On the Power of Biased Derivative Information for Nonconvex Stochastic Optimization

author names withheld

Editor: Under Review for COLT 2026

Abstract

We consider the problem of finding δ-stationary points, i.e., x such that $||\nabla F(x)|| \leq \delta$, for smooth, non-convex objectives, where the derivative oracles are not only stochastic but also biased. In the first-order setting, we provide lower bounds for finding an $O((\epsilon + B^2)^{1/2})$-stationary point, where B is a bound on the gradient bias, which nearly matches the upper bound in Ajalloeian and Stich (2020). We then establish bias-dependent lower bounds for algorithms that use higher-order derivative information for finding ϵ-stationary points, and to complement these lower bounds, we develop p^{th}-order trust-region based methods that, under small enough bias and for $p \to \infty$, achieves the $O(\epsilon^{-3})$ query complexity in the lower bound, thereby providing insight to the benefits of higher-order information in the presence of bias. We further improve upon the bias restriction for the first-order derivative through a variance-reduction scheme, while still maintaining the property that taking $p \to \infty$ recovers the known worst-case $O(\epsilon^{-3})$ query complexity.

Keywords: Biased Stochastic Oracles, High-Order Derivative Information, Variance Reduction

1. Introduction

For a function $F : \mathbb{R}^d \to \mathbb{R}$ which is at least $(p+1)$-times differentiable, has Lipschitz continuous derivatives, and has bounded suboptimality Δ such that $F(0) - \inf_x F(x) \leq \Delta$, we focus on the task of finding an ϵ-stationary point: that is, $x \in \mathbb{R}^d$ such that

$$\|\nabla F(x)\| \leq \epsilon$$

for some precision parameter $\epsilon > 0$. Finding such stationary points is a task that has been explored in numerous previous works (e.g. Carmon et al. (2019a), Carmon et al. (2019b)) and serves as a natural proxy for finding approximate local optima.

When working with smooth, but potentially nonconvex functions, finding global optima has been shown to be intractable. In fact, it was shown that for functions F whose p derivatives are all smooth, the worst case oracle complexity of finding a point x such that $f(x) \leq f(x^*) + \epsilon$ scales at least as $(1/\epsilon)^{d/p}$, where d is the dimensionality of the problem (Nemirovsky and Yudin (1983)). Therefore, we naturally turn to finding local optima whose gradient norm is sufficiently small. Moreover, just like Nemirovsky and Yudin (1983), we refer to oracle complexity as the number of queries to derivative oracles, where the i^{th} order derivative oracle returns an i^{th} derivative estimate of F at a query point x.

There have been a variety of work that has studied the oracle complexity of finding ϵ-stationary points (i.e $\mathbb{E}||\nabla F(x)|| \leq \epsilon$). In Ghadimi and Lan (2013), the authors derive an $O(\epsilon^{-4})$ oracle complexity bound for using first order methods (SGD) to find an ϵ-stationary point. This first order

complexity bound was improved in Fang et al. (2019) to $O(\epsilon^{-3.5})$, with the additional assumption that the stochastic gradient $\nabla F(x, \xi)$ was L-smooth. One can also further improve this complexity bound to $O(\epsilon^{-3})$ if the noisy gradient satisfy a mean-squared smoothness property (Arjevani et al., 2019). Moreover, after incorporating access to a stochastic Hessian $\nabla^2 F(x, \xi)$, we also get a complexity bound $O(\epsilon^{-3.5})$, while also relaxing the smoothness assumption of the stochastic gradient (Tripuraneni et al., 2017). There has also been several works which employ variance reduction (e.g. Fang et al. (2018), Zhou et al. (2018)), some of which use methods like hessian-vector products to compute better representations of the gradient and have an even better $O(\epsilon^{-3})$ oracle complexity. These works (as well as more empirical-based works such as Kingma and Ba (2015), Luo et al. (2025), Chen et al. (2023)) are part of a broader set of works that assume stochastic and unbiased derivative oracles, where for all derivatives $i = 1, \ldots, p$, we have that

$$\mathbb{E}[\widehat{\nabla}^i F(x, \xi)] = \nabla^i F(x)$$

and

$$\mathbb{E}||\widehat{\nabla}^i F(x, \xi) - \nabla^i F(x)||_{\text{op}} \leq \sigma_i^2$$

for some set of variance parameters $\sigma_1, \ldots, \sigma_p$ and where the noise ξ drawn from some distribution P_ξ is a random variable. In Arjevani et al. (2020), the authors extended the analysis done in previous works by introducing a so-called "elbow effect", which essentially represents the fact that even when using higher order derivatives beyond the second order in the unbiased but stochastic oracle setting, the worst case oracle complexity still scales as $O(\epsilon^{-3})$.

However, in many cases, even the assumption that the derivative oracles are unbiased may be too strong. In machine learning, for instance, it is often intractable to compute an unbiased representation of a higher-order derivative, and therefore, we consider a setting where the derivative oracles are both stochastic and biased. For example, in Liu et al. (2024), the authors presents an algorithm called Sophia, which uses an unbiased estimation of the Hessian diagonal for large language model pretraining tasks. Notice that we can formulate the computation of an unbiased representation of the Hessian diagonal as the algorithm querying a second order oracle that is biased (as we only think about the main diagonal) and stochastic (since we compute an unbiased representation of this diagonal. In addition, when considering distributed optimization, where the data is split among multiple workers, delayed gradient or higher order derivative updates introduce bias into the weight updates (Lin et al. (2018), Beznosikov et al. (2023)). Moreover, there have been other works that have began to consider settings where the derivative oracles are both stochastic and biased (Adil et al. (2025), Demidovich et al. (2024), Ajalloeian and Stich (2020)).

In our work, we assume that our derivative oracles are stochastic and biased, where for all derivatives $i = 1, \ldots, p$, we have that

$$\widehat{\nabla}^i F(x, \xi, b) = \nabla^i F(x) + \xi_i(x, z) + b_i(x)$$

where

$$\mathbb{E}_{z \sim P_z}[\xi_i(x, z)] = 0, \mathbb{E}||\xi_i(x, z)||_{\text{op}}^2 \leq \sigma_i^2, ||b_i(x)||_{\text{op}} \leq B_i$$

for some set of variance parameters $\sigma_1, \ldots, \sigma_p$ and bias parameters $B_1, \ldots, B_p$. We not only build upon existing analyses for first and second methods in the stochastic and biased setting, but also extend this exploration to the higher order setting, demonstrating that there are advantages to appealing to higher order information in a stochastic and biased environment.

1.1. Our Main Contributions

We build on previous works in stochastic nonconvex optimization, by now considering a setting where our derivative oracles are biased as well stochastic. Below, we outline our results in our paper.

First Order Extension. In Ajalloeian and Stich (2020), the authors considered derived the following upper bound for the oracle complexity for finding iterates $\{x_t\}$ such that $\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}||\nabla F(x^{(t)})||^2 = O(\epsilon + B_1^2)$:

$$O(\frac{\Delta L_1}{\epsilon + B_1^2} + \frac{\Delta L_1 \sigma_1^2}{\epsilon^2 + B_1^4})$$

equivalent to finding an $O((\epsilon + B_1^2)^{\frac{1}{2}})$ stationary point, where L_1 is the Lipschitz constant of ∇F. We derive a lower bound that matches the stochastic term $(\frac{L_1\sigma_1^2}{\epsilon^2+B_1^4})$of the above upper bound up to constant factors to demonstrate that the upper bound is quite tight:

$$O(\frac{\Delta L_1 \sigma_1^2}{\Delta_0 \ell_0^2 \ell_1 (\epsilon^2 + B_1^4)})$$

where we define ℓ_0, Δ_0, ℓ_1 in Appendix C where we show the proof of this lower bound. This raises a question as to how derivative orders $p \geq 2$ would behave when these oracles were biased and stochastic.

Higher Order Lower Bound. To understand how algorithms that use derivative orders $p \geq 2$ would behave in the worst case scenario, we derive the following worst case oracle complexity for finding ϵ-stationary points:

$$\Omega(1) \cdot \frac{(\sigma_1^2 + B_1^2)\Delta}{\epsilon^3} \cdot \min_{q' \in \{1,\ldots,p\}, q \in \{2,\ldots,p\}} \min\{(\frac{\ell_0^2(\sigma_q^2 + B_q^2)}{2\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4\epsilon^2 B_1)})^{\frac{1}{2(q-1)}}, (\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}$$

We then wondered if this lower bound was tight, and developed algorithms to try and match these lower bounds as closely as possible. Building on this, we further show that this lower bound is nearly tight by developing algorithm . . .

Minibatch Higher Order Derivative (MHOD) Estimation. We develop an algorithm where each derivative estimate D^i can be computed as an average of n_i calls to the i^{th} derivative oracle:

$$D^i(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} \tilde{\nabla}^i F(x, \xi_j, b_i)$$

where $\tilde{\nabla}^i F$ is a biased and stochastic i^{th} derivative oracle. At each step, we solve the following subproblem:

$$x_{t+1} = \underset{y:||y-x_t||\leq\eta}{\arg\min} \sum_{i=1}^{p} \frac{1}{i!} D^{(i)}[y - x_t]^i + \frac{M}{(p+1)!}||y - x_t||^{p+1}$$

for some $M \geq 8L_p$. Assuming that $B_1 \leq O(\epsilon^2)$ and $B_i \leq O(\epsilon^{\frac{2(p-1)}{p}})$ for all $i \geq 2$, we have the following oracle complexity bound to find an ϵ-stationary point:

$$O(\frac{\sigma_1^2 \Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{(\max_{2\leq i\leq p} \sigma_i)^2}{\epsilon^{\frac{3p-1}{p}}})$$

where as $p \to \infty$, we recover the ϵ^{-3} dependence given by previously known lower bounds.

Variance Reduction Based Derivative Estimation. Given numerous previous works which show the advantages of using variance reduction in derivative estimation, we also utilize variance reduction based techniques with hopes of improving the bias restrictions and the oracle complexity bound for finding ϵ-stationary points. Adopting a similar scheme to Arjevani et al. (2020), we find that if $B_i \leq O(\epsilon^{\frac{2(p-1)}{p}})$ for all $i \geq 1$, we have the following oracle complexity bound to find an ϵ-stationary point:

$$O(\frac{\Delta(\max_{i\geq 1} \sigma_i)^2}{\epsilon^{\frac{3p+1}{p}}} + \frac{\Delta}{\epsilon^{\frac{2p+1}{p}}})$$

thereby reflecting a slightly weaker bias restriction for the first order term and the property that the ϵ^{-3} dependence is recovered as $p \to \infty$.

1.2. Additional Related Work

We briefly discuss additional related works that give some more broader context for our work. First, we discuss several known rates for finding ϵ-stationary points for nonconvex objectives, where the oracles are deterministic (i.e. noiseless and unbiased). First, an improvement $O(\epsilon^{-\frac{7}{4}})$ to the previously known $O(\epsilon^{-2})$ query complexity for first order methods was achieved in Carmon et al. (2017) through incorporating second order information and assuming that the Hessian is Lipschitz continuous, where the lower bound for deterministic algorithms that only rely on first and second order information as $\Omega(\epsilon^{-\frac{12}{7}})$ (Carmon et al., 2019b). Furthermore, the authors highlight how this query complexity can be improved by appealing to higher order information, as we do in our work as well. In particular, when using p^{th} order oracles (assuming that all p derivatives are Lipschitz continuous), we get an oracle complexity of $O(\epsilon^{(-1-\frac{1}{p})})$, thereby yielding an $O(\epsilon^{-1})$ complexity as $p \to \infty$. Moreover, in Agarwal et al. (2017), the authors present a method that uses cubic regularization to achieve an $\tilde{O}(\epsilon^{-\frac{7}{4}})$ oracle complexity for finding an ϵ-stationary point x that also satisfies $\nabla^2 f(x) \succeq -\epsilon^{\frac{1}{2}} I$.

1.3. Paper Organization

We formally go over our problem setup in Section 2, including the function class assumptions and oracle setup. In Section 3, we present the lower bounds for both the first and higher order settings.

In Section 4, we present the algorithms that use minibatch-higher order derivative estimation (1) and variance reduction based derivative estimation (3). We conclude the paper in Section 5, providing some potential avenues for future work, and prove theorem 3 in appendices A, theorem 4 in appendix B, and theorems 1 and 2 in appendix C.

Notation. For some $1 \leq i \leq p$, let $\nabla^i F$ refer to the i^{th} derivative of a function $F \in \mathcal{C}^p$, where $\mathcal{C}^p$ denotes the set of p times differentiable, continuous functions. For all i, $[\nabla^i F(x)]_{j_1,\ldots,j_i} = \frac{\partial^i F}{\partial x_{j_1} \ldots \partial x_{j_i}}$. For matrices A and tensors T, $||\cdot||_{\text{op}}$ denotes the operator norm, and unless otherwise specified $||\cdot||$ refers to the operator norm. For a symmetric tensor T, we let $||T||_{\text{op}} = \sup_{||v||=1} |\langle T, v, \ldots, v\rangle|$.

2. Setup

We study the task of finding ϵ-approximate stationary points. Here, we go over the framework for our analysis, which includes the function class and the oracle setup. In this paper, we consider p^{th} order optimization algorithms which access the function $F \in \mathcal{F}_p(\Delta, L_{1:p})$ and its' derivatives through a stochastic and biased p^{th} order oracle $(O_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$.

2.1. Function Class

We consider the same framework as Arjevani et al. (2020) and restate it for completeness. We consider p-times differentiable functions satisfying standard regularity conditions and let

$$\mathcal{F}_p(\Delta, L_{1:p}) = \{F : \mathbb{R}^d \to \mathbb{R} : \ ||\nabla^q F(x) - \nabla^q F(y)|| \leq L_q ||x - y|| \ \ \forall x, y \in \mathbb{R}^d, q \in [p]\}$$

and where for all $F \in \mathcal{F}_p(\Delta, L_{1:p})$, we have that

$$F(0) - \inf_x F(x) \leq \Delta$$

where $L_{1:p} = (L_1, \ldots, L_p)$ represents the Lipschitz constants of the first to p^{th} order derivatives $\nabla^p F$ with respect to the operator norm.

2.2. Oracles

For a function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, we consider a class of biased and stochastic p^{th} order oracles, defined by a distribution P_z over a measurable set $\mathcal{Z}$ and an estimator

$$O_F^p(x, z, b) := (\tilde{F}(x, z, b_0), \tilde{\nabla} F(x, z, b_1), \ldots, \tilde{\nabla}^p F(x, z, b_p))$$

where $\{\tilde{\nabla}^q F(x, z, b_q)\}_{q=0}^p$ are biased and stochastic estimators for their respective derivatives. For all x and $q \in [p]$, we have that $\tilde{\nabla}^q F(x, z, b) = \nabla F^q(x) + \xi_q(x, z) + b_q(x)$, where $\mathbb{E}_{z \sim P_z}[\xi_q(x, z)] = 0$, $\mathbb{E}||\xi_q(x, z)||^2 \leq \sigma_q^2$, and $||b_q(x)|| \leq B_q$.

Given variance parameters $\sigma_{1:p}$ and bias parameters $B_{1:p}$, we define the oracle class $\mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$ to be the set of all biased and stochastic p^{th} order oracles such that the conditions above hold.

3. Lower Bounds

We first consider the scenario of finding an $O(f(\epsilon) + g(B_1, \ldots, B_p))$ stationary point, where f and g are positive functions of the precision parameter ϵ and the bias terms respectively. In Ajalloeian and Stich (2020), the authors present the following upper bound on the number of oracle queries for finding iterates $\{x_t\}$ where $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla F(x^{(t)})||^2 = O(\epsilon + B_1^2)$ (equivalently finding $x \in \{x_t\}$ such that $||\nabla F(x)|| = O((\epsilon + B_1^2)^{\frac{1}{2}})$:

$$O(\frac{\Delta L_1}{\epsilon + B_1^2} + \frac{\Delta L_1 \sigma_1^2}{\epsilon^2 + B_1^4})$$

In theorem 1, we derive a nearly matching lower bound that matches the stochastic term of the upper bound. In fact, when the stochastic term dominates, the lower bound matches the provided upper bound exactly (up to constant factors).

Theorem 1 *When $p = 1$, there exists $F \in \mathcal{F}_1(\Delta, L_1)$ and $(O_F^1, P_z) \in \mathcal{O}_1(F, \sigma_1, B_1)$, such that for any first-order zero-respecting algorithm and $\epsilon < \frac{\sqrt{2}}{4}$, the minimum number of queries to obtain a $(\epsilon + B_1^2)^{\frac{1}{2}}$-stationary point with constant probability is bounded below by*

$$\Omega(\frac{\Delta L_1 \sigma_1^2}{\Delta_0 \ell_0^2 \ell_1 (\epsilon^2 + B_1^4)})$$

where Δ_0, ℓ_0, ℓ_1 are defined in Appendix C.

Given the nearly matching upper and lower bounds in the first order setting, one natural question was the following: Can we derive analogous lower and upper bounds in the higher order setting? In Theorem 2, we derive a lower bound for finding in ϵ-stationary point using biased and stochastic derivatives using derivatives $1, \ldots, p$.

Theorem 2 *For all $p \in \mathbb{N}$, Δ, $L_{1:p}$, $\sigma_{1:p} > 0$, $B_{1:p} > 0$, and $\epsilon < \frac{1}{4}$, there exists $F \in \mathcal{F}_p(\Delta, L_{1:p})$ and $(O_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$, such that for any pth-order zero-respecting algorithm, the number of queries to obtain ϵ-stationary point with constant probability is bounded below by*

$$\Omega(1) \cdot \frac{(\sigma_1^2 + B_1^2)\Delta}{\epsilon^3} \cdot \min_{q' \in \{1,\ldots,p\}, q \in \{2,\ldots,p\}} \min\{(\frac{\ell_0^2(\sigma_q^2 + B_q^2)}{2\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4\epsilon^2 B_1)})^{\frac{1}{2(q-1)}}, (\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}$$

where $\{\ell_q\}$ is defined in Appendix C.

Proof Here we provide a brief proof sketch, deferring the full proof to Appendix C. Note that this proof sketch applies to theorem 1 as well.

We use a specific hard function (3) and a specific expression to estimate the derivative of that function (5) and show that this collection of derivative estimators forms a probability-ρ chain (for some $0 \leq \rho \leq 1$) where the following properties hold:

$$\Pr(\exists x | \operatorname{prog}(\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)) = \operatorname{prog}_{\frac{1}{4}}(x) + 1) \leq \rho \tag{1}$$

$$\Pr(\exists x | \operatorname{prog}(\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)) = \operatorname{prog}_{\frac{1}{4}}(x) + i) = 0 \tag{2}$$

or all $i > 1$. In other words, every oracle query can "discover" at most one new coordinate of x, thereby providing a lower bound on the number of queries needed to make sufficient progress (see appendix C for a more rigorous definition of "progress"). We then show that the derivative estimators $\{\tilde{\nabla}^q F_T(x, z, b)\}$ form a probability-ρ zero-chain and set our $F_T^*(x) = \alpha F_T(\beta x)$ and solve for the constants α, β based on conditions in appendix C. The proof sketch of theorem 1 is similar to that of theorem 2. ■

As in Arjevani et al. (2020), we consider the same "hard" function F from Carmon et al. (2019a), where for a fixed $T \geq 0$ and $x \in \mathbb{R}^T$:

$$F_T(x) = -\Psi(1)\Phi(1) + \sum_{i=2}^{T} [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)] \tag{3}$$

where

$$\Psi(x) = \begin{cases} 0, & \text{if } x \leq \frac{1}{2} \\ \exp(1 - \frac{1}{(2x-1)^2}), & \text{if } x > \frac{1}{2}, \end{cases}, \Phi(x) = \sqrt{e} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \tag{4}$$

We introduce the following derivative estimator (for some $0 \leq \rho \leq 1$) accounting for the fact that the oracles are stochastic *and biased*:

$$[\tilde{\nabla}^q F_T(x, z, b)]_i = (1 + \mathbf{1}\{i > \text{prog}_{\frac{1}{4}}(x)\}(\frac{z}{\rho} - 1)) \cdot (\nabla_i^q F_T(x) + b_i^q(x)) \tag{5}$$

This function F_T is such that any zero-respecting algorithm that uses a p^{th} order stochastic and biased oracle must make at least $\Omega(T/\rho)$ oracle queries to make the gradient small, and thus, we can use it to lower bound the overall oracle complexity for biased and stochastic p^{th} order oracles. We use this same function to prove theorem 1 as well. The analysis for both of these theorems as well as more details about this lower-bound framework is presented in Appendix C.

4. Upper Bounds

Given these lower bounds, we develop various algorithms to see if there exist any upper bounds which match the lower bound. In Algorithm 1, we minimize a regularized p^{th} order model of the function F at each step of our algorithm.

Algorithm 1 **M**inibatch-**H**igher **O**rder **D**erivative Estimation **(MOHD)**

Input: Derivative order p, Biased and stochastic oracle $(O_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$ for $F \in \mathcal{F}_p(\Delta, L_{1:p})$, Precision parameter ϵ, Initial parameter $x^{(0)}$

1: Find constants $\{C_i\}_{i=1}^p$ such that (for all n and all $t \geq 1$): $\mathbb{E}||D_t^{(i)} - \nabla F^{(i)}(x^{(t)})||_{\text{op}}^{\frac{p+1}{p}} \leq 2^{1/p} \cdot ((\frac{C_i \cdot \sigma_i^2}{n})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{p}})$

2: Set $M = 8L_p$, $\eta = \epsilon^{\frac{1}{p}}$, $T = \left\lceil \frac{2(p+1)!\Delta}{M\eta^{p+1}} \right\rceil$, $n_1 = \left\lceil \sigma_1^2 C_1 \cdot (\frac{128(p+1)! \cdot A_p}{M\eta^{p+1} - 128(p+1)! \cdot A_p B_1^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} \right\rceil$

3: Let $A_p = \frac{(p!)^{\frac{p+1}{p}}}{4M^{\frac{1}{p}}} + 2(\frac{2p!}{M})^{\frac{1}{p}}$, $A_p' = \frac{(p!)^{\frac{p+1}{p}}}{4M^{\frac{1}{p}}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}$, $\sigma = \max_{2 \leq i \leq p} \sigma_i$, $B = \max_{2 \leq i \leq p} B_i$, $C = \max_{2 \leq i \leq p} C_i$

4: Set $n = \left\lceil \sigma^2 C \cdot (\frac{128(p+1)! \cdot (p-1) A_p'}{M\eta^{\frac{p^2-1}{p}} - 128(p+1)! \cdot (p-1) A_p' B^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} \right\rceil$

5: **for** $t = 1$ **to** T **do**

6: Query the first order oracle n_1 times at $x^{(t)}$ and compute

$$D^1(x^{(t)}) = \frac{1}{n_1} \sum_{j=1}^{n_1} \tilde{\nabla} F(x^{(t)}, z^{(t,j)}, b_1), \quad z^{(t,j)} \sim P_z$$

7: For $i \in \{2, \ldots, p\}$, query the i^{th} order oracle n times at $x^{(t)}$ and compute

$$D^{(i)}(x^{(t)}) = \frac{1}{n} \sum_{j=1}^{n} \tilde{\nabla}^i F(x^{(t)}, z^{(t,j)}, b_i), \quad z^{(t,j)} \sim P_z$$

Set the next point $x^{(t+1)}$ as

$$x^{(t+1)} = \underset{y: ||y - x^{(t)}|| \leq \eta}{\arg\min} \sum_{i=1}^{p} \frac{1}{i!} D^{(i)}[y - x^{(t)}]^i + \frac{M}{(p+1)!} ||y - x^{(t)}||^{p+1}$$

8: **end for**

Output: $\hat{x}$ chosen uniformly at random from $\{x^{(t)}\}_{t=1}^T$

In lemma 5, we show that there do exists constants $\{C_i\}$ such that the condition on line (1) holds. In theorem 3, we analyze the oracle complexity of this algorithm for reaching an ϵ-stationary point, and observe that if the bias amounts are small enough, we get a total oracle complexity of $O(\frac{1}{\epsilon^{\frac{3p+1}{p}}} + \frac{1}{\epsilon^{\frac{3p-1}{p}}})$, implying that as $p \to \infty$, we recover the known oracle complexity lower bound of $O(\frac{1}{\epsilon^3})$. Here, we realize the benefits of using higher order information in biased and stochastic oracle settings.

Theorem 3 *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, biased and stochastic p-order oracles in $\mathcal{O}(F, \sigma_{1:p})$ such that $B_1 = O(\epsilon^2)$ and $B_i = O(\epsilon^{\frac{2(p-1)}{p}})$ for all $2 \leq i \leq p$, with probability at least $\frac{5}{8}$, Algorithm*

1 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \leq \epsilon$ and performs at most

$$O(\frac{\sigma_1^2 \Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{(\max_{2 \leq i \leq p} \sigma_i)^2 \Delta}{\epsilon^{\frac{3p-1}{p}}})$$

queries to the stochastic and biased derivative oracles.

Proof Here we outline a proof sketch of theorem 3, deferring the full proof to Appendix A. We first show that (lemma 7) for all $M \geq 8L_p$ and $0 \leq \eta < 1$, we have that

$$F(x) - F(y) > F(x) - F(y) > \frac{M}{8(p+1)!}||y-x||^{p+1} - (\frac{2p!}{M})^{\frac{1}{p}} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}$$
$$- \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^i(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot \eta^{\frac{p+1}{p}}$$

where

$$y \in \underset{z:||z-x|| \leq \eta}{\arg\min}\ m_x(z)$$

and

$$m_x(y) = F(x) + \langle D^{(1)}, y - x \rangle + \sum_{i=2}^{p} \frac{1}{i!} D^{(i)}[y-x]^i + \frac{M}{(p+1)!}||y-x||^{p+1}$$

We extend this lemma to consider the case where $D^{(i)}$ are random variables in lemma 9 and then develop an expression for the probability of reaching an ϵ-stationary point, remarking that under the following bias conditions, we get the ϵ-stationary point with probability at least $\frac{5}{8}$. ■

4.1. Variance Reduction

Given the many works that have demonstrated the advantages of using variance reduction for derivative estimation, we investigate the potential advantages of using variance reduction in a biased setting as well. Many previous works have primarily relied on recursive variance reduction (e.g. Fang et al. (2018)) to compute cheap estimators of the gradient $\nabla F(x^{(t)})$. In our implementation of recursive variance reduction, we build on that of Arjevani et al. (2020) by estimating $\nabla^i F(x^{(t)}) - \nabla^i F(x^{(t+1)})$ by averaging $\nabla^{i+1} F$-vector products for all $i \in [p]$, instead of just doing this with the gradient. To derive this estimator, we first note that for all i, it holds that (by the Fundamental Theorem of Calculus) for all x, x': $\nabla^i F(x) - \nabla^i F(x') = \int_0^1 \nabla^{i+1} F(xt + x'(1-t))(x - x')dt$. Now, to approximate this integral, we construct the following estimator for $\nabla^i F$, where K is chosen to be proportional to $||x - x'||^2$:

$$\tilde{\nabla}^i F = \frac{1}{K}\sum_{k=0}^{K-1} \tilde{\nabla}^{i+1} F(x \cdot (1 - \frac{k}{K}) + x' \cdot \frac{k}{K}, z^{(i)}, b_i)(x - x')$$

We reset the derivative estimators according to a defined probability metric b and dynamically set the batch size proportional to the difference between the current iterate and the previous iterate squared and incorporate this recursive variance reduction approach for all p derivatives. In theorem 4, we analyze the oracle complexity of this algorithm for finding an ϵ-stationary point.

Algorithm 2 **H**igher-**O**rder **R**ecursive **V**ariance **R**eduction **(HO-RVR)**

Input: Precision parameter ϵ, probability b, current iterate x, previous iterate x_{prev}, derivative order i, derivative estimate with respect to x_{prev}, D^i_{prev}, Biased and stochastic oracle $(O^p_F, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$ for $F \in \mathcal{F}_p(\Delta, L_{1:p})$.

Set

$$K = \left\lceil \frac{5(\sigma_{i+1}^2 + L_{i+1}\epsilon)}{b\epsilon^2} \cdot ||x - x_{\text{prev}}||^2 \right\rceil$$

Set $n = \left\lceil \frac{5\sigma_i^2}{\epsilon^2} \right\rceil$

Sample $C \sim \text{Bernoulli}(b)$.

if C is 1 **or** D^i_{prev} is None **then**

Query the i^{th} order oracle n times at x and set

$$D^{(i)} = \frac{1}{n}\sum_{j=1}^{n} \tilde{\nabla}^i F(x, z^{(j)}, b_i), \quad z^{(j)} \sim P_z$$

else

For $k \in \{0, \ldots, K\}$, set

$$x^{(k)} = \frac{k}{K}x + (1 - \frac{k}{K})x_{\text{prev}}$$

Query the i^{th} order oracle at the points $\{x^{(k)}\}_{k=0}^{K-1}$ and set

$$D^{(i)} = D^{(i)}_{\text{prev}} + \sum_{k=1}^{K} \tilde{\nabla}^{i+1} F(x^{(k-1)}, z^{(k)}, b_{i+1}), \quad z^{(k)} \sim P_z$$

end if

Output: $D^{(i)}$

Theorem 4 *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, biased and stochastic p-order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$ such that $B_i = O(\epsilon^{\frac{2(p-1)}{p}})$ for all $1 \le i \le p$, with probability at least $\frac{3}{4}$, Algorithm 3 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \le \epsilon$ and performs at most*

$$O(\frac{(\max_{1\le i\le p} \sigma_i)^2 \Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{\Delta}{\epsilon^{\frac{2p+1}{p}}})$$

queries to the stochastic and biased derivative oracles.

Proof We provide a brief sketch of the proof, deferring the full proof to appendix B. In lemma 11, we prove that for all $1 \le i \le p$ and all timesteps $t \ge 1$,

$$\mathbb{E}||D^{(i)}(x^{(t)}) - \nabla^i F(x^{(t)})||^2 \le 4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}}$$

Algorithm 3 **H**igher-**O**rder **R**ecursive **V**ariance **R**eduction **D**erivative Estimation **(HO-RVR-D)**

Input: Precision parameter ϵ, Biased and stochastic oracle $(O_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$ for $F \in \mathcal{F}_p(\Delta, L_{1:p})$, derivative order p .

1: Pick b such that $0 < b \leq 1$.

2: Find M_1 such that

$$\frac{16(p+1)!}{M_1 b^{p+1}}\left(\frac{(p!)^{\frac{p+1}{p}}}{4M_1^{\frac{1}{p}}} + 2\left(\frac{2p!}{M_1}\right)^{\frac{1}{p}}\right) \cdot \left(\frac{66}{5} + O(1)\right) \leq \frac{1}{8}$$

3: Find M_2 such that

$$\frac{16(p+1)! \cdot (p-1)}{M_2 b^{\frac{p^2-1}{p}}}\left(\frac{(p!)^{\frac{p+1}{p}}}{4M_2^{\frac{1}{p}}} + \left(\frac{2p \cdot p!}{M_2}\right)^{\frac{1}{p}}\right) \cdot \left(\frac{66}{5} + O(1)\right) \leq \frac{1}{8}$$

4: Set $M = \max(M_1, M_2, 8L_p), \eta = b\epsilon^{\frac{1}{p}}, T = \left\lceil \frac{2(p+1)! \cdot \Delta}{M\eta^{p+1}} \right\rceil$

5: Set $x^{(0)} = x^{(1)} = 0$, $D^{(i)} =$ None for $i \in \{1, \ldots, p\}$

6: **for** $t = 1$ **to** T **do**

7: $\quad D_t^{(i)} =$ **HO-RVR**(ϵ, b, $x^{(t)}$, $x^{(t-1)}$, $D_{t-1}^{(i)}$)

8: $\quad$ Set the next point $x^{(t+1)}$ as

$$x^{(t+1)} = \underset{y:||y-x^{(t)}|| \leq \eta}{\arg\min} \sum_{i=1}^{p} \frac{1}{i!} D_t^{(i)}[y - x^{(t)}]^i + \frac{M}{(p+1)!}||y - x^{(t)}||^{p+1}$$

9: **end for**

Output: $\hat{x}$ chosen uniformly at random from $\{x^{(t)}\}_{t=1}^{T}$

where $B = \max_{1 \leq i \leq p} B_i$. Then we combine this result with the probability expression in theorem 3 (located in appendix A) to prove the theorem. ■

Surprisingly, we find that applying recursive variance reduction does not improve the $O(\frac{1}{\epsilon^{\frac{3p+1}{p}}} + \frac{1}{\epsilon^{\frac{3p-1}{p}}})$ oracle complexity from Algorithm 1 and only reduces the allowable bias for the first order term. It is interesting, however, that with this scheme, we can actually have $O(\epsilon^{\frac{2(p-1)}{p}})$ bias for the gradient term as opposed to the $O(\epsilon^2)$ gradient bias requirement for the MHOD algorithm.

5. Conclusion

This paper extends the settings of deterministic derivative oracles and stochastic but unbiased oracles to consider derivative oracles that are both stochastic and biased. We provide a nearly matching first order lower bound to complement the first order upper bound that is provided in this stochastic and biased scenario in Ajalloeian and Stich (2020). We further extend this lower bound for algorithms that use second order derivative information or higher for finding ϵ-stationary points, and

developed a p^{th} order-regularized trust region higher order algorithm that as $p \to \infty$, approaches the ϵ^{-3} dependence in the lower bound under certain bias constructions. We slightly improve on one of these bias constraint by using a variance-reduction based scheme by maintaining the ϵ^{-3} dependence as $p \to \infty$.

There are several opportunities for future work that arises from the results in this work. First, our higher-order upper bounds only match the ϵ dependence of the corresponding lower bound in the limit as $p \to \infty$, which leaves open the possibility of a stronger upper bound. Second, it would be interesting to try and develop an algorithm that implements a variance-reduction based scheme to try and loosen the bias restrictions on the first order term further, as well as reduce the bias amount needed for the higher order terms as well.

References

Deeksha Adil, Brian Bullins, Aaron Sidford, and Chenyi Zhang. Balancing gradient and hessian queries in non-convex optimization. *Advances in Neural Information Processing Systems*, 2025.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. *Symposium on Theory of Computing*, 2017.

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients. *International Conference on Machine Learning, Workshop on "Beyond First Order Methods in ML Systems"*, 2020.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 2019.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. *Conference on Learning Theory*, 2020.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 2023.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. *International Conference on Machine Learning*, 2017.

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

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: First-order methods. *Mathematical Programming*, 2019b.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 2023.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36, 2024.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *Advances in Neural Information Processing Systems*, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. *Conference on Learning Theory*, 2019.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 2013.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwith for distributed training. *International Conference on Learning Representations*, 2018.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *International Conference on Learning Representations*, 2024.

Xinyu Luo, Cedar Site Bai, Bolian Li, Petros Drineas, Ruqi Zhang, and Brian Bullins. Stacey: Promoting stochastic steepest descent via accelerated ℓ_p-smooth nonconvex optimization. *International Conference on Machine Learning*, 2025.

Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 2014.

A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. *Problem Complexity and Method Efficiency in Optimization*, 1983.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in Neural Information Processing Systems*, 2017.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 2018.

Appendix A. Proof of Theorem 3

We start by defining and proving some important lemmas.

Lemma 5 *For any integer $p \geq 1$, there exists a d-dependent, n_i-independent constant $C \geq 0$ such that*

$$\mathbb{E}||D_t^{(i)} - \nabla F^i(x^{(t)})||_{\text{op}}^{\frac{p+1}{p}} \leq 2^{\frac{1}{p}} \cdot ((\frac{C \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{p}})$$

for all $t \geq 1$.

Proof

$$\begin{aligned}
&\mathbb{E}[||D_t^{(i)} - \nabla F^i(x^{(t)})||^{\frac{p+1}{p}}] \\
&= \mathbb{E}[||D_t^{(i)} - \nabla F^i(x^{(t)}) - b_i(x^{(t)}) + b_i(x^{(t)})||^{\frac{p+1}{p}}] \\
&\leq 2^{1/p} \cdot (\mathbb{E}[||D_t^{(i)} - \nabla F^i(x^{(t)}) - b_i(x^{(t)})||^{\frac{p+1}{p}}] + \mathbb{E}[||b_i(x^{(t)})||^{\frac{p+1}{p}}]) \\
&\leq 2^{1/p} \cdot (\mathbb{E}[||D_t^{(i)} - \nabla F^i(x^{(t)}) - b_i(x^{(t)})||^{\frac{p+1}{p}}] + B_i^{\frac{p+1}{p}})
\end{aligned}$$

Now, we can say that for any $r \in [1, 2]$

$$\begin{aligned}
&\mathbb{E}[||D_t^{(i)} - \nabla F^i(x^{(t)}) - b_i(x^{(t)})||_{\text{op}}^r] \\
&= \mathbb{E}[||\frac{1}{n_i}\sum_{j=1}^{n_i} \tilde{\nabla}^i F(x^{(t)}, z^{(t,j)}) - \nabla F^i(x^{(t)}) - b_i(x^{(t)})||_{\text{op}}^r] \\
&\leq (\mathbb{E}[||\frac{1}{n_i}\sum_{j=1}^{n_i} \tilde{\nabla}^i F(x^{(t)}, z^{(t,j)}) - \nabla F^i(x^{(t)}) - b_i(x^{(t)})||_{\text{op}}^2])^{r/2} \\
&\leq (\frac{C \cdot \sigma_i^2}{n_i})^{r/2}
\end{aligned}$$

which follows from Lyapunov's inequality and the proof below. Thus, we have a final bound of

$$2^{1/p} \cdot ((\frac{C \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{p}})$$

■

Lemma 6 *Given $A_i \in \mathbb{R}^{d_1 \times \cdots \times d_m}$, where $d_1 = \ldots = d_m = d$, and $\mathbb{E}[A_i] = B$ and $\mathbb{E}[||A_i - B||^2] \leq \sigma^2$, we have that*

$$\mathbb{E}[||\frac{1}{n}\sum_{i=1}^{n} A_i - B||_{\text{op}}^2] \leq \frac{C \cdot \sigma^2}{n}$$

for some d-dependent, n-independent constant $C \geq 0$.

Proof Let $X_i = A_i - B$ and observe that

$$\begin{aligned}
&\mathbb{E}[||\sum_{i=1}^{n} X_i||_{\text{op}}^2] \leq \mathbb{E}_{X,X'}[||\sum_{i=1}^{n} X_i - X_i'||_{\text{op}}^2] \\
&= \mathbb{E}_{X,X',\epsilon}[||\sum_{i=1}^{n} \epsilon_i (X_i - X_i')||_{\text{op}}^2] \\
&\leq 4\mathbb{E}_{X,\epsilon}[||\sum_{i=1}^{n} \epsilon_i X_i||_{\text{op}}^2]
\end{aligned}$$

where $(X_i')_{i=1}^n$ is a sequence of independent copies of $(X_i)_{i=1}^n$ and $(\epsilon_i)_{i=1}^n$ is a sequence of Rademacher random variables. Now, take S such that $S \subset \{1, \ldots, m\}$, where $|S| = \lfloor m/2 \rfloor$. We define

$$Z_i \in \mathbb{R}^{(\prod_{k \in S} d_k) \times (\prod_{k \in S^c} d_k)}$$

to be a flattened version of X_i. Let $D = \min\{\prod_{k \in S} d_k, \prod_{k \in S^c} d_k)$, so in this case, $D = d^{\lfloor m/2 \rfloor}$. We now prove that for any p, there exists d-dependent constants C_1, C_2, C_3 such that

$$||X_i||_{\text{op}} \leq C_2 \cdot ||Z_i||_{S_{2p}} \leq C_2 C_1 C_3^2 \cdot D^{\frac{1}{2p}} \cdot ||X_i||_{\text{op}}$$

We note that

$$\begin{aligned}
&||X_i||_{\text{op}} \\
&= \sup_{||u^{(1)}||=1 \ \ldots \ ||u^{(m)}||=1} \langle X_i, u^{(1)} \otimes \ldots \otimes u^{(m)} \rangle \\
&= \sup_{||a_1||=||b_1||=1} \langle Z_i, a_1 b_1^T \rangle \\
&\leq \sup_{||a||=||b||=1} \langle Z_i, ab^T \rangle \\
&= ||Z_i||_{\text{op}} \leq C_2 \cdot ||Z_i||_2 \leq C_2 \cdot \sigma_{\max}(Z_i)
\end{aligned}$$

for some constant C_2, since due to norm equivalence in finite dimensional spaces, there exists constants C_1, C_2 such that $C_1 \cdot ||Z_i||_2 \leq ||Z_i||_{\text{op}} \leq C_2 \cdot ||Z_i||_2$. Also,

$$a_1 = \bigotimes_{k \in S} u^{(k)}, b_1 = \bigotimes_{k \notin S} u^{(k)}$$

Now

$$\begin{aligned}
&C_2 \cdot \sigma_{\max}(Z_i) \\
&\leq C_2 (\sum_{j=1}^{D} \sigma_j^{2p}(Z_i))^{\frac{1}{2p}} \\
&= C_2 \cdot ||Z_i||_{S_{2p}} \\
&\leq C_2 \cdot (D \cdot \sigma_{\max}^{2p}(Z_i))^{\frac{1}{2p}} \\
&= C_2 \cdot D^{\frac{1}{2p}} \cdot \sigma_{\max}(Z_i) \\
&\leq C_2 C_1 \cdot D^{\frac{1}{2p}} \cdot ||Z_i||_{\text{op}}
\end{aligned}$$

Since

$$||Z_i||_{\text{op}} = \sup_{||a||=||b||=1} a^T Z_i b$$

expand a and b in their orthonormal bases as follows:

$$a = \sum_{\alpha=1}^{d^{\lfloor m/2 \rfloor}} a_\alpha e_\alpha, b = \sum_{\beta=1}^{d^{\lceil m/2 \rceil}} b_\beta f_\beta$$

which implies that

$$\begin{aligned}
&|a^T Z_i b| \\
&= |\sum_{\alpha,\beta} a_\alpha b_\beta \langle X_i, e_\alpha \otimes f_\beta \rangle| \\
&\leq \sum_{\alpha,\beta} |a_\alpha| \cdot |b_\beta| \cdot |\langle X_i, e_\alpha \otimes f_\beta \rangle| \\
&\leq \sum_{\alpha,\beta} |a_\alpha| \cdot |b_\beta| \cdot ||X_i||_{\text{op}} \\
&\leq ||X_i||_{\text{op}} \cdot ||a||_1 \cdot ||b||_1 \\
&\leq C_3^2 \cdot ||X_i||_{\text{op}}
\end{aligned}$$

due to Cauchy-Schwarz inequality and since $||x||_1 \leq C_3 \cdot ||x||$ for some constant C_3 for all x. This implies that

$$D^{\frac{1}{2p}} \cdot ||Z_i||_{\text{op}} \leq D^{\frac{1}{2p}} \cdot C_3^2 \cdot ||X_i||_{\text{op}}$$

which proves the inequality. Now, letting $p = 1$, we have that

$$\mathbb{E}_{X,\epsilon}[||\sum_{i=1}^n \epsilon_i X_i||_{\text{op}}^2] \leq C_2^2 \cdot \mathbb{E}_{X,\epsilon}[||\sum_{i=1}^n \epsilon_i Z_i||_{S_2}^2]$$

By Matrix-Khintchine inequality (Mackey et al., 2014), we have that

$$\begin{aligned}
&(\mathbb{E}[\sum_{i=1}^n ||\epsilon_i Z_i||_{S_{2p}}^{2p}])^{1/p} \\
&\leq (2p-1) \cdot ||(\sum_{i=1}^n Z_i^2)^{1/2}||_{S_{2p}}^2 \\
&= (2p-1) \cdot ||\sum_{i=1}^n Z_i^2||_{S_{2p}} \\
&\leq (2p-1) \cdot \sum_{i=1}^n ||Z_i||_{S_{2p}}^2 \\
&\leq (2p-1) \cdot D^{1/p} C_2^2 C_1^2 \cdot \sum_{i=1}^n ||Z_i||_{\text{op}}^2 \\
&\leq (2p-1) \cdot D^{1/p} C_2^2 C_1^2 C_3^4 \cdot \sum_{i=1}^n ||X_i||_{\text{op}}^2
\end{aligned}$$

which implies that

$$\mathbb{E}_{X,\epsilon}[||\sum_{i=1}^{n}\epsilon_i Z_i||^2] \leq DC_2^2C_1^2C_3^4 \cdot \sum_{i=1}^{n}\mathbb{E}[||X_i||_{\text{op}}^2] \leq d^{m/2}C_2^2C_1^2C_3^4 \cdot n\sigma^2$$

Thus, after normalizing the final upper bound is

$$\frac{4d^{m/2}C_2^2C_1^2C_3^4 \cdot \sigma^2}{n}$$

which finishes the proof. ■

Lemma 7 *Let*

$$m_x(y) = F(x) + \langle D^{(1)}, y - x\rangle + \sum_{i=2}^{p}\frac{1}{i!}D^{(i)}(x)[y-x]^i + \frac{M}{(p+1)!}||y-x||^{p+1}$$

and let $y \in \arg\min_{z:||z-x||\leq\eta} m_x(z)$*. Then, we have that*

$$F(x) - F(y) > \frac{M}{8(p+1)!}||y-x||^{p+1} - (\frac{2p!}{M})^{\frac{1}{p}} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}$$
$$- \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^i(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot \eta^{\frac{p+1}{p}}$$

for all $M \geq 8L_p$ *and* $0 \leq \eta < 1$*.*

Proof We have that $F(y) - F(x)$

$$\leq F(x) + \langle \nabla F(x), y-x\rangle + \sum_{i=2}^{p}\frac{1}{i!}\nabla^i F(x)[y-x]^i + \frac{L_p}{(p+1)!}||y-x||^{p+1} - F(x)$$
$$\leq m_x(y) + \langle \nabla F(x) - D^{(1)}(x), y-x\rangle + \sum_{i=2}^{p}\frac{1}{i!}(\nabla^i F(x) - D^{(i)}(x))[y-x]^i + \frac{L_p - M}{(p+1)!}||y-x||^{p+1} - m_x(x)$$
$$\leq \langle \nabla F(x) - g, y-x\rangle + \sum_{i=2}^{p}\frac{1}{i!}(\nabla^i F(x) - D^{(i)}(x))[y-x]^i + \frac{L_p - M}{(p+1)!}||y-x||^{p+1}$$
$$\leq -\frac{7M}{8(p+1)!}||y-x||^{p+1} + ||\nabla F(x) - g|| \cdot ||y-x||$$
$$+ \sum_{i=2}^{p}\frac{1}{i!}||\nabla^i F(x)[y-x,:,\ldots,:] - D^{(i)}(x)[y-x,:,\ldots,:]||_{\text{op}} \cdot ||y-x||$$

since $||y-x|| \leq \eta$ and since $\eta < 1$, we have that $||y-x||^{i-1} \leq ||y-x||$ for $i \geq 2$. By Young's inequality, we have that

$$||\nabla F(x) - D^{(1)}(x)|| \cdot ||y-x|| \leq ((\frac{2p!}{M})^{\frac{1}{p}} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \cdot \frac{p}{p+1}) + (\frac{||y-x||^{p+1}}{(p+1)} \cdot \frac{M}{2p!})$$
$$= (\frac{2p!}{M})^{\frac{1}{p}}(\frac{p \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}}{p+1}) + \frac{M||y-x||^{p+1}}{2(p+1)!}$$

and

$$
\begin{aligned}
&||\nabla^i F(x)[y-x,:,\dots,:] - D^{(i)}(x)[y-x,:,\dots,:]||_{\text{op}} \cdot ||y-x|| \\
&\leq (\frac{2p\cdot p!}{M})^{\frac{1}{p}} \frac{p\cdot ||\nabla^i F(x)[y-x,:,\dots,:] - D^{(i)}(x)[y-x,:,\dots,:]||_{\text{op}}^{\frac{p+1}{p}}}{p+1} + \frac{M||y-x||^{p+1}}{(p+1)\cdot(2\cdot p!)} \\
&= (\frac{2p\cdot p!}{M})^{\frac{1}{p}} \frac{p\cdot ||\nabla^i F(x)[y-x,:,\dots,:] - D^{(i)}(x)[y-x,:,\dots,:]||_{\text{op}}^{\frac{p+1}{p}}}{p+1} + \frac{M||y-x||^{p+1}}{2p\cdot(p+1)!}
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&-\frac{7M}{8(p+1)!}||y-x||^{p+1} + ||\nabla F(x) - D^{(1)}(x)|| \cdot ||y-x|| \\
&+\sum_{i=2}^{p}\frac{1}{i!}||\nabla^i F(x)[y-x,:,\dots,:] - D^{(i)}(x)[y-x,:,\dots,:]||_{\text{op}} \cdot ||y-x|| \\
&\leq -\frac{7M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - g||^{\frac{p+1}{p}} + \frac{M||y-x||^{p+1}}{2(p+1)!} \\
&+\sum_{i=2}^{p}\frac{1}{i!}\cdot[(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \frac{p\cdot ||\nabla^i F(x)[y-x,:,\dots,:] - D^{(i)}(x)[y-x,:,\dots,:]||_{\text{op}}^{\frac{p+1}{p}}}{p+1} + \frac{M||y-x||^{p+1}}{2p\cdot(p+1)!}] \\
&\leq -\frac{3M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \\
&+\sum_{i=2}^{p}\frac{1}{i!}[(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \frac{p\cdot ||\nabla^i F(x) - D^{(i)}(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot ||y-x||^{\frac{p+1}{p}}}{p+1} + \frac{M\cdot||y-x||^{p+1}}{2p\cdot(p+1)!}] \\
&< -\frac{3M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \\
&+\sum_{i=2}^{p}\frac{1}{2}[(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \frac{p\cdot ||\nabla^i F(x) - D^{(i)}(x)||^{\frac{p+1}{p}} \cdot ||y-x||^{\frac{p+1}{p}}}{p+1} + \frac{M\cdot||y-x||^{p+1}}{2p\cdot(p+1)!}] \\
&< -\frac{3M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \\
&+\frac{1}{2}\sum_{i=2}^{p}[(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^{(i)}(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot ||y-x||^{\frac{p+1}{p}} + \frac{M\cdot||y-x||^{p+1}}{2p\cdot(p+1)!}] \\
&< -\frac{M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \\
&+\frac{1}{2}\sum_{i=2}^{p}(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^{(i)}(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot \eta^{\frac{p+1}{p}} \\
&< -\frac{M}{8(p+1)!}||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} + \frac{1}{2}\sum_{i=2}^{p}(\frac{2p\cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^{(i)}(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot \eta^{\frac{p+1}{p}}
\end{aligned}
$$

which proves the lemma. ■

Lemma 8 *Under the same setting as lemma 7, we have that*

$$\mathbf{1}[||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p] \leq \frac{1}{\eta^p}||y-x||^p + \frac{p!}{M\eta^p}(||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)||_{\text{op}} \cdot \eta^i)$$

Proof We have that

$$\begin{aligned}
&||\nabla F(y)|| \\
&\leq ||\nabla F(y) - \sum_{i=0}^{p-1}\frac{1}{i!}\nabla^{i+1}F(x)[y-x]^i|| + ||\sum_{i=0}^{p-1}\frac{1}{i!}\nabla^{i+1}F(x)[y-x]^i|| \\
&\leq \frac{L_p}{p!}||y-x||^p + ||\nabla F(x) + \sum_{i=1}^{p-1}\frac{1}{i!}\nabla^{i+1}F(x)[y-x]^i|| \\
&\leq \frac{L_p}{p!}||y-x||^p + ||\nabla F(x) - D^{(1)}(x)|| + ||\sum_{i=1}^{p-1}\frac{1}{i!}[\nabla^{i+1}F(x)[y-x]^i - D^{(i+1)}(x)[y-x]^i]|| \\
&+ ||D^{(1)}(x) + \sum_{i=1}^{p-1}\frac{1}{i!}D^{(i+1)}[y-x]^i|| \\
&\leq \frac{L_p}{p!}||y-x||^p + ||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot ||y-x||^i \\
&+ ||D^{(1)}(x) + \sum_{i=1}^{p-1}\frac{1}{i!}D^{(i+1)}[y-x]^i|| \\
&\leq \frac{L_p+M}{p!}||y-x||^p + ||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot ||y-x||^i \\
&\leq \frac{L_p+M}{p!}||y-x||^p + ||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i
\end{aligned}$$

since under first order optimality conditions for $y \in \arg\min_z m_x(z)$, we have that

$$D^{(1)}(x) + \sum_{i=1}^{p-1}\frac{1}{i!}D^{(i+1)}[y-x]^i + \frac{M}{p!}||y-x||^{p+1}(x-y) = 0$$

We now have that

$$\begin{aligned}
&||y-x||^p \geq \frac{p!}{L_p+M}(||\nabla F(y)|| - ||\nabla F(x) - D^{(1)}(x)|| - \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i) \\
&\geq \min\{\eta^p, \frac{p!}{L_p+M}(||\nabla F(y)|| - ||\nabla F(x) - D^{(1)}(x)|| - \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i))\} \\
&\geq \min\{\eta^p, \frac{p!}{L_p+M}||\nabla F(y)||\} - \frac{p!}{L_p+M}||\nabla F(x) - D^{(1)}(x)|| - \frac{p!}{L_p+M}\sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i
\end{aligned}$$

and since $L_p \leq \frac{M}{8}$ and $\frac{M}{L_p+M} < 1$, we have that

$$M||y-x||^p \geq \min\{M\eta^p, \frac{Mp!}{L_p+M}||\nabla F(y)||\} - \frac{Mp!}{L_p+M}||\nabla F(x) - D^{(1)}(x)||$$
$$- \frac{Mp!}{L_p+M}\sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i$$
$$> \min\{M\eta^p, \frac{8p!}{9}||\nabla F(y)||\} - p!(||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i)$$

which means that

$$\min\{M\eta^p, \frac{8p!}{9}||\nabla F(y)||\} < M||y-x||^p + p!(||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i)$$

which then implies that (since for all $a, b \geq 0$, $a\mathbf{1}[b \geq a] \leq \min\{a, b\}$)

$$M\eta^p \cdot \mathbf{1}[||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p] \leq M||y-x||^p + p!(||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i)$$
$$\implies \mathbf{1}[||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p] \leq \frac{1}{\eta^p}||y-x||^p + \frac{p!}{M\eta^p}(||\nabla F(x) - D^{(1)}(x)|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{(i+1)}(x)|| \cdot \eta^i)$$

which proves the lemma. ■

Lemma 9 *Consider the same setting as lemma 7, but let the derivative estimates D^i be random variables. Then, we have that*

$$\mathbb{E}[F(x) - F(y)] \geq \frac{M\eta^{p+1}}{16(p+1)!} \Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p+1}}] \cdot [(\frac{\sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p+1}}) \sum_{i=2}^{p}[(\frac{\sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

Proof First, we note that

$$\begin{aligned}
&\mathbf{1}[||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p] \\
&\leq (\frac{1}{\eta^p}||y-x||^p + \frac{p!}{M\eta^p}(||\nabla F(x) - g|| + \sum_{i=1}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i))^{\frac{p+1}{p}} \\
&\leq \frac{2^{1/p}}{\eta^{p+1}}||y-x||^{p+1} + 2^{1/p} \cdot (\frac{p!}{M\eta^p})^{\frac{p+1}{p}} \cdot (\sum_{i=0}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i)^{\frac{p+1}{p}} \\
&\leq \frac{2^{1/p}}{\eta^{p+1}}||y-x||^{p+1} + 2^{1/p} \cdot (\frac{p!}{M\eta^p})^{\frac{p+1}{p}} \cdot (\sum_{i=0}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i)^{\frac{p+1}{p}} \\
&< \frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{2(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot (\sum_{i=0}^{p-1}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i)^{\frac{p+1}{p}} \\
&\leq \frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{2(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot p^{1/p} \cdot \sum_{i=0}^{p-1}||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}} \\
&< \frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{4(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot \sum_{i=0}^{p-1}||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}
\end{aligned}$$

where we used the fact that for any $a_i \geq 0$, we have that

$$(\sum_{i=1}^{n} a_i)^{\frac{p+1}{p}} \leq n^{\frac{1}{p}} \sum_{i=1}^{n} a_i^{\frac{p+1}{p}}$$

which follows from an application of Hölder's inequality. We also used the fact that for all $p \geq 1$, $p^{1/p} < 2$. Taking expectations on each side, we have that

$$\mathbb{E}[||y-x||^{p+1}] \geq \frac{\eta^{p+1}}{2} \Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - \frac{2(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}} \cdot \sum_{i=0}^{p-1}\mathbb{E}[||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}]$$

From the earlier lemma, we know that

$$
\begin{aligned}
&\mathbb{E}[F(x) - F(y)] \\
&> \frac{M}{8(p+1)!}||y-x||^{p+1} - (\frac{2p!}{M})^{\frac{1}{p}} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}} \\
&- \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot ||\nabla^i F(x) - D^i(x)||_{\text{op}}^{\frac{p+1}{p}} \cdot \eta^{\frac{p+1}{p}} \\
&\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - \frac{2(p!)^{\frac{p+1}{p}}}{\sqrt[p]{M} \cdot 8(p+1)!} \cdot \sum_{i=0}^{p-1}\mathbb{E}[||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}] \\
&- (\frac{2p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}] - \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \cdot \eta^{\frac{p+1}{p}} \\
&\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - \frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} \cdot \sum_{i=0}^{p-1}\mathbb{E}[||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}] \\
&- (\frac{2p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}] - \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \cdot \eta^{\frac{p+1}{p}}
\end{aligned}
$$

We then have that this expression

$$
\begin{aligned}
&= \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot \mathbb{E}[||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}] \\
&- \frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} \cdot \sum_{i=1}^{p-1}\mathbb{E}[||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}] - \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \cdot \eta^{\frac{p+1}{p}} \\
&= \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot \mathbb{E}[||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}] \\
&- \frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} \cdot \sum_{i=2}^{p}\mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{(i-1)(p+1)}{p}}] - \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \cdot \eta^{\frac{p+1}{p}}
\end{aligned}
$$

Since $\eta \leq 1$, we can further lower bound the expression by

$$\frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot \mathbb{E}[||\nabla F(x) - g||^{\frac{p+1}{p}}]$$
$$- \frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} \cdot \sum_{i=2}^{p} \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{(p+1)}{p}}] - \frac{1}{2}\sum_{i=2}^{p}(\frac{2p \cdot p!}{M})^{\frac{1}{p}} \cdot \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \cdot \eta^{\frac{p+1}{p}}$$
$$= \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot \mathbb{E}[||\nabla F(x) - g||^{\frac{p+1}{p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} \mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot 2^{1/p}[(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} 2^{1/p} \cdot [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

where we used the fact that $2^{1/p} \leq 2$ for all $p \geq 1$. ■

We are now ready to prove theorem 3 which we repeat here for the reader's convenience.

Theorem 10 *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, biased and stochastic p-order oracles in $\mathcal{O}(F, \sigma_{1:p})$ such that $B_1 = O(\epsilon^2)$ and $B_i = O(\epsilon^{\frac{2(p-1)}{p}})$ for all $2 \leq i \leq p$, with probability at least $\frac{5}{8}$, Algorithm 1 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \leq \epsilon$ and performs at most*

$$O(\frac{\sigma_1^2\Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{(\max_{2\leq i\leq p}\sigma_i)^2\Delta}{\epsilon^{\frac{3p-1}{p}}})$$

queries to the stochastic and biased derivative oracles.

Proof Using lemma 9, we have that:

$$\mathbb{E}[F(x^{(t)}) - F(x^{(t+1)})]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(x^{(t+1)})|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

and telescoping this recurrence implies that

$$\mathbb{E}[F(x^{(1)}) - F(x^{(T+1)})]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!} \cdot T \cdot (\frac{1}{T}\sum_{t=1}^{T} \Pr(||\nabla F(x^{(t+1)})|| \geq \frac{9M}{8p!}\eta^p)) - T[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- T\eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$
$$= \frac{M\eta^{p+1}}{16(p+1)!} \cdot T \cdot (\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p)) - T[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$- T\eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

which implies that

$$\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) \leq \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}}[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$+ \frac{16(p+1)!}{M\eta^{p+1}}\eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$
$$= \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}}[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}] \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}]$$
$$+ \frac{16(p+1)!}{M\eta^{\frac{p^2-1}{p}}}(\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

Now, from our setting of T, we immediately conclude that

$$\frac{16(p+1)!}{M\eta^{p+1}T}\Delta \leq \frac{1}{8}$$

Since

$$A_p = \frac{(p!)^{\frac{p+1}{p}}}{4M^{1/p}} + 2(\frac{2p!}{M})^{\frac{1}{p}}$$

and

$$A'_p = \frac{(p!)^{\frac{p+1}{p}}}{4M^{1/p}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}$$

From our setting of

$$n_1 = \lceil \sigma_1^2 C_1 \cdot (\frac{128(p+1)! \cdot A_p}{M\eta^{p+1} - 128(p+1)! \cdot A_p B_1^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} \rceil$$

we have that

$$(\frac{\sigma_1^2}{n_1})^{\frac{p+1}{2p}} \leq (\frac{1}{C_1})^{\frac{p+1}{2p}} \cdot (\frac{M\eta^{p+1}}{128(p+1)! \cdot A_p} - B_1^{\frac{p+1}{2p}})$$

which then means that

$$\frac{16(p+1)!}{M\eta^{p+1}} A_p \cdot [(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}] \leq \frac{1}{8}$$

Since

$$\sigma = \max_{2 \leq i \leq p} \sigma_i, B = \max_{2 \leq i \leq p} B_i, C = \max_{2 \leq i \leq p} C_i$$

From our setting of

$$n = n_i = \lceil \sigma^2 C \cdot (\frac{128(p+1)! \cdot (p-1)A'_p}{M\eta^{\frac{p^2-1}{p}} - 128(p+1)! \cdot (p-1)A'_p B^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} \rceil$$

for all $2 \leq i \leq p$ we have that

$$(\frac{\sigma^2}{n_i}) \leq \frac{M\eta^{\frac{p^2-1}{p}}}{128(p+1)! \cdot (p-1)A'_p} - B^{\frac{p+1}{2p}}$$

which implies that

$$\frac{16(p+1)!}{M\eta^{\frac{p^2-1}{p}}} A'_p \sum_{i=2}^{p} [(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}] \leq \frac{1}{8}$$

which implies that

$$\Pr(||\nabla F(\hat{x})|| \leq \frac{9M}{8p!}\epsilon) \geq \frac{5}{8}$$

proving the first part of theorem 3. Note that running the same argument with $\epsilon \leftarrow \frac{8p!}{9M}\epsilon$ will recover the ϵ upper bound exactly. Let M_i denote the total number of oracle queries for $\nabla^i F$ and

let $M = \sum_{i=1}^{p} M_i$ denote the total number of oracle queries. We have that

$$
\begin{aligned}
&\mathbb{E}[M] \\
&= M_1 + \sum_{i=2}^{p} \mathbb{E}[M_i] \\
&= Tn_1 + T\sum_{i=2}^{p} n_i \\
&\leq T(\sigma_1^2 C_1 \cdot (\frac{128(p+1)! \cdot A_p}{M\eta^{p+1} - 128(p+1)! \cdot A_p B_1^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} + 1) \\
&+ T\sum_{i=2}^{p} \sigma^2 C \cdot (\frac{128(p+1)! \cdot (p-1)A'_p}{M\eta^{\frac{p^2-1}{p}} - 128(p+1)! \cdot (p-1)A'_p B^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} + 1 \\
&\leq T(\sigma_1^2 C_1 \cdot (\frac{128(p+1)! \cdot A_p}{M\epsilon^{\frac{p+1}{p}} - 128(p+1)! \cdot A_p B_1^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}}) \\
&+ T\sum_{i=2}^{p} \sigma^2 C \cdot (\frac{128(p+1)! \cdot (p-1)A'_p}{M\epsilon^{\frac{p^2-1}{p^2}} - 128(p+1)! \cdot (p-1)A'_p B^{\frac{p+1}{2p}}})^{\frac{2p}{p+1}} + Tp \\
&\leq T \cdot O(\frac{\sigma_1^2}{\epsilon^2}) + T \cdot O(\frac{\sigma^2}{\epsilon^{\frac{2(p-1)}{p}}}) + Tp \\
&\leq (\frac{2(p+1)!\Delta}{M\epsilon^{\frac{p+1}{p}}} + 1) \cdot O(\frac{\sigma_1^2}{\epsilon^2}) + (\frac{2(p+1)!\Delta}{M\epsilon^{\frac{p+1}{p}}} + 1) \cdot O(\frac{\sigma^2}{\epsilon^{\frac{2(p-1)}{p}}}) + (\frac{2(p+1)!\Delta}{M\epsilon^{\frac{p+1}{p}}} + 1) \cdot O(1) \\
&\leq O(\frac{\sigma_1^2\Delta}{\epsilon^{2+\frac{p+1}{p}}} + \frac{\sigma^2\Delta}{\epsilon^{\frac{3p-1}{p}}} + \frac{\sigma_1^2}{\epsilon^2} + \frac{\sigma^2\Delta}{\epsilon^{\frac{2(p-1)}{p}}} + \frac{\Delta}{\epsilon^{\frac{p+1}{p}}}) \\
&\leq O(\frac{\sigma_1^2\Delta}{\epsilon^{2+\frac{p+1}{p}}} + \frac{(\max_{2\leq i\leq p}\sigma_i)^2\Delta}{\epsilon^{\frac{3p-1}{p}}} + \frac{\sigma_1^2}{\epsilon^2} + \frac{(\max_{2\leq i\leq p}\sigma_i)^2\Delta}{\epsilon^{\frac{2(p-1)}{p}}} + \frac{\Delta}{\epsilon^{\frac{p+1}{p}}})
\end{aligned}
$$

which (after Markov's inequality and a union bound) finishes the proof of theorem 3. ■

Appendix B. Proof of Theorem 4

In this section, we prove theorem 4. Given an objective $F \in \mathcal{F}_p(\Delta, L_{1:p})$ and a stochastic and biased p^{th} order oracle in $\mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$, in order to obtain derivative estimates $D^{(1)}(x^{(t)}), \ldots, D^{(p)}(x^{(t)})$ for each timestep $t \geq 1$ and each derivative order $i \in [p]$, we do the following (adapted from Arjevani et al. (2020)):

$$
\begin{aligned}
x^{(t)} &= \boldsymbol{A}^{(t)}(D_0^{(i)}, \ldots, D_{t-1}^{(i)}), b^{(t)} = \boldsymbol{B}^{(t)}(r^{(t-1)}) \\
D_t^{(i)} &= RVR(\epsilon, b^{(t)}, x^{(t)}, x^{(t-1)}, D_{t-1}^{(i)})
\end{aligned}
$$

where $\boldsymbol{A}^{(t)}$ and $\boldsymbol{B}^{(t)}$ are measurable mappings of the optimization algorithm and $\{r^{(t)}\}$ is an independent sequence of random seeds. In our setting, $b^{(t)}$ is fixed as a constant b for all $t \geq 1$. Theorem

4 holds for any sequence of queries such that $\mathcal{G}^{(t)} = \sigma(\{D_j^{(i)}, r^{(j)}\}_{j<t})$, where we have that $b^{(t)}$ is independent of $\mathcal{G}^{(t-1)}$ and $D_{t-1}^{(i)}$ for some filtration $\mathcal{G}^{(t)}$.

Lemma 11 *Let $B = \max_{1\leq i\leq p} B_i$. Then, we have that*

$$\mathbb{E}[||D^i(x^{(t)}) - \nabla^i F(x^{(t)})||^2] \leq 4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}}$$

for all $1 \leq i \leq p$ and all $t \geq 1$.

Proof First, we note that

$$\begin{aligned}
&\mathbb{E}||D^i(x^{(1)}) - \nabla^i F(x^{(1)})||^2 \\
&= \mathbb{E}[||b_i(x^{(1)}) + \frac{1}{n_1}\sum_{j=1}^{n_1}\epsilon_1(x^{(1)}, z^{(1,j)})||^2] \\
&\leq 2B_i^2 + \frac{2}{n_1^2}\sum_{j=1}^{n_1}||\epsilon_1(x^{(1)}, z^{(1,j)})||^2 \leq 2B_i^2 + \frac{2\sigma_1^2}{n_1} \leq 2B_i^2 + \frac{2\epsilon^2}{5}
\end{aligned}$$

Let $e^{(t)} = D_t^i(x^{(t)}) - \nabla F^i(x^{(t)})$, and we have that

$$\mathbb{E}[||e^{(t)}||^2|b^{(t)}] = b^{(t)} \cdot \mathbb{E}[||e^{(t)}||^2|C^{(t)} = 1] + (1 - b^{(t)}) \cdot \mathbb{E}[||e^{(t)}||^2|C^{(t)} = 0]$$

where

$$\mathbb{E}[||e^{(t)}||^2|C^{(t)} = 1] \leq 2B_i^2 + \frac{2\epsilon^2}{5}$$

We now say that

$$\begin{aligned}
&\mathbb{E}[||e^{(t)}||^2|C^{(t)} = 0] \\
&\leq \mathbb{E}[||e^{(t-1)} + \mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]||^2] + \mathbb{E}[||\psi^{(t)} - \mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]||^2] \\
&\leq \mathbb{E}[(1 + \frac{2}{b^{(t)}}) \cdot ||e^{(t-1)}||^2] + \mathbb{E}[(1 + \frac{2}{b^{(t)}}) \cdot ||\mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]||^2] + \mathbb{E}[||\psi^{(t)} - \mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]||^2]
\end{aligned}$$

where the first step follows from the fact that $\mathcal{G}^{(t)}$ is a measurable set, and the second step is by Young's inequality. Above, we have that

$$\psi^{(t)} = e^{(t)} - e^{(t-1)} = \sum_{k=1}^{K^{(t)}} \tilde{\nabla}^{i+1} F(x^{(t,k-1)}, z^{(t,k)}, b_i)(x^{(t,k)} - x^{(t,k-1)}) - \nabla^i F(x^{(t)}) + \nabla^i F(x^{(t-1)})$$

We can calculate that

$$\begin{aligned}
&\mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}] \\
&= \sum_{k=1}^{K^{(t)}} (\nabla^{i+1} F(x^{(t,k-1)}) + b_{i+1}(x^{(t,k-1)}))(x^{(t,k)} - x^{(t,k-1)}) - \nabla^i F(x^{(t)}) + \nabla^i F(x^{(t-1)})
\end{aligned}$$

which implies that

$$
\begin{aligned}
&||\mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]|| \\
&\leq \sum_{k=1}^{K^{(t)}} ||(\nabla^i F(x^{(t,k)}) - \nabla^i F(x^{(t,k-1)}) - \nabla^{i+1} F(x^{(t,k-1)})(x^{(t,k)} - x^{(t,k-1)})|| \\
&+ \sum_{k=1}^{K^{(t)}} ||b_{i+1}(x^{(t,k-1)})|| \cdot ||x^{(t,k)} - x^{(t,k-1)}|| \\
&\leq K^{(t)} \cdot \frac{L_{i+1}}{2} \cdot (\frac{||x^{(t)} - x^{(t-1)}||}{K^{(t)}})^2 + B_{i+1}||x^{(t)} - x^{(t-1)}|| \\
&\leq \frac{b^{(t)}\epsilon}{10} + B_{i+1}\eta \\
&= \frac{b^{(t)}\epsilon}{10} + B_{i+1} b^{(t)} \epsilon^{\frac{1}{p}}
\end{aligned}
$$

since we set

$$\eta = b^{(t)} \epsilon^{\frac{1}{p}}$$

We can also derive that

$$
\begin{aligned}
&\mathbb{E}[||\psi^{(t)} - \mathbb{E}[\psi^{(t)}|\mathcal{G}^{(t)}]||^2] \\
&= \frac{1}{(K^{(t)})^2} \sum_{k=1}^{K^{(t)}} \mathbb{E}[||(\tilde{\nabla}^{i+1} F(x^{(t,k-1)}, z^{(t,k)}) - \nabla^{i+1} F(x^{(t,k-1)}) - b_{i+1}(x^{(t,k-1)})(x^{(t)} - x^{(t-1)})||^2 |\mathcal{G}^{(t)}] \\
&\leq \frac{1}{(K^{(t)})^2} \sum_{k=1}^{K^{(t)}} \mathbb{E}[||(\tilde{\nabla}^{i+1} F(x^{(t,k-1)}, z^{(t,k)}) - \nabla^{i+1} F(x^{(t,k-1)}) - b_{i+1}(x^{(t,k-1)})||_{\text{op}}^2 |\mathcal{G}^{(t)}] \cdot ||x^{(t)} - x^{(t-1)}||^2 \\
&\leq \sigma_{i+1}^2 \frac{||x^{(t)} - x^{(t-1)}||^2}{K^{(t)}} \leq b^{(t)} \frac{\epsilon^2}{5}
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&\mathbb{E}||e^{(t)}||^2 \\
&\leq \mathbb{E}[b^{(t)} \cdot (2B_i^2 + \frac{2\epsilon^2}{5}) + (1 - b^{(t)})(1 + \frac{b^{(t)}}{2})||e^{(t-1)}||^2 + (1 - b^{(t)})(1 + \frac{2}{b^{(t)}})(\frac{b^{(t)}\epsilon}{10} + B_{i+1} b^{(t+1)} \epsilon^{\frac{1}{p}})^2 + b^{(t)} \frac{\epsilon^2}{5}] \\
&\leq \mathbb{E}[b^{(t)} \cdot (2B^2 + \frac{2\epsilon^2}{5})] + (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[3b^{(t)}(\epsilon + B\epsilon^{\frac{1}{p}})^2 + b^{(t)} \frac{\epsilon^2}{5}] \\
&\leq (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}] \cdot (2B^2 + \frac{2\epsilon^2}{5} + 6\epsilon^2 + 6B\epsilon^{\frac{2}{p}} + \frac{\epsilon^2}{5}) \\
&= (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}] \cdot (2B^2 + \frac{33}{5}\epsilon^2 + 6B\epsilon^{\frac{2}{p}}) \\
&= (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \frac{\mathbb{E}[b^{(t)}]}{2} \cdot (4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}})
\end{aligned}
$$

which implies that

$$\begin{aligned}
&\mathbb{E}||e^{(t)}||^2 \\
&\leq (4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}}) - (2B^2 + \frac{64}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}}) \prod_{s=2}^{t}(1 - \frac{1}{2}\mathbb{E}[b^{(s)}]) \\
&\leq 4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}}
\end{aligned}$$

since $1 - \frac{1}{2}\mathbb{E}[b^{(s)}] \leq 1$ ■

Now, we prove theorem 4. We state this theorem here again for the reader's convenience.

Theorem 12 *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, biased and stochastic p-order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$ such that $B_i = O(\epsilon^{\frac{2(p-1)}{p}})$ for all $1 \leq i \leq p$, with probability at least $\frac{3}{4}$, Algorithm 3 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \leq \epsilon$ and performs at most*

$$O(\frac{\sigma^2\Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{\Delta}{\epsilon^{\frac{2p+1}{p}}})$$

queries to the stochastic and biased derivative oracles.

Proof Observe that from the proof of theorem 3 (with A_p and A'_p defined as before), we can say that

$$\begin{aligned}
&\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) \leq \frac{16(p+1)!\cdot\Delta}{M\eta^{p+1}T} + \frac{16(p+1)!\cdot A_p}{M\eta^{p+1}}\mathbb{E}[||\nabla^1 F(x) - D^1(x)||^{\frac{p+1}{p}}] \\
&+ \frac{16(p+1)!}{M\eta^{\frac{p^2-1}{p}}}A'_p \cdot \sum_{i=2}^{p}\mathbb{E}[||\nabla^i F(x) - D^i(x)||^{\frac{p+1}{p}}] \\
&\leq \frac{16(p+1)!\cdot\Delta}{M\eta^{p+1}T} + \frac{16(p+1)!\cdot A_p}{M\eta^{p+1}}\mathbb{E}[||\nabla^1 F(x) - D^1(x)||^2]^{\frac{p+1}{2p}} \\
&+ \frac{16(p+1)!}{M\eta^{\frac{p^2-1}{p}}}A'_p \cdot \sum_{i=2}^{p}\mathbb{E}[||\nabla^i F(x) - D^i(x)||^2]^{\frac{p+1}{2p}} \\
&\leq \frac{16(p+1)!\cdot\Delta}{M\eta^{p+1}T} + \frac{16(p+1)!\cdot A_p}{M\eta^{p+1}}(4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}})^{\frac{p+1}{2p}} \\
&+ \frac{16(p+1)!\cdot(p-1)}{M\eta^{\frac{p^2-1}{p}}}A'_p \cdot (4B^2 + \frac{66}{5}\epsilon^2 + 12B\epsilon^{\frac{2}{p}})^{\frac{p+1}{2p}} \\
&\leq \frac{16(p+1)!\cdot\Delta}{M\eta^{p+1}T} + \frac{16(p+1)!\cdot A_p}{M\eta^{p+1}}(4B^{\frac{p+1}{p}} + \frac{66}{5}\epsilon^{\frac{p+1}{p}} + 12B^{\frac{p+1}{2p}}\epsilon^{\frac{p+1}{p^2}}) \\
&+ \frac{16(p+1)!\cdot(p-1)A'_p}{M\eta^{\frac{p^2-1}{p}}} \cdot (4B^{\frac{p+1}{p}} + \frac{66}{5}\epsilon^{\frac{p+1}{p}} + 12B^{\frac{p+1}{2p}}\epsilon^{\frac{p+1}{p^2}}) \\
&\leq \frac{1}{8} + \frac{16(p+1)!\cdot A_p}{Mb^{p+1}}(\frac{4B^{\frac{p+1}{p}}}{\epsilon^{\frac{p+1}{p}}} + \frac{66}{5} + \frac{12B^{\frac{p+1}{2p}}}{\epsilon^{\frac{p^2-1}{p^2}}}) + \frac{16(p+1)!\cdot(p-1)A'_p}{Mb^{\frac{p^2-1}{p}}}(\frac{4B^{\frac{p+1}{p}}}{\epsilon^{\frac{p^2-1}{p^2}}} + \frac{66}{5}\epsilon^{\frac{p+1}{p^2}} + \frac{12B^{\frac{p+1}{2p}}}{\epsilon^{\frac{(p-2)(p+1)}{p^2}}})
\end{aligned}$$

Based on our setting of M, it is clear that we need $B = O(\epsilon^{\frac{2(p-1)}{p}})$ to achieve ϵ-stationarity with constant probability, and so we have that

$$\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) = \Pr(||\nabla F(\hat{x})|| \geq \frac{9Mb^p}{8p!}\epsilon) \leq \frac{3}{8}$$

which proves the first part of the theorem. Note that running the same argument with $\epsilon \leftarrow \frac{8p!}{9Mb^p}\epsilon$ will recover the ϵ upper bound exactly. Let M_i denote the total number of oracle queries for $\nabla^i F$, m_i denote the number of oracle queries in each pass, and let $M = \sum_{i=1}^{p+1} M_i$ denote the total number of oracle queries. We have that

$$\begin{aligned}
&\mathbb{E}[M] \\
&= \sum_{i=1}^{p+1} \mathbb{E}[M_i] \\
&= T\sum_{i=1}^{p+1} \Pr(C=1)\mathbb{E}[m_i|C=1] + \Pr(C=0)\mathbb{E}[m_i|C=0] \\
&= T\sum_{i=1}^{p} bn_i + (1-b)K_i \\
&\leq T\sum_{i=1}^{p} b(\frac{5\sigma_i^2}{\epsilon^2}+1) + (1-b)(\frac{5(\sigma_{i+1}^2 + L_{i+1}\epsilon)}{b\epsilon^2}+1)
\end{aligned}$$

Letting $\sigma = \max_i \sigma_i$, we can upper bound by

$$\begin{aligned}
&T\sum_{i=1}^{p} b(\frac{5\sigma^2}{\epsilon^2}+1) + (1-b)(\frac{5(\sigma^2 + L_{i+1}\epsilon)}{b\epsilon^2}+1) \\
&\leq T\sum_{i=1}^{p} \frac{5b^2\sigma^2 + 5\sigma^2 + 5L_{i+1}\epsilon}{b\epsilon^2} + 2 \\
&\leq T \cdot O(\frac{\sigma^2}{\epsilon^2} + \frac{1}{\epsilon}) \\
&\leq (\frac{2(p+1)!\Delta}{M\epsilon^{\frac{p+1}{p}}} + 1) \cdot O(\frac{\sigma^2}{\epsilon^2} + \frac{1}{\epsilon}) \\
&\leq O(\frac{\sigma^2\Delta}{\epsilon^{\frac{3p+1}{p}}} + \frac{\Delta}{\epsilon^{\frac{2p+1}{p}}})
\end{aligned}$$

which finishes the proof. ■

Appendix C. Proof of Theorem 2 and Theorem 1

We first go over some notational conventions used in this section. These are the same as in Arjevani et al. (2020), but we repeat them here for the reader's convenience. Given a p^{th} order tensor $T \in$

$\mathbb{R}^{d\times\ldots\times d}$, we define the support of T as the following:

$$\operatorname{supp}(T) = \{i \in [d], T_i \neq 0\}$$

where T_i is the $(p-1)$ order subtensor denoted by $[T_i]_{j_1,\ldots,j_{p-1}} = T_{i,j_1,\ldots,j_{p-1}}$. For a tuple of tensors $\mathcal{T} = (T^{(1)}, T^{(2)}, \ldots)$, we define

$$\operatorname{supp}(\mathcal{T}) = \bigcup_i \operatorname{supp}(T_i)$$

Moreover, given $x \in \mathbb{R}^d$, let

$$\operatorname{prog}_\alpha(x) = \max\{i \geq 0, |x_i| > \alpha\}$$

which represents the highest index of x whose entry is at least α from zero. Notice that for any $\alpha_1, \alpha_2 \in [0, 1)$ such that $\alpha_1 < \alpha_2$, we have that $\operatorname{prog}_{\alpha_2}(x) < \operatorname{prog}_{\alpha_1}(x)$. For a tensor T, we define $\operatorname{prog}(T) = \max\{\operatorname{supp}\{T\}\}$ which represents the highest index in $\operatorname{supp}\{T\}$, and naturally for a collection of tensors $\mathcal{T} = \{T^{(i)}\}$, we define $\operatorname{prog}(\mathcal{T}) = \max_i \operatorname{prog}(T^{(i)})$.

Definition 13 *A biased and stochastic algorithm A is zero-respecting if for any function F and pth-order oracle O_F^p, the iterates $\{x^{(t)}\}$ satisfy*

$$\operatorname{supp}(x^{(t)}) \subseteq \bigcup_{i<t} \operatorname{supp}(O_F^p(x^{(i)}, z^{(i)}, b^{(i)}))$$

for all $t \in \mathbb{N}$.

Definition 14 *A collection of derivative estimators $\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)$ for a function F form a probability-ρ zero-chain if*

$$\Pr(\exists x | \operatorname{prog}(\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)) = \operatorname{prog}_{\frac{1}{4}}(x) + 1) \leq \rho$$

and

$$\Pr(\exists x | \operatorname{prog}(\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)) = \operatorname{prog}_{\frac{1}{4}}(x) + i) = 0$$

for all $i > 1$.

Lemma 15 *Let $\tilde{\nabla}^1 F(x, z, b), \ldots, \tilde{\nabla}^p F(x, z, b)$ be a collection of probability-ρ zero-chain derivative estimators for $F : \mathbb{R}^T \to \mathbb{R}$, and let $O_F^p(x, z, b) = (\tilde{\nabla}^q F(x, z, b))_{q\in\{1,\ldots,p\}}$. Let $\{x^{(t)}_{A[O_F]}\}$ be a sequence of queries produced by algorithm A interacting with O_F^p. Then, with probability at least $1 - \delta$,*

$$\operatorname{prog}(x^{(t)}) < T$$

for all

$$t \leq \frac{T - \log(1/\delta)}{2\rho}$$

Proof Proved in (Arjevani et al., 2020). ■

Define

$$F_T(x) = -\Psi(1)\Phi(1) + \sum_{i=2}^{T} \Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)]$$

where

$$\Psi(x) = 0, x \leq 1/2$$
$$\Psi(x) = \exp(1 - \frac{1}{(2x-1)^2}), x > \frac{1}{2}$$

and

$$\Phi(x) = \sqrt{e} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

Lemma 16 *For F_T, the following properties hold:*

- $F_T(0) - \inf_x F_T(x) \leq \Delta_0 T$, *where* $\Delta_0 = 12$
- *For all* $p \geq 1$, *the* p^{th} *order derivatives of* F_t *are* ℓ_p*-Lipschitz continuous, where* $\ell_p \leq \exp(\frac{5}{2}p \log p + cp)$ *for some* $c < \infty$
- *For all* $x \in \mathbb{R}^T$, $p \in \mathbb{N}$, *and* $1 \leq i \leq T$, *we have that* $||\nabla_i^p F_T(x)||_{\text{op}} \leq \ell_{p-1}$
- *For all* $x \in \mathbb{R}^T$, $p \in \mathbb{N}$, $\text{prog}(\nabla^q F_T(x)) \leq \text{prog}_{\frac{1}{2}}(x) + 1$
- *For all* $x \in \mathbb{R}^T$, *if* $\text{prog}_1(x) < T$, *then* $||\nabla F_T(x)|| \geq |\nabla_{\text{prog}_1(x)+1} F_T(x)| > 1$

Proof Follows from (Carmon et al., 2019a) and (Arjevani et al., 2020) ■

Define the derivative estimators used to be

$$[\tilde{\nabla}^q F_T(x,z)]_i = (1 + \mathbf{1}\{i > \text{prog}_{\frac{1}{4}}(x)\}(\frac{z}{\rho} - 1)) \cdot (\nabla_i^q F_T(x) + b_i^q(x))$$

where $b_i^q(x) = 0$ for all $i > \text{prog}_{1/4}(x) + 1$, $||b_i^q(x)|| \leq B_q$, and $z \sim$ Bernoulli(ρ).

Lemma 17 *The derivative estimators $\tilde{\nabla}^q F_T$ form a probability-ρ zero-chain and satisfy:*

$$\mathbb{E}[||\tilde{\nabla}^q F_T(x,z) - \nabla^q F_T(x)||^2] \leq \frac{2\ell_{q-1}^2(1-\rho)}{\rho} + 2B_q^2$$

Proof First, we prove that these derivative estimators form a probability-ρ chain. First, by the definition of F_T and b_i, we can immediately conclude that $[\tilde{\nabla}^q F_T(x,z)]_i = 0$ for all $i > \text{prog}_{\frac{1}{4}}(x) + 1$. Now, when $i = \text{prog}_{\frac{1}{4}}(x) + 1$, we have that $[\tilde{\nabla}^q F_T(x,z)]_i = \frac{z}{p} \cdot (\nabla_i^q F_T(x) + b_i(x))$. if $z = 0$

(with probability $1-\rho$), then we have that $[\tilde{\nabla}^q F_T(x,z)]_i = 0$. So, the first condition also follows. Now, we can say that

$$\begin{aligned}
&\mathbb{E}[||\tilde{\nabla}^q F_T(x,z) - \nabla^q F_T(x)||^2] \\
&\leq 2\mathbb{E}[||\tilde{\nabla}^q F_T(x,z) - \overline{\nabla}^q F_T(x,z)||^2] + 2\mathbb{E}[||\overline{\nabla}^q F_T(x,z) - \nabla^q F_T(x,z)||^2] \\
&\leq 2||b_q(x)||^2 + \frac{2\ell_{q-1}^2(1-\rho)}{\rho} \\
&\leq 2B_q^2 + \frac{2\ell_{q-1}^2(1-\rho)}{\rho}
\end{aligned}$$

where $\overline{\nabla}$ is an unbiased but stochastic derivative estimator. ■

Now, we prove theorem 2. We let $F_T^* = \alpha F_T(\beta x)$ for some constants α, β. With probability at least $\frac{3}{4}$, we have that $\text{prog}(x^{(t)}_{A[O_F^p]}) < T$ for all $t \leq \frac{T-2}{2\rho}$. Since $\text{prog}_1(x) \leq \text{prog}(x)$, we have that

$$\mathbb{E}||\nabla F_T^*(x^{(t)}_{A[O_F^p]})|| = \alpha\beta\mathbb{E}||\nabla F_T(x^{(t)}_{A[O_F^p]})|| \geq \frac{\alpha\beta}{2}$$

and that

$$\mathbb{E}||\tilde{\nabla}^q F_T^*(x,z) - \nabla^q F_T^*(x,z)||^2 \leq \alpha^2\beta^{2q}(\frac{2\ell_{q-1}^2(1-\rho)}{\rho} + 2B_q^2)$$

We now set constants such that

- $\alpha\Delta_0 T \leq \Delta$
- $\alpha\beta^{q+1}\ell_q \leq L_q$
- $\frac{\alpha\beta}{2} \geq \epsilon$
- $\alpha^2\beta^{2q}(\frac{2\ell_{q-1}^2(1-\rho)}{\rho} + 2B_q^2) \leq 2\sigma_q^2 + 2B_q^2 \implies \alpha^2\beta^{2q}(\frac{\ell_{q-1}^2(1-\rho)}{\rho} + B_q^2) \leq \sigma_q^2 + B_q^2$

First, we let $\alpha = 2\epsilon/\beta$. Now, we set

$$\rho = \min\{\frac{\alpha^2\beta^2\ell_0^2}{\sigma_1^2 + B_1^2 - \alpha^2\beta^2 B_1^2}, 1\}$$

which means that we can say that

$$\begin{aligned}
&\alpha^2\beta^{2q}(\frac{\ell_{q-1}^2(1-\rho)}{\rho} + B_q^2) \\
&\leq \alpha^2\beta^{2q}(\frac{\ell_{q-1}^2}{\rho} + B_q^2) \\
&\leq \alpha^2\beta^{2q}(\frac{\ell_{q-1}^2(\sigma_1^2 + B_1^2 - \alpha^2\beta^2 B_1^2)}{\alpha^2\beta^2\ell_0^2} + B_q^2) \\
&\leq \frac{\beta^{2(q-1)}\ell_{q-1}^2(\sigma_1^2 + B_1^2 - \alpha^2\beta^2 B_1^2)}{\ell_0^2} + \alpha^2\beta^{2q}B_q^2 \\
&= \frac{\beta^{2(q-1)}\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4\epsilon^2 B_1^2)}{\ell_0^2} + \alpha^2\beta^{2q}B_q^2
\end{aligned}$$

We now set

$$\frac{\beta^{2(q-1)}\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1^2)}{\ell_0^2}+\alpha^2\beta^{2q}B_q^2\leq\sigma_q^2+B_q^2$$

We solve for β such that

$$\frac{\beta^{2(q-1)}\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1^2)}{\ell_0^2}\leq\frac{\sigma_q^2+B_q^2}{2}$$

and

$$\alpha^2\beta^{2q}B_q^2\leq\frac{\sigma_q^2+B_q^2}{2}$$

From these two conditions and the Lipschitz condition, we set

$$\beta=\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{\sigma_q^2+B_q^2}{8\epsilon^2B_q^2})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}$$

which implies that we can set

$$\beta=\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}$$

and we also set

$$T=\lfloor\frac{\Delta}{\alpha\Delta_0}\rfloor=\lfloor\frac{\Delta\beta}{2\Delta_0\epsilon}\rfloor$$

which means that we have that (assuming $\epsilon\leq\sqrt{2}/4$ and $T\geq5$)

$$\begin{aligned}
&\frac{T-2}{2\rho}=\frac{1}{2\rho}(\lfloor\frac{\Delta\beta}{2\Delta_0\epsilon}\rfloor-2)\\
&\geq\frac{1}{2\rho}\cdot\frac{\Delta\beta}{4\Delta_0\epsilon}\\
&\geq\frac{2(\sigma_1^2+B_1^2-4\epsilon^2B_1^2)}{\alpha^2\beta^2\ell_0^2}\cdot\frac{\Delta}{4\Delta_0\epsilon}\cdot\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}\\
&\geq\frac{2(\sigma_1^2+B_1^2-4\epsilon^2B_1^2)}{4\epsilon^2\ell_0^2}\cdot\frac{\Delta}{4\Delta_0\epsilon}\cdot\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}\\
&\geq\frac{(\sigma_1^2+B_1^2-4\epsilon^2B_1^2)\Delta}{8\epsilon^3\ell_0^2\Delta_0}\cdot\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}\\
&\geq\frac{(\sigma_1^2+0.5B_1^2)\Delta}{8\epsilon^3\ell_0^2\Delta_0}\cdot\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4\epsilon^2B_1)})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2\epsilon\ell_{q'}})^{\frac{1}{q'}}\}
\end{aligned}$$

We now prove the first order lower bound result, with the corresponding upper bound presented in Ajalloeian and Stich (2020). The analysis follows in similar vein to that presented above, but we go through it again for completeness.

Remark 18 *Our oracle model is the same as in Ajalloeian and Stich (2020), when setting* $M = m = 0$, $\sigma = \sigma_1$, $\zeta = B_1$, *and finding a point* x *where* $||\nabla F(x)|| = O((\epsilon + B_1^2)^{1/2})$

Proof The setting of constants M, m, σ, ζ follows from definition 1, assumption 3, and assumption 4 in Ajalloeian and Stich (2020). The last point follows from the fact that in theorem 4 of Ajalloeian and Stich (2020), the goal was to have iterates $\{x_t\}$ such that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(x^{(t)})||^2 = O(\epsilon + B_1^2)$$

which we can equivalently express as

$$||\nabla F(\hat{x})|| = O((\epsilon + B_1^2)^{\frac{1}{2}})$$

where $\hat{x}$ is drawn uniformly from $\{x_t\}$. ■

We now begin to prove theorem 1. Using the same function F_T as earlier and letting $F_T^* = \alpha F_T(\beta x)$, we set constants such that

- $\alpha\Delta_0 T \leq \Delta$
- $\alpha\beta^2\ell_1 \leq L_1$
- $\frac{\alpha\beta}{2} \geq (\epsilon + B_1^2)^{\frac{1}{2}}$
- $\alpha^2\beta^2(\frac{\ell_0^2(1-\rho)}{\rho} + B_1^2) \leq \sigma_1^2 + B_1^2$

Carrying out similar algebraic procedures as before, we set

- $\alpha = 2(\epsilon + B_1^2)^{\frac{1}{2}}/\beta$
- $\rho = \frac{4(\epsilon+B_1^2)\ell_0^2}{\sigma_1^2+B_1^2-4(\epsilon+B_1^2)B_1^2}$
- $\beta = \frac{L_1}{2(\epsilon+B_1^2)^{\frac{1}{2}}\ell_1}$
- $T = \lfloor \frac{\Delta\beta}{2\Delta_0(\epsilon+B_1^2)^{\frac{1}{2}}} \rfloor$

We have that with probability at least $\frac{3}{4}$ (assuming $\epsilon < \frac{1}{4}$ and $T \geq 5$),

$$\begin{aligned}
&\frac{T-2}{2\rho} = \frac{1}{2\rho}(\lfloor \frac{\Delta\beta}{2\Delta_0(\epsilon+B_1^2)^{\frac{1}{2}}} - 2 \rfloor) \\
&\geq \frac{1}{2\rho}\cdot\frac{\Delta\beta}{4\Delta_0(\epsilon+B_1)^{\frac{1}{2}}} \\
&\geq \frac{\sigma_1^2+B_1^2-4(\epsilon+B_1^2)B_1^2}{8(\epsilon+B_1^2)\ell_0^2}\cdot\frac{\Delta}{4\Delta_0(\epsilon+B_1^2)^{\frac{1}{2}}}\cdot\frac{L_1}{2(\epsilon+B_1^2)^{\frac{1}{2}}\ell_1} \\
&= \frac{\Delta L_1(\sigma_1^2+(1-4\epsilon)B_1^2-4B_1^4)}{64\Delta_0\ell_0^2\ell_1(\epsilon+B_1^2)^2}
\end{aligned}$$

Now, since $\epsilon < \frac{1}{4}$, we can continue to lower bound this expression as follows:

$$\begin{aligned}
&\frac{\Delta L_1(\sigma_1^2 - 4B_1^4)}{64\Delta_0\ell_0^2\ell_1(\epsilon + B_1^2)^2} \\
&\geq \frac{\Delta L_1\sigma_1^2}{64\Delta_0\ell_0^2\ell_1(\epsilon + B_1^2)^2} - \frac{4\Delta L_1}{64\Delta_0\ell_0^2\ell_1} \\
&\geq O(\frac{\Delta L_1\sigma_1^2}{\Delta_0\ell_0^2\ell_1(\epsilon + B_1^2)^2} \\
&= O(\frac{\Delta L_1\sigma_1^2}{\Delta_0\ell_0^2\ell_1(\epsilon^2 + B_1^4)})
\end{aligned}$$

which finishes the proof.