(\cdot, z)$ under truncations of x are only necessary in the next section. We also note that the requirement (12) implies that any F for which g is an unbiased gradient estimator must itself be a (robust) zero-chain.

The next lemma formalizes the idea that any zero-respecting algorithm interacting with a probabilistic zero-chain requires many rounds to discover all coordinates.

Lemma 1 *Let $g(x, z)$ be a probability-p zero-chain gradient estimator for $F : \mathbb{R}^T \to \mathbb{R}$, and let O be any oracle with $O_F(x, z) = (F(x), g(x, z))$. Let $\left\{x^{(t,k)}_{A[O_F]}\right\}$ be the queries of any $A \in \mathcal{A}_{zr}(K)$ interacting with O_F. Then, with probability at least $1 - \delta$,*

$$\max_{k\in[K]} \mathrm{prog}_0\left(x^{(t,k)}_{A[O_F]}\right) < T, \quad \textit{for all } t \leq \frac{T - \log(1/\delta)}{2p}.$$

The intuition behind Lemma 1 is that any zero-respecting algorithm must activate coordinates in sequence, and must wait at least $\Omega(1/p)$ rounds between activations on average, leading to a total waiting time of $\Omega(T/p)$ rounds. The proof below makes this intuition formal; note that throughout the proof we use that prog_α is non-increasing in α.

Proof For brevity, we omit the subscript $\mathsf{A}[\mathsf{O}_F]$ from $\left\{x^{(t,k)}_{\mathsf{A}[\mathsf{O}_F]}\right\}$. Let

$$\pi^{(t)} = \max_{i \le t} \max_{k \in [K]} \text{prog}_0(x^{(i,k)}) = \max\{j \le T \,|\, x^{(i,k)}_j \ne 0 \text{ for some } i \le t, k \in [K]\}$$

and

$$T_p := \left\lfloor \frac{T - \log(1/\delta)}{2p} \right\rfloor,$$

so the lemma is equivalent to $\mathbb{P}(\pi^{(T_p)} \ge T) \le \delta$.

Recall that $z^{(1)}, z^{(2)}, \ldots$ is the sequence of oracle randomness values, and for every $t \ge 1$ define the binary random variable

$$B^{(t)} := \mathbb{1}\{\exists x : \text{prog}_0\Big(g(x; z^{(t)})\Big) = 1 + \text{prog}_{\frac{1}{4}}(x)\}. \tag{13}$$

Note that $\{B^{(t)}\}_{t \ge 1}$ are i.i.d. Bernoulli with probability of success at most p due to Definition 2. Moreover, they are independent of any randomization in the algorithm A.

With notation, we have that, for every t and $k \in [K]$,

$$\text{prog}_0(x^{(t,k)}) \overset{(i)}{\le} \max_{s<t} \max_{k' \in [K]} \text{prog}_0\Big(g(x^{(s,k')}, z^{(s)})\Big) \overset{(ii)}{\le} \max_{s<t}\{B^{(s)} + \pi^{(s)}\}, \tag{14}$$

where (i) follows from Definition 1 of zero-respecting algorithms and (ii) follows from the definition of $B^{(s)}$ and Eq. (12), which together imply that $\text{prog}_0(g(x, z^{(s)})) \le B^{(s)} + \text{prog}_{\frac{1}{4}}(x) \le B^{(s)} + \text{prog}_0(x)$ for all x and all s. From the bound (14) it follows by straightforward induction that

$$\pi^{(t)} \le \sum_{s<t} B^{(s)}$$

for all $t \ge 1$. We can therefore control deviations of $\pi^{(t)}$ with the Chernoff method:

$$\begin{aligned}\mathbb{P}(\pi^{(T_p)} \ge T) &\le \mathbb{P}\left(\sum_{s<T_p} B^{(s)} \ge T\right) = \mathbb{P}\Big(e^{\sum_{s<T_p} B^{(s)}} \ge e^T\Big) \le e^{-T}\,\mathbb{E}\, e^{\sum_{s<T_p} B^{(s)}} \\ &\overset{(\star)}{\le} e^{-T}(1 - p + pe)^{T_p} \le e^{2pT_p - T} \le \delta, \end{aligned} \tag{15}$$

where $(\star)$ uses the fact that $\{B^{(s)}\}$ are i.i.d. Bernoulli with $\mathbb{P}(B^{(s)} \ne 0) \le p$. □

3.2 Lower bound for the bounded variance setting

Lemma 1 suggests a natural lower bound strategy:

i. Construct a function $F \in \mathcal{F}(\Delta, L)$ whose gradients are large for all $x \in \mathbb{R}^T$ with $\text{prog}_0\big(x^{(i)}\big) < T$.
ii. Construct g, a probability-p zero chain gradient estimator for F.

Together with Lemma 1, these steps guarantee that any zero-respecting algorithm interacting with g will take at least $\Omega(T/p)$ rounds to make the gradient of F small. We first execute our strategy for the bounded variance setting (2).

We choose the underlying function F to be the construction of [15]. For each $T \in \mathbb{N}$, we define

$$F_T(x) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T}\big[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\big], \tag{16}$$

where the component functions Ψ and Φ are

$$\Psi(t) = \begin{cases} 0, & t \le 1/2, \\ \exp\Big(1 - \frac{1}{(2t-1)^2}\Big), & t > 1/2. \end{cases} \quad \text{and} \quad \Phi(t) = \sqrt{e}\int_{-\infty}^{t} e^{-\frac{1}{2}\tau^2} d\tau. \tag{17}$$

The function F_T is a (deterministic) zero-chain, and has large gradient unless all coordinates are large ($\text{prog}_1(x) \ge T$). We enumerate all the relevant properties of F_T in the following.

Lemma 2 (Carmon et al. [15]) *The function F_T satisfies:*

1. $F_T(0) - \inf_x F_T(x) \le \Delta_0 \cdot T$, *where* $\Delta_0 = 12$.
2. *The gradient of F_T is ℓ_1-Lipschitz continuous, where* $\ell_1 = 152$.
3. *For all* $x \in \mathbb{R}^T$, $\|\nabla F_T(x)\|_\infty \le \gamma_\infty$, *where* $\gamma_\infty = 23$.
4. *For all* $x \in \mathbb{R}^T$, $\text{prog}_0(\nabla F_T(x)) \le \text{prog}_{\frac{1}{2}}(x) + 1$.
5. *For all* $x \in \mathbb{R}^T$ *and* $i = \text{prog}_{\frac{1}{2}}(x)$, $\nabla F(x) = \nabla F(x_{\le 1+i})$ *and* $[\nabla F(x)]_{\le i} = [\nabla F(x_{\le i})]_{\le i}$.
6. *For all* $x \in \mathbb{R}^T$, *if* $\text{prog}_1(x) < T$ *then* $\|\nabla F_T(x)\| \ge |\nabla_{\text{prog}_1(x)+1} F_T(x)| > 1$.

Parts 1–3 of the lemma follow from [15, Lemma 3] and its proof; we derive the precise value $\ell_1 = 152$ in Appendix A.1. Parts 4 and 5 follow from [15, Observation 3] and part 6 is [15, Lemma 2].

We now turn to the construction of a probabilistic zero-chain for F_T. The main technical difficulty in the construction lies in keeping the variance of the stochastic gradient function bounded and, in particular, independent of the dimension T. Indeed, consider a naive construction that when queried at point x, returns 0 with probability $1 - p$ and returns $\frac{1}{p} \cdot \nabla F_T(x)$ with probability p. While this is clearly a probability-p zero-chain, the variance at point x is $\Omega\big(\|\nabla F_T(x)\|_2^2/p\big)$, which can be as large as T/p. As we let the dimension T depend polynomially on $1/\epsilon$, removing this dimension

dependence from the variance is critical for making the oracle belong to $\mathcal{O}(K, \sigma^2)$ after rescaling.

Our key observation is that, since $\|\nabla F_T(x)\|_\infty \le 23$ by Lemma 2.3, we can keep the variance bounded if, instead of deleting all coordinates uniformly, we delete only a single important coordinate. Since our goal is to construct a probabilistic zero-chain, and since F_T is itself a deterministic zero-chain, a natural choice of coordinate is $\mathrm{prog}_{\frac{1}{4}}(x) + 1$. This leads to the following stochastic gradient function:

$$[g_T(x, z)]_i := \nabla_i F_T(x) \cdot \left(1 + \mathbb{1}\{i > \mathrm{prog}_{\frac{1}{4}}(x)\}\left(\frac{z}{p} - 1\right)\right), \tag{18}$$

where $z \sim \mathrm{Bernoulli}(p)$. Note that for all $i > \mathrm{prog}_{\frac{1}{4}}(x) + 1$, $\nabla_i F_T(x) = 0$, so only the specific coordinate $\mathrm{prog}_{\frac{1}{4}}(x) + 1$ is noisy.

Lemma 3 *The stochastic gradient estimator g_T is a probability-p zero-chain, is unbiased for ∇F_T, and has variance*

$$\mathbb{E}\|g_T(x, z) - \nabla F_T(x)\|^2 \le \varsigma^2 \cdot \frac{1-p}{p} \quad \textit{for all } x \in \mathbb{R}^T, \ \textit{where } \varsigma = 23.$$

Proof First, we observe that $\mathbb{E}[g_T(x, z)] = \nabla F_T(x)$ for all $x \in \mathbb{R}^T$ by the definition (18) and the fact $\mathbb{E}[\frac{z}{p}] = 1$, so any O with $\mathsf{O}_{F_T}(x, z) = (F_T(x), g_T(x, z))$ is indeed a stochastic first-order oracle.

Second, we argue that the probability-p zero-chain property (Definition 2) holds. Recall that $\mathrm{prog}_\alpha(x)$ is non-increasing in α, so $\mathrm{prog}_{\frac{1}{4}}(x) \ge \mathrm{prog}_{\frac{1}{2}}(x)$. Therefore, by Lemma 2.4, $[g_T(x, z)]_i = \nabla_i F_T(x) = 0$ for all $i > \mathrm{prog}_{\frac{1}{4}}(x) + 1$, implying that $\mathrm{prog}_0(g_T(x, z)) \le 1 + \mathrm{prog}_{\frac{1}{4}}(x)$ for all $x \in \mathbb{R}^T$ and $z \in \{0, 1\}$. Moreover, for $x' := x_{\le 1 + \mathrm{prog}_{\frac{1}{4}}(x)}$, Lemma 2.5 implies that $\nabla F(x) = \nabla F(x')$, and since $\mathrm{prog}_{\frac{1}{4}}(x) = \mathrm{prog}_{\frac{1}{4}}(x')$ we conclude that $g_T(x, z) = g_T(x', z)$ for all x and z, giving Eq. (12).

To show that Eq. (11) holds, note that for $i \ge 1 + \mathrm{prog}_{\frac{1}{4}}(x)$ and $z = 0$, the RHS of Eq. (18) is 0. Consequently, $g_T(x, 0) = [\nabla F_T(x)]_{\le \mathrm{prog}_{\frac{1}{4}}(x)}$ for all $x \in \mathbb{R}^T$, implying that $\mathrm{prog}_0(g_T(x, 0)) \le \mathrm{prog}_{\frac{1}{4}}(x)$ and (by Lemma 2.5) that $g_T(x, 0) = g_T\big(x_{\le \mathrm{prog}_{\frac{1}{4}}(x)}, 0\big)$. Since $\mathbb{P}(z = 0) = 1 - p$, we obtain Eq. (11) and establish the probabilistic zero-chain property.

Finally, we bound the variance. Note that the error term $g_T(x, z) - \nabla F_T(x)$ is non-zero only in the coordinate $i_x := \mathrm{prog}_{\frac{1}{4}}(x) + 1$. Therefore,

$$\begin{aligned}\mathbb{E}\|g_T(x, z) - \nabla F_T(x)\|^2 &= \left|\nabla_{i_x} F_T(x)\right|^2 \mathbb{E}\left(\frac{z}{p} - 1\right)^2 \\ &\le \frac{\|\nabla F_T(x)\|_\infty^2 (1-p)}{p} \le \frac{23^2 (1-p)}{p},\end{aligned}$$

where the last inequality follows from Lemma 2.3. □

With the construction in hand, we prove our first lower bound.

Theorem 1 *There exist numerical constants $c, c' > 0$ such that for all $L, \Delta, \sigma^2 > 0$ and $\epsilon \leq c'\sqrt{L\Delta}$,*

$$\mathfrak{m}_\epsilon^{zr}(K, \Delta, L, \sigma^2) \geq c \cdot \left(\frac{\Delta L \sigma^2}{\epsilon^4} + \cdot \frac{\Delta L}{\epsilon^2} \right).$$

Constructions of dimension $d = O\left(\frac{\Delta L}{\epsilon^2}\right)$ realize the lower bound.

Before giving the proof, let us make a few remarks.

- The bound is tight, in that it matches (up to a numerical constant) the convergence rate for SGD (which is zero-respecting) [30, Eq. (2.13)]. Note that the restriction that $\epsilon \leq c'\sqrt{L\Delta}$ is without loss of generality, since for $\epsilon > c'\sqrt{L\Delta}$ we have $\|\nabla F(0)\| = O(\epsilon)$ for all functions $F \in \mathcal{F}(\Delta, L)$, so an ϵ-stationary point is trivial to find.
- When $\epsilon \leq \sigma$, the optimal complexity $\Theta(\frac{\Delta L \sigma^2}{\epsilon^4})$ is the product of the first-order oracle complexity for the noiseless setting, which is $\Theta(\frac{\Delta L}{\epsilon^2})$ [15], and the sample complexity of estimating a single gradient to precision ϵ, which is $\Theta(\frac{\sigma^2}{\epsilon^2})$. This is the first setting we are aware of where the *product* of these respective complexities characterizes the stochastic first-order complexity. Contrast to the convex setting, where the complexity scales with the sum [27].
- The lower bound does not depend on K, meaning that additional batch queries cannot by themselves improve on the rate obtained by SGD. While at first glance this may seem like a strange consequence of the zero-respecting assumption, we will show that the same holds true for arbitrary algorithms, provided the dimension is sufficiently large.

Proof of Theorem 1 Let Δ_0, ℓ_1 and ς be the numerical constants in Lemmas 2.1, 2.2 and 3, respectively. Given accuracy parameter ϵ, initial suboptimality Δ, smoothness parameter L and variance parameter σ^2, we define

$$F_T^\star(x) = \frac{L\lambda^2}{\ell_1} F_T\left(\frac{x}{\lambda}\right), \text{ where } \lambda = \frac{\ell_1}{L} \cdot 2\epsilon,$$
$$\text{and } T = \left\lfloor \frac{\Delta}{\Delta_0(L\lambda^2/\ell_1)} \right\rfloor = \left\lfloor \frac{L\Delta}{\Delta_0 \ell_1 (2\epsilon)^2} \right\rfloor,$$

where we assume $T \geq 3$, or equivalently $\epsilon \leq \sqrt{\frac{L\Delta}{12\Delta_0\ell_1}}$. Let

$$g_T^\star(x, z) = \frac{L\lambda}{\ell_1} \cdot g_T(x/\lambda, z) = 2\epsilon \cdot g_T(x/\lambda, z)$$

denote the corresponding scaled stochastic gradient function. Now, by Lemmas 2.1 and 2.2, we have that $F_T^\star$ is $\frac{L}{\ell_1} \cdot \ell_1 = L$-smooth and has initial suboptimality bounded

by Δ. Likewise, by Lemma 3,

$$\mathbb{E}\|g_T^\star(x,z) - \nabla F_T^\star(x)\|^2 = \left(\frac{L\lambda}{\ell_1}\right)^2 \mathbb{E}\left\|g_T^\star\left(\frac{x}{\lambda}, z\right) - \nabla F_T^\star\left(\frac{x}{\lambda}\right)\right\|^2 \le \frac{(2\varsigma\epsilon)^2(1-p)}{p}.$$

Therefore, setting $\frac{1}{p} = \frac{\sigma^2}{(2\varsigma\epsilon)^2} + 1$ guarantees a variance bound of σ^2.

Next, Let O be any oracle in $\mathcal{O}(K, \sigma^2)$ for which $\mathsf{O}_{F_T^\star}(x,z) = (F_T^\star(x), g_T^\star(x,z))$. Instantiating Lemma 1 for $\delta = 1/2$, we have that with probability at least $1/2$, $\max_{k\in[K]} \mathrm{prog}_0\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(t,k)}\right) < T$ for all $t \le (T-1)/2p$ and $k \in [K]$. Now, by Lemma 2.6, for every $x \in \mathbb{R}^T$ such that $\mathrm{prog}_0(x) < T$, it holds that

$$\left\|\nabla F_T^\star(x)\right\| = \frac{L\lambda}{\ell_1}\left\|\nabla F_T\left(\frac{x}{\lambda}\right)\right\| > \frac{L\lambda}{\ell_1} = 2\epsilon,$$

So with probability at least $1/2$, we have for all $t \le (T-1)/2p$ and $k \in [K]$ that $\left\|\nabla F_T^\star(x_{\mathsf{A}[\mathsf{O}_F]}^{(t,k)})\right\| > 2\epsilon$. Therefore,

$$\mathbb{E}\left\|\nabla F_T^\star\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(t,k)}\right)\right\| > \epsilon, \tag{19}$$

by which it follows that

$$\mathfrak{m}_\epsilon^{\mathsf{zr}}(K,\Delta,L,\sigma^2) > \frac{T-1}{2p} \ge \left(\left\lfloor \frac{L\Delta}{4\Delta_0\ell_1\epsilon^2}\right\rfloor - 1\right)\left(\frac{\sigma^2}{2(2\varsigma\epsilon)^2} + \frac{1}{2}\right) \ge \frac{1}{2^6\ell_1\Delta_0\varsigma^2}\cdot\frac{L\Delta\sigma^2}{\epsilon^4} + \frac{1}{2^4\ell_1\Delta_0}\cdot\frac{L\Delta}{\epsilon^2},$$

where the last inequality uses that $\lfloor x\rfloor - 1 \ge x/2$ whenever $x \ge 3$. □

3.3 Lower bound for the mean-squared smooth setting

We now turn to lower bounds for the mean-squared smooth setting. Here, we must ensure that in addition to the variance constraint, our stochastic gradient function satisfies the mean-squared smoothness constraint (4). This requires a more sophisticated construction than before, as the use of the indicator function $\mathbb{1}\{i > \mathrm{prog}_{\frac{1}{4}}(x)\}$ makes the stochastic gradient g_T discontinuous. Indeed, let $x = (1, 1/4 - \delta, 0)$ and $y = (1, 1/4, 0)$. Then $\mathrm{prog}_{\frac{1}{4}}(x) = 1 < 2 = \mathrm{prog}_{\frac{1}{4}}(y)$, and for any $\delta \in (0, 1/2)$ we have

$$\mathbb{E}_z\|g_T(x,z)-g_T(y,z)\|^2 \geq \mathbb{E}_z\left|[g_T(x,z)]_2-[g_T(y,z)]_2\right|^2$$
$$=(1-p)|\Psi(1)\Phi'(1/4-\delta)|^2+p\left|\tfrac{1}{p}\Psi(1)\Phi'(1/4-\delta)-\Psi(1)\Phi'(1/4)\right|^2,$$

which does not approach zero as $\delta \to 0$.

To overcome this issue, we replace the indicator $\mathbb{1}\{i > \text{prog}_{\frac{1}{4}}(x)\}$ with a smooth surrogate. Let $\Gamma : \mathbb{R} \to \mathbb{R}$ be any smooth non-decreasing Lipschitz function with $\Gamma(t) = 0$ for all $t \leq 1/4$ and $\Gamma(t) = 1$ for all $t \geq 1/2$. For each i, we define the following smoothed version of $\mathbb{1}\{i > \text{prog}_{\frac{1}{4}}(x)\}$:

$$\Theta_i(x) := \Gamma\left(1-\left(\sum_{k=i}^{T}\Gamma^2(|x_k|)\right)^{1/2}\right) = \Gamma\big(1-\big\|\Gamma\big(|x_{\geq i}|\big)\big\|\big), \tag{20}$$

where $\Gamma\big(|x_{\geq i}|\big)$ is a shorthand for a vector with non-zero entries $\Gamma(|x_i|), \Gamma(|x_{i+1}|), \ldots, \Gamma(|x_T|)$. Observe that Θ_i indeed acts as a smoothed indicator: We have $\Theta_i(x) = 1$ for all $i > \text{prog}_{\frac{1}{4}}(x)$ and $\Theta_i(x) = 0$ for all $i \leq \text{prog}_{\frac{1}{2}}(x)$, and therefore

$$\mathbb{1}\{i > \text{prog}_{\frac{1}{4}}(x)\} \leq \Theta_i(x) \leq \mathbb{1}\{i > \text{prog}_{\frac{1}{2}}(x)\}.$$

We define a new stochastic gradient function $\bar{g}_T$ by replacing the indicator function in g_T with the smoothed indicator Θ_i:

$$[\bar{g}_T(x,z)]_i := \nabla_i F_T(x)\cdot v_i(x,z),$$
$$\text{where } \; v_i(x,z) := 1+\Theta_i(x)\left(\frac{z}{p}-1\right), \tag{21}$$

and $z \sim \text{Bernoulli}(p)$. To fully specify the construction, we take

$$\Gamma(t) = \frac{\int_{1/4}^{t}\Lambda(\tau)d\tau}{\int_{1/4}^{1/2}\Lambda(\tau')d\tau'},$$
$$\text{where } \; \Lambda(t) = \begin{cases} 0, & t \leq \frac{1}{4} \text{ or } t \geq \frac{1}{2}, \\ \exp\left(-\frac{1}{100\left(t-\frac{1}{4}\right)\left(\frac{1}{2}-t\right)}\right), & \frac{1}{4} < t < \frac{1}{2}. \end{cases} \tag{22}$$

This is simply an integrated bump function construction; see Fig. 1.

Observation 1 The function Γ satisfies

1. $\Gamma(t) = 0$ for all $t \in (-\infty, 1/4]$.
2. $\Gamma(t) = 1$ for all $t \in [1/2, \infty)$.
3. $\Gamma \in \mathcal{C}^\infty$, with $0 \leq \Gamma'(t) \leq 6$ and $|\Gamma''(t)| \leq 128$ for all $t \in \mathbb{R}$.

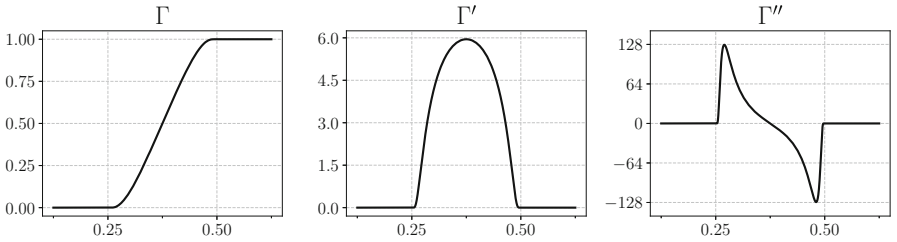

Fig. 1 The construction Γ in Eq. (22) and its derivatives; Observation 1 is evident

With these properties established, we prove the following mean-squared smooth analogue of Lemma 3.

Lemma 4 *The stochastic gradient estimator $\bar{g}_T$ is a probability-p zero-chain, is unbiased for ∇F_T, and satisfies*

$$\mathbb{E}\,\|\bar{g}_T(x,z) - \nabla F_T(x)\|^2 \le \frac{\varsigma^2(1-p)}{p}$$
$$\text{and } \mathbb{E}\,\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2 \le \frac{\bar{\ell}_1^2}{p}\|x-y\|^2, \tag{23}$$

for all $x, y \in \mathbb{R}^T$, where $\varsigma = 23$ and $\bar{\ell}_1 = 328$.

We defer the proof of Lemma 4 to Appendix A.2. The proofs for the probability-p zero-chain property and variance bound are similar to Lemma 3. For the mean-squared smooth property, we show that for any x, the vector $\delta(x,z) = \bar{g}_T(x,z) - \nabla F_T(x)$ has at most one non-zero coordinate, given by $\text{prog}_{\frac{1}{2}}(x) + 1$. If we denote $i_x = \text{prog}_{\frac{1}{2}}(x) + 1$ and $i_y = \text{prog}_{\frac{1}{2}}(y) + 1$, then we can bound $\mathbb{E}\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2$ by first appealing to smoothness of F_T, and then using the Lipschitz property of Θ_i to bound $\mathbb{E}\left|\delta_{i_x}(x,z) - \delta_{i_x}(y,z)\right|^2$ and $\mathbb{E}\left|\delta_{i_y}(x,z) - \delta_{i_y}(y,z)\right|^2$.

Our lower bound for the mean-squared smooth setting now follows from another simple scaling argument.

Theorem 2 *There exist numerical constants $c, c' > 0$ such that for all $\bar{L}, \Delta, \sigma^2 > 0$ and $\epsilon \le c'\sqrt{\bar{L}\Delta}$,*

$$\bar{\mathfrak{m}}_\epsilon^{zr}(K, \Delta, \bar{L}, \sigma^2) \ge c \cdot \left(\frac{\Delta\bar{L}\sigma}{\epsilon^3} + \frac{\Delta\bar{L}}{\epsilon^2} + \frac{\sigma^2}{\epsilon^2}\right).$$

Constructions of dimension $d = O(1 + \frac{\Delta\bar{L}}{\sigma\epsilon})$ realize the lower bound.

Theorem 2 is tight, since the upper bounds for SPIDER [25] and SNVRG[3] [56] match it up to constants. As with Theorem 1, the restriction $\epsilon \le O(\sqrt{\bar{L}\Delta})$ is essentially

[3] The iterates of SPIDER and SNVRG are a linear combination of previously computed gradients, and therefore these algorithms are zero-respecting.

without loss of generality. Theorem 2 leaves open the possibility that there exists an algorithm that achieves $O(\epsilon^{-3})$ in the mean-squared smooth setting using $K = 1$; see Sect. 6 for further discussion.

We defer the proof of Theorem 2 to Appendix A.3, as it is very similar to that of Theorem 1. In particular, it uses the same scaling argument and replaces L with roughly $\bar{L}\epsilon/\sigma$. This results in the final instance scaled as $F_T^\star(x) \propto \epsilon\sigma\bar{L}^{-1}F_T(\frac{\bar{L}x}{\sigma})$. The new scaling introduces an additional restriction that $\epsilon \leq O(\frac{\Delta\bar{L}}{\sigma})$. When this does not hold, one has $\frac{\Delta\bar{L}\sigma}{\epsilon^3} \geq c \cdot \frac{\sigma^2}{\epsilon^2}$, and the claimed lower bound follows from a standard estimation lower bound (see Lemma 10 in Appendix A.1).

4 Lower bounds for randomized algorithms

We now extend our lower bound construction for zero-respecting algorithms into a lower bound for arbitrary, potentially randomized algorithms. Our main theorem provides optimal lower bounds on the minimax complexities (7) and (8) for the bounded variance and mean-squared smooth settings.

Theorem 3 *There exist numerical constants $c, c' > 0$ such that for all $L, \Delta, \sigma^2 > 0$ and $\epsilon \leq c'\sqrt{L\Delta}$,*

$$\mathfrak{m}_\epsilon^{rand}(K, \Delta, L, \sigma^2) \geq c \cdot \left(\frac{\Delta L\sigma^2}{\epsilon^4} + \frac{\Delta L}{\epsilon^2}\right), \tag{24}$$

and for all $\bar{L} > 0$ and $\epsilon \leq c'\sqrt{\bar{L}\Delta}$, we have

$$\bar{\mathfrak{m}}_\epsilon^{rand}(K, \Delta, \bar{L}, \sigma^2) \geq c \cdot \left(\frac{\Delta\bar{L}\sigma}{\epsilon^3} + \frac{\Delta\bar{L}}{\epsilon^2} + \frac{\sigma^2}{\epsilon^2}\right). \tag{25}$$

Constructions of dimension $d = O\left(\frac{\Delta L}{\epsilon^2}\log\frac{K\Delta L\sigma^2}{\epsilon^4}\right)$ realize the lower bound (24), and constructions of dimension $d = O\left(1 + \frac{\Delta\bar{L}}{\sigma\epsilon}\log\frac{K\Delta\bar{L}\sigma}{\epsilon^3}\right)$ realize the lower bound (25).

In the remainder of the section we outline the proof of Theorem 3; we defer all formal proofs to Appendix B. Our approach is to lift the single hard instance developed in the previous section to a distribution over functions such that for any randomized algorithm a random function drawn from this distribution is hard with high probability. This approach closely follows [15, 51], though the analysis differs in a few technical points.

Given a function $F(x)$ and a gradient estimator $g(x, z)$, we define the rotated instance

$$\tilde{F}_U(x) := F(U^\top x), \quad \text{and} \quad \tilde{g}_U(x, z) := Ug(U^\top x, z),$$

where $U \in \mathsf{Ortho}(d, T) := \{U \in \mathbb{R}^{d\times T} | U^\top U = I_T\}$ is a matrix with orthogonal unit norm columns. For any such U we define an oracle for the rotated function according to

$$\mathsf{O}_{\tilde{F}_U}(x, z) = \tilde{g}_U(x, z). \tag{26}$$

When U is drawn uniformly from $\mathsf{Ortho}(d, T)$, any algorithm interacting with $\mathsf{O}_{\tilde{F}_U}$ produces queries $\{x^{(t,k)}\}$ such that the sequence $\{U^\top x^{(t,k)}\}$ behaves essentially like the queries of a zero-respecting algorithm interacting with of O_F. More precisely, for sufficiently large d we can guarantee that every entry of $U^\top x^{(t,k)}$ that is significantly far from zero (say, with absolute value $> 1/4$) is in the support of a previous oracle response $g(U^\top x^{(t',k')}, z^{(t')})$ for some $t' < t$ and $k' \in [K]$. This follows because oracle responses provide essentially no information on coordinates outside that support, and therefore, these coordinates of $U^\top x^{(t,k)}$ behave roughly as coordinates of a spherically uniform vector in dimension $d - t$, and we can obtain a high probability bound on their magnitude that scales as $\|x^{(t,k)}\|/\sqrt{d}$; the precise argument requires careful handling of the information leaked at each step. By assuming that the queries are bounded and choosing sufficiently large d, we guarantee that coordinates outside the support are smaller than $1/4$ and therefore that the zero-respecting structure obtains. Combining this structure with Definition 2 of probabilistic zero-chains implies control over $\mathrm{prog}_{\frac{1}{4}}(U^\top x^{(t,k)})$, as we state formally in the following generalization of Lemma 1, whose proof we provide in Appendix B.1.

Lemma 5 *Let $F : \mathbb{R}^T \to \mathbb{R}$ and let $g : \mathbb{R}^T \times \mathcal{Z} \to \mathbb{R}^T$ be probability-p zero chain. Let $R > 0$, $\delta \in (0, 1)$, and $\mathsf{A} \in \mathcal{A}_{rand}(K)$ be any algorithm that produces queries with norm bounded by R. Additionally let $d \geq \lceil T + 32R^2 \log \frac{2KT^2}{p\delta} \rceil$, U be uniform on $\mathsf{Ortho}(d, T)$, and $\mathsf{O}_{\tilde{F}_U}$ be as in (26). Then with probability at least $1 - \delta$,*[4]

$$\max_{k\in[K]} \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(t,k)}_{\mathsf{A}[\mathsf{O}_{\tilde{F}_U}]}) < T \;\; \textit{for all} \;\; t \leq \frac{T - \log\frac{2}{\delta}}{2p}. \tag{27}$$

Applying Lemma 5 to the hard instance $(F_T, \bar{g}_T)$ defined in Eq. (18) and (21) provides the lower bound we want, but restricted to algorithms with bounded iterates. To handle unbounded iterates, we follow [15] and compose the construction with a soft projection to a ball centered at the origin. Our final (unscaled) construction is

$$\widehat{F}_{T,U}(x) = F_T(U^\top \rho(x)) + \frac{\eta}{2}\|x\|^2, \;\; \text{where } \rho(x) = \frac{x}{\sqrt{1 + \|x\|^2/R^2}},$$
$$R = 230\sqrt{T}, \;\text{and } \eta = 1/5. \tag{28}$$

The corresponding stochastic gradient estimator is

$$\widehat{g}_{T,U}(x, z) = J(x)^\top U\, \bar{g}_T(U^\top \rho(x), z) + \eta \cdot x, \tag{29}$$

where $J(x) = \left[\frac{\partial \rho_i(x)}{\partial x_j}\right]_{i,j}$ is the Jacobian of ρ. The next lemma shows that this new construction is difficult for any algorithm in $\mathcal{A}_{\mathsf{rand}}$. The lemma has two components:

[4] The event holds with probability at least $1 - \delta$ with respect to the random choice of U and the oracle seeds $\{z^{(t)}\}$, even when conditioned over any randomness in A.

First, since the iterates always satisfy $\|\rho(x^{(t,k)})\| \leq R$, we can apply Lemma 5 to this sequence to control progress. Second, the additional regularization term in (28) ensures that we cannot make the gradient small by increasing the norm, so low progress indeed implies large gradient.

Lemma 6 *Let O be any oracle with $O_{\widehat{F}_{T,U}}(x,z) = \widehat{g}_{T,U}(x,z)$, where $\widehat{F}_{T,U}$ is the compressed and rotated hard instance (28) and $\widehat{g}_{T,U}$ is the corresponding probability-p zero chain (29). Let $\delta \in (0,1)$, $d \geq \lceil (32\cdot 230^2+1)T \log \frac{2KT^2}{p\delta} \rceil$, and U be uniformly distributed on $\mathsf{Ortho}(d,T)$. Then for any $A \in \mathcal{A}_{rand}(K)$, with probability at least $1-\delta$,*

$$\min_{k\in[K]} \left\| \nabla \widehat{F}_{T,U}(x^{(t,k)}_{A[O_{\widehat{F}_{T,U}}]}) \right\| \geq \frac{1}{2} \text{ for all } t \leq \frac{T - \log\frac{2}{\delta}}{2p}. \tag{30}$$

(See Appendix B.2 for a proof.)

All that remains is to verify that the final constructions (28) and (29) still satisfy the various boundedness properties required for the lower bound. The following bounds are a consequence of a generic result about rotation and soft projection, which we prove in Appendix B.3.

Lemma 7 *The function $\widehat{F}_{T,U}$ and stochastic gradient function $\widehat{g}_{T,U}$ satisfy the following properties for all $U \in \mathsf{Ortho}(d,T)$.*

1. $\widehat{F}_{T,U}(0) - \inf_x \widehat{F}_{T,U}(x) \leq \Delta_0 T$, *where* $\Delta_0 = 12$.
2. *The first derivative of $\widehat{F}_{T,U}$ is ℓ_1-Lipschitz continuous, where* $\ell_1 = 155$.
3. $\mathbb{E}\left\|\widehat{g}_{T,U}(x,z) - \nabla \widehat{F}_{T,U}(x)\right\|^2 \leq \frac{\varsigma^2(1-p)}{p}$ *for all* $x \in \mathbb{R}^d$, *where* $\varsigma = 23$.
4. $\mathbb{E}\|\widehat{g}_{T,U}(x,z) - \widehat{g}_{T,U}(y,z)\|^2 \leq \frac{\bar{\ell}_1^2}{p} \cdot \|x-y\|^2$ *for all* $x, y \in \mathbb{R}^d$, *where* $\bar{\ell}_1 = 336$.

5 Extensions

While Theorem 3 constitutes our main technical result, implying lower bounds for methods using stochastic first-order information, it is interesting to extend the bounds to allow more sophisticated querying strategies and more informative oracles.

5.1 Statistical learning oracles

To this point, our assumptions on the stochastic gradient function $g(x,z)$ concern only its first and second moments (requirements (2) and (4)). Yet the oracles in statistical learning and stochastic approximation problems often have the common structural property that $g(x,z)$ is the gradient of a function. Here we show that this property does not improve the worst-case complexity of stochastic optimization. Specifically, we consider oracles specified by a function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ for which

$$F(x) = \mathbb{E}[f(x,z)] \quad \text{and} \quad g(x,z) = \nabla_x f(x,z). \tag{31}$$

All of the lower bounds in this paper extend to this setting, at the cost of a slightly more involved construction. The idea is the same as in the preceding construction, but

to construct a valid function $f(x, z)$ with $F(x) = \mathbb{E}[f(x, z)]$ we apply the smoothed progress function to the function value for F_T rather than the gradient. Letting $\mathcal{Z} = \{0, 1\}$ be the oracle seed space, we define

$$f_T(x, z) = -\Psi(1)\Phi(x_1)v_1(x, z) + \sum_{i=2}^{T} \left[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right] v_i(x, z), \tag{32}$$

where $v_i(x, z)$ the smoothed indicator (21) and the random seed $z \sim \text{Bernoulli}(p)$. It is immediate that $\mathbb{E}[f_T(x, z)] = F_T(x)$. The stochastic gradient function $\nabla_x f_T(x, z)$ has a similar form to our previous construction $\bar{g}_T(x, z)$, but with nuisance terms arising from the gradient of the soft progress function itself. The thrust of the analysis for the new construction is to show that these nuisance terms do not spoil the key properties of $\bar{g}_T$.

Lemma 8 *The stochastic gradient function $\nabla_x f_T$ is a probability-p zero-chain, is unbiased for ∇F_T, and there exist numerical constants ς and $\bar{\ell}_1$ independent of p and T such that*

$$\mathbb{E}\|\nabla f_T(x, z) - \nabla F_T(x)\|^2 \le \frac{\varsigma^2}{p} \tag{33}$$

and

$$\mathbb{E}\|\nabla f_T(x, z) - \nabla f_T(y, z)\|^2 \le \frac{\bar{\ell}_1^2}{p}\|x - y\|^2 \tag{34}$$

for all $x, y \in \mathbb{R}^T$.

We prove Lemma 8 in Appendix C. With the lemma in hand, all that is required to prove the $\Omega(\frac{\Delta L \sigma^2}{\epsilon^4})$ lower bound for the bounded variance setting and the $\Omega(\frac{\Delta L \sigma}{\epsilon^3} + \frac{\sigma^2}{\epsilon^2})$ lower bound for the mean-squared smooth setting is to compose the instance with a rotation and soft projection as in (28), then rescale as in Theorem 3. This leads to the following result.

Proposition 1 *Theorem 3 holds (with different numerical constants) even when restricting the oracle class to statistical learning-type stochastic gradient functions of the form (31).*

5.2 Active oracles

Our main results consider a model in which the algorithm performs batches of K simultaneous queries, but the random seed z is drawn i.i.d. once per batch. Another stronger model allows *active* oracles, where the queries consist of both a point x and a seed z [22, 25, 32, 43, 44, 50, 56]. Active oracles are essential to finite-sum optimization problems where $F(x) = \sum_{i=1}^{n} f_i(x)$ and are more general than our K-query oracles,

since a randomized algorithm can simulate a K-query oracle using an active oracle by drawing $z \sim P_z$ and querying $(x^{(1)}, z), \ldots, (x^{(K)}, z)$. For *convex* finite-sum minimization problems, stochastic oracles are significantly weaker than active oracles [6]. Nevertheless, in this section we show that our ϵ^{-4} lower bound for zero-respecting algorithms (Theorem 1) extends to active oracles, even with additional finite-sum structure ($\mathcal{Z}$ is finite, P_z is uniform). We believe further extensions for randomized algorithms, mean-squared smooth gradient estimators and statistical learning oracles are straightforward, but we omit them for brevity.

The precise active oracle model we consider is as follows: at round i, the algorithm proposes a point $x^{(i)}$ and seed $z^{(i)}$ and receives an oracle response $\mathsf{O}_F(x^{(i)}, z^{(i)}) = g(x^{(i)}, z^{(i)})$. As before, we assume that, when $z^{(i)} \sim P_z$, the stochastic gradients are unbiased and have variance bounded by σ^2, and we allow the algorithm to know the distribution P_z.

The key step in converting our basic probabilistic zero-chain construction (18) to achieve a lower bound for the active finite-sum setting is to allow for independent randomness in each of the chain coordinates; this safeguards against algorithms that "abuse" the active oracle by repeatedly querying the same (informative) value of z. More formally, we take $\{0,1\}^T$ to be the oracle seed space and consider the stochastic gradient function $g_T^{\text{coord}} : \mathbb{R}^T \times \{0,1\}^T \to \mathbb{R}^T$,

$$\left[g_T^{\text{coord}}(x, z)\right]_i := \nabla_i F_T(x) \cdot \left(1 + \mathbb{1}\{i > \text{prog}_{\frac{1}{4}}(x)\}\left(\frac{z_i}{p} - 1\right)\right); \tag{35}$$

the only difference compared to the passive construction (18) is that the seed $z = (z_1, \ldots, z_T)$ is now a vector of T bits, and we use the ith bit only for coordinate i of the stochastic gradient function. If we draw the bits of z i.i.d. from a Bernoulli(p) distribution, then g_T^{coord} is unbiased for ∇F_T and satisfies the variance bound in Lemma 3.

The next step is to convert the distribution over $z \in \{0,1\}^T$ into a uniform distribution over a larger set, so that the instance has finite-sum structure. To do so, we assume without loss of generality that $p = 1/N$ for $N \in \mathbb{N}$ (we can always round $1/p = c \cdot \sigma^2/\epsilon^2$ appropriately). We choose $\mathcal{Z} = \{1, \ldots, N^T\}$ as the seed space and define $\zeta : \mathcal{Z} \to \{0,1\}^T$ as

$$\zeta_j(k) := \mathbb{1}\{\text{the } j \text{ th digit of } k \text{ in the } N \text{ -ary basis is } 0\}.$$

To obtain the hard active oracle construction, we take

$$g_\pi(x; i) := g_T^{\text{coord}}(x, \zeta(\pi(i))),$$

where π is any permutation of N^T elements. Note that for any choice of the permutation π, the random function $g_\pi(\cdot; i)$ with i uniform in $\mathcal{Z}$ has the same distribution as $g_T^{\text{coord}}(\cdot; z)$ with the elements of z i.i.d. Bernoulli(p), and therefore g_π is also unbiased for ∇F_T and satisfies the variance bound in Lemma 3. By choosing π to be a random permutation, the active oracle corresponding to g_π satisfies a progress bound analogous to Lemma 1.

Lemma 9 *Let $\delta \in (0,1)$, let $N, T > 1$ be integers, let π be a random permutation of N^T elements and consider the active oracle $\mathsf{O}^{\pi}_{F_T}(x,i) = g_{\pi}(x;i)$. Let $\{x^{(i)}\}$ be the iterates of any zero-respecting algorithm interacting with $\mathsf{O}^{\pi}_{F_T}$. Then, for $p = 1/N$, with probability at least $1-\delta$ over the random choice of π,*

$$\mathrm{prog}_0\left(x^{(t)}\right) < T, \quad \textit{for all } \; t \le \frac{T - \log(1/\delta)}{4p}.$$

We prove Lemma 9 in Appendix C.2 and sketch the intuition behind the result here. Let $(x^{(1)}, i^{(1)}), \ldots, (x^{(t)}, i^{(t)})$ be the algorithm's queries and $g^{(1)}, \ldots, g^{(t)}$ be the oracle responses up to some iteration t. Let $\gamma = \max_{t' < t} \mathrm{prog}_0(g^{(t')})$, so that $\mathrm{prog}_0(x^{(t)}) \le \gamma$ by the zero-respecting assumption. For an algorithm to guarantee $\mathrm{prog}_0(g^{(t)}) = 1+\gamma$ (and thereby make progress in x), the $(1+\gamma)$th coordinate of $\zeta(\pi(i^{(t)}))$ must be 1. The key observation is that the algorithm's previous queries provide very little information on $\zeta_{1+\gamma}(\pi(\cdot))$. In particular, we argue that after $t-1$ queries, the most we can possibly know is a set of $t-1$ indices i for which $\zeta_{1+\gamma}(\pi(i)) = 0$. Since all other indices are identically distributed, any query $i^{(t)}$ has probability at most $N^{T-1}/(N^T - (t-1))$ of satisfying $\zeta_{1+\gamma}(\pi(i^{(t)})) = 1$. Since $t \le T/p < N^T/2$, the probability of making a unit of progress at any iteration is no more that $2/N = 2p$, which gives the result via the same arguments that prove Lemma 3.

Using the same scaling arguments as in the proof of Theorem 1, Lemma 9 implies an analogous lower bound for the active setting. However, the distributional complexity we now lower bound is slightly different, because we randomize over the choice of oracles instead of choosing a fixed oracle. Consequently, we let the supremum in Eq. (7) be over all distributions P_{O} on $\mathcal{O}(K, \sigma^2)$, and take the expectation also with respect to a draw of $\mathsf{O} \sim P_{\mathsf{O}}$. (For zero-respecting lower bounds, we still replace $\mathcal{A}_{\mathsf{rand}}(K)$ with $\mathcal{A}_{\mathsf{zr}}(K)$ and it still suffices to consider point masses for P_F).

Proposition 2 *Theorem 1 also holds in the active oracle model, with the above complexity measure, finite $\mathcal{Z}$, and uniform P_z.*

This lower bound has the following implication on minimax complexity: For every zero-respecting algorithm there exists a "hard" active oracle (corresponding to some permutation of the coordinates) for a scaled version of F_T such that finding an ϵ-stationary point requires at least $\Omega(\epsilon^{-4})$ iterations.

Using the techniques of Sect. 4 we can lift these results to finite sum active oracle lower bounds for randomized algorithms. Moreover, the "different bit per coordinate" approach extends straightforwardly the mean-square smooth construction (21) as well as the "statistical learning" construction (32).

The set $\mathcal{Z}$ in the lower bounds described above is very large—since N scales as σ^2/ϵ^2 and T is polynomial in $1/\epsilon$, the cardinality $|\mathcal{Z}| = N^T$ is super-exponential in $1/\epsilon$. Designing lower bound constructions with smaller cardinality $|\mathcal{Z}| = n$ remains an open problem. We note that for the mean-square smooth setting, the smallest possible value for n is $\Omega(\sigma^2/\epsilon^2)$, since for $n = o(\sigma^2/\epsilon^2)$ the upper bound $O(\sqrt{n}\bar{L}\Delta\epsilon^{-2})$ attained by SPIDER [25] will be smaller than the desired n-independent lower bound $\Omega(\bar{L}\Delta\sigma\epsilon^{-3})$. We also remark that [25] prove a lower bound of $\Omega(\sqrt{n}\bar{L}\Delta\epsilon^{-2})$ for active oracles, but their construction does not keep the variance σ^2 bounded.

5.3 Oracles with zero-order information

Our lower bounds continue to hold for oracles of the form $\mathsf{O}_F(x,z) = (f(x,z), g(x,z))$ that provide, in addition to the gradient estimator $g(x,z)$, a function value estimator $f(x,z) \in \mathbb{R}$ satisfying $\mathbb{E}f(x,z) = F(x)$ for all x. Indeed, it is straightforward to extend the notion of a probabilistic zero-chain to such oracle by adding the requirement $f(x,z) = f(x_{\leq \mathrm{prog}_{\frac{1}{4}}(x)}, z)$ to the probability bound (11), i.e., requiring that

$$\mathbb{P}\big(\forall x : \mathrm{prog}_0(g(x,z)) \leq \mathrm{prog}_{\frac{1}{4}}(x) \text{ and } \mathsf{O}_F(x,z) = \mathsf{O}_F(x_{\leq \mathrm{prog}_{\frac{1}{4}}(x)}, z)\big) \geq 1-p,$$

and analogously extending the requirement (12) to

$$\mathbb{P}\big(\forall x : \mathrm{prog}_0(g(x,z)) \leq 1 + \mathrm{prog}_{\frac{1}{4}}(x) \text{ and } \mathsf{O}_F(x,z) = \mathsf{O}_F(x_{\leq 1+\mathrm{prog}_{\frac{1}{4}}(x)}, z)\big) = 1.$$

With this modified definition, it is straightforward to confirm that that Lemma 5 continues to hold for stochastic zero-order information. Moreover, the "statistical learning" oracle construction f_T in Eq. (32) satisfies the modified probabilistic zero-chain definition for $\mathsf{O}_F(x,z) = (f_T(x,z), \nabla_x f_T(x,z))$; we show this in the proof of Lemma 8. Calculating the variance of f_T and following the re-scaling argument in the proof of Theorem 3 reveals that the variance $\mathbb{E}(f(x,z) - F(x))^2$ of the resulting hard instance is of the order $\sigma^2\epsilon^2 L^{-2}$ in the basic smooth setting, and $\sigma^4 \bar{L}^{-2}$ in the mean-squared smooth setting. Therefore, the conclusion of Theorem 3 continues to hold even for oracles with quite accurate stochastic zero-order information.

It is possible to take one step further and extend our lower bounds to oracles that provide *exact* zero-order information, i.e., that return $\mathsf{O}_F(x,z) = (F(x), g(x,z))$. For zero-respecting algorithms this is trivial, since for such oracles the proof of 1 goes through unchanged, and consequently Theorems 1 and 2 hold as well. It is possible to also extend these lower bounds to general randomized algorithm, but only with higher-dimensional constructions. Specifically, in an earlier manuscript of this paper[5] we show lower bounds for oracles with exact zero-order information matching those of Theorem 3 but with domain dimension $\widetilde{O}\big(K\Delta^2 L^2 \sigma^2 \epsilon^{-6}\big)$ in the smooth case (compared to $\widetilde{O}\big(\Delta L \epsilon^{-2}\big)$ in Theorem 3) and $\widetilde{O}\big(K\Delta^2 \bar{L}^2 \epsilon^{-4}\big)$ in the mean-squared smooth case (compared to $\widetilde{O}\big(\Delta \bar{L} \sigma^{-1} \epsilon^{-1}\big)$ in Theorem 3). The difference in dimensionalities stems from a difference in proof strategies for Lemma 5: our earlier proof worked with a more relaxed notion of probabilistic zero-chains which allowed for exact zero-order information, but required random projection to a much higher-dimensional space.

At least part of that higher dimension, namely the linear dependence on K, is a *necessary* cost for obtaining lower bounds valid against exact zero-order oracles. To see why this is so, note that—for any smooth function F—using $K = d+1$ parallel exact function value queries we can compute ∇F at a single point to arbitrarily high precision via finite differences, thus simulating a noiseless gradient oracle for F. Therefore, any instance with dimension sublinear in K cannot show a lower bound

[5] Available at arxiv.org/abs/1912.02365v2.

better than the noiseless optimal rate of $\Theta(\Delta L \epsilon^{-2})$. In particular, such hard instances cannot exhibit the additional complexity incurred by noisy gradient oracles.

6 Discussion

We have established tight lower bounds on the stochastic first-order complexity of finding stationary points for non-convex functions, with and without mean-squared smoothness. We hope that the basic ideas behind our lower bound constructions will find further use in non-convex stochastic optimization. A few natural open questions and future directions along these lines are as follows.

*Lower bounds for mean-squared smooth oracles with a single query*In the mean-squared smooth setting, all known algorithms that achieve the optimal $O(\epsilon^{-3})$ oracle complexity (SPIDER [25], SNVRG [56]) require $K = 2$ simultaneous queries. With $K = 1$, the best result known for the mean-squared smooth setting is still the standard $O(\Delta L \sigma^2 \epsilon^{-4})$ rate obtained by SGD. However, under additional higher-order smoothness assumptions, perturbed SGD can achieve convergence $O(\epsilon^{-3.5})$ with $K = 1$ [26]. It remains an open question whether any algorithm can achieve complexity scaling as ϵ^{-3} when $K = 1$, or whether the ϵ^{-4} rate of SGD is optimal.

Lower bounds under additional oracle assumptions Rather than assuming a mean-squared smooth oracle, one can make the stronger assumption that the stochastic gradient function $g(\cdot, z)$ is smooth almost surely, or assume that the error $\|g(x, z) - \nabla F(x)\|$ is bounded by σ almost surely. We are not aware of any algorithms that leverage such stronger assumptions, and yet extending our lower bounds to handle them seems non-trivial. Resolving the importance of these assumptions therefore remains an interesting topic for future work.

Lower bounds for higher-order algorithms Our results resolve the complexity of finding first-order stationary points with stochastic first-order methods, but we have not addressed the oracle complexity of other basic non-convex stochastic optimization problems, such as finding first-order stationary points with higher-order smoothness (possibly with stochastic access to Hessian, Hessian vector-products, or other higher-order derivatives) or finding second-order stationary points. Building on our work, [8] provide upper and lower bounds for finding first- and second-order stationary points using stochastic pth-order gradient information. In particular, they show that when the objective is second-order smooth, an algorithm using stochastic gradients and stochastic Hessian-vector products can find an ϵ-stationary point using order ϵ^{-3} queries. They also show lower bounds proving that this is unimprovable, even when using pth order derivative information for any $p \geq 2$, and even when the objective is pth order smooth.

Acknowledgements Part of this work was completed while the authors were visiting the Simons Institute for the Foundations of Deep Learning program. We thank Ayush Sekhari, Ohad Shamir, Aaron Sidford and Karthik Sridharan for several helpful discussions. YC was supported by the Stanford Graduate Fellowship. JCD acknowledges support from NSF CAREER award 1553086, the Sloan Foundation, and ONR-YIP N00014-19-1-2288. DF was supported by NSF TRIPODS award #1740751. BW was supported by the Google PhD Fellowship program. Division of Computing and Communication Foundations (Grant Number 1553086

Appendix

A Proofs from Section 3

A.1 Basic technical results

Before proving the main results from Sect. 3, we first state two self-contained technical results that will be used in subsequent proofs. The first result bounds component functions Ψ and Φ and gives the calculation for the parameter ℓ_1 in Lemma 2.2.

Observation 2 The functions Ψ and Φ in (17) and their derivatives satisfy

$$0 \le \Psi \le e, \quad 0 \le \Psi' \le \sqrt{54/e}, \quad |\Psi''| \le 32.5, \quad 0 \le \Phi \le \sqrt{2\pi e}, \\ 0 \le \Phi' \le \sqrt{e} \text{ and } |\Phi''| \le 1. \tag{36}$$

Proof of Lemma 2.2 We note that the Hessian of F_T is tridiagonal. Consequently, for any $x \in \mathbb{R}^d$,

$$\begin{aligned} \|\nabla^2 F_T(x)\|_{\text{op}} &\le \max_{i\in[T]} |\nabla^2_{i,i} F_T(x)| + \max_{i\in[T]} |\nabla^2_{i,i+1} F_T(x)| + \max_{i\in[T]} |\nabla^2_{i+1,i} F_T(x)| \\ &\overset{(i)}{\le} \sup_{z\in\mathbb{R}} |\Phi''(z)| \sup_{z\in\mathbb{R}} |\Psi(z)| + \sup_{z\in\mathbb{R}} |\Phi(z)| \sup_{z\in\mathbb{R}} |\Psi''(z)| \\ &\quad + 2 \sup_{z\in\mathbb{R}} |\Phi'(z)| \sup_{z\in\mathbb{R}} |\Psi'(z)| \overset{(ii)}{\le} 152, \end{aligned}$$

where (i) is a direct calculation using the definition (16) of F_T and (ii) follows from (36). □

The second result is an $\Omega(\frac{\sigma^2}{\epsilon^2})$ lower bound on the sample complexity of finding stationary points whenever $\epsilon \le O(\sqrt{\Delta L})$. This result handles an edge case in the proof of Theorem 2. A similar lower bound appeared in [27], but the result we prove here is slightly stronger because it holds even for dimension $d = 1$.

Lemma 10 *There exists a number $c_0 > 0$ such that for any number of simultaneous queries K, dimension d and $\epsilon \le \sqrt{\frac{\bar{L}\Delta}{8}}$, we have*

$$\bar{\mathfrak{m}}^{zr}_\epsilon(K, \Delta, \bar{L}, \sigma^2) \ge \bar{\mathfrak{m}}^{rand}_\epsilon(K, \Delta, \bar{L}, \sigma^2) \ge c_0 \cdot \frac{\sigma^2}{\epsilon^2}. \tag{37}$$

Our approach for proving Lemma 10 is as follows. Given a dimension d, we construct a function $F : \mathbb{R}^d \to \mathbb{R}$, a family of distributions P_z, and a family of functions $f(x, z)$ for which $F(x) = \mathbb{E}_z[f(x, z)]$, and for which the initial suboptimality, variance, and mean-squared smoothness are bounded by Δ, σ^2 and $\bar{L}$, respectively. We then prove a lower bound in the *global stochastic model* in which at round t the oracle returns the full function $f(\cdot, z^{(t)})$, rather than just its value and derivatives at the queried point. The global stochastic model is more powerful than the K-query

stochastic first-order model (with $g(x, z) = \nabla_x f(x, z)$) for every value of K, so this will imply the claimed result as a special case.

Lemma 11 *Whenever $\epsilon \leq \sqrt{\frac{\bar{L}\Delta}{8}}$, the number of samples required to obtain an ϵ-stationary point in the global stochastic model defined above is $\Omega(1) \cdot \frac{\sigma^2}{\epsilon^2}$.*

Proof of Lemma 11 The proof follows standard arguments used to derive information-theoretic lower bounds for statistical estimation [31, 54].

We consider a family of functions $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ given by

$$f(x, z) = \frac{\bar{L}}{2}\left(\|x\|^2 - 2zx_1 + r^2\right), \tag{38}$$

where $r \in (0, \sqrt{2\Delta/\bar{L}})$ is a fixed parameter. We take P_z to have the form $P_z^s := \mathcal{N}(rs, \frac{\sigma^2}{\bar{L}^2})$, where $s \in \{-1, 1\}$, and let $\theta_s := (rs, 0, \ldots, 0) \in \mathbb{R}^d$. Then, when $P_z = P_z^s$, we have $F_s(x) := \mathbb{E}_z[f(x, z)] = \frac{\bar{L}}{2}\|x - \theta_s\|^2$, and furthermore for any $x, y \in \mathbb{R}^d$ we have

$$\mathbb{E}_z[\|\nabla_x f(x, z) - \nabla F_s(x)\|^2] = \bar{L} \cdot \mathbb{E}_z[(z - rs)^2] = \sigma^2,$$

and

$$\mathbb{E}_z[\|\nabla_x f(x, z) - \nabla_x f(y, z)\|^2] = \bar{L}^2 \cdot \|x - y\|^2.$$

Note that F_s is indeed an $\bar{L}$-smooth, and has initial suboptimality at $x^{(0)} = 0$ bounded as $F_s(0) - \inf_{x \in \mathbb{R}^d} F_s(x) = \bar{L}r^2/2 \leq \Delta$.

Now, we provide a distribution over the underlying instance by drawing S uniformly from $\{\pm 1\}$, and consider any algorithm that takes as input samples $z_1, \ldots, z_T \sim P_z^S$, and returns iterate $\hat{x}$. To bound the expected norm of the gradient at $\hat{x}$ (over the randomness of the oracle, the randomness of the algorithm, and the choice of the underlying instance S), we define $\hat{S} := \arg\min_{s' \in \{1,-1\}} \|\nabla F_{s'}(\hat{x})\|$, with ties broken arbitrarily. Observe that we have

$$\mathbb{E}[\|\nabla F_S(\hat{x})\|] \overset{(i)}{\geq} r\bar{L}\mathbb{P}\left(\|\nabla F_S(\hat{x})\| \geq r\bar{L}\right) \overset{(ii)}{\geq} r\bar{L}\mathbb{P}(\hat{S} \neq S), \tag{39}$$

where (i) follows by Markov's inequality and (ii) follows because when $\hat{S} \neq S$, the definition of $\hat{S}$ implies

$$\begin{aligned} 2 \cdot \|\nabla F_S(\hat{x})\| &\geq \inf_{x \in \mathbb{R}^d} \{\|\nabla F_{-1}(x)\| + \|\nabla F_1(x)\|\} \\ &= \bar{L} \cdot \inf_{x \in \mathbb{R}^d} \{\|x - \theta_1\| + \|x - \theta_{-1}\|\} \geq \bar{L}\|\theta_1 - \theta_{-1}\| = 2r\bar{L}. \end{aligned}$$

Next, for $s \in \{\pm 1\}$ let $\mathbb{P}_s = \mathcal{N}^{\otimes T}(rs, \frac{\sigma^2}{\bar{L}^2})$ denote the law of $(z_1, \ldots, z_T)$ conditioned on $S = s$. We have

$$\begin{aligned}
\mathbb{P}(\hat{S} \neq S) = 1 - \mathbb{P}(\hat{S} = S) &\geq 1 - \frac{1}{2} \sup_{A \text{ is measurable}} \{\mathbb{P}_1(A) + \mathbb{P}_{-1}(A^c)\} \\
&= \frac{1}{2} - \frac{1}{2} \sup_{A \text{ is measurable}} \{\mathbb{P}_1(A) - \mathbb{P}_{-1}(A)\} \\
&= \frac{1}{2} \{1 - \|\mathbb{P}_1 - \mathbb{P}_{-1}\|\} \\
&\geq \frac{1}{2} \left\{1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_1 || \mathbb{P}_{-1})}\right\} \\
&= \frac{1}{2} \left(1 - \frac{r\bar{L}\sqrt{T}}{\sigma}\right),
\end{aligned}$$

where the penultimate step follows by Pinsker's inequality and the last step uses that $\mathbb{P}_s = \mathcal{N}^{\otimes T}(rs, \frac{\sigma^2}{\bar{L}^2})$. Combining this lower bound with (39) yields

$$\mathbb{E}[\|F_S(\hat{x})\|] \geq \frac{r\bar{L}}{2} \left(1 - \frac{r\bar{L}\sqrt{T}}{\sigma}\right).$$

Finally, setting $r = \min\{\frac{\sigma}{2\bar{L}\sqrt{T}}, \sqrt{\frac{2\Delta}{\bar{L}}}\}$, implies

$$\begin{aligned}
\max\left\{\mathbb{E}[\|F_1(\hat{x})\|], \mathbb{E}[\|F_{-1}(\hat{x})\|]\right\} &\geq \frac{1}{2}\left(\mathbb{E}[\|F_1(\hat{x})\|] + \mathbb{E}[\|F_{-1}(\hat{x})\|]\right) \\
&= \mathbb{E}[\|F_S(\hat{x})\|] \geq \min\{\frac{\sigma}{8\sqrt{T}}, \sqrt{\frac{\bar{L}\Delta}{8}}\}.
\end{aligned}$$

Stated equivalently, whenever $\epsilon \leq \sqrt{\bar{L}\Delta/8}$, there exists $s \in \{-1, 1\}$ such that the number of oracle calls T required to ensure $\mathbb{E}[\|\nabla F_s(\hat{x})\|] \leq \epsilon$ satisfies

$$T \geq \frac{\sigma^2}{64\epsilon^2},$$

concluding the proof. □

A.2 Proof of Lemma 4

First, we note that $\mathbb{E}[\nu_i(x, z)] = 1$ for all x and i, and therefore $\mathbb{E}[\bar{g}_T(x, z)] = \nabla F_T(x)$. To show the probabilistic zero-chain property, note that, due to Observation 1.1, we have $\nu_i(x, z) = \nu_i(x_{\leq \mathrm{prog}_{\frac{1}{4}}(x)}, z)$ for all x and z. Moreover, for $i \geq 1 + \mathrm{prog}_{\frac{1}{4}}(x)$ we have $\Gamma(|x_{\geq i}|) = 0$ and therefore $\Theta_i(x) = \Gamma(1) = 1$ and $\nu_i(x, 0) = 0$.

With these observation, the proof of the zero-chain property is analogous its proof in Lemma 3: for all x and z we have $\text{prog}_0(\bar{g}_T(x,z)) \le 1 + \text{prog}_{\frac{1}{4}}(x)$ (from Lemma 2.4) and $\bar{g}_T(x,z) = \bar{g}_T\left(x_{\le 1+\text{prog}_{\frac{1}{4}}(x)}, z\right)$ (from Lemma 2.5 and $\nu_i(x,z) = \nu_i(x_{\le \text{prog}_{\frac{1}{4}}(x)}, z)$), giving Eq. (12); for $z = 0$ we further have $\text{prog}_0(\bar{g}_T(x,z)) \le \text{prog}_{\frac{1}{4}}(x)$ (from $\nu_i(x,0) = 0$) and $\bar{g}_T(x,z) = \bar{g}_T\left(x_{\le \text{prog}_{\frac{1}{4}}(x)}, z\right)$ (from Lemma 2.5 and $\nu_i(x,z) = \nu_i(x_{\le \text{prog}_{\frac{1}{4}}(x)}, z)$), giving Eq. (11).

To bound the variance of the gradient estimator we observe that for all $i \le \text{prog}_{\frac{1}{2}}(x)$, $\|\Gamma(|x_{\ge i}|)\| \ge \Gamma(1/2) = 1$ and therefore $\Theta_i(x) = 0$ and $\nu_i(x,z) = 1$, so that

$$[\bar{g}_T(x,z)]_i = \nabla_i F_T(x) \ \ \forall i \le \text{prog}_{\frac{1}{2}}(x).$$

On the other hand, Lemma 2.4 gives us that

$$[\bar{g}_T(x,z)]_i = \nabla_i F_T(x) = 0 \ \ \forall i > 1 + \text{prog}_{\frac{1}{2}}(x).$$

We conclude that $\delta(x,z) = \bar{g}_T(x,z) - \nabla F_T(x)$ has at most a single nonzero entry in coordinate $i_x = \text{prog}_{\frac{1}{2}}(x) + 1$. Moreover, for every i

$$\delta_i(x,z) = \nabla_i F_T(x)(\nu_i(x,z) - 1) = \nabla_i F_T(x)\Theta_i(x)\left(\frac{z}{p} - 1\right).$$

Therefore,

$$\mathbb{E}\|\bar{g}_T(x,z) - \nabla F_T(x)\|^2 = \mathbb{E}\delta_i^2(x,z) = \left|\nabla_{i_x} F_T(x)\right|^2 \Theta_i^2(x)\,\frac{1-p}{p} \le \frac{(1-p)23^2}{p},$$

where the final transition used Lemma 2.3 and $\Theta_i^2(x) \le 1$ for all x and i, establishing the variance bound in (23) with $\varsigma = 23$.

To bound $\mathbb{E}\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2$, we use that $\mathbb{E}[\delta(\cdot,z)] = 0$ and that $\delta(\cdot,z)$ has at most one nonzero coordinate to write

$$\begin{aligned}\mathbb{E}\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2 &= \mathbb{E}\|\delta(x,z) - \delta(y,z)\|^2 + \|\nabla F_T(x) - \nabla F_T(y)\|^2 \\ &= \sum_{i\in\{i_x,i_y\}} \mathbb{E}(\delta_i(x,z) - \delta_i(y,z))^2 + \|\nabla F_T(x) - \nabla F_T(y)\|^2,\end{aligned} \tag{40}$$

where $i_y = \text{prog}_{\frac{1}{2}}(y) + 1$ is the nonzero index of $\delta(y,z)$. For any $i \le T$, we have

$$\begin{aligned}&\mathbb{E}(\delta_i(x,z) - \delta_i(y,z))^2 \\ &\quad= (\nabla_i F_T(x)\Theta_i(x) - \nabla_i F_T(y)\Theta_i(y))^2\,\mathbb{E}\left(\frac{z}{p} - 1\right)^2\end{aligned}$$

$$= (\nabla_i F_T(x)(\Theta_i(x) - \Theta_i(y)) + (\nabla_i F_T(x) - \nabla_i F_T(y))\Theta_i(y))^2 \frac{1-p}{p}$$
$$\leq \left(2(\nabla_i F_T(x))^2(\Theta_i(x) - \Theta_i(y))^2 + 2(\nabla_i F_T(x) - \nabla_i F_T(y))^2\Theta_i^2(y)\right)\frac{1}{p}\,.$$

By Observation 1.3, Γ_i is 6-Lipschitz. Since the Euclidean norm $\|\cdot\|$ is 1-Lipschitz, we have

$$|\Theta_i(x) - \Theta_i(y)| \leq 6\left|\,\left\|\Gamma(|x_{\geq i}|)\right\| - \left\|\Gamma(|y_{\geq i}|)\right\|\,\right| \leq 6\left\|\,\Gamma(|x_{\geq i}|) - \Gamma(|y_{\geq i}|)\,\right\|$$
$$\leq 6^2\left\|\,|x_{\geq i}| - |y_{\geq i}|\,\right\| \leq 6^2\|x-y\|.$$

That is, Θ_i is 6^2-Lipschitz. Since $\Theta_i^2(y) \leq 1$ and $(\nabla_i F_T(x))^2 \leq 23^2$ by Lemma 2.3, we have

$$(\delta_i(x,z) - \delta_i(y,z))^2 \leq \frac{(23\cdot 6)^2\|x-y\|^2 + 2(\nabla_i F_T(x) - \nabla_i F_T(y))^2}{p}$$

for all i. Substituting back into (40) we obtain

$$\mathbb{E}\|\bar{g}_T(x,z) - \bar{g}_T(y,z)\|^2 \leq \frac{2\cdot(23\cdot 6)^2\|x-y\|^2 + 2\|\nabla F_T(x) - \nabla F_T(y)\|^2}{p}$$
$$+ \|\nabla F_T(x) - \nabla F_T(y)\|^2.$$

Recalling that $\|\nabla F_T(x) - \nabla F_T(y)\| \leq \ell_1\|x-y\|$ by Lemma 2.2, establishes the mean-square smoothness bound in (23) with $\bar{\ell}_1 = \sqrt{2\cdot(\varsigma\cdot 6)^2 + 3\ell_1^2}$.

A.3 Proof of Theorem 2

Let Δ_0, ℓ_1, ς and $\bar{\ell}_1$ be the numerical constants in Lemmas 2.1, 2.2 and 4, respectively. Let the accuracy parameter ϵ, initial suboptimality Δ, mean-squared smoothness parameter $\bar{L}$, and variance parameter σ^2 be fixed, and let $L \leq \bar{L}$ be specified later. We rescale F_T as in the proof of Theorem 1,

$$F_T^\star(x) = \frac{L\lambda^2}{\ell_1} F_T\left(\frac{x}{\lambda}\right),$$
$$\text{where}\quad \lambda = \frac{\ell_1}{L}\cdot 2\epsilon, \quad \text{and}\quad T = \left\lfloor \frac{\Delta}{\Delta_0(L\lambda^2/\ell_1)} \right\rfloor = \left\lfloor \frac{L\Delta}{\Delta_0\ell_1(2\epsilon)^2} \right\rfloor.$$

This guarantees that $F_T^\star \in \mathcal{F}(\Delta, L)$ and that the corresponding scaled gradient estimator $g_T^\star(x,z) = (L\lambda/\ell_1)\bar{g}_T(x/\lambda, z)$ is such that every zero respecting algorithm A interacting with $\mathsf{O}_{F_T^\star}(x,z) = (F_T^\star(x), g_T^\star(x,z))$ satisfies

$$\mathbb{E}\left\|\nabla F_T^\star\left(x_{\mathsf{A}[\mathsf{O}_F]}^{(t,k)}\right)\right\| > \epsilon,$$

for all $t \leq (T-1)/2p$ and $k \in [K]$. It remains to choose p and L such that $\mathsf{O}_{F_T^\star}$ belongs to $\mathcal{O}(K, \sigma^2, \bar{L})$. As in the proof of Theorem 1, setting $\frac{1}{p} = \frac{\sigma^2}{(2\varsigma\epsilon)^2} + 1$ and using Lemma 4 guarantees a variance bound of σ^2. Moreover, by Lemma 4 we have

$$\begin{aligned}\mathbb{E}\|g_T^\star(x,z) - g_T^\star(y,z)\|^2 &= \left(\frac{L\lambda}{\ell_1}\right)^2 \mathbb{E}\left\|\bar{g}_T\left(\frac{x}{\lambda}, z\right) - \bar{g}_T\left(\frac{y}{\lambda}, z\right)\right\|^2 \\ &\leq \left(\frac{L\lambda}{\ell_1}\right)^2 \frac{\bar{\ell}_1^2}{p}\left\|\frac{x}{\lambda} - \frac{y}{\lambda}\right\|^2 \\ &= \left(\frac{\bar{\ell}_1 L}{\ell_1\sqrt{p}}\right)^2 \|x-y\|^2.\end{aligned}$$

Therefore, taking

$$L = \frac{\ell_1}{\bar{\ell}_1}\bar{L}\sqrt{p} = \frac{\ell_1}{\bar{\ell}_1}\min\left\{\frac{2\varsigma\epsilon}{\sigma}, 1\right\}\bar{L} \leq \bar{L}$$

guarantees membership in the oracle class and implies the lower bound

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{zr}}(K, \Delta, \bar{L}, \sigma^2) > \frac{T-1}{2p} = \left(\left\lfloor \frac{\bar{L}\Delta\sqrt{p}}{4\bar{\ell}_1\Delta_0\epsilon^2}\right\rfloor - 1\right)\frac{1}{2p}.$$

We consider the cases $\frac{\bar{L}\Delta\sqrt{p}}{4\bar{\ell}_1\Delta_0\epsilon^2} \geq 3$ and $\frac{\bar{L}\Delta\sqrt{p}}{4\bar{\ell}_1\Delta_0\epsilon^2} < 3$ separately. In the former case (which is the more interesting one), we use $\lfloor x \rfloor - 1 \geq x/2$ for $x \geq 3$ and the setting of p to write

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{zr}}(K, \Delta, \bar{L}, \sigma^2) \geq \frac{\bar{L}\Delta}{16\bar{\ell}_1\Delta_0\epsilon^2\sqrt{p}} \geq \frac{1}{64\bar{\ell}_1\Delta_0\varsigma}\cdot\frac{\bar{L}\Delta\sigma}{\epsilon^3} + \frac{1}{32\bar{\ell}_1\Delta_0}\cdot\frac{\bar{L}\Delta}{\epsilon^2}. \quad (41)$$

Moreover, we choose $c' = 12\bar{\ell}_1\Delta_0$ so that $\epsilon \leq \sqrt{\frac{\bar{L}\Delta}{12\bar{\ell}_1\Delta_0}} \leq \sqrt{\frac{\bar{L}\Delta}{8}}$ holds. By Lemma 11,

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{zr}}(K, \Delta, \bar{L}, \sigma^2) > c_0 \cdot \frac{\sigma^2}{\epsilon^2}, \quad (42)$$

where c_0 is a universal constant (this lower bound holds for any value of d). Together, the bounds (41) and (42) imply the desired result when $\frac{\bar{L}\Delta\sqrt{p}}{4\bar{\ell}_1\Delta_0\epsilon^2} \geq 3$.

Finally, we consider the edge case $\frac{\bar{L}\Delta\sqrt{p}}{4\bar{\ell}_1\Delta_0\epsilon^2} < 3$. We note that the assumption $\epsilon \leq \sqrt{\frac{\bar{L}\Delta}{12\bar{\ell}_1\Delta_0}}$ precludes the option that $p = 1$ in this case. Therefore we must have $\frac{\bar{L}\Delta\varsigma}{2\bar{\ell}_1\Delta_0\sigma\epsilon} < 3$ or, equivalently, $\frac{\sigma^2}{\epsilon^2} > \frac{\varsigma}{6\bar{\ell}_1\Delta_0}\cdot\frac{\bar{L}\Delta\sigma}{\epsilon^3}$. Thus, in this case the bound (42) implies (41) up to a constant, concluding the proof.

B Proofs from Section 4

B.1 Proof of Lemma 5

The proof combines the techniques of the proofs of Lemma 1 and random-projection based lower bounds on the sequential oracle complexity of optimization [15, 51] in their refined form [13, 23] which yields low-dimensional hard instances.

Let us adopt the shorthand $x^{(i)} := x^{(i)}_{\mathsf{A}[\mathsf{O}_{\tilde{F}_U}]}$, which we recall is defined via

$$x^{(i)}_{\mathsf{A}[\mathsf{O}_{\tilde{F}_U}]} = \mathsf{A}^{(i)}\Big(r, \mathsf{O}_{\tilde{F}_U}\big(x^{(1)}_{\mathsf{A}[\mathsf{O}_{\tilde{F}_U}]}, z^{(1)}\big), \ldots, \mathsf{O}_{\tilde{F}_U}\big(x^{(i-1)}_{\mathsf{A}[\mathsf{O}_{\tilde{F}_U}]}, z^{(i-1)}\big)\Big),$$

where r is the algorithm's random seed. Following the proof strategy of Lemma 1, we define

$$\begin{aligned}\pi^{(t)} &= \max_{i\le t}\max_{k\in[K]} \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(i,k)}) \\ &= \max\left\{j \le T \,|\, |\langle u^{(j)}, x^{(i,k)}\rangle| \ge \tfrac{1}{4} \text{ for some } i \le t, k \in [K]\right\}\end{aligned}$$

and

$$B^{(t)} := \mathbb{1}\{\exists x : \mathrm{prog}_0\Big(g(x; z^{(t)})\Big) = 1 + \mathrm{prog}_{\frac{1}{4}}(x)\}, \tag{43}$$

recalling that, due to Definition 2, $\{B^{(t)}\}_{t\ge 1}$ are i.i.d. Bernoulli with probability of success at most p, and are independent of any randomization in the algorithm A. With the shorthand

$$C^{(t)} := \sum_{s<t} B^{(s)}$$

We additionally define, for every $t \ge 1$, the event

$$\mathfrak{E}^{(t)} := \bigcap_{s\le t}\{\pi^{(s)} \le C^{(s)}\}.$$

Writing

$$T_p := \left\lfloor \frac{T - \log(2/\delta)}{2p} \right\rfloor,$$

the claim of the lemma is equivalent to the statement that $\mathbb{P}(\pi^{(T_p)} \ge T) \le \delta$. Since

$$\mathbb{P}(\pi^{(T_p)} \ge T) \le \mathbb{P}\Big(\big[\mathfrak{E}^{(T_p)}\big]^c \text{ or } C^{(T_p)} \ge T\Big) \le \mathbb{P}\Big(\big[\mathfrak{E}^{(T_p)}\big]^c\Big) + \mathbb{P}\Big(C^{(T_p)} \ge T\Big),$$

it suffices to show that both

$$\mathbb{P}\Big(\big[\mathfrak{E}^{(T_p)}\big]^c\Big) \le \frac{\delta}{2}. \tag{44}$$

and

$$\mathbb{P}\Big(C^{(T_p)} \ge T\Big) \le \frac{\delta}{2}. \tag{45}$$

The bound (45) follows identically to Eq. (15) in the proof of Lemma 1, and so the remainder of the proof consists of establishing the bound (44).

We begin by rewriting the event $\big[\mathfrak{E}^{(T_p)}\big]^c$ as follows

$$\begin{aligned}
\big[\mathfrak{E}^{(T_p)}\big]^c &= \bigcup_{t \le T_p} \Big\{\pi^{(t)} > C^{(t)}\Big\} \cap \mathfrak{E}^{(t-1)} \\
&= \bigcup_{t \le T_p} \Big\{\max_{k \in [K]} \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(t,k)}) > C^{(t)}\Big\} \cap \mathfrak{E}^{(t-1)} \\
&= \bigcup_{t \le T_p} \Big\{\exists k \in [K],\, j > C^{(t)} : |\langle u^{(j)}, x^{(t,k)}\rangle| \ge \frac{1}{4}\Big\} \cap \mathfrak{E}^{(t-1)}.
\end{aligned}$$

Define the σ-field

$$\mathcal{F} := \sigma\Big(z^{(1)}, \ldots, z^{(T_p)}, r\Big),$$

where we recall that r is the algorithm's random seed, and note that $C^{(t)} \in \mathcal{F}$ for all $t \le T_p$. Conditioning on $\mathcal{F}$ and applying the union bound, we have

$$\mathbb{P}\Big(\big[\mathfrak{E}^{(T_p)}\big]^c \;\Big|\; \mathcal{F}\Big) \le \sum_{t \le T_p} \sum_{k \in [K]} \sum_{j=1+C^{(t)}}^{T} \mathbb{P}\Big(|\langle u^{(j)}, x^{(t,k)}\rangle| > \frac{1}{4}\,,\ \mathfrak{E}^{(t-1)} \;\Big|\; \mathcal{F}\Big). \tag{46}$$

Therefore, to establish the bound (44) and with it the result, it suffices to prove that the probability $\mathbb{P}\big(|\langle u^{(j)}, x^{(t,k)}\rangle| > \frac{1}{4}\,,\ \mathfrak{E}^{(t-1)} \;\big|\; \mathcal{F}\big) \le \frac{\delta}{2T_pKT}$ for every $t \le T_p$, $k \in [K]$ and $j > C^{(t)}$. To do so, we leverage probabilistic zero-chain property in order show the following.

Lemma 12 *Fix $t \ge 1$, and condition on $\mathcal{F}$. If $\mathfrak{E}^{(t-1)}$ holds, then for every $s \le t$ and $k \in [K]$, $x^{(s,k)}$ is measurable with respect to $u^{(1)}, \ldots, u^{(C^{(s)})}$.*

Proof Throughout the proof, we adopt the shorthand $U_{\le c}$ for $[u^{(1)}; \ldots u^{(c)}; 0, \ldots, 0]$, i.e., a version of U where the last $T - c$ columns are replaced with zeros. We also recall the notation $x_{\le i}$ for the replacement of all but the first i coordinates of x with zeros.

The crux of the proof is the following claim: for any $s < t$ and $k \in [K]$, if $x^{(s,k)}$ is measurable w.r.t. $U_{\leq C^{(s)}}$ and $\mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s,k)}) \leq C^{(s)}$, then the oracle response to query $x^{(s,k)}$ is measurable w.r.t. $U_{\leq C^{(s+1)}}$. To see why this holds, let $g^{(s,k)} = g(U^\top x^{(s,k)}, z^{(s)})$ and note that Definition 2 of the probabilistic zero chain, along with definition (43) of the sequence $(B^{(t)})$, implies that

$$\begin{aligned}\mathrm{prog}_0(g^{(s,k)}) \leq B^{(s)} + \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s,k)}) \ \text{ and } \ g^{(s,k)} \\ = g\left(\left[U^\top x^{(s,k)}\right]_{\leq B^{(s)} + \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s,k)})}, z^{(s)}\right).\end{aligned}$$

The assumption $\mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s,k)}) \leq C^{(s)}$ implies that $B^{(s)} + \mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s,k)}) \leq B^{(s)} + C^{(s)} = C^{(s+1)}$, and—noting that $[U^\top v]_{\leq c} = U^\top_{\leq c} v$—we consequently have

$$\mathrm{prog}_0(g^{(s,k)}) \leq C^{(s+1)} \ \text{ and } \ g^{(s,k)} = g(U^\top_{\leq C^{(s+1)}} x^{(s,k)}, z^{(s)}).$$

Therefore, the oracle response to query $x^{(s,k)}$ has the form

$$\begin{aligned}\mathsf{O}_{\tilde{F}_U}(x^{(s,k)}, z^{(s)}) = \tilde{g}_U(x^{(s,k)}, z^{(s)}) = U g^{(s,k)} \\ = U_{\leq \mathrm{prog}_0(g^{(s,k)})} g^{(s,k)} = U_{\leq C^{(s+1)}} g(U^\top_{\leq C^{(s+1)}} x^{(s,k)}, z^{(s)}),\end{aligned}$$

so that if $x^{(s,k)}$ is measurable w.r.t. $U_{\leq C^{(s)}}$, then $\mathsf{O}_{\tilde{F}_U}(x^{(s,k)}, z^{(s)})$ is measurable w.r.t. $U_{\leq C^{(s+1)}}$.

From here the lemma follows by straightforward induction. The base case $t = 1$ is trivial, since the algorithm's initial queries do not depend on U. For the induction step, fix s and suppose that $x^{(s',k')}$ is measurable w.r.t. $U_{\leq C^{(s')}}$ for all $s' < s \leq t$ and $k' \in [K]$. That $\mathfrak{E}^{(t-1)}$ holds implies that $\mathrm{prog}_{\frac{1}{4}}(U^\top x^{(s',k')}) \leq C^{(s')}$ for all $s' < t$, and hence by the discussion above we conclude that the oracle's responses to all queries at iterations $1, \ldots, s-1$ are measurable w.r.t. $U_{\leq C^{(s)}}$. Since $x^{(s,k)}$ is a (measurable) function of r and the oracle responses up to iteration s, we conclude that it is measurable w.r.t. $U_{\leq C^{(s)}}$ as well. □

From Lemma 12, we conclude that there exists a function $\mathsf{f}^{(t,k)} : (\mathbb{R}^d)^{C^{(t)}} \to \{x \in \mathbb{R}^d \,|\, \|x\| \leq R\}$ (implicitly also dependent on $\mathcal{F}$), such that $x^{(t,k)} = \mathsf{f}^{(t,k)}(u^{(1)}, \ldots, u^{(C^{(t)})})$. Consequently,

$$\begin{aligned}&\mathbb{P}\left(|\langle u^{(j)}, x^{(t,k)}\rangle| > \frac{1}{4}\,,\ \mathfrak{E}^{(t-1)} \,\middle|\, \mathcal{F}\right) \\ &= \mathbb{P}\left(|\langle u^{(j)}, \mathsf{f}^{(t,k)}(u^{(1)}, \ldots, u^{(C^{(t)})})\rangle| > \frac{1}{4}\,,\ \mathfrak{E}^{(t-1)} \,\middle|\, \mathcal{F}\right) \\ &\leq \mathbb{P}\left(|\langle u^{(j)}, \mathsf{f}^{(t,k)}(u^{(1)}, \ldots, u^{(C^{(t)})})\rangle| > \frac{1}{4} \,\middle|\, \mathcal{F}\right).\end{aligned}$$

Conditional on $\mathcal{F}$ and $u^{(1)}, \ldots, u^{(C^{(t)})}$, we have that $\mathsf{f}^{(t,k)}(u^{(1)}, \ldots, u^{(C^{(t)})})$ is a fixed vector with norm at most R, while for every $j > C^{(t)}$, the vector $u^{(j)}$ is uniformly distributed on the $(d - C^{(t)})$-dimensional unit sphere. Therefore, concentration of measure on the sphere (see, e.g., Lemma 2.2 of [9]) gives

$$\begin{aligned}&\mathbb{P}\left(|\langle u^{(j)}, \mathsf{f}^{(t,k)}(u^{(1)}, \ldots, u^{(C^{(t)})})\rangle| > \frac{1}{4} \,\middle|\, \mathcal{F}\right) \\ &\quad \le 2\exp\left\{-\frac{1}{2}\cdot\left(\frac{1}{4R}\right)^2 \cdot (d-T)\right\} \le \frac{\delta}{2T_p K T},\end{aligned}$$

where the last transition follows from our choice of $d - T \ge 32R^2 \log\frac{2T^2K}{p\delta} \ge 32R^2 \log\frac{4T_pKT}{\delta}$. Substituting back into Eq. (46), we obtain the bound (44) and conclude the proof of Lemma 5.

B.2 Proof of Lemma 6

Before proving Lemma 6 we first list the relevant continuity properties of the compression function

$$\rho(x) = \frac{x}{\sqrt{1 + \|x\|^2/R^2}}.$$

Lemma 13 *Let* $J(x) = \frac{\partial \rho}{\partial x}(x) = \frac{I - \rho(x)\rho(x)^\top/R^2}{\sqrt{1+\|x\|^2/R^2}}$. *For all* $x, y \in \mathbb{R}^d$ *we have*

$$\begin{aligned}&\|J(x)\|_{\mathrm{op}} = \frac{1}{\sqrt{1 + \|x\|^2/R^2}} \le 1, \ \|\rho(x) - \rho(y)\| \le \|x - y\|, \\ &\text{and } \|J(x) - J(y)\|_{\mathrm{op}} \le \frac{3}{R}\|x - y\|. \end{aligned} \tag{47}$$

Proof of Lemma 13 Note that $\|\rho(x)\| \le R$ and therefore $0 \preceq I - \rho(x)\rho(x)^\top/R^2 \preceq I$. Consequently, we have $\|J(x)\|_{\mathrm{op}} = (1 + \|x\|^2/R^2)^{-1/2} \le 1$. The guarantee $\|\rho(x) - \rho(y)\| \le \|x - y\|$ follows immediately by Taylor's theorem. For the last statement, define $h(t) = \frac{1}{\sqrt{1+t^2}}$, and note that $|h(t)|, \left|h'(t)\right| \le 1$. By triangle inequality and the aforementioned boundedness and Lipschitzness properties of h, we have

$$\begin{aligned}&\|J(x) - J(y)\|_{\mathrm{op}} \\ &\quad \le h(\|y\|/R) \cdot \left\|\rho(x)\rho(x)^\top/R^2 - \rho(y)\rho(y)^\top/R^2\right\|_{\mathrm{op}} \\ &\qquad + \left\|I - \rho(x)\rho(x)^\top/R^2\right\|_{\mathrm{op}} \cdot |h(\|x\|/R) - h(\|y\|/R)|\end{aligned}$$

$$\leq \left\|\rho(x)\rho(x)^\top/R^2 - \rho(y)\rho(y)^\top/R^2\right\|_{\mathrm{op}} + \left\|I - \rho(x)\rho(x)^\top/R^2\right\|_{\mathrm{op}} \cdot |\|x\|/R - \|y\|/R|.$$

For the first term, observe that for any x, y we have $\|x\|, \|y\| \leq 1$, we have $\|xx^\top - yy^\top\|_{\mathrm{op}} \leq 2\|x-y\|$; this follows because for any $\|v\| = 1$, we have $\|(xx^\top - yy^\top)v\| \leq \|x-y\||\langle v, x\rangle| + \|y\||\langle v, x-y\rangle| \leq 2\|x-y\|$. Since $\|\rho(x)/R\| \leq 1$, it follows that

$$\left\|\rho(x)\rho(x)^\top/R^2 - \rho(y)\rho(y)^\top/R^2\right\|_{\mathrm{op}} \leq \frac{2}{R}\|x-y\|.$$

For the second term, we again use that $\|\rho(x)\| \leq R$ to write

$$\left\|I - \rho(x)\rho(x)^\top/R^2\right\|_{\mathrm{op}} \cdot |\|x\|/R - \|y\|/R| \leq \frac{1}{R}|\|x\| - \|y\|| \leq \frac{1}{R}\|x-y\|.$$

□

Proof of Lemma 6 The argument here is essentially identical to [15, Lemma 5]. Define $y^{(i)} = (y^{(i,1)}, \ldots, y^{(i,K)})$, where $y^{(i,k)} = \rho(x^{(i,k)})$. Observe that for each i and k, the oracle response $(\widehat{F}_{T,U}(x^{(i,k)}), \widehat{g}_{T,U}(x^{(i,k)}, z^{(i)}))$ is a measurable function of $x^{(i,k)}$ and $(\tilde{F}_{T,U}(y^{(i,k)}), \tilde{g}_{T,U}(y^{(i,k)}, z^{(i)}))$. Consequently, we can regard the sequence $y^{(1)}, \ldots, y^{(T)}$ as realized by some algorithm in $\mathcal{A}_{\mathsf{rand}}(K)$ applied to an oracle with $\mathsf{O}_{\tilde{F}_{T,U}}(y, z) = (\tilde{F}_{T,U}(y), \tilde{g}_{T,U}(y, z))$. Lemma 5 then implies that as long as $d \geq \lceil(32 \cdot 230^2 + 1)T \log \frac{2KT^2}{p\delta}\rceil \geq \lceil T + 32R^2 \log \frac{2KT^2}{p\delta}\rceil$, we have that with probability at least $1 - \delta$,

$$\max_{k\in[K]} \mathrm{prog}_{\frac{1}{4}}(U^\top \rho(x^{(i,k)})) = \max_{k\in[K]} \mathrm{prog}_{\frac{1}{4}}(U^\top y^{(i,k)}) < T, \tag{48}$$

as long as $i \leq (T - \log(2/\delta))/2p$.

We now show that the gradient must be large for all of the iterates. Let i and k be fixed. We first consider the case where $\|x^{(i,k)}\| \leq R/2$. Observe that (48) implies that $\mathrm{prog}_1(U^\top y^{(i,k)}) < T$ and so by Lemma 2.6, if we set $j = \mathrm{prog}_1(U^\top y^{(i,k)}) + 1$, we have

$$|\langle u^{(j)}, y^{(i,k)}\rangle| < 1 \quad \text{and} \quad \left|\left\langle u^{(j)}, \nabla \tilde{F}_{T,U}(y^{(i,k)})\right\rangle\right| \geq 1. \tag{49}$$

Now, observe that we have

$$\left\langle u^{(j)}, \nabla \widehat{F}_{T,U}(x^{(i,k)})\right\rangle = \left\langle u^{(j)}, J(x)^\top \nabla \tilde{F}_{T,U}(y^{(i,k)})\right\rangle + \eta\left\langle u^{(j)}, x^{(i,k)}\right\rangle.$$

Using that $J(x) = \frac{I - \rho(x)\rho(x)^\top/R^2}{\sqrt{1+\|x\|^2/R^2}}$, this is equal to

$$\frac{\left\langle u^{(j)}, \nabla \tilde{F}_{T,U}(y^{(i,k)})\right\rangle}{\sqrt{1+\|x^{(i,k)}\|^2/R^2}} - \frac{\left\langle u^{(j)}, y^{(i,k)}\right\rangle\left\langle y^{(i,k)}, \tilde{F}_{T,U}(y^{(i,k)})\right\rangle/R^2}{\sqrt{1+\|x^{(i,k)}\|^2/R^2}} + \eta\left\langle u^{(j)}, y^{(i,k)}\right\rangle\sqrt{1+\|x^{(i,k)}\|^2/R^2}.$$

Since $\|y^{(i,k)}\| \le \|x^{(i,k)}\| \le R/2$, this implies

$$\left|\left\langle u^{(j)}, \nabla \widehat{F}_{T,U}(x^{(i,k)})\right\rangle\right| \ge \frac{2}{\sqrt{5}}\left|\left\langle u^{(j)}, \nabla \tilde{F}_{T,U}(y^{(i,k)})\right\rangle\right| - \left|\left\langle u^{(j)}, y^{(i,k)}\right\rangle\right|\left(\frac{\left\|\tilde{F}_{T,U}(y^{(i,k)})\right\|}{2R} + \eta\frac{\sqrt{5}}{2}\right).$$

By Lemma 2 we have $\|\tilde{F}_{T,U}(y^{(i,k)})\| \le 23\sqrt{T}$. At this point, the choice $\eta = 1/5$, $R = 230\sqrt{T}$, as well as (49) imply that $\left|\left\langle u^{(j)}, \nabla \widehat{F}_{T,U}(x^{(i,k)})\right\rangle\right| \ge \frac{2}{\sqrt{5}} - \left(\frac{1}{20} + \frac{1}{2\sqrt{5}}\right) \ge \frac{1}{2}$.

Next, we handle the case where $\|x^{(i,k)}\| > R/2$. Here, we have

$$\left\|\nabla \widehat{F}_{T,U}(x^{(i,k)})\right\| \ge \eta\left\|x^{(i,k)}\right\| - \left\|J(x^{(i,k)})\right\|_{\mathrm{op}}\left\|\nabla \tilde{F}_{T,U}(y^{(i,k)})\right\| \ge \frac{R}{10} \ge \sqrt{T}.$$

where the second inequality uses that $\|J(x^{(i,k)})\|_{\mathrm{op}} \le \frac{1}{\sqrt{1+\|x^{(i,k)}\|^2/R^2}} \le 2/\sqrt{5}$ which follows from Lemma 13 and $\|x^{(i,k)}\| > R/2$. □

B.3 Proof of Lemma 7

To establish Lemma 7 we first prove a generic result showing that composition with the compression function ρ and an orthogonal transformation U never significantly hurts the regularity requirements in our lower bounds. In the following, we use the notation $a \vee b := \max\{a, b\}$.

Lemma 14 *Let $F : \mathbb{R}^T \to \mathbb{R}$ be an arbitrary twice-differentiable function with $\|\nabla F(x)\| \le \ell_0$ and $\|\nabla F(x) - \nabla F(y)\| \le \ell_1 \cdot \|x - y\|$, and let $g(x, z)$ and a random variable $z \sim P_z$ satisfy for all $x, y \in \mathbb{R}^T$,*

$$\mathbb{E}[g(x,z)] = \nabla F(x), \quad \mathbb{E}\|g(x,z) - F(x)\|^2 \le \sigma^2, \quad \text{and} \quad \mathbb{E}\|g(x,z) - g(y,z)\|^2 \le \bar{L}^2\|x-y\|^2. \tag{50}$$

Let $R \ge \ell_0 \vee 1$, $d \ge T$, and $U \in \mathsf{Ortho}(d, T)$. Then the functions

$$\widehat{F}_U(x) = F(U^\top \rho(x)) \quad \text{and} \quad \widehat{g}_U(x,z) = J(x)^\top U g(U^\top \rho(x), z)$$

satisfy the following properties.

1. $\widehat{F}_U(0) - \inf_x \widehat{F}_U(x) \leq F(0) - \inf_x F(x)$.
2. *The first derivative of* $\widehat{F}_U$ *is* $(\ell_1 + 3)$*-Lipschitz continuous.*
3. $\mathbb{E}\left\|\widehat{g}_U(x,z) - \nabla \widehat{F}_U(x)\right\|^2 \leq \sigma^2$ *for all* $x \in \mathbb{R}^d$.
4. $\mathbb{E}\|\widehat{g}_U(x,z) - \widehat{g}_U(y,z)\|^2 \leq (\bar{L}^2 + 9\sigma^2 + 9)\|x - y\|^2$ *for all* $x, y \in \mathbb{R}^d$.

Proof of Lemma 14 Property 1 is immediate, since the range of ρ is a subset of $\mathbb{R}^T$. For property 2, we use the triangle inequality along with Lemma 13 and the assumed smoothness properties of F as follows:

$$\begin{aligned}
&\left\|\nabla \widehat{F}_U(x) - \nabla \widehat{F}_U(y)\right\| \\
&\quad \leq \left\|J(x)^\top U \nabla F(U^\top \rho(x)) - J(x)^\top U \nabla F(U^\top \rho(y))\right\| \\
&\qquad + \left\|J(x) U \nabla F(U^\top \rho(y)) - J(y) U \nabla F(U^\top \rho(y))\right\| \\
&\quad \leq \left\|\nabla F(U^\top \rho(x)) - \nabla F(U^\top \rho(y))\right\| + \left\|\nabla F(U^\top \rho(y))\right\| \cdot \|J(x) - J(y)\|_{\mathrm{op}} \\
&\quad \leq \ell_1 \cdot \|\rho(x) - \rho(y)\| + \ell_0 \cdot \|J(x) - J(y)\| \\
&\quad \leq \left(\ell_1 + \frac{3\ell_0}{R}\right)\|x - y\|.
\end{aligned}$$

For the variance bound (property 3), observe that we have

$$\begin{aligned}
\mathbb{E}\left\|\widehat{g}_U(x,z) - \nabla \widehat{F}_U(x)\right\|^2 &= \mathbb{E}\left\|J(x)^\top U g(U^\top \rho(x), z) - J(x)^\top U \nabla F(U^\top \rho(x))\right\|^2 \\
&\leq \mathbb{E}\left[\left\|J(x)^\top U\right\|_{\mathrm{op}}^2 \cdot \left\|g(U^\top \rho(x), z) - \nabla F(U^\top \rho(x))\right\|^2\right] \\
&\leq \mathbb{E}\left\|g(U^\top \rho(x), z) - \nabla F(U^\top \rho(x))\right\|^2 \leq \sigma^2.
\end{aligned}$$

Here the second inequality follows from (47) and the fact that $U \in \mathsf{Ortho}(d, T)$, and the third inequality follows because the variance bound in (50) holds uniformly for all points in the domain $\mathbb{R}^T$ (in particular, those in the range of $x \mapsto U^\top \rho(x)$).

Lastly, to prove property 4 we first invoke the triangle inequality and the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

$$\begin{aligned}
&\mathbb{E}\|\widehat{g}_U(x,z) - \widehat{g}_U(y,z)\|^2 \\
&\quad = \mathbb{E}\left\|J(x)^\top U g(U^\top \rho(x), z) - J(y)^\top U g(U^\top \rho(y), z)\right\|^2 \\
&\quad \leq 2\mathbb{E}\left\|J(x)^\top U g(U^\top \rho(x), z) - J(x)^\top U g(U^\top \rho(y), z)\right\|^2 \\
&\qquad + 2\mathbb{E}\left\|\left(J(x)^\top - J(y)^\top\right) U g(U^\top \rho(x), z)\right\|^2
\end{aligned}$$

For the first term, we use the Jacobian operator norm bound from (47) and the assumed mean-squared smoothness of g:

$$\begin{aligned}
&\mathbb{E}\left\| J(x)^\top U g(U^\top \rho(x), z) - J(x)^\top U g(U^\top \rho(y), z)\right\|^2 \\
&\quad \le \mathbb{E}\left\| g(U^\top \rho(x), z) - g(U^\top \rho(y), z)\right\|^2 \\
&\quad \le \bar{L}^2 \mathbb{E}\|\rho(x) - \rho(y)\|^2 \\
&\quad \le \bar{L}^2 \mathbb{E}\|x - y\|^2.
\end{aligned}$$

For the second term, we use the Jacobian Lipschitzness from (47):

$$\mathbb{E}\left\| \left(J(x)^\top - J(y)^\top\right) U g(U^\top \rho(x), z)\right\|^2 \le \frac{9}{R^2}\|x - y\|^2 \cdot \mathbb{E}\left\| g(U^\top \rho(x), z)\right\|^2$$

We now use the assumed Lipschitzness of F and variance bound for g:

$$\begin{aligned}
\mathbb{E}\|g(U^\top \rho(x), z)\|^2 = \mathbb{E}\|g(U^\top \rho(x), z) - \nabla F(U^\top \rho(x))\|^2 \\
+ \|\nabla F(U^\top \rho(x))\|^2 \le \sigma^2 + \ell_0^2.
\end{aligned}$$

Putting everything together, we have

$$\mathbb{E}\|\widehat{g}_U(x, z) - \widehat{g}_U(y, z)\|^2 \le \left(\bar{L}^2 + 9\sigma^2/R^2 + 9\ell_0^2/R^2\right) \cdot \|x - y\|^2.$$

□

Proof of Lemma 7 For property 1, observe that $\widehat{F}_{T,U}(0) = F_T(0)$, and

$$\min_x \widehat{F}_{T,U}(x) \ge \min_x F_T(U^\top \rho(x)) \ge \min_x F_T(U^\top x) \ge \min_x F_T(x).$$

For properties 2, 3, and 4 we observe from Lemma 14 that $\widehat{F}_{T,U}$ and $\widehat{g}_{T,U}$, ignoring the quadratic regularization term, satisfy the same smoothness, variance, and mean-squared smoothness bounds as in Lemma 2/Lemma 4/Lemma 8 up to constant factors. The additional regularization term in (28) leads to an additional $\eta = 1/5$ factor in the smoothness and mean-squared-smoothness. □

B.4 Proof of Theorem 3

We prove the lower bound for the bounded variance and mean-squared smooth settings in turn. The proofs follow the same outline as the proofs of Theorems 1 and 2, relying on Lemmas 6 and 7 rather than Lemmas 1 and 4, respectively. Throughout, let Δ_0, ℓ_1, ς and $\bar{\ell}_1$ be the numerical constants in Lemma 7. *Bounded variance setting setting* Given accuracy parameter ϵ, initial suboptimality Δ, smoothness parameter L and variance parameter σ^2, we define for each $U \in \mathsf{Ortho}(d, T)$ a scaled instance

$$F^\star_{T,U}(x) = \frac{L\lambda^2}{\ell_1}\widehat{F}_{T,U}\left(\frac{x}{\lambda}\right), \text{ where } \lambda = \frac{\ell_1}{L}\cdot 4\epsilon,$$
$$\text{and } T = \left\lfloor \frac{\Delta}{\Delta_0(L\lambda^2/\ell_1)} \right\rfloor = \left\lfloor \frac{L\Delta}{\ell_1\Delta_0(4\epsilon)^2} \right\rfloor. \tag{51}$$

We assume $T \geq 4$, or equivalently $\epsilon \leq \sqrt{\frac{L\Delta}{64\ell_1\Delta_0}}$. Let $g^\star_T(x,z)$ denote the corresponding scaled version of the stochastic gradient function $\widehat{g}_{T,U}$. Now, by Lemma 7, we have that $F^\star_{T,U} \in \mathcal{F}(\Delta, L)$ and moreover,

$$\mathbb{E}\|g^\star_{T,U}(x,z) - \nabla F^\star_{T,U}(x)\|^2 = \left(\frac{L\lambda}{\ell_1}\right)^2$$
$$\mathbb{E}\left\|\widehat{g}_{T,U}\left(\frac{x}{\lambda},z\right) - \nabla\widehat{F}_{T,U}\left(\frac{x}{\lambda}\right)\right\|^2 \leq \frac{(4\varsigma\epsilon)^2(1-p)}{p}.$$

Therefore, setting $\frac{1}{p} = \frac{\sigma^2}{(4\varsigma\epsilon)^2} + 1$ guarantees a variance bound of σ^2.

Next, Let O be an oracle for which $\mathsf{O}_{F^\star_{T,U}}(x,z) = (F^\star_{T,U}(x), g^\star_{T,U}(x,z))$ for all $U \in \mathsf{Ortho}(d,T)$. Observe that for any $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}(K)$, we may regard the sequence $\left\{x^{(i,k)}_{\mathsf{A}[\mathsf{O}_{F^\star_{T,U}}]}/\lambda\right\}$ as queries made by an algorithm $\mathsf{A}' \in \mathcal{A}_{\mathsf{rand}}(K)$ interacting with the unscaled oracle $\mathsf{O}_{\widehat{F}_{T,U}}(x,z) = (\widehat{F}_{T,U}(x), \widehat{g}_{T,U}(x,z))$. Instantiating Lemma 6 for $\delta = \frac{1}{2}$, we have that w.p. at least $\frac{1}{2}$, $\min_{k\in[K]}\left\|\nabla\widehat{F}_{T,U}\left(\frac{1}{\lambda}x^{(t,k)}_{\mathsf{A}[\mathsf{O}_{F^\star_{T,U}}]}\right)\right\| > \frac{1}{2}$ for all $t \leq \frac{T-2}{2p}$. Therefore,

$$\mathbb{E}\min_{k\in[K]}\left\|\nabla F^\star_{T,U}\left(x^{(t,k)}_{\mathsf{A}[\mathsf{O}_{F^\star_{T,U}}]}\right)\right\| = \frac{L\lambda}{\ell_1}\cdot\mathbb{E}\min_{k\in[K]}\left\|\nabla\widehat{F}_{T,U}\left(\tfrac{1}{\lambda}x^{(t,k)}_{\mathsf{A}[\mathsf{O}_{F^\star_{T,U}}]}\right)\right\| \geq \frac{L\lambda}{4\ell_1} = \epsilon, \tag{52}$$

by which it follows that

$$\mathfrak{m}^{\mathsf{rand}}_\epsilon(K,\Delta,L,\sigma^2) > \frac{T-2}{2p} = \left(\left\lfloor\frac{L\Delta}{16\ell_1\Delta_0\epsilon^2}\right\rfloor - 2\right)\frac{1}{2p} \geq \frac{1}{2^{11}\ell_1\Delta_0\varsigma}\cdot\frac{L\Delta\sigma^2}{\epsilon^4} + \frac{1}{2^7\ell_1\Delta_0}\cdot\frac{L\Delta}{\epsilon^2},$$

where the second inequality uses that $\lfloor x\rfloor - 2 \geq x/4$ whenever $x \geq 4$.

Mean-squared smooth setting We use the scaling (51), choose $p = \min\left\{(4\varsigma\epsilon)^2/\sigma^2, 1\right\}$ as above, and let

$$L = \frac{\ell_1}{\bar{\ell}_1}\bar{L}\sqrt{p} = \frac{\ell_1}{\bar{\ell}_1}\min\{\frac{4\varsigma\epsilon}{\sigma}, 1\}\bar{L} \leq \bar{L}.$$

Using Lemma 7 and the calculation from the proof of Theorem 2, this setting guarantees that $\mathsf{O}_{F^\star_{T,U}}(x,z)$ is in the class $\mathcal{O}(K,\sigma^2,\bar{L})$. Consequently, the inequality (52) implies the lower bound

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{rand}}(K, \Delta, \bar{L}, \sigma^2) > \frac{T-2}{2p} = \left(\left\lfloor \frac{\bar{L}\Delta\sqrt{p}}{16\bar{\ell}_1\Delta_0\epsilon^2} \right\rfloor - 1\right)\frac{1}{2p}.$$

When $\frac{\bar{L}\Delta\sqrt{p}}{16\bar{\ell}_1\Delta_0\epsilon^2} \geq 4$, we have $T \geq 4$ and (53) along with $\lfloor x \rfloor - 2 \geq x/4$ for $x \geq 4$ gives

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{rand}}(K, \Delta, \bar{L}, \sigma^2) \geq \frac{\bar{L}\Delta}{2^7\bar{\ell}_1\Delta_0\epsilon^2\sqrt{p}} \geq \frac{1}{2^{10}\bar{\ell}_1\Delta_0\varsigma} \cdot \frac{\bar{L}\Delta\sigma}{\epsilon^3} + \frac{1}{2^8\bar{\ell}_1\Delta_0} \cdot \frac{\bar{L}\Delta}{\epsilon^2}. \quad (53)$$

Moreover, we choose c' so that $\epsilon \leq \sqrt{\frac{\bar{L}\Delta}{64\bar{\ell}_1\Delta_0}} \leq \sqrt{\frac{\bar{L}\Delta}{8}}$ holds. Lemma 11 then gives the lower bound

$$\bar{\mathfrak{m}}_\epsilon^{\mathsf{rand}}(K, \Delta, \bar{L}, \sigma^2) > c_0 \cdot \frac{\sigma^2}{\epsilon^2}, \quad (54)$$

for a universal constant c_0. Together, the bounds (53) and (54) imply the desired result when $\frac{\bar{L}\Delta\sqrt{p}}{16\bar{\ell}_1\Delta_0\epsilon^2} \geq 4$. As we argue in the proof of Theorem 2, in the complementary case $\frac{\bar{L}\Delta\sqrt{p}}{16\bar{\ell}_1\Delta_0\epsilon^2} < 4$, the bound (54) dominates (53), and consequently the result holds there as well.

C Proofs from Section 5

C.1 Statistical learning oracles

To prove the mean-squared smoothness properties of the construction (32) we must first argue about the continuity of $\nabla\Theta_i$, where $\Theta_i : \mathbb{R}^T \to \mathbb{R}$ is the "soft indicator" function given by

$$\Theta_i(x) := \Gamma\left(1 - \left(\sum_{k=i}^{T} \Gamma^2(|x_k|)\right)^{1/2}\right) = \Gamma\big(1 - \big\|\Gamma\big(|x_{\geq i}|\big)\big\|\big).$$

Lemma 15 *For all $i \geq j$, $\nabla_i\Theta_j(x)$ is well-defined with*

$$\nabla_i\Theta_j(x) = \begin{cases} -\Gamma'(1 - \|\Gamma(|x_{\geq j}|)\|) \cdot \frac{\Gamma(|x_i|)}{\|\Gamma(|x_{\geq j}|)\|} \cdot \Gamma'(|x_i|) \cdot \operatorname{sgn}(x_i), \\ i \geq j \text{ and } \|\Gamma(|x_{\geq j}|)\| > 0, \, 0, \text{ otherwise.} \end{cases} \quad (55)$$

Moreover, Θ_j satisfies the following properties:

1. $\|\nabla\Theta_j(x)\| \leq 6^2$.
2. $\|\nabla\Theta_j(x) - \nabla\Theta_j(y)\| \leq 10^4 \cdot \|x - y\|$.

Proof of Lemma 15 First, we verify that the function $x_i \mapsto \|\Gamma(|x_{\geq j}|)\|$ is differentiable everywhere for each i. From here it follows from Observation 1 that $\Theta_j(x)$ is differentiable, and (55) follows from the chain rule. Let $i \geq j$, and let $a = \sqrt{\sum_{k\geq j, k\neq i} \Gamma^2(|x_k|)}$. Then $\|\Gamma(|x_{\geq j}|)\| = \sqrt{a^2 + \Gamma^2(|x_i|)}$. This function is clearly differentiable with respect to x_i when $a > 0$, and when $a = 0$ it is equal to $\Gamma(|x_i|)$, which is also differentiable.

Property 1 follows because for all j,

$$\|\nabla\Theta_j(x)\| \leq \frac{6}{\|\Gamma(|x_{\geq j}|)\|} \cdot \sqrt{\sum_{i\geq j} \left(\Gamma(|x_i|)\Gamma'(|x_i|)\right)^2} \leq 6^2, \tag{56}$$

where we have used Observation 1.3.

To prove Property 2, we restrict to the case $j = 1$ so that $x_{\geq j} = x$ and subsequently drop the '$\geq j$' subscript to simplify notation; the case $j > 1$ follows as an immediate consequence. Define $\mu(x) \in \mathbb{R}^T$ via $\mu_i(x) = \Gamma(|x_i|)\Gamma'(|x_i|)\text{sgn}(x_i)$. Assume without loss of generality that $0 < \|\Gamma(|x|)\| \leq \|\Gamma(|y|)\|$. By triangle inequality, we have

$$\begin{aligned}\|\nabla\Theta_1(x) - \nabla\Theta_1(y)\| \leq{}& \left|\Gamma'(1 - \|\Gamma(|x|)\|) - \Gamma'(1 - \|\Gamma(|y|)\|)\right| \cdot \frac{\|\mu(x)\|}{\|\Gamma(|x|)\|} \\ &+ \Gamma'(1 - \|\Gamma(|x|)\|) \cdot \left\| \frac{\mu(x)}{\|\Gamma(|x|)\|} - \frac{\mu(y)}{\|\Gamma(|y|)\|} \right\|.\end{aligned}$$

To proceed, we state some useful facts, all of which follow from Observation 1.3:

1. Γ is 6-Lipschitz.
2. Γ' is 128-Lipschitz, and in particular $\Gamma'(1 - \|\Gamma(|x|)\|) \leq 128 \cdot \|\Gamma(|x|)\|$ (since $\Gamma'(1) = 0$).
3. $\|\mu(x)\| \leq 6 \cdot \|\Gamma(|x|)\|$ for all x.
4. $\|\mu(x) - \mu(y)\| \leq (128 \cdot 1 + 6^2) \cdot \|x - y\| = 164 \cdot \|x - y\|$ for all x, y.

Using the first, second, and third facts, we bound the first term as

$$\begin{aligned}&\frac{\|\mu(x)\|}{\|\Gamma(|x|)\|} \cdot \left|\Gamma'(1 - \|\Gamma(|x|)\|) - \Gamma'(1 - \|\Gamma(|y|)\|)\right| \\ &\quad\leq 6 \left|\Gamma'(1 - \|\Gamma(|x|)\|) - \Gamma'(1 - \|\Gamma(|y|)\|)\right| \\ &\quad\leq 128 \cdot 6 \left|\|\Gamma(|x|)\| - \|\Gamma(|y|)\|\right| \\ &\quad\leq 128 \cdot 6^2 \left|\|x\| - \|y\|\right| \\ &\quad\leq 5000 \|x - y\|.\end{aligned}$$

For the second term, we apply the second fact and the triangle inequality to upper bound by

$$\begin{aligned}
&\Gamma'(1-\|\Gamma(|x|)\|)\cdot\left\|\frac{\mu(x)}{\|\Gamma(|x|)\|}-\frac{\mu(y)}{\|\Gamma(|y|)\|}\right\| \\
&\le 128\|\Gamma(|x|)\|\cdot\left\|\frac{\mu(x)}{\|\Gamma(|x|)\|}-\frac{\mu(y)}{\|\Gamma(|y|)\|}\right\| \\
&\le 128\frac{\|\Gamma(|x|)\|}{\|\Gamma(|y|)\|}\cdot\|\mu(x)-\mu(y)\|+128\|\Gamma(|x|)\|\|\mu(x)\|\cdot\left|\frac{1}{\|\Gamma(|x|)\|}-\frac{1}{\|\Gamma(|y|)\|}\right|.
\end{aligned}$$

Using the fourth fact and the assumption that $\|\Gamma(|x|)\| \le \|\Gamma(|y|)\|$, we have

$$\frac{\|\Gamma(|x|)\|}{\|\Gamma(|y|)\|}\cdot\|\mu(x)-\mu(y)\| \le 164\|x-y\|.$$

Using the third fact and $\|\Gamma(|x|)\| \le \|\Gamma(|y|)\|$, we have

$$\begin{aligned}
&\|\Gamma(|x|)\|\|\mu(x)\|\cdot\left|\frac{1}{\|\Gamma(|x|)\|}-\frac{1}{\|\Gamma(|y|)\|}\right| \\
&\le 6\|\Gamma(|x|)\|^2\cdot\left|\frac{1}{\|\Gamma(|x|)\|}-\frac{1}{\|\Gamma(|y|)\|}\right| \\
&= 6\frac{\|\Gamma(|x|)\|}{\|\Gamma(|y|)\|}\cdot|\|\Gamma(|x|)\|\|\Gamma(|y|)\||\le 6^2\|x-y\|.
\end{aligned}$$

Gathering all of the constants, this establishes that

$$\|\nabla\Theta_1(x)-\nabla\Theta_1(y)\| \le 10^4\cdot\|x-y\|.$$

□

We are now ready to prove Lemma 8. For ease of reference, we restate the construction (32):

$$f_T(x,z) = -\Psi(1)\Phi(x_1)v_1(x,z) + \sum_{i=2}^{T}\left[\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right]v_i(x,z),$$

where

$$v_i(x,z) := 1+\Theta_i(x)\left(\frac{z}{p}-1\right).$$

Proof of Lemma 8 To begin, we introduce some shorthand. Define

$$\begin{aligned}
H(s,t) &= \Psi(-s)\Phi(-t) - \Psi(s)\Phi(t), \\
h_1(s,t) &= \Psi(-s)\Phi'(-t) + \Psi(s)\Phi'(t), \\
h_2(s,t) &= \Psi'(-s)\Phi(-t) + \Psi'(s)\Phi(t).
\end{aligned}$$

The gradient of the noiseless hard function F_T can then be written as

$$\nabla_i F_T(x) = -h_1(x_{i-1}, x_i) - h_2(x_i, x_{i+1}).$$

Next, define

$$g_i(x, z) = -h_1(x_{i-1}, x_i) \cdot v_i(x, z) - h_2(x_i, x_{i+1}) \cdot v_{i+1}(x, z). \tag{57}$$

With these definitions, we have the expression

$$\nabla_i f_T(x, z) = g_i(x, z) + \left(\frac{z}{p} - 1\right) \sum_{j=1}^{i} H(x_{j-1}, x_j) \cdot \nabla_i \Theta_j(x). \tag{58}$$

We begin by noting that ∇f_T is unbiased for F_T: Since $\mathbb{E}[v_i(x, z)] = 1$ for all i and $\mathbb{E}(\frac{z}{p} - 1) = 1$, it follows immediately from (58) that $\mathbb{E}[\nabla f_T(x, z)] = \nabla F(x)$.

Next, we show that ∇f_T is a probability-p zero chain. with an argument analogous to the 4. First, we claim that $[\nabla f_T(x, z)]_i = 0$, for all x, z and $i > 1 + \text{prog}_{\frac{1}{4}}(x)$, yielding $\text{prog}_0(\nabla f_T(x, z)) \leq 1 + \text{prog}_{\frac{1}{4}}(x)$. Since $|x_{i-1}|, |x_i| < 1/4$, it follows from (57) that $g_i(x, z) = 0$ and from (55) that $\nabla_i \Theta_j(x) = 0$ for all j, establishing the first claim. Now, consider the case $i = \text{prog}_{\frac{1}{4}}(x) + 1$ and $z = 0$. Here (since $|x_i| < 1/4$) we still have $\nabla_i \Theta_j(x) = 0$ for all j, so $\nabla_i f_T(x, z) = g_i(x, z)$. Since $\Gamma(\left|x_{\geq i}\right|) = \Gamma(\left|x_{\geq i+1}\right|) = 0$, we have $v_i(x, 0) = v_{i+1}(x, 0) = 0$, so $g_i(x, z) = 0$. It follows immediately that $\text{prog}_0(\nabla f_T(x, 0)) \leq \text{prog}_{\frac{1}{4}}(x)$ for all x. Finally, examining the definition (32) of f_T, it is straightforward to verify that $f_T(y, z) = f_T(y_{\leq 1+\text{prog}_{\frac{1}{4}}(x)}, z)$ for all y in a neighborhood of x, and all x and z. This implies $f_T(x, z) = f_T(x_{\leq 1+\text{prog}_{\frac{1}{4}}(x)}, z)$ and, via differentiation $\nabla f_T(x, z) = \nabla f_T(x_{\leq 1+\text{prog}_{\frac{1}{4}}(x)}, z)$. Similarly, one has $f_T(y, 0) = f_T(y_{\leq \text{prog}_{\frac{1}{4}}(x)}, 0)$ for y in a neighborhood of x, concluding the proof of the probabilistic zero-chain property.

To bound the variance and mean-squared smoothness of ∇f_T, we begin by analyzing the sparsity pattern of the error vector

$$\delta(x, z) := \nabla f_T(x, z) - \nabla F_T(x, z).$$

Let $i_x = \text{prog}_{\frac{1}{2}}(x) + 1$. Observe that if $j < i_x$, we have $\|\Gamma(\left|x_{\geq j}\right|)\| \geq \Gamma(\left|x_{i_x-1}\right|) \geq \Gamma(1/2) = 1$, and so $\Gamma'(1 - \|\Gamma(\left|x_{\geq j}\right|)\|) = 0$ and consequently $\nabla_i \Theta_j(x) = 0$ for all i. Note also that if $j > i_x$, we have $H(x_{j-1}, x_j) = 0$. We conclude that (58) simplifies to

$$\nabla_i f_T(x, z) = g_i(x, z) + \left(\frac{z}{p} - 1\right) \cdot H(x_{i_x-1}, x_{i_x}) \cdot \nabla_i \Theta_{i_x}(x). \tag{59}$$

As in Lemma 4, we have $v_i(x, z) = 1$ for all $i < i_x$ and $g_i(x, z) = \nabla_i F_T(x) = 0$ for all $i > i_x$. Thus, using the expression (57) along with (59), we have

$$\delta_i(x,z) = \left(\frac{z}{p} - 1\right) H(x_{i_x-1}, x_{i_x}) \cdot \nabla_i \Theta_{i_x}(x) - \left(\frac{z}{p} - 1\right) \begin{cases} h_2(x_{i_x-1}, x_{i_x}) \cdot \Theta_{i_x}(x), & i = i_x - 1, \\ h_1(x_{i_x-1}, x_{i_x}) \cdot \Theta_{i_x}(x), & i = i_x, \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

It follows immediately that the variance can be bounded as

$$\mathbb{E}\|\nabla f_T(x,z) - \nabla F_T(z)\|^2 \le \frac{2}{p} H(x_{i_x-1}, x_{i_x})^2 \|\nabla \Theta_{i_x}(x)\|^2 + \frac{2}{p} h_1^2(x_{i_x-1}, x_{i_x}) \cdot \Theta_{i_x}(x)^2 + \frac{2}{p} h_2^2(x_{i_x-1}, x_{i_x}) \cdot \Theta_{i_x}(x)^2.$$

From (56) we have $\|\nabla \Theta_{i_x}(x)\| \le 6^2$, and from (36) we have $|H(x,y)| \le 12$, so the first term contributes at most $\frac{2 \cdot 144 \cdot 6^4}{p}$. Since $|\Theta_i(x)| \le 1$, Lemma 2 implies that the second and third term together contribute at most $\frac{4 \cdot 23^2}{p}$. To conclude, we may take

$$\mathbb{E}\|\nabla f_T(x,z) - \nabla F_T(z)\|^2 \le \frac{\varsigma^2}{p},$$

where $\varsigma \le 10^3$.

To bound the mean-squared smoothness $\mathbb{E}\|\nabla f_T(x,z) - \nabla f_T(y,z)\|^2$, we first use that $\mathbb{E}[\delta(x,z)] = 0$, which implies

$$\mathbb{E}\|\nabla f_T(x,z) - \nabla f_T(y,z)\|^2 = \mathbb{E}\|\delta(x,z) - \delta(y,z)\|^2 + \|F_T(x) - F_T(y)\|^2.$$

We have $\|\nabla F_T(x) - \nabla F_T(y)\| \le \ell_1 \|x - y\|$ by Lemma 2.2. For the other term, we use the sparsity pattern of $\delta(x,z)$ established in (60) along with the fact that $\mathbb{E}\left(\frac{z}{p} - 1\right)^2 \le \frac{1}{p}$ to show

$$\begin{aligned} &\mathbb{E}\|\delta(x,z) - \delta(y,z)\|^2 \\ &\le \frac{3}{p} \underbrace{\sum_{i \in \{i_x, i_y\}} (h_1(x_{i-1}, x_i) \cdot \Theta_i(x) - h_1(y_{i-1}, y_i) \cdot \Theta_i(y))^2}_{=:\mathcal{E}_1} \\ &\quad + \frac{3}{p} \underbrace{\sum_{i \in \{i_x, i_y\}} (h_2(x_{i-1}, x_i) \cdot \Theta_i(x) - h_2(y_{i-1}, y_i) \cdot \Theta_i(y))^2}_{=:\mathcal{E}_2} \\ &\quad + \frac{3}{p} \underbrace{\sum_{i=1}^{T} \left(H(x_{i_x-1}, x_{i_x}) \cdot \nabla_i \Theta_{i_x}(x) - H(y_{i_y-1}, y_{i_y}) \cdot \nabla_i \Theta_{i_y}(y)\right)^2}_{=:\mathcal{E}_3}, \end{aligned}$$

where $i_y = \text{prog}_{\frac{1}{2}}(y) + 1$.

We bound $\mathcal{E}_1$ and $\mathcal{E}_2$ using similar arguments to Lemma 4. Focusing on $\mathcal{E}_1$, and letting $i \in \{i_x, i_y\}$ be fixed, we have

$$\begin{aligned}&(h_1(x_{i-1}, x_i) \cdot \Theta_i(x) - h_1(y_{i-1}, y_i) \cdot \Theta_i(y))^2 \\ &\leq 2(h_1(x_{i-1}, x_i) - h_1(y_{i-1}, y_i))^2 \Theta_i(x)^2 + 2(\Theta_i(x) - \Theta_i(y))^2 h_1(y_{i-1}, y_i)^2.\end{aligned}$$

Note that by Lemma 15, (i) Θ_i is 6^2 Lipschitz and $\Theta_i \leq 1$ and (ii) h_1 is 23-Lipschitz and $|h_1| \leq 5$ (from Observation 2 and Lemma 2). Consequently,

$$\mathcal{E}_1 \leq 2 \cdot 10^5 \cdot \|x - y\|^2.$$

Since h_2 is 23-Lipschitz and has $|h_2| \leq 20$, an identical argument also yields that

$$\mathcal{E}_2 \leq 5 \cdot 10^6 \cdot \|x - y\|^2.$$

To bound $\mathcal{E}_3$, we use the earlier observation that for all i and $j \neq i_x$ we have $H(x_{j-1}, x_j)\nabla_i \Theta_j(x) = 0$, and likewise that $H(y_{j-1}, y_j)\nabla_i \Theta_j(y) = 0$ for all $j \neq i_y$. This allows us to write

$$\begin{aligned}\mathcal{E}_3 &= \sum_{i=1}^{T} \left(\sum_{j \in \{i_x, i_y\}} H(x_{j-1}, x_j) \cdot \nabla_i \Theta_j(x) - H(y_{j-1}, y_j) \cdot \nabla_i \Theta_j(y) \right)^2 \\ &\leq 2 \sum_{j \in \{i_x, i_y\}} \sum_{i=1}^{T} \big(H(x_{j-1}, x_j) \cdot \nabla_i \Theta_j(x) - H(y_{j-1}, y_j) \cdot \nabla_i \Theta_j(y) \big)^2.\end{aligned}$$

Letting $j \in \{i_x, i_y\}$ be fixed, we upper bound the inner summation as

$$\begin{aligned}&\sum_{i=1}^{T} \big(H(x_{j-1}, x_j) \cdot \nabla_i \Theta_j(x) - H(y_{j-1}, y_j) \cdot \nabla_i \Theta_j(y) \big)^2 \\ &\quad \leq 2 \sum_{i=1}^{T} \big(H(x_{j-1}, x_j) \cdot (\nabla_i \Theta_j(x) - \nabla_i \Theta_j(y)) \big)^2 \\ &\qquad + \big((H(x_{j-1}, x_j) - H(y_{j-1}, y_j)) \cdot \nabla_i \Theta_j(y) \big)^2 \\ &\quad = 2H(x_{j-1}, x_j)^2 \|\nabla \Theta_j(x) - \nabla \Theta_j(y)\|^2 \\ &\qquad + 2(H(x_{j-1}, x_j) - H(y_{j-1}, y_j))^2 \|\nabla \Theta_j(y)\|^2.\end{aligned}$$

We may now upper bound this quantity by applying the following basic results:

1. $H(x_{j-1}, x_j) \leq 12$ by (36).
2. $\big|H(x_{j-1}, x_j) - H(y_{j-1}, y_j)\big| \leq 20\|x - y\|$, by (36).
3. $\|\nabla \Theta_j(y)\| \leq 6^2$ by Lemma 15.1.

4. $\|\nabla\Theta_j(x) - \nabla\Theta_j(y)\| \le 10^4 \cdot \|x - y\|$, by Lemma 15.2.

It follows that $\mathcal{E}_3 \le 3 \cdot 10^{10} \cdot \|x - y\|^2$. Collecting the bounds on $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, this establishes that

$$\mathbb{E}\|\nabla f_T(x, z) - \nabla f_T(y, z)\|^2 \le \frac{\bar{\ell}_1^2}{p} \cdot \|x - y\|^2.$$

with $\bar{\ell}_1 \le \sqrt{10^{11} + \ell_1^2}$. □

C.2 Active oracles

Proof of Lemma 9 Denoting

$$\mathcal{G}^{(t)} = \sigma(r, g^{(1)}, \ldots, g^{(t)}) \text{ and } \gamma^{(t)} := \max_{i \le t} \text{prog}_0(g^{(i)}),$$

we see that the equality $\mathbb{P}(\gamma^{(t)} - \gamma^{(t-1)} \notin \{0, 1\} | \mathcal{G}^{(t-1)}) = 0$ holds for our setting as well. Moreover, we claim that

$$\mathbb{P}(\gamma^{(t)} - \gamma^{(t-1)} = 1 | \mathcal{G}^{(t-1)}) \le 2p. \tag{61}$$

Given the bound (61), the remainder of the proof is identical to that of Lemma 1, with $2p$ replacing p. To see why (61) holds, let $(x^{(1)}, i^{(1)}), \ldots, (x^{(t)}, i^{(t)}) \in \mathcal{G}^{(t-1)}$ denote the sequence of queries made by the algorithm. We first observe that, by the construction of g_π, we have $\gamma^{(t)} = 1 + \gamma^{(t-1)}$ only if $\zeta_{1+\gamma^{(t-1)}}(\pi(i^{(t)})) = 1$. Therefore,

$$\mathbb{P}(\gamma^{(t)} - \gamma^{(t-1)} = 1 | \mathcal{G}^{(t-1)}) \le \mathbb{P}\big(\zeta_{1+\gamma^{(t-1)}}(\pi(i^{(t)})) = 1 | \mathcal{G}^{(t-1)}\big). \tag{62}$$

Next, let $b \in \{0, 1\}^{N^T}$ denote a (random) vector whose ith entry is $b_i := \zeta_{1+\gamma^{(t-1)}}(\pi(i))$. The vector b has N^{T-1} elements equal to 1 and its distribution is permutation invariant. Note that, by construction, the vector b is independent of $\{\zeta_j(\pi(i))\}_{j \ne 1+\gamma^{(t-1)}, i \in N^T}$. Consequently, the gradient estimates $g^{(1)}, \ldots, g^{(t-1)}$ depend on b only through their $(1 + \gamma^{(t-1)})$th coordinate, which for iterate $t' \le t - 1$ is

$$g^{(t')}_{1+\gamma^{(t-1)}} = \Big[\nabla_{1+\gamma^{(t-1)}} F_T(x^{(t')})\Big] b_{i^{(t')}}.$$

From this expression we see that $g^{(t')}$ depends on b only for index queries in the set

$$S^{(t-1)} := \{i^{(t')} | t' < t \text{ and } \nabla_{1+\gamma^{(t-1)}} F_T(x^{(t')}) \ne 0\} \in \mathcal{G}^{(t-1)}.$$

Moreover, for every $i \in S^{(t-1)}$ we have that $b_i = 0$, because otherwise there exists $t' < t$ such that $g^{(t')}_{1+\gamma^{(t-1)}} \ne 0$ which gives the contradiction $\gamma^{(t-1)} \ge \gamma^{(t')} \ge \text{prog}_0(g^{(t')}) \ge 1 + \gamma^{(t-1)} > \gamma^{(t-1)}$. In conclusion, we have for every $i \in N^T$

$$\mathbb{P}\big(\zeta_{1+\gamma^{(t-1)}}(\pi(i)) = 1|\mathcal{G}^{(t-1)}\big) = \mathbb{P}\big(b_i = 1|b_j = 0\ \forall j \in S^{(t-1)}\big) = \begin{cases} \frac{N^T-1}{N^T-|S^{(t-1)}|} & i \notin S^{(t-1)} \\ 0\text{otherwise}, \end{cases} \tag{63}$$

where the last equality follows from the permutation invariance of b.

Combining the observations above with the fact that $|S^{(t-1)}| \leq t-1 \leq \frac{T}{4p} \leq \frac{1}{4}NT \leq \frac{1}{2}N^T$ gives the desired result (61), since

$$\begin{aligned} \mathbb{P}(\gamma^{(t)} - \gamma^{(t-1)} &= 1|\mathcal{G}^{(t-1)}) \overset{(62)}{\leq} \mathbb{P}\big(\zeta_{1+\gamma^{(t-1)}}(\pi(i^{(t)}))\\ &= 1|\mathcal{G}^{(t-1)}\big) \overset{(63)}{\leq} \frac{N^T-1}{N^T-t} \leq \frac{2}{N} = 2p. \end{aligned}$$

We remark that the argument above depends crucially on using a different bit for every coordinate. Indeed, had we instead used the original construction g_T in Eq. (18) and set $g_\pi(x;i) = g_T(\zeta_1(\pi(i)))$, an algorithm that queried roughly N random indices would find an index $i^\star$ such that $\zeta_1(\pi(i^\star)) = 1$ and could then continue to query it exclusively, achieving a unit of progress at every query. This would decrease the lower bound from $\Omega(T/p) = \Omega(NT)$ to $\Omega(N+T)$. □

References

1. Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory **5**(58), 3235–3249 (2012)
2. Allen-Zhu, Z.: How to make the gradients small stochastically: even faster convex and nonconvex SGD. In: Advances in Neural Information Processing Systems, pp. 1165–1175 (2018a)
3. Allen-Zhu, Z.: Natasha 2: Faster non-convex optimization than SGD. In: Advances in Neural Information Processing Systems, pp. 2675–2686, (2018b)
4. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. In International conference on machine learning, pp. 699–707 (2016)
5. Allen-Zhu, Z., Li. Y.: Neon2: finding local minima via first-order oracles. In: Advances in Neural Information Processing Systems, pp. 3716–3726 (2018)
6. Arjevani, Y.: Limitations on variance-reduction and acceleration schemes for finite sums optimization. In: Advances in Neural Information Processing Systems, pp. 3540–3549 (2017)
7. Arjevani, Y., Shamir, O.: Dimension-free iteration complexity of finite sum optimization problems. In: Advances in Neural Information Processing Systems, pp. 3540–3548 (2016)
8. Arjevani, Y., Carmon, Y., Duchi, J.C., Foster, D.J., Sekhari, A., Sridharan, K.: Second-order information in non-convex stochastic optimization: power and limitations. In: Conference on Learning Theory, pp. 242–299. PMLR (2020)
9. Ball, K.: An elementary introduction to modern convex geometry. In Levy, S. (ed Flavors of Geometry, pp. 1–58. MSRI Publications (1997)
10. Bottou, L., Bousquet, O.: The tradeoffs of large scale learning. In Advances in neural information processing systems, pp. 161–168 (2008)
11. Bottou, L., Curtis, F., Nocedal, J.: Optimization methods for large-scale learning. SIAM Rev. **60**(2), 223–311 (2018)
12. Braun, G., Guzmán, C., Pokutta, S.: Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. IEEE Trans. Inf. Theory **63**(7), 4709–4724 (2017)
13. Bubeck, S., Jiang, Q., Lee, Y.T., Li, Y., Sidford, A.: Complexity of highly parallel non-smooth convex optimization. In: Advances in Neural Information Processing Systems 32 (2019)

14. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In Proceedings of the 34th International Conference on Machine Learning, pp. 654–663 (2017)
15. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Math. Progr. **184**(1), 71–120 (2019)
16. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points II: First-order methods. Math. Progr. **185**(1), 315–355 (2021)
17. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. Siam J. Opt. **20**(6), 2833–2852 (2010)
18. Cartis, C., Gould, N.I., Toint, P.L.: Complexity bounds for second-order optimality in unconstrained optimization. J. Complex. **28**(1), 93–108 (2012)
19. Cartis, C., Gould, N.I., Toint, P.L.: How much patience to you have?: a worst-case perspective on smooth noncovex optimization. Optima **88**, 1–10 (2012)
20. Cartis, C., Gould, N.I., Toint, P.L.: Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. arXiv preprint arXiv:1709.07180, (2017)
21. Cutkosky, A., Orabona, F.: Momentum-based variance reduction in non-convex SGD. Adv. Neural Inf. Process. Syst. (2019)
22. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems 27, (2014)
23. Diakonikolas, J., Guzmán, C.: Lower bounds for parallel and randomized convex optimization. In: Proceedings of the Thirty Second Annual Conference on Computational Learning Theory (2019)
24. Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. arXiv preprint arXiv:1910.01845 (2019)
25. Fang, C., Li, C.J., Lin, Z., Zhang, T.: Spider: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Advances in Neural Information Processing Systems, pp. 689–699 (2018)
26. Fang, C., Lin, Z., Zhang, T.: Sharp analysis for nonconvex SGD escaping from saddle points. In: Beygelzimer, A., Hsu, D., (eds) Proceedings of the Thirty-Second Conference on Learning Theory, vol. 99, pp. 1192–1234. PMLR (2019)
27. Foster, D.J., Sekhari, A., Shamir, O., Srebro, N., Sridharan, K., Woodworth, B.: The complexity of making the gradient small in stochastic convex optimization. In: Proceedings of the Thirty-Second Conference on Learning Theory, pp. 1319–1345 (2019)
28. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points: online stochastic gradient for tensor decomposition. In: Conference on Learning Theory, pp. 797–842 (2015)
29. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems, pp. 2973–2981 (2016)
30. Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM J. Opt. **23**(4), 2341–2368 (2013)
31. LeCam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**(1), 38–53 (1973)
32. Lei, L., Ju, C., Chen, J., Jordan, M.I.: Non-convex finite-sum optimization via SCSG methods. In: Advances in Neural Information Processing Systems, pp. 2348–2358 (2017)
33. Ma, C., Wang, K., Chi, Y., Chen, Y.: Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. Found. Comput. Math. (2019). https://doi.org/10.1007/s10208-019-09429-9
34. Murty, K.G., Kabadi, S.N.: Some np-complete problems in quadratic and nonlinear programming. Math. progr. **39**(2), 117–129 (1987)
35. Nemirovski, A.: On parallel complexity of nonsmooth convex optimization. J. Complex. **10**(4), 451–463 (1994)
36. Nemirovski, A., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley (1983)
37. Nesterov, Y.: Introductory Lectures of Convex Optimization. Kluwer Academic Publishers (2004)
38. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Math. Progr. **108**(1), 177–205 (2006)
39. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $o(1/k^2)$. Sov. Math. Dokl. **27**(2), 372–376 (1983)

40. Nocedal, J., Wright, S.: Numerical Optimization. Springer Science & Business Media (2006)
41. Raginsky, M., Rakhlin, A.: Information-based complexity, feedback and dynamics in convex programming. IEEE Trans. Inf. Theory **57**(10), 7036–7056 (2011)
42. Reddi, S.J., Hefny, A., Sra, S., Poczos, B., Smola, A.: Stochastic variance reduction for nonconvex optimization. In: International Conference on Machine Learning, pp. 314–323 (2016)
43. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems 24 (2011)
44. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
45. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. Found. Comput. Math. **18**(5), 1131–1198 (2018)
46. Traub, J.F., Wasilkowski, G.W., Woźniakowski H.: Information-Based Complexity. (1988)
47. Tripuraneni, N., Stern, M., Jin, C., Regier, J., Jordan, M.I.: Stochastic cubic regularization for fast nonconvex optimization. In: Advances in Neural Information Processing Systems, pp. 2899–2908 (2018)
48. Vavasis, S.A.: Black-box complexity of local minimization. SIAM J. Opt. **3**(1), 60–80 (1993)
49. Wang, Z., Ji, K., Zhou, Y., Liang, Y., Tarokh, V.: Spiderboost: a class of faster variance-reduced algorithms for nonconvex optimization. arXiv preprint arXiv:1810.10690, (2018)
50. Woodworth, B., Srebro, N.: Tight complexity bounds for optimizing composite objectives. In: Advances in Neural Information Processing Systems, pp. 3639–3647 (2016)
51. Woodworth, B., Srebro, N.: Lower bound for randomized first order convex optimization. arXiv preprint, arXiv:1709.03594 (2017)
52. Xu, Y., Rong, J., Yang, T.: First-order stochastic algorithms for escaping from saddle points in almost linear time. In: Advances in Neural Information Processing Systems, pp. 5530–5540 (2018)
53. Yao, A.C.-C.: Probabilistic computations: toward a unified measure of complexity. In 18th Annual Symposium on Foundations of Computer Science, pp. 222–227. IEEE (1977)
54. Yu, B.: Assouad, Fano, and Le Cam. In Festschrift for Lucien Le Cam, pp. 423–435. Springer (1997)
55. Zhou, D., Gu, Q.: Lower bounds for smooth nonconvex finite-sum optimization. In: International Conference on Machine Learning (2019)
56. Zhou, D., Xu, P., Gu, Q.: Stochastic nested variance reduction for nonconvex optimization. In: Advances in Neural Information Processing Systems, pp. 3925–3936. Curran Associates Inc. (2018)
57. Zhou, D., Xu, P., Gu, Q.: Stochastic nested variance reduction for nonconvex optimization. J. Mach. Learn. Res. (2020)