On the Limits of Biased Derivative Information for Nonconvex Stochastic Optimization

Anant Shyam
Department of Computer Science
Purdue University
West Lafayette, IN 47907
shyama@purdue.edu

Brian Bullins
Department of Computer Science
Purdue University
West Lafayette, IN 47907
bbullins@purdue.edu

Abstract

We consider the problem of finding δ-stationary points, i.e., $x \in \mathbb{R}^d$ 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 tight lower bounds for finding an $O((\epsilon + B^2)^{1/2})$-stationary point, where B is a bound on the gradient bias, matching the upper bounds of Ajalloeian and Stich (2020). We then establish bias-dependent lower bounds for algorithms that use higher-order derivative information for finding $O(\epsilon + B)$-stationary points, where B is a bound on the maximum bias for all derivatives. To complement these lower bounds, we develop trust-region based methods that, for certain ranges of bias, provide guarantees that match the corresponding lower bounds. We further improve upon the oracle complexity in high bias settings through a higher order variance reduction scheme, in particular demonstrating the benefits, in some cases, of using higher-order derivative information, while such improvements are known to be unattainable for unbiased settings.

1 Introduction

For a smooth function $F : \mathbb{R}^d \to \mathbb{R}$ which has Lipschitz continuous derivatives, and has bounded suboptimality Δ such that $F(0) - \inf_{x \in \mathbb{R}^d} F(x) \leq \Delta$, 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 stationary points is a task that has been explored in numerous previous works (e.g., [10, 11]) 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 [23]. Therefore, we naturally turn to finding local optima whose gradient norm is sufficiently small. Moreover, just like [23], 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 has been a collection of work that has studied the oracle complexity of finding ϵ-stationary points (i.e $\mathbb{E}||\nabla F(x)|| \leq \epsilon$). In [17], 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 [16] to $O(\epsilon^{-3.5})$, with the additional assumption that the stochastic gradient

Preprint.

$\nabla F(x,\xi)$ was L-smooth. One can also further improve this complexity bound to $O(\epsilon^{-3})$ if the noisy gradient satisfies a mean-squared smoothness property [6].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 [25]. There has also been several works which employ variance reduction (e.g. [15, 26]), 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 are part of a broader literature that assume access to stochastic and unbiased derivative oracles, where for all derivatives $i = 1, \ldots, p$, we have that

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

for some set of variance parameters $\sigma_1, \ldots, \sigma_p$ and where the noise ξ is drawn from some distribution P_ξ.

However, in many settings it becomes necessary to relax these unbiased derivative assumptions, for example in distributed [14, 4, 5], gradient-free [24], and bandit convex optimization [18], thus highlighting the need to better understand the limits of working with biased derivative information.

In this paper, we establish the limits of biased derivative information for both first and higher-order settings. We first complement the first-order upper bound in [3] by providing a matching lower bound. We then show how to handle biased and stochastic *high-order* derivative information, through providing corresponding higher order lower bounds and developing higher order trust-region and variance-reduction based algorithms to complement these lower bounds. We show that, unlike in the *unbiased* case [7], appealing to higher order information beyond second-order information *can* offer benefits for certain ranges of bias. In particular, we consider the following oracle model, where for all derivatives $i = 1, \ldots, p$, we have that

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

where

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

for some set of variance parameters $\sigma_1, \ldots, \sigma_p$ and bias parameters $B_1, \ldots, B_p$.

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 the main contributions of this paper.

First-Order Lower Bound. In [3], the authors derived the following upper bound for the oracle complexity for finding iterates $\{x_t\}$ such that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla F(x^{(t)})||^2 = O(\epsilon + B_1^2)$:

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

which is 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 matching lower bound that matches the provided upper bound up to constant factors to demonstrate that the upper bound is tight. This is shown in Theorem 1.

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 $\epsilon + \max_i B_i$-stationary points:

$$\Omega(1) \cdot \frac{(\sigma_1^2 - 4(\epsilon+B)^2 B_1^2)\Delta}{32(\epsilon+B)^3 \ell_0^2 \Delta_0} \cdot \min_{q' \in \{1,\ldots,p\}, q \in \{2,\ldots,p\}} \min\{(\frac{\ell_0^2(\sigma_q^2 + B_q^2)}{2\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4(\epsilon+B)^2 B_1^2)})^{\frac{1}{2(q-1)}}, (\frac{\sigma_q^2 + B_q^2}{8(\epsilon+B)^2 B_q^2})^{\frac{1}{2(q-1)}}, (\frac{L_{q'}}{2(\epsilon+B)\ell_{q'}})^{\frac{1}{q'}}\}$$

where $B = \max_i B_i$. This is shown in Theorem 2. To check if these lower bounds were tight, we developed algorithms to try and match these lower bounds as closely as possible.

Minibatch Derivative 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:

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

for some $M \geq 8L_p$ where L_p is the Lipschitz constant of $\nabla^p F$. We found that as $p \to \infty$, the oracle complexity for finding an $O(\epsilon + \max_i B_i)$ stationary point worsens, so we run the above scheme for $p = 2$. See Theorem 3 for a more detailed description.

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. We develop an improved oracle complexity bound for finding $O(\epsilon + \max_i B_i)$ stationary points in the constant bias setting, as well as provide a bound for finding a $O((\epsilon^2 + (\max_i B_i)^2)^{\frac{1}{2}}(\epsilon + \max_i B_i))$ stationary point for the high bias setting. Furthermore, unlike the previous setting, we do see benefits for appealing to higher order derivative information for the constant bias setting. See Theorem 4 for a more detailed description.

1.2 Additional Related Works

Biased Gradient Methods. Here, we briefly discuss some prior work that has been done relating to biased gradient methods. In [19], the authors proposed a biased SGD algorithm and analyzed the sample complexities for convex and nonconvex objectives. In [8], the authors explore a balance between biased and unbiased estimation of the gradient to resolve the tradeoff between the cost and benefit of computing an unbiased derivative estimation. In [13, 21], the authors also present algorithms under biased gradient estimation. In [12], the authors present an analysis of biased SGD is convex and nonconvex settings, under weaker assumptions than many previous works in this area.

Deterministic Oracles. 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 [9] 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 is $\Omega(\epsilon^{-\frac{12}{7}})$ [11]. Furthermore, the authors highlight how this query complexity can be improved by appealing to higher order information. 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 [2], the authors present a method that uses a cubic regularized Newton step 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 introduce our problem setup in section 2, including the function class and derivative oracle properties. In section 3, we present the lower bounds for both the first and higher order settings. In section 4, we present algorithms that use minibatch-based derivative estimation (1) and variance reduction based derivative estimation (3). We conclude the paper in section 5. We prove Theorem 3 in appendix 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 $||T||_{\text{op}} = \sup_{||v||=1} |\langle T, v, \ldots, v\rangle|$. We also let $B = \max_{1\le i\le p} B_i$.

2 Setup and Background

2.1 Function Class

We consider smooth, differentiable functions in the following function class:

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

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

$$F(0) - \inf_{x\in\mathbb{R}^d} F(x) \le \Delta$$

2.2 Oracles

For such a function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, we consider a class of biased and stochastic derivative oracles for $\nabla^1 F, \ldots, \nabla^p F$ 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 $\tilde{\nabla}^q F(x, z, b_q)$ is a biased and stochastic estimate for $\nabla^q F(x)$. 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 $\mathbb{E}_{z\sim P_z}[\xi_q(x, z)] = 0$, $\mathbb{E}||\xi_q(x, z)||^2 \le \sigma_q^2$, and $||b_q(x)|| \le B_q$. 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 [3], 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 matching lower bound that matches the aforementioned upper bound.

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 (definition 1) where $\epsilon < \frac{1}{4}$ and $B_1 \le O(1)$, the minimum number of queries to obtain a $(\epsilon + B_1^2)^{\frac{1}{2}}$ stationary point with constant probability is bounded below by*

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

Given the matching upper and lower bounds in the first order setting, one natural question was whether one could observe analogous lower and upper bounds in higher order settings. In Theorem 2, we derive a lower bound for finding an $O(\epsilon + B)$ stationary point using derivatives $1, \ldots, p$ for $B \le \frac{\sqrt{3}}{2}\sigma_1$. The function F that we will use is a rescaling of the function F_T defined on line 1. In Lemma 10, we prove that for all $x, y \in \mathbb{R}^d$, there exists a constant $C \ge 0$ such that $||\nabla F_T(x)|| \le C\sqrt{T}$. Since we scale this function as the following $F_T^*(x) = \alpha F_T(\beta x)$ for constants α, β, we have that $||\nabla F_T(x)|| \le C_1\sqrt{T}$ for a different C_1. For $B \ge C_1\sqrt{T} - \epsilon$, we need zero oracle queries to reach such the given stationarity condition, as for any $x \in \mathbb{R}^d$, we have that $||\nabla F(x)|| \le C_1\sqrt{T}$, so the lower bound is vacuous for this scenario.

Theorem 2. *For all $p \geq 2$, Δ, $L_{1:p}$, $\sigma_{1:p} > 0$, $\epsilon < \sqrt{\sigma_1}$, and $B \leq \frac{\sqrt{3}}{2}\sigma_1$, 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 p^{th} order zero-respecting algorithm (definition 1), the number of queries to obtain a point an $\epsilon + \max_i B_i$ stationary point with constant probability is bounded below by*

$$\Omega(1) \cdot \frac{(\sigma_1^2 - 4(\epsilon+B)^2 B_1^2)\Delta}{32(\epsilon+B)^3 \ell_0^2 \Delta_0} \cdot$$
$$\min_{q' \in \{1,\dots,p\}, q \in \{2,\dots,p\}} \min\{(\frac{\ell_0^2(\sigma_q^2 + B_q^2)}{2\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4(\epsilon+B)^2 B_1^2)})^{\frac{1}{2(q-1)}}, (\frac{\sigma_q^2 + B_q^2}{8(\epsilon+B)^2 B_q^2})^{\frac{1}{2(q-1)}}, (\frac{L_{q'}}{2(\epsilon+B)\ell_{q'}})^{\frac{1}{q'}}\}$$

Proof Sketch. Here, we provide a brief proof sketch, deferring the full proofs to appendix C. We outline the proof of Theorem 2, as the proof of Theorem 1 will follow a similar argument to this one. We first introduce the following "hard" function [10] where for a fixed $T \geq 0$ and $x \in \mathbb{R}^T$:

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

where

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

and we let ℓ_p represent the Lipschitz constant of the p^{th} derivative of F_T. We then show how to construct the following series of derivative estimators for this function (for all derivatives $i \in \{1, \dots, p\}$):

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

which crucially account for the fact that the oracles are not only stochastic *but also biased*. If T is a k-dimensional tensor, then T_i is a $k-1$-dimensional subtensor where $[T_i]_{j_1,\dots,j_{k-1}} = T_{i,j_1,\dots,j_k}$ and $\text{prog}_\alpha(x) = \max\{i \geq 0, |x_i| > \alpha\}$, representing the highest index of $x \in \mathbb{R}^d$ that is at least α away from zero. We carefully devised the above construction such that this collection of derivative estimators formed a probability-ρ zero-chain (for some $0 \leq \rho \leq 1$) where the following properties hold:

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

for all $i > 1$. Through this zero-chain construction (as is well studied in [1, 10, 11, 6]), we enforce that every oracle query can reveal information about at most one new coordinate, thereby requiring any algorithm to make sequential progress, which yields a lower bound on the number of queries to make sufficient progress (see appendix C for a more formal explanation). Given that we have shown that $\{\tilde{\nabla}^q F_T(x, z, b)\}$ form a probability-ρ zero-chain, we used a scaled version of F_T, parametrized by two constants α and β:

$$F_T^*(x) = \alpha F_T(\beta x)$$

where we solved for α and β to enforce the required suboptimality, noise, bias, and higher order Lipschitz conditions. □

4 Upper Bounds

4.1 Minibatch-Based Derivative Estimation

Given these lower bounds, we develop various algorithms to how the upper bounds compared. In Algorithm 1, we minimize a regularized second order model of our objective F at each iteration. In Theorem 3, we prove an upper bound for the oracle complexity of Algorithm 1 for reaching an $O(\epsilon + B)$ stationary point.

Algorithm 1 Biased and **S**tochastic **C**ubic-**R**egularized **T**rust **R**egion

Require: Biased and stochastic oracle $(O_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$ for $F \in \mathcal{F}_p(\Delta, L_{0:p})$, 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: $M \leftarrow 8L_p$, $\eta \leftarrow \min\{(\epsilon + \max_i B_i)^{\frac{1}{p}}, 1 - \epsilon\}$

3: $A \leftarrow \frac{16(p+1)!}{M} \max\{1, \frac{(p!)^{\frac{p+1}{p}}}{4M^{\frac{1}{p}}} + 2(\frac{2p!}{M})^{\frac{1}{p}}, \frac{(p!)^{\frac{p+1}{p}}}{4M^{\frac{1}{p}}} + 2(\frac{2p \cdot p!}{M})^{\frac{1}{p}}\}$

4: $T \leftarrow \lceil \frac{8A\Delta}{\eta^{p+1}} \rceil$

5: Pick n_1 such that

$$\max\{\frac{C_1 \cdot \sigma_1^2}{(\frac{\eta^{p+1}}{8A} - B_1^{\frac{p+1}{2p}})^{\frac{2p}{p+1}}}, 1\} \leq n_1 \leq \frac{(\epsilon + \max_i B_i)^{\frac{p+1}{p}} (\max_i \sigma_i)^2}{\epsilon^3 (1 + \max_i B_i)^{\frac{p+1}{p}}}$$

6: For all $2 \leq i \leq p$, pick n_i such that

$$\max\{\frac{C_i \sigma_i^2}{(\frac{\eta^{\frac{p^2-1}{p}}}{8Ap} - B_i^{\frac{p+1}{2p}})^{\frac{2p}{p+1}}}, 1\} \leq n_i \leq \frac{(\epsilon + \max_i B_i)^{\frac{p+1}{p}} (\max_i \sigma_i)^2}{\epsilon^3 (1 + \max_i B_i)^{\frac{p+1}{p}}}$$

7: **for** $t = 0$ to $T - 1$ **do**

8: **for** $i = 1$ to p **do**

9: Query the i^{th} order oracle n_i times at $x^{(t)}$ and compute

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

10: **end for**

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

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

12: **end for**

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

Theorem 3. *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, where $p \geq 2$, $\epsilon > 0$, with biased and stochastic p^{th}-order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$ where $B \geq \Omega(\epsilon^{\frac{3p}{3p+1}})$, with probability at least $\frac{5}{8}$, Algorithm 1 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \leq O(\epsilon + B)$ and performs at most*

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

queries to the stochastic and biased derivative oracles.

Here, the fact that we require $B \geq \Omega(\epsilon^{\frac{3p}{3p+1}})$ is due to requirements of the algorithm in terms of batch size and that the existence of constants $\{C_i\}$ is guaranteed by Lemma 1. Note in the unbiased case (i.e $B = 0$), we recover the $O(\epsilon^{-3})$ guarantee known from [7]. Moreover, as $p \to \infty$, the oracle complexity worsens, thereby implying that using derivative information beyond the second order does not help in this scenario. Therefore, in the minibatch derivative estimation setting, setting $p = 2$ yields optimal oracle complexity as compared to any $p > 2$. Interestingly, however, when using variance reduction, we do realize benefits to appealing to higher order derivative information for certain ranges of bias.

Proof Sketch. Here, we provide a brief proof sketch of Theorem 3, deferring the full proof to appendix A. In Lemma 1, we prove that there do exist constants C_i that satisfy the condition on line 1 of Algorithm 1. We then derive a lower bound for $F(x) - F(y)$ in Lemma 3 for all $M \geq 8L_p$ and $0 \leq \eta < 1$, where $x \in \mathbb{R}^d$ and $y \in \text{argmin}_{z:||z-x|| \leq \eta} m_x(z)$, where m_x represents the regularized p^{th} order model of F around x using biased and stochastic derivatives:

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

We then extend this Lemma to the case where $D^{(i)}$ are random variables in Lemma 5, and then (through using an intermediary Lemma 6)) prove an upper bound of $\frac{3}{8}$ on the probability of reaching a point $\hat{x}$ such that $\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}(\epsilon + B))$, which completes the proof. □

Below, we also provide the comparison to the lower bound in Theorem 2 in terms of the oracle complexity.

Bias Regime	Lower Bound (Theorem 2)	Upper Bound (Theorem 3)
$B \leq O(\epsilon)$	$\Omega(\epsilon^{-3})$	$O(\epsilon^{-3})$
$B = \Theta(1)$	$\Omega(1)$	$O(\epsilon^{-3})$
$B = \Theta(\epsilon^{-q}), q > 0$	Trivial	$O(\epsilon^{\frac{q(p+1)-3p}{p}} + \epsilon^{-3})$

This demonstrates that the lower bound in Theorem 2 is tight for $B \leq O(\epsilon)$. We work on improving these bounds through a variance reduction scheme (Algorithm 3), succeed in improving upon the upper bound for the $B = \Theta(1)$ case to $O(\epsilon^{-2})$, and leave further improvements for future work. We note that if one knew apriori that $B = \Theta(\epsilon^{-q})$ for $q > 0$, then one can directly return the starting iterate $x^{(0)}$, so in theory, one could match the lower bound for this setting as well. We do not assume such knowledge in our algorithm.

4.2 Variance Reduction

Given the numerous prior 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. [15]) to compute cheap estimators of the gradient $\nabla F(x^{(t)})$. In our implementation of recursive variance reduction, we build on that of [7] 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 a variance-reduction based algorithm (Algorithm 2) for finding an $O(\epsilon + B)$ stationary point and a $O((\epsilon^2 + B^2)^{\frac{1}{2}}(\epsilon + B))$ stationary point.

Theorem 4. *For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, with biased and stochastic p^{th} order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$, with probability at least $\frac{5}{8}$, Algorithm 3 returns a point $\hat{x}$ such that:*

- *If $B = \Theta(1)$, then $||\nabla F(\hat{x})|| \leq O(\epsilon + B)$ with at most*

$$O(\frac{\Delta(\max_i \sigma_i)^2(\epsilon + B)^{\frac{1}{p}} + (\max_i \sigma_i)^2}{\epsilon^2} + \frac{\Delta(\epsilon + B)^{\frac{1}{p}} + 1}{\epsilon})$$

queries to the stochastic and biased derivative oracles.

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

Require: 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

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

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

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

end if

return $D^{(i)}$

- *If $B > \Omega(1)$, then $||\nabla F(\hat{x})|| \leq O((\epsilon^2 + B^2)^{\frac{1}{2}}(\epsilon + B))$ with at most*

$$O(\frac{(\max_i \sigma_i)^2}{\epsilon^2(\epsilon + B)^{\frac{p+1}{p}}} + \frac{1}{\epsilon(\epsilon + B)^{\frac{p+1}{p}}})$$

queries to the stochastic and biased derivative oracles.

Proof Sketch. Here, we provide a brief proof sketch, deferring the full proof to appendix B. In Lemma 7, we prove that

$$\mathbb{E}||D^{(i)}(x^{(t)}) - \nabla^i F(x^{(t)})||^2 \leq 4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{\frac{2}{p}} + 18B^2$$

thereby establishing a bound on the difference in the derivative estimate versus the true derivative for all derivative orders. We then derive an upper bound for $\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p)$ in terms of B, ϵ, and the hyperparameters of Algorithm 3 in Lemma 8. Plugging the parameters in gives an upper bound of $\frac{3}{8}$ for $\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p)$, which finishes the proof. □

Notably, from Theorem 4, one can observe the following important conclusions:

- Unlike the purely minibatch-based approach for derivative estimation (utilized in Algorithm 1), appealing to higher order information does provide advantages in the event where $B = \Theta(1)$.
- In the $B = \Theta(1)$ setting, we have a significantly improved oracle complexity of $O(\epsilon^{-2})$ compared to the $O(\epsilon^{-3})$ complexity from Algorithm 1.

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 .

Pick b such that $0 < b \leq 1$, let $B = \max_i B_i$
Let $X = 4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{\frac{2}{p}} + 18B^{\frac{2}{p}}$
Set $\eta = \min\{(\epsilon + B)^{\frac{1}{p}}, 1 - \epsilon\}$
Let $A = \max(16(p+1)!, 2(p+1)! \cdot [(p!)^{\frac{p+1}{p}} + 8 \cdot (2p!)^{\frac{1}{p}}], 2(p+1)! \cdot [(p!)^{\frac{p+1}{p}} + 4(2p \cdot p!)^{\frac{1}{p}}])$
Set $M = \max\{(\frac{8AX^{\frac{p+1}{2p}}}{\eta^{p+1}})^{\frac{p}{p+1}}, (\frac{8Ap \cdot X^{\frac{p+1}{2p}}}{\eta^{\frac{p^2-1}{p}}})^{\frac{p}{p+1}}, (\epsilon + B)^{\frac{-p-2}{p}}, 8L_p)$
Set $T = \lceil \frac{8A\Delta}{M\eta^{p+1}} \rceil$
Set $x^{(0)} = x^{(1)} = 0$, $D^{(i)} =$ None for $i \in \{1, \ldots, p\}$
for $t = 1$ to T **do**
    $D_t^{(i)} =$ **HO-RVR**(ϵ, b, $x^{(t)}$, $x^{(t-1)}$, $D_{t-1}^{(i)}$)
    Set the next point $x^{(t+1)}$ as

$$x^{(t+1)} = \operatorname*{argmin}_{y:||y-x^{(t)}|| \leq \eta} \sum_{i=1}^{p} \frac{1}{i!} D_t^{(i)}[y - x^{(t)}]^i + \frac{M}{(p+1)!}||y - x^{(t)}||^{p+1}$$

end for
return $\hat{x}$ chosen uniformly at random from $\{x^{(t)}\}_{t=2}^{T+1}$

- Consider the scenario where $B > \Omega(1)$. Unlike the previous scenario, appealing to higher order derivatives does not yield a better oracle complexity. Our intuition is to why this is the case, is if $B > \Omega(1)$, the stationary guarantee is quite weak, implying that there may be little to no difference between appealing to higher order information and not doing so in terms of oracle complexity. Moreover, for the $p = 2$ case, we have an improved oracle complexity from the variance reduction based scheme in Algorithm 2 compared to Algorithm 1 for all settings of $\{B_i\}$ such that $B > \Omega(1)$.

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 matching first order lower bound to complement the first order upper bound that is provided in this stochastic and biased scenario in [3]. We further extend this lower bound for algorithms that use second order derivative information or higher for finding $O(\epsilon + B)$ stationary points. Then, to complement these lower bounds, we developed trust region based methods, that under certain bias regimes, matches the corresponding lower bound up to constant factors. We then improved upon these algorithms by incorporating a higher order variance reduction scheme, which improves the oracle complexity for certain ranges of bias, and in some cases, reveals advantages of appealing to higher order derivative information.

With regards to opportunities for future work, our upper bound in Theorem 3 only matched the corresponding higher order lower bound in Theorem 2 for $O(\epsilon)$ bias, which leaves open the possibility of a stronger upper bound. Moreover, it would be interesting to consider additional cases where appealing to higher order derivative information would be beneficial for algorithms relying on biased and stochastic oracle access.

LLM Usage. LLMs were used as an assistive tool for checking the algebraic steps of our proofs. All algorithms, theorem statements, and final proofs were developed and verified by the authors.

References

[1] Deeksha Adil, Brian Bullins, Arun Jambulapati, and Aaron Sidford. Convex optimization with p-norm oracles. In *37th International Conference on Algorithmic Learning Theory*, 2026. URL `https://openreview.net/forum?id=gZax9QXNyR`.

[2] 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.

[3] 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.

[4] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 440–445, 2017.

[5] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[6] 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.

[7] 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.

[8] Jia Bi and Steve R. Gunn. A stochastic gradient method with biased estimation for faster nonconvex optimization. *arXiv*, 2019.

[9] 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.

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

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

[12] 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.

[13] Derek Driggs, Jingwei Liang, and Carola-Bibiane Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 2022.

[14] Nikoli Dryden, Tim Moon, Sam Ade Jacobs, and Brian Van Essen. Communication quantization for data-parallel training of deep neural networks. In *2016 2nd Workshop on machine learning in hpc environments (MLHPC)*, pages 1–8. IEEE, 2016.

[15] 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.

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

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

[18] Xiaowei Hu, LA Prashanth, András György, and Csaba Szepesvari. (bandit) convex optimization with biased noisy gradient oracles. In *Artificial Intelligence and Statistics*, pages 819–828. PMLR, 2016.

[19] Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 2020.

[20] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2011.

[21] Yin Liu and Sam Davanloo Tajbakhsh. Stochastic optimization algorithms for problems with controllable biased oracles. *arXiv*, 2026.

[22] 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.

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

[24] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[25] 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.

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

A Appendix A

Lemma 1. *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, we have that

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

By Matrix-Khintchine inequality [22], we have that

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

Taking $p = 1$, we have that

$$\mathbb{E}_\epsilon[\sum_{i=1}^n ||\epsilon_i Z_i||_{S_2}^2] \leq D C_1^2 C_3^4 \cdot \sum_{i=1}^n ||X_i||_{\text{op}}^2$$

and when taking expectation with respect to X, we have that

$$\begin{aligned}
&\mathbb{E}[||\sum_{i=1}^{n}\epsilon_i X_i||_{\text{op}}^2] \\
&\leq C_2^2 \cdot DC_1^2C_3^4 \cdot \sum_{i=1}^{n}\mathbb{E}||X_i||_{\text{op}}^2 \\
&\leq DC_2^2C_1^2C_3^4\sigma^2 \\
&\leq d^{m/2}C_2^2C_1^2C_3^4\sigma^2
\end{aligned}$$

Putting everything together and normalizing gives a final bound of

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

proving the claim. □

Lemma 2. *For all integers $p \geq 1$, for all $i \in \{1, \ldots, p\}$ and all $t \geq 1$, there exists a d-dependent, n_i-independent constant $C \geq 0$ such that*

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

Proof. First, we can say that

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

Notice that for any $r \in [1, 2]$, we can have that

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

where we have used Lyapunov's inequality and the result from lemma 1. Thus, we have a final bound of

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

which finishes the proof. □

Lemma 3. *Given a function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, 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 $y \in \text{argmin}_{z:||z-x|| \leq \eta} m_x(z)$ for $0 \leq \eta < 1$. Then, for all $M \geq 8L_p$, we have that

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

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

$$||\nabla^i F(x)[y-x,:,\ldots,:] - D^{(i)}(x)[y-x,:,\ldots,:]||_{\text{op}} \cdot ||y-x||$$

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

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

which implies that

$$- \frac{7M}{8(p+1)!} ||y-x||^{p+1} + ||\nabla F(x) - D^{(1)}(x)|| \cdot ||y-x||$$

$$+ \sum_{i=2}^{p} \frac{1}{i!} ||\nabla^i F(x)[y-x,:,\ldots,:] - D^{(i)}(x)[y-x,:,\ldots,:]||_{\text{op}} \cdot ||y-x||$$

$$\leq -\frac{7M}{8(p+1)!} ||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - g||^{\frac{p+1}{p}} + \frac{M||y-x||^{p+1}}{2(p+1)!}$$

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

$$\leq -\frac{3M}{8(p+1)!} ||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}$$

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

$$< -\frac{3M}{8(p+1)!} ||y-x||^{p+1} + (\frac{2p!}{M})^{\frac{1}{p}} \cdot \frac{p}{p+1} \cdot ||\nabla F(x) - D^{(1)}(x)||^{\frac{p+1}{p}}$$

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

We can further bound this expression by

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

which finishes the proof. □

Lemma 4. *Given a function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, let $y \in \operatorname{argmin}_{z:||z-x||\leq\eta} m_x(z)$, where*

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

for $M \geq 8L_p$ and $0 \leq \eta < 1$. It holds 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 $y \in \operatorname{argmin}_z m_x(z)$, 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

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

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

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

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 5. *We now consider the setting where the derivative estimates D^i are random variables. For all $F \in \mathcal{F}_p(\Delta, L_{1:p})$, it holds 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

$$\mathbf{1}[||\nabla F(y)|| \geq \frac{9M}{8p!}\eta^p]$$

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

$$\leq \frac{2^{1/p}}{\eta^{p+1}}||y-x||^{p+1} + 2^{1/p} \cdot (\frac{p!}{M\eta^p})^{\frac{p+1}{p}} \cdot (\sum_{i=0}^{p-1}\frac{1}{i!}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i)^{\frac{p+1}{p}}$$

$$< \frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{2(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot (\sum_{i=0}^{p-1}||\nabla^{i+1}F(x) - D^{i+1}(x)|| \cdot \eta^i)^{\frac{p+1}{p}}$$

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 can now continue to bound the above expression by

$$\frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{2(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot p^{1/p} \cdot \sum_{i=0}^{p-1} ||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}$$
$$< \frac{2}{\eta^{p+1}}||y-x||^{p+1} + \frac{4(p!)^{\frac{p+1}{p}}}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot \sum_{i=0}^{p-1} ||\nabla^{i+1}F(x) - D^{i+1}(x)||^{\frac{p+1}{p}} \cdot \eta^{\frac{i(p+1)}{p}}$$

where we used the fact that for all $p \geq 1$, $p^{1/p} < 2$. Taking expectations on each side, we have that

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

We also have 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}} \\
&= \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}} \\
&\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 \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}}
\end{aligned}$$

where the last step follows from the fact that $\eta \leq 1$. We further lower bound this expression by

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

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

Lemma 6. *Let $F \in \mathcal{F}_p(\Delta, L_{1:p})$ be given. Then, if the derivative estimates D^i are random variables, it holds that*

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

Proof. From lemma 5, we have that

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

Telescoping this recurrence from $t = 1$ to T gives

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

which implies that

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

which finishes the proof. □

Theorem 5. *(Theorem 3 restated). For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, where $p \geq 2$, $\epsilon > 0$, with biased and stochastic p^{th}-order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$ where $\max_i B_i \geq \Omega(\epsilon^{\frac{3p}{3p+1}})$, with probability at least $\frac{5}{8}$, Algorithm 1 returns a point $\hat{x}$ such that $||\nabla F(\hat{x})|| \leq O(\epsilon + \max_i B_i)$ and performs at most*

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

queries to the stochastic and biased derivative oracles.

Proof. From lemma 6, we have that

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

Let

$$A = \max(\frac{16(p+1)!}{M}, \frac{16(p+1)!}{M}[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + 2(\frac{2p!}{M})^{\frac{1}{p}}], \frac{16(p+1)!}{M}[\frac{(p!)^{\frac{p+1}{p}}}{4\sqrt[p]{M}} + (\frac{2p \cdot p!}{M})^{\frac{1}{p}}])$$

Therefore, we have the following upper bound:

$$\frac{A\Delta}{\eta^{p+1}T} + \frac{A}{\eta^{p+1}}[(\frac{C_1 \cdot \sigma_1^2}{n_1})^{\frac{p+1}{2p}} + B_1^{\frac{p+1}{2p}}] + \frac{A}{\eta^{\frac{p^2-1}{p}}}\sum_{i=2}^{p}[(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}]$$

From our choice of n_1 (where such an n_1 exists due to $\max_i B_i \geq \Omega(\epsilon^{\frac{3p}{3p+1}})$ such that

$$\max\{\frac{C_1 \cdot \sigma_1^2}{(\frac{\eta^{p+1}}{8A} - B_1^{\frac{p+1}{2p}})^{\frac{2p}{p+1}}}, 1\} \leq n_1 \leq \frac{(\epsilon + \max_i B_i)^{\frac{p+1}{p}}(\max_i \sigma_i)^2}{\epsilon^3(\max_i B_i + 1)^{\frac{p+1}{p}}}$$

we have that

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

From our choice of $T = \lceil \frac{8A\Delta}{\eta^{p+1}} \rceil$ we have that

$$\frac{A\Delta}{\eta^{p+1}T} \leq \frac{1}{8}$$

From our choice of n_i (for all $i \geq 2$, which exists due to $\max_i B_i \geq \Omega(\epsilon^{\frac{3p}{3p+1}})$) such that

$$\max\{C_i\sigma_i^2(\frac{\eta^{\frac{p^2-1}{p}}}{8Ap} - B_i^{\frac{p+1}{2p}})^{\frac{-2p}{p+1}}, 1\} \leq n_i \leq \frac{(\epsilon + \max_i B_i)^{\frac{p+1}{p}}(\max_i \sigma_i)^2}{\epsilon^3(\max_i B_i + 1)^{\frac{p+1}{p}}}$$

it holds that

$$\frac{A}{\eta^{\frac{p^2-1}{p}}}[(\frac{C_i \cdot \sigma_i^2}{n_i})^{\frac{p+1}{2p}} + B_i^{\frac{p+1}{2p}}] \leq \frac{1}{8p}$$

which implies that

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

Therefore, we have that

$$\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}(\epsilon + \max_i B_i)) \leq \frac{3}{8}$$

which implies that

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

Now, we give a bound on the oracle complexity. Let M be the total number of oracle queries that we make. In every iteration, we query the i^{th} derivative oracle n_i times which yields that

$$\begin{aligned}\mathbb{E}[M] &= T\sum_{i=1}^{p} n_i \\ &\leq (\frac{8A\Delta}{\eta^{p+1}} + 1)\sum_{i=1}^{p} n_i\end{aligned}$$

Substituting the upper bound for n_i, we can further bound this expression by

$$\begin{aligned}&(\frac{8A\Delta}{\eta^{p+1}} + 1)\sum_{i=1}^{p}\frac{(\epsilon + \max_j B_j)^{\frac{p+1}{p}} \cdot (\max_i \sigma_i)^2}{\epsilon^3(\max_j B_j + 1)^{\frac{p+1}{p}}} \\ &\leq O(\frac{\Delta}{(\epsilon + \max_j B_j)^{\frac{p+1}{p}}} \cdot \frac{(\epsilon + \max_j B_j)^{\frac{p+1}{p}}(\max_i \sigma_i)^2}{\epsilon^3(\max_j B_j + 1)^{\frac{p+1}{p}}} + \frac{(\epsilon + \max_j B_j)^{\frac{p+1}{p}}(\max_i \sigma_i)^2}{\epsilon^3(\max_j B_j + 1)^{\frac{p+1}{p}}}) \\ &= O(\frac{\Delta(\max_i \sigma_i)^2}{\epsilon^3(\max_j B_j + 1)^{\frac{p+1}{p}}} + \frac{(\epsilon + \max_j B_j)^{\frac{p+1}{p}}(\max_i \sigma_i)^2}{\epsilon^3(\max_j B_j + 1)^{\frac{p+1}{p}}})\end{aligned}$$

which finishes the proof. □

B Appendix B

Lemma 7. *Let $F \in \mathcal{F}_p(\Delta, L_{1:p})$. For any biased and stochastic oracle in $\mathcal{O}_p(F, \sigma_{1:p}, B_{1:p})$, let $\{D^i(x^{(t)})\}$ represent the sequence of i^{th} derivative iterates generated by Algorithm 2. Let $B = \max_{1 \le i \le p} B_i$. Then, we have that*

$$\mathbb{E}[||D^i(x^{(t)}) - \nabla^i F(x^{(t)})||^2] \le 4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p}$$

for all $1 \le i \le p$ and all $t \ge 1$.

Proof. We can first say that

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

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

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

We now say that

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

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

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

which implies that

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

We can also derive that

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

Combining both of these inequalities together, we have that

$$\begin{aligned}
&\mathbb{E}||e^{(t)}||^2 \\
&= b^{(t)}(2B_i^2 + \frac{2\epsilon^2}{5}) + (1 - b^{(t)}) \cdot \mathbb{E}[||e^{(t)}||^2|C^{(t)} = 0] \\
&\leq b^{(t)}(2B_i^2 + \frac{2\epsilon^2}{5}) + \mathbb{E}[(1 - b^{(t)})(1 + \frac{2}{b^{(t)}})||e^{(t-1)}||^2 + (1 - b^{(t)})(1 + \frac{2}{b^{(t)}})(\frac{b^{(t)}\epsilon}{10} + B_{i+1}b^{(t)}(\epsilon + \max_i B_i)^{\frac{1}{p}})^2] \\
&+ \mathbb{E}[(1 - b^{(t)}) \cdot \frac{b^{(t)}\epsilon^2}{5}] \\
&\leq \mathbb{E}[b^{(t)} \cdot (2B^2 + \frac{2\epsilon^2}{5})] + (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[3b^{(t)} \cdot (\epsilon + B(\epsilon + B)^{1/p})^2 + b^{(t)}\frac{\epsilon^2}{5}] \\
&\leq (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}](2B^2 + \frac{3\epsilon^2}{5} + 3(\epsilon + B(\epsilon + B)^{1/p})^2) \\
&\leq (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}] \cdot (2B^2 + \frac{3\epsilon^2}{5} + 3(\epsilon + B\epsilon^{1/p} + B^{1/p})^2) \\
&\leq (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}] \cdot (2B^2 + \frac{3\epsilon^2}{5} + 9\epsilon^2 + 9B^2\epsilon^{2/p} + 9B^{2/p}) \\
&= (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \mathbb{E}[b^{(t)}] \cdot (2B^2 + \frac{48\epsilon^2}{5} + 9B^2\epsilon^{2/p} + 9B^{2/p}) \\
&= (1 - \frac{\mathbb{E}[b^{(t)}]}{2})||e^{(t-1)}||^2 + \frac{\mathbb{E}[b^{(t)}]}{2} \cdot (4B^2 + \frac{96\epsilon^2}{5} + 18B^2\epsilon^{2/p} + 18B^{2/p})
\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}||e^{(t)}||^2 \\
&\leq (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p}) - (2B^2 + \frac{94}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})\prod_{s=2}^{t}(1 - \frac{b^{(s)}}{2}) \\
&\leq 4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p}
\end{aligned}$$

which finishes the proof. □

Lemma 8. *It holds that*

$$\begin{aligned}
&\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) \leq \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&+ \frac{16p(p+1)!}{M\eta^{\frac{p^2-1}{p}}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}
\end{aligned}$$

Proof. From Lemma 5, we have that

$$\mathbb{E}[F(x^{(t)}) - F(x^{(t+1)})]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(x^{(t+1)})|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot \mathbb{E}[||\nabla F(x^{(t)}) - D^{(1)}(x^{(t)})||^{\frac{p+1}{p}}]$$
$$- \eta^{\frac{p+1}{p}}(\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}) \sum_{i=2}^{p} \mathbb{E}[||\nabla^i F(x^{(t)}) - D^i(x^{(t)})||^{\frac{p+1}{p}}]$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(x^{(t+1)})|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (\mathbb{E}[||\nabla F(x^{(t)}) - D^{(1)}(x^{(t)})||^2])^{\frac{p+1}{2p}}$$
$$- \eta^{\frac{p+1}{p}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot \sum_{i=2}^{p} (\mathbb{E}[||\nabla^{(i)} F(x^{(t)}) - D^{(i)}(x^{(t)})||^2])^{\frac{p+1}{2p}}$$
$$\geq \frac{M\eta^{p+1}}{16(p+1)!}\Pr(||\nabla F(x^{(t+1)})|| \geq \frac{9M}{8p!}\eta^p) - [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{\frac{2}{p}} + 18B^{\frac{2}{p}})^{\frac{p+1}{2p}}$$
$$- p\eta^{\frac{p+1}{p}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{\frac{2}{p}} + 18B^{\frac{2}{p}})^{\frac{p+1}{2p}}$$

Telescoping this recurrence from $t = 1$ to T gives

$$\mathbb{E}[F(x^{(1)}) - F(x^{(T+1)})]$$
$$\geq \frac{M\eta^{p+1}T}{16(p+1)!}\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) - T \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}$$
$$- p\eta^{\frac{p+1}{p}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}$$

which implies that

$$\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) \leq \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}$$
$$+ \frac{16p(p+1)!}{M\eta^{\frac{p^2-1}{p}}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}$$

which finishes the proof. □

Theorem 6. *Theorem (4 restated). For any function $F \in \mathcal{F}_p(\Delta, L_{1:p})$, with biased and stochastic p^{th} order oracles in $\mathcal{O}(F, \sigma_{1:p}, B_{1:p})$, with probability at least $\frac{5}{8}$, Algorithm 3 returns a point $\hat{x}$ such that:*

- *If $\max_i B_i = \Theta(1)$, then $||\nabla F(\hat{x})|| \leq O(\epsilon + \max_i B_i)$ with at most*

$$O(\frac{\Delta(\max_i \sigma_i)^2(\epsilon + B)^{\frac{1}{p}} + (\max_i \sigma_i)^2}{\epsilon^2} + \frac{\Delta(\epsilon + B)^{\frac{1}{p}} + 1}{\epsilon})$$

  queries to the stochastic and biased derivative oracles.

- *If $\max_i B_i > \Omega(1)$, then $||\nabla F(\hat{x})|| \leq O((\epsilon^2 + B^2)^{\frac{1}{2}}(\epsilon + B))$ with at most*

$$O(\frac{(\max_i \sigma_i)^2}{\epsilon^2(\epsilon + B)^{\frac{p+1}{p}}} + \frac{1}{\epsilon(\epsilon + B)^{\frac{p+1}{p}}})$$

  queries to the stochastic and biased derivative oracles.

where $B = \max_i B_i$.

Proof. From Lemma 8, we have that

$$
\begin{aligned}
&\Pr(||\nabla F(\hat{x})|| \geq \frac{9M}{8p!}\eta^p) \\
&\leq \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + (\frac{2p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&+ \frac{16p(p+1)!}{M\eta^{\frac{p^2-1}{p}}} \cdot [\frac{(p!)^{\frac{p+1}{p}}}{8\sqrt[p]{M}} + \frac{1}{2}(\frac{2p \cdot p!}{M})^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&= \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{16(p+1)!}{M\eta^{p+1}}[\frac{(p!)^{\frac{p+1}{p}} + 8 \cdot (2p!)^{\frac{1}{p}}}{8M^{\frac{1}{p}}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&+ \frac{16p(p+1)!}{M\eta^{\frac{p^2-1}{p}}} \cdot [\frac{(p!)^{\frac{p+1}{p}} + 4(2p \cdot p!)^{\frac{1}{p}}}{8M^{\frac{1}{p}}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&= \frac{16(p+1)!}{M\eta^{p+1}T}\Delta + \frac{2(p+1)!}{M^{\frac{p+1}{p}}\eta^{p+1}}[(p!)^{\frac{p+1}{p}} + 8 \cdot (2p!)^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&+ \frac{2p(p+1)!}{M^{\frac{p+1}{p}}\eta^{\frac{p^2-1}{p}}} \cdot [(p!)^{\frac{p+1}{p}} + 4(2p \cdot p!)^{\frac{1}{p}}] \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}
\end{aligned}
$$

Letting

$$A = \max(16(p+1)!, 2(p+1)! \cdot [(p!)^{\frac{p+1}{p}} + 8 \cdot (2p!)^{\frac{1}{p}}], 2(p+1)! \cdot [(p!)^{\frac{p+1}{p}} + 4(2p \cdot p!)^{\frac{1}{p}}])$$

we get an upper bound of

$$
\begin{aligned}
&\frac{A\Delta}{M\eta^{p+1}T} + \frac{A}{M^{\frac{p+1}{p}}\eta^{p+1}} \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \\
&+ \frac{A}{M^{\frac{p+1}{p}}\eta^{\frac{p^2-1}{p}}} \cdot (4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}}
\end{aligned}
$$

Let $X = 4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p}$. Since we set

$$M = \max\{(\frac{8AX^{\frac{p+1}{2p}}}{\eta^{p+1}})^{\frac{p}{p+1}}, (\frac{8Ap \cdot X^{\frac{p+1}{2p}}}{\eta^{\frac{p^2-1}{p}}})^{\frac{p}{p+1}}, (\epsilon + B)^{\frac{-p-2}{p}}, 8L_p)$$

we have that

$$\frac{A}{M^{\frac{p+1}{p}}\eta^{p+1}}(4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \leq \frac{1}{8}$$

and

$$\frac{pA}{M^{\frac{p+1}{p}}\eta^{\frac{p^2-1}{p^2}}}(4B^2 + \frac{96}{5}\epsilon^2 + 18B^2\epsilon^{2/p} + 18B^{2/p})^{\frac{p+1}{2p}} \leq \frac{1}{8}$$

and through setting

$$T = \lceil \frac{8A\Delta}{M\eta^{p+1}} \rceil > \frac{8A\Delta}{M\eta^{p+1}}$$

we have that

$$\frac{A\Delta}{M\eta^{p+1}T} < \frac{1}{8}$$

Notice that if $B = \Theta(1)$, then $X = \Theta(1)$, we have that

$$\frac{5}{8} \leq \Pr(||\nabla F(\hat{x})|| \leq \frac{9M}{8p!}\eta^p) \leq \Pr(||\nabla F(\hat{x})|| \leq O(1) \cdot (\epsilon + B))$$

On the other hand, if $B > \Omega(1)$, then $X = O(B^2 + \epsilon^2)$, which implies that $M = \max(O(\frac{(B^2+\epsilon^2)^{1/2}}{\eta^p}), O(\frac{(B^2+\epsilon^2)^{\frac{1}{2}}}{\eta^{p-1}}), O(1), O(1))$. Since $B > \Omega(1)$, $1-\epsilon < (\epsilon+B)^{\frac{1}{p}}$, so $\eta = 1-\epsilon$. Therefore, $M = O((B^2+\epsilon^2)^{\frac{1}{2}})$. We then have that

$$\frac{5}{8} \leq \Pr(||\nabla F(\hat{x})|| \leq \frac{9M}{8p!}\eta^p) \leq \Pr(||\nabla F(\hat{x})|| \leq O((\epsilon^2+B^2)^{\frac{1}{2}}(\epsilon+B))$$

Let M_i be the number of oracle queries for derivative order i, and let $M = \sum_{i=1}^{p+1} M_i$ be the total number of oracle queries. With regards to the oracle complexity, we have that

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

Letting $\sigma = \max_i \sigma_i$, we upper bound the expression above by

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

We again analyze the following cases: $B = \Theta(1)$ and $B > \Omega(1)$. If $B = \Theta(1)$, then

$$\frac{8A\Delta}{M\eta^{p+1}} \leq \frac{8A\Delta}{(\epsilon+B)^{-\frac{1}{p}}\eta^{p+1}} \leq \frac{8A\Delta}{(\epsilon+B)^{\frac{-p-2}{p}}(\epsilon+B)^{\frac{p+1}{p}} \cdot O(1)} = O(1) \cdot 8A\Delta \cdot (\epsilon+B)^{\frac{1}{p}}$$

Therefore,

$$T \cdot O(\frac{\sigma^2}{\epsilon^2} + \frac{1}{\epsilon}) \leq O(\frac{\Delta\sigma^2(\epsilon+B)^{\frac{1}{p}} + \sigma^2}{\epsilon^2} + \frac{\Delta(\epsilon+B)^{\frac{1}{p}} + 1}{\epsilon})$$

If $B > \Omega(1)$, then

$$\frac{8A\Delta}{M\eta^{p+1}} = \frac{8A\Delta}{M(\epsilon+B)^{\frac{p+1}{p}}} \leq \frac{8A\Delta}{8L_p \cdot (\epsilon+B)^{\frac{p+1}{p}}} \leq O(\frac{\Delta}{L_p \cdot (\epsilon+B)^{\frac{p+1}{p}}})$$

and so we have that

$$T \cdot O(\frac{\sigma^2}{\epsilon^2} + \frac{1}{\epsilon}) \leq O(\frac{\sigma^2}{\epsilon^2(\epsilon+B)^{\frac{p+1}{p}}} + \frac{1}{\epsilon(\epsilon+B)^{\frac{p+1}{p}}})$$

which finishes the proof. □

C Appendix C

C.1 Preliminaries

We first introduce some important notational conventions used throughout this section. Given a p^{th} order tensor $T \in \mathbb{R}^{d \times \ldots, \times d}$, we define the support of T as the following:

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

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 1. *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 2. *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 9. *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 Lemma 16 of [7] □

Definition 3. *Let*

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

where

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

Lemma 10. *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* $||\nabla F_T(x)|| \geq |\nabla_{\text{prog}_1(x)+1} F_T(x)| > 1$
- *For all* $x, y \in \mathbb{R}^d$, *there exists a constant* $C \geq 0$ *such that* $||F(x) - F(y)|| \leq C\sqrt{T}||x - y||$

Proof. The first five statements follow from Lemma 2, Lemma 3, and Observation 3 of [10] and section G.1.1. of [7]. We now prove the last statement coordinate-wise by considering three separate cases for a coordinate i: $1 < i < T$, $i = 1$, and $i = T$. First, let $1 < i < T$. So, we have that $\partial_i F_T(x)$

$$= \frac{\partial}{\partial x_i}(\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)) + \frac{\partial}{\partial x_i}(\Psi(-x_i)\Phi(-x_{i+1}) - \Psi(x_i)\Phi(x_{i+1}))$$
$$= -\Psi(-x_{i-1})\Phi'(-x_i) - \Psi(x_{i-1})\Phi'(x_i) - \Psi'(-x_i)\Phi(-x_{i+1}) - \Psi'(x_i)\Phi(x_{i+1})$$

Observe that by construction of Ψ, we have that $|\Psi(x)| \leq e$ and $|\Psi'(x)| \leq 5$. Also, observe that

$$\Phi(x) = \sqrt{e}\int_{-\infty}^{x} e^{-t^2/2} \leq \sqrt{e}\int_{-\infty}^{\infty} e^{-t^2/2} = \sqrt{2\pi e}$$

and that $|\Phi'(x)| \leq \sqrt{e}$. Therefore, it holds that

$$|\partial_i F_T(x)| \leq e\sqrt{e} + e\sqrt{e} + 5\sqrt{2\pi e} + 5\sqrt{2\pi e} \leq 51$$

One can derive similar constants for the $i = 1$ and $i = T$ case. Let C be the maximum of 51 and these constants. We have that

$$||\nabla F_T(x)||_2 = (\sum_{i=1}^{T} |\partial_i F_T(x)|^2)^{\frac{1}{2}} \leq (\sum_{i=1}^{T} C^2)^{\frac{1}{2}} = C\sqrt{T}$$

The statement follows by applying norm equivalence in finite dimensional spaces.

□

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

Lemma 11. *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 follows. Let $\overline{\nabla}^q F(x)$ be a stochastic but unbiased estimator of $\nabla^q F(x)$. We then have that

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

which finishes the proof. □

C.2 Theorem 1 Proof

Remark 1. *Our oracle model is the same as in [3], 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 [3]. The last point follows from the fact that in Theorem 4 of [3], 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

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

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

Theorem 7. *(Theorem 1 restated). 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 (definition 1) where $\epsilon < \frac{1}{4}$ and $B_1 \leq O(1)$, the minimum number of queries to obtain a $(\epsilon + B_1^2)^{\frac{1}{2}}$ stationary point with constant probability is bounded below by*

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

Proof. We let $F_T^* = \alpha F_T(\beta x)$ for some constants α, β which we set in this proof. 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

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

for all q. Notice that by construction of F_T^*, we have that

- $F_T^*(0) - \inf_x F_T^*(x) = \alpha(F_T(0) - \inf_x F_T(\alpha x)) \leq \alpha\Delta_0 T$
- $||\nabla^2 F_T^*(x)|| = \alpha\beta^{q+1}||\nabla^2 F_T(\beta x)|| \leq \alpha\beta^2\ell_1$
- $||\nabla F_T^*(x)|| \geq \alpha\beta||\nabla F_T(x)|| \geq \frac{\alpha\beta}{2}$

We also have that

$$\mathbb{E}||\tilde{\nabla} F_T^*(x,z) - \nabla F_T^*(x,z)||^2 \leq \alpha^2\beta^2(\frac{2\ell_0^2(1-\rho)}{\rho} + 2B_1^2)$$

We set constants such that

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

First, let $\alpha = 2(\epsilon + B_1^2)^{\frac{1}{2}}/\beta$. We then set

$$\rho = \min\{\frac{\alpha^2\beta^2\ell_0^2}{\sigma_1^2 + B_1^2 - \alpha^2\beta^2 B_1^2}, 1\} = \min\{\frac{4(\epsilon + B_1^2)\ell_0^2}{\sigma_1^2 + B_1^2 - 4(\epsilon + B_1^2)B_1^2}, 1\}$$

With this choice of ρ, it's easy to check that

$$\alpha^2\beta^2(\frac{\ell_0^2(1-\rho)}{\rho} + B_1^2) \leq \sigma_1^2 + B_1^2$$

To satisfy the Lipschitz condition, we set

$$\beta = \frac{L_1}{2(\epsilon + B_1^2)^{\frac{1}{2}}\ell_1}$$

and we set

$$T = \lfloor \frac{\Delta}{\alpha\Delta_0} \rfloor = \lfloor \frac{\Delta\beta}{2\Delta_0(\epsilon + B_1^2)^{\frac{1}{2}}} \rfloor$$

By Lemma 9, we have that

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

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

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

Now, considering the case where the derivative oracles are biased but not stochastic (i.e $\sigma_1 = 0$), we have the following conditions:

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

Again, we let $\alpha = 2(\epsilon + B_1^2)^{\frac{1}{2}}/\beta$. We then set

$$\rho = \min\{\frac{\alpha^2\beta^2\ell_0^2}{B_1^2(1 - \alpha^2\beta^2)}, 1\} = \min\{\frac{4(\epsilon + B_1^2)\ell_0^2}{B_1^2 - 4B_1^2(\epsilon + B_1^2)}, 1\}$$

We again set

$$\beta = \frac{L_1}{2(\epsilon + B_1^2)^{\frac{1}{2}}\ell_1}$$

and

$$T = \lfloor \frac{\Delta}{\alpha \Delta_0} \rfloor = \lfloor \frac{\Delta\beta}{2(\epsilon + B_1^2)^{\frac{1}{2}}\Delta_0} \rfloor$$

By Lemma 9, we have that

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

since for all $B_1 \leq O(1)$, $\rho = \Theta(1)$. When we further lower bound this expression, we have that

$$\Omega(\frac{\Delta\beta}{\Delta_0(\epsilon + B_1^2)^{\frac{1}{2}}}) \geq \Omega(\frac{\Delta L_1}{\Delta_0(\epsilon + B_1^2)\ell_1}) = \Omega(\frac{\Delta L_1}{\epsilon + B_1^2})$$

Putting the two lower bound expressions together (as in [20]) yields the matching lower bound:

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

□

C.3 Theorem 2 Proof

Theorem 8. *(Theorem 2 restated). For all $p \geq 2$, Δ, $L_{1:p}$, $\sigma_{1:p} > 0$, $\epsilon < \sqrt{\sigma_1}$, and $\max_i B_i \leq \frac{\sqrt{3}}{2}\sigma_1$, 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 p^{th} order zero-respecting algorithm, the number of queries to obtain a point an $\epsilon + \max_i B_i$ stationary point with constant probability is bounded below by*

$$\Omega(1) \cdot \frac{(\sigma_1^2 - 4(\epsilon + B)^2 B_1^2)\Delta}{32(\epsilon + B)^3 \ell_0^2 \Delta_0} \cdot$$
$$\min_{q' \in \{1,\ldots,p\}, q \in \{2,\ldots,p\}} \min\{(\frac{\ell_0^2(\sigma_q^2 + B_q^2)}{2\ell_{q-1}^2(\sigma_1^2 + B_1^2 - 4(\epsilon + B)^2 B_1^2)})^{\frac{1}{2(q-1)}}, (\frac{\sigma_q^2 + B_q^2}{8(\epsilon + B)^2 B_q^2})^{\frac{1}{2(q-1)}}, (\frac{L_{q'}}{2(\epsilon + B)\ell_{q'}})^{\frac{1}{q'}}\}$$

Proof. We perform a similar argument to that for the proof of Theorem 1, except now accounting for the higher order Lipschitz constraints. We have that

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

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

- $\alpha\Delta_0 T \leq \Delta$
- $\alpha\beta^{q+1}\ell_q \leq L_q$
- $\frac{\alpha\beta}{2} \geq \epsilon + \max_j B_j$

- $\alpha^2\beta^{2q}(\frac{2\ell_{q-1}^2(1-\rho)}{\rho}+2B_q^2)\leq 2\sigma_q^2+2B_q^2 \implies \alpha^2\beta^{2q}(\frac{\ell_{q-1}^2(1-\rho)}{\rho}+B_q^2)\leq \sigma_q^2+B_q^2$

First, let

$$\alpha=\frac{2(\epsilon+\max_j B_j)}{\beta}$$

Now, we set

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

This implies that

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

where $B=\max_j B_j$. Letting

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

and solving for β such that

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

and

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

and the L_q-condition holds, yields

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

Setting

$$T=\lfloor\frac{\Delta}{\alpha\Delta_0}\rfloor=\lfloor\frac{\Delta\beta}{2\Delta_0(\epsilon+B)}\rfloor$$

We now have that (assuming $T\geq 5$)

$$\begin{aligned}
&\frac{T-2}{2\rho}=\frac{1}{2\rho}(\lfloor\frac{\Delta\beta}{2\Delta_0(\epsilon+B)}\rfloor-2)\\
&\geq\frac{1}{2\rho}\cdot\frac{\Delta\beta}{4\Delta_0(\epsilon+B)}\\
&\geq\frac{\sigma_1^2+B_1^2-\alpha^2\beta^2B_1^2}{2\alpha^2\beta^2\ell_0^2}\cdot\frac{\Delta}{4\Delta_0(\epsilon+B)}\cdot\\
&\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4(\epsilon+B)^2B_1^2)})^{\frac{1}{2(q-1)}},(\frac{\sigma_q^2+B_q^2}{8(\epsilon+B)^2B_q^2})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2(\epsilon+B)\ell_{q'}})^{\frac{1}{q'}}\}\\
&\geq\frac{(\sigma_1^2-4(\epsilon+B)^2B_1^2)\Delta}{32(\epsilon+B)^3\ell_0^2\Delta_0}\cdot\\
&\min_{q'\in\{1,\ldots,p\},q\in\{2,\ldots,p\}}\min\{(\frac{\ell_0^2(\sigma_q^2+B_q^2)}{2\ell_{q-1}^2(\sigma_1^2+B_1^2-4(\epsilon+B)^2B_1^2)})^{\frac{1}{2(q-1)}},(\frac{\sigma_q^2+B_q^2}{8(\epsilon+B)^2B_q^2})^{\frac{1}{2(q-1)}},(\frac{L_{q'}}{2(\epsilon+B)\ell_{q'}})^{\frac{1}{q'}}\}
\end{aligned}$$

which finishes the proof.

□