Second-Order Information in Non-Convex Stochastic Optimization: Power and Limitations

Yossi Arjevani* YOSSIA@NYU.EDU
Yair Carmon† YAIRC@STANFORD.EDU
John C. Duchi† JDUCHI@STANFORD.EDU
Dylan J. Foster‡ DYLANF@MIT.EDU
Ayush Sekhari§ AS3663@CORNELL.EDU
Karthik Sridharan§ KS999@CORNELL.EDU
*New York University, †Stanford University, ‡Massachusetts Institute of Technology, §Cornell University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We design an algorithm which finds an ϵ-approximate stationary point (with $\|\nabla F(x)\| \le \epsilon$) using $O(\epsilon^{-3})$ stochastic gradient and Hessian-vector products, matching guarantees that were previously available only under a stronger assumption of access to multiple queries with the same random seed. We prove a lower bound which establishes that this rate is optimal and—surprisingly—that it cannot be improved using stochastic pth order methods for any $p \ge 2$, even when the first p derivatives of the objective are Lipschitz. Together, these results characterize the complexity of non-convex stochastic optimization with second-order methods and beyond. Expanding our scope to the oracle complexity of finding (ϵ, γ)-approximate second-order stationary points, we establish nearly matching upper and lower bounds for stochastic second-order methods. Our lower bounds here are novel even in the noiseless case.

Keywords: Stochastic optimization, non-convex optimization, second-order methods, variance reduction, Hessian-vector products.

1. Introduction

Let $F : \mathbb{R}^d \to \mathbb{R}$ have Lipschitz continuous gradient and Hessian, and consider the task of finding an (ϵ, γ)-second-order stationary point (SOSP), that is, $x \in \mathbb{R}^d$ such that

$$\|\nabla F(x)\| \le \epsilon \quad \text{and} \quad \nabla^2 F(x) \succeq -\gamma I. \tag{1}$$

This task plays a central role in the study of non-convex optimization: for functions satisfying a weak strict saddle condition [20], exact SOSPs (with $\epsilon = \gamma = 0$) are local minima, and therefore the condition (1) serves as a proxy for approximate local optimality.[1] Moreover, for a growing set of non-convex optimization problems arising in machine learning, SOSPs are in fact *global minima* [20, 21, 35, 25]. Consequently, there has been intense recent interest in the design of efficient algorithms for finding approximate SOSPs [23, 2, 11, 17, 36, 38, 18].

1. However, it is NP-Hard to decide whether a SOSP is a local minimum or a high-order saddle point [28].

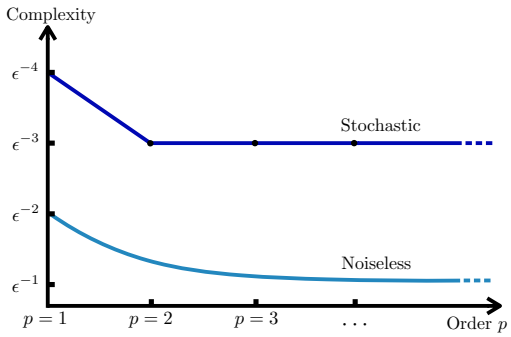

Figure 1: The "elbow effect:" for stochastic oracles, optimal complexity sharply improves from ϵ^{-4} for $p = 1$ to ϵ^{-3} for $p = 2$, but has no further improvement for $p > 2$. Noiseless complexity begins at ϵ^{-2} for $p = 1$ and smoothly approaches ϵ^{-1} as the derivative order $p \to \infty$.

Method	Requires $\widehat{\nabla^2 F}$?	Complexity bound	Additional assumptions
SGD [22]	No	$O(\epsilon^{-4})$	
Restarted SGD [18]	No	$O(\epsilon^{-3.5})$	$\widehat{\nabla F}$ Lipschitz almost surely
Subsampled regularized Newton [36]	Yes	$O(\epsilon^{-3.5})$	
Recursive variance reduction [e.g., 17]	No	$O(\epsilon^{-3})$	Mean-squared smoothness, Sim. queries (see Appendix C)
Hessian-vector recursive VR (Algorithm 2)	Yes	$O(\epsilon^{-3})$	None
Subsampled Newton with VR (Algorithm 3)	Yes	$O(\epsilon^{-3})$	None

Table 1: Comparison of guarantees for finding ϵ-stationary points (i.e., $\mathbb{E}\|\nabla F(x)\| \le \epsilon$) for a function F with Lipschitz gradient and Hessian. See Table 2 for detailed comparison.

In stochastic approximation tasks—particularly those motivated by machine learning—access to the objective function is often restricted to stochastic estimates of its gradient; for each query point $x \in \mathbb{R}^d$ we observe $\widehat{\nabla F}(x, z)$, where $z \sim P_z$ is a random variable such that

$$\mathbb{E}\left[\widehat{\nabla F}(x, z)\right] = \nabla F(x) \quad \text{and} \quad \mathbb{E}\left\|\widehat{\nabla F}(x, z) - \nabla F(x)\right\|^2 \le \sigma_1^2. \tag{2}$$

This restriction typically arises due to computational considerations (when $\widehat{\nabla F}(\cdot, z)$ is much cheaper to compute than $\nabla F(\cdot)$, as in empirical risk minimization or Monte Carlo simulation), or due to fundamental online nature of the problem at hand (e.g., when x represents a routing scheme and z represents traffic on a given day). However, for many problems with additional structure, we have access to extra information. For example, we often have access to stochastic second-order information in the form of a Hessian estimator $\widehat{\nabla^2 F}(x, z)$ satisfying

$$\mathbb{E}\left[\widehat{\nabla^2 F}(x, z)\right] = \nabla^2 F(x) \quad \text{and} \quad \mathbb{E}\left\|\widehat{\nabla^2 F}(x, z) - \nabla^2 F(x)\right\|_{\text{op}}^2 \le \sigma_2^2. \tag{3}$$

In this paper, we characterize the extent to which the stochastic Hessian information (3), as well as higher-order information, contributes to the efficiency of finding first- and second-order stationary points. We approach this question from the perspective of *oracle complexity* [29], which measures efficiency by the number of queries to estimators of the form (2)—and possibly (3)—required to satisfy the condition (1).

1.1. Our Contributions

We provide new upper and lower bounds on the stochastic oracle complexity of finding ϵ-stationary points and (ϵ, γ)-SOSPs. In brief, our main results are as follows.

- **Finding ϵ-stationary points: The elbow effect.** We propose a new algorithm that finds an ϵ-stationary point ($\gamma = \infty$) with $O(\epsilon^{-3})$ stochastic gradients and stochastic Hessian-vector

products. We furthermore show that this guarantee is not improvable via a complementary $\Omega(\epsilon^{-3})$ lower bound. All previous algorithms achieving $O(\epsilon^{-3})$ complexity require "multi-point" queries, in which the algorithm can query stochastic gradients at multiple points for the same random seed. Moreover, we show that $\Omega(\epsilon^{-3})$ remains a lower bound for stochastic pth-order methods for all $p \geq 2$ and hence—in contrast to the deterministic setting—the optimal rates for higher-order methods exhibit an "elbow effect"; see Figure 1.

- (ϵ, γ)**-stationary points: Improved algorithm and nearly matching lower bound.** We extend our algorithm to find (ϵ, γ)-stationary points using $O(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-5})$ stochastic gradient and Hessian-vector products, and prove a nearly matching $\Omega(\epsilon^{-3} + \gamma^{-5})$ lower bound.

In the remainder of this section we overview our results in greater detail. Unless otherwise stated, we assume F has both Lipschitz gradient and Hessian. To simplify the overview, we focus on dependence on ϵ^{-1} and γ^{-1} while keeping the other parameters—namely the initial optimality gap $F(x^{(0)}) - \inf_{x \in \mathbb{R}^d} F(x)$, the Lipschitz constants of ∇F and $\nabla^2 F$, and the variances of their estimators—held fixed. Our main theorems give explicit dependence on these parameters.

1.1.1. FIRST-ORDER STATIONARY POINTS ($\gamma = \infty$)

We first describe our developments for the task of finding ϵ-approximate first-order stationary points (satisfying (1) with $\gamma = \infty$), and subsequently extend our results to general γ. The reader may also refer to Table 1 for a succinct comparison of upper bounds.

Variance reduction via Hessian-vector products: A new gradient estimator. Using stochastic gradients and stochastic Hessian-vector products as primitives, we design a new variance-reduced gradient estimator. Plugging it into standard stochastic gradient descent (SGD), we obtain an algorithm that returns a point $\widehat{x}$ satisfying $\mathbb{E}\left\|\nabla F(\widehat{x})\right\| \leq \epsilon$ and requires $O(\epsilon^{-3})$ stochastic gradient and HVP queries in expectation. In comparison, vanilla SGD requires $O(\epsilon^{-4})$ queries [22], and the previously best known rate under our assumptions was $O(\epsilon^{-3.5})$, by both cubic-regularized Newton's method and a restarted variant of SGD [36, 18].

Our approach builds on a line of work by Fang et al. [17], Zhou et al. [39], Wang et al. [37], Cutkosky and Orabona [16] that also develop algorithms with complexity $O(\epsilon^{-3})$, but require a "multi-point" oracle in which algorithm can query the stochastic gradient at multiple points for the same random seed. Specifically, in the n-point variant of this model, the algorithm can query at the set of points $(x_1, \ldots, x_n)$ and receive

$$\widehat{\nabla F}(x_1, z), \ldots, \widehat{\nabla F}(x_n, z), \quad \text{where} \quad z \overset{\text{i.i.d.}}{\sim} P_z, \tag{4}$$

and where the estimator $\widehat{\nabla F}(x, z)$ is unbiased and has bounded variance in the sense of (2). The aforementioned works achieve $O(\epsilon^{-3})$ complexity using $n = 2$ simultaneous queries, while our new algorithm achieves the same rate using $n = 1$ (i.e., z is drawn afresh at each query), but using stochastic Hessian-vector products in addition to stochastic gradients. However, we show in Appendix C that under the statistical assumptions made in these works, the two-point stochastic gradient oracle model is *strictly* stronger than the single-point stochastic gradient/Hessian-vector product oracle we consider here. On the other hand, unlike our algorithm, these works do not require Lipschitz Hessian.

The algorithms that achieve complexity $O(\epsilon^{-3})$ using two-point queries work by estimating gradient differences of the form $\nabla F(x) - \nabla F(x')$ using $\widehat{\nabla F}(x, z) - \widehat{\nabla F}(x', z)$ and applying

recursive variance reduction [31]. Our primary algorithmic contribution is a second-order stochastic estimator for $\nabla F(x) - \nabla F(x')$ which avoids simultaneous queries while maintaining comparable error guarantees. To derive our estimator, we note that $\nabla F(x) - \nabla F(x') = \int_0^1 \nabla^2 F(xt + x'(1-t))(x - x')dt$, and use K queries to the stochastic Hessian estimator (3) to numerically approximate this integral.[2] Specifically, our estimator takes the form

$$\frac{1}{K}\sum_{k=0}^{K-1} \widehat{\nabla^2 F}\big(x \cdot (1 - \tfrac{k}{K}) + x' \cdot \tfrac{k}{K}, z^{(i)}\big)(x - x'), \tag{5}$$

where $z^{(i)} \overset{\text{i.i.d.}}{\sim} P_z$. Unlike the usual estimator $\widehat{\nabla F}(x, z) - \widehat{\nabla F}(x', z)$, the estimator (5) is biased. Nevertheless, we show that *choosing K dynamically* according to $K \propto \|x - x'\|^2$ provides adequate control over both bias and variance while maintaining the desired query complexity. Combining the integral estimator (5) with recursive variance reduction, we attain $O(\epsilon^{-3})$ complexity.

Demonstrating the power of second-order information. For functions with Lipschitz gradient and Hessian, we prove an $\Omega(\epsilon^{-3.5})$ lower bound on the minimax oracle complexity of algorithms for finding stationary points using *only* stochastic gradients (2).[3] This lower bound is an extension of the results of Arjevani et al. [8], who showed that for functions with Lipschitz gradient but *not* Lipschitz Hessian, the optimal rate is $\Theta(\epsilon^{-4})$ using *only* stochastic gradients (2). Together with our new $O(\epsilon^{-3})$ upper bound, this lower bound reveals that stochastic Hessian-vector products offer an $\Omega(\epsilon^{-0.5})$ improvement in the oracle complexity for finding stationary points in the single-point query model. This contrasts the noiseless optimization setting, where finite gradient differences can approximate Hessian-vector products arbitrarily well, meaning these oracle models are equivalent.

Demonstrating the limitations of higher-order information ($p > 2$). For algorithms that can query both stochastic gradients and stochastic Hessians, we prove a lower bound of $\Omega(\epsilon^{-3})$ on the oracle complexity of finding an expected ϵ-stationary point. This proves that our $O(\epsilon^{-3})$ upper bound is optimal in the leading order term in ϵ, despite using only stochastic Hessian-vector products rather than full stochastic Hessian queries.

Notably, our $\Omega(\epsilon^{-3})$ lower bound extends to settings where stochastic higher-order oracles are available, i.e, when the first p derivatives are Lipschitz and we have bounded-variance estimators $\{\widehat{\nabla^q F}(\cdot,\cdot)\}_{q \le p}$. The lower bound holds for any finite p, and thus, as a function of the oracle order p, the minimax complexity has an elbow (Figure 1): for $p = 1$ the complexity is $\Theta(\epsilon^{-4})$ [8] while for all $p \ge 2$ it is $\Theta(\epsilon^{-3})$. This means that smoothness and stochastic derivatives beyond the second-order cannot improve the leading term in rates of convergence to stationarity, establishing a fundamental limitation of stochastic high-order information. This highlights another contrast with the noiseless setting, where pth order methods enjoy improved complexity for every p [12].

As we discuss in Appendix C, for multi-point stochastic oracles (4), the rate $O(\epsilon^{-3})$ is attainable even without stochastic Hessian access. Moreover, our $\Omega(\epsilon^{-3})$ lower bound for stochastic pth order oracles holds even when multi-point queries are allowed. Consequently, when viewed through the lens of worst-case oracle complexity, our lower bounds show that even stochastic Hessian information is not helpful in the multi-point setting.

2. More precisely, our estimator (5) only requires stochastic Hessian-vector products, whose computation is often roughly as expensive as that of a stochastic gradient [33].
3. We formally prove our results for the structured class of *zero-respecting algorithms* [12]; the lower bounds extend to general randomized algorithms via similar arguments to Arjevani et al. [8].

1.1.2. SECOND-ORDER STATIONARY POINTS

Upper bounds for general γ. We incorporate our recursive variance-reduced Hessian-vector product-based gradient estimator into an algorithm that combines SGD with negative curvature search. Under the slightly stronger (relative to (3)) assumption that the stochastic Hessians have almost surely bounded error, we prove that—with constant probability—the algorithm returns an (ϵ, γ)-SOSP after performing $O(\epsilon^{-3} + \epsilon^{-2}\gamma^{-2} + \gamma^{-5})$ stochastic gradient and Hessian-vector product queries.

A lower bound for finding second-order stationary points. We prove a minimax lower bound which establishes that the stochastic second-order oracle complexity of finding (ϵ, γ)-SOSPs is $\Omega(\epsilon^{-3} + \gamma^{-5})$. Consequently, the algorithms we develop have optimal worst-case complexity in the regimes $\gamma = O(\epsilon^{2/3})$ and $\gamma = \Omega(\epsilon^{0.5})$. Compared to our lower bounds for finding ϵ-stationary points, proving the $\Omega(\gamma^{-5})$ lower bound requires a more substantial modification of the constructions of [12] and [8]. In fact, our lower bound is new even in the noiseless regime (i.e., $\sigma_1 = \sigma_2 = 0$), where it becomes $\Omega(\epsilon^{-1.5} + \gamma^{-3})$; this matches the guarantee of the cubic-regularized Newton's method [30] and consequently characterizes the optimal rate for finding approximate SOSPs using noiseless second-order methods.

1.2. Further related work

We briefly survey additional upper and lower complexity bounds related to our work and place our results within their context. The works of Monteiro and Svaiter [27], Arjevani et al. [9], Agarwal and Hazan [1] delineate the second-order oracle complexity of *convex* optimization in the noiseless setting; [7] treat the finite-sum setting.

For functions with Lipschitz gradient and Hessian, oracle access to the Hessian significantly accelerates convergence to ε-approximate global minima, reducing the complexity from $\Theta(\varepsilon^{-0.5})$ to $\Theta(\varepsilon^{-2/7})$. However, since the hard instances for first-order convex optimization are quadratic [29, 6, 34], assuming Lipschitz continuity of the Hessian does not improve the complexity if one only has access to a first-order oracle. This contrasts the case for finding ϵ-approximate stationary points of *non-convex* functions with noiseless oracles. There, Lipschitz continuity of the Hessian improves the first-order oracle complexity from $\Theta(\epsilon^{-2})$ to $O(\epsilon^{-1.75})$, with a lower bound of $\Omega(\epsilon^{-12/7})$ for deterministic algorithms [10, 13]. Additional access to full Hessian further improves this complexity to $\Theta(\epsilon^{-1.5})$, and for pth-order oracles with Lipschitz pth derivative, the complexity further improves to $\Theta(\epsilon^{-(1+\frac{1}{p})})$ [12]; see Figure 1.

1.3. Paper organization

We formally introduce our notation and oracle model in Section 2. Section 3 contains our results concerning the complexity of finding ϵ-first-order stationary points: algorithmic upper bounds (Section 3.1) and algorithm-independent lower bounds (Section 3.2). Following a similar outline, Section 4 describes our upper and lower bounds for finding (ϵ, γ)-SOSPs. In Appendix A, we discussion directions for future research. Additional technical comparison with related work is given in Appendix B and C, and proofs are given in Appendix D through Appendix H.

Notation. We let $\mathcal{C}^p$ denote the class of p-times differentiable real-valued functions, and let $\nabla^q F$ denote the qth derivative of a given function $F \in \mathcal{C}^p$ for $q \in \{1, \ldots, p\}$. Given a function

$F \in \mathcal{C}^1$, we let $\nabla_i F(x) := [\nabla F(x)]_i = \frac{\partial}{\partial x_i} F(x)$. When $F \in \mathcal{C}^2$ is twice differentiable, we define, $\nabla^2_{ij} f(x) := \left[\nabla^2 f(x)\right]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$, and similarly define $[\nabla^p f(x)]_{i_1,i_2,\ldots,i_p} = \frac{\partial^p}{\partial x_{i_1} \cdots \partial x_{i_p}} f(x)$ for pth-order derivatives. For a vector $x \in \mathbb{R}^d$, $\|x\|$ denotes the Euclidean norm and $\|x\|_\infty$ denotes the ℓ_∞ norm. For matrices $A \in \mathbb{R}^{d \times d}$, $\|A\|_{\mathrm{op}}$ denotes the operator norm. More generally, for symmetric pth order tensors T, we define the operator norm via $\|T\|_{\mathrm{op}} = \sup_{\|v\|=1} |\langle T, v^{\otimes p} \rangle|$, and we let $T[v^{(1)}, \ldots, v^{(p)}] = \left\langle T, v^{(1)} \otimes \cdots \otimes v^{(p)} \right\rangle$. Note that for a vector $x \in \mathbb{R}^d$ the operator norm $\|x\|_{\mathrm{op}}$ coincides with the Euclidean norm $\|x\|$. We let $\mathbb{S}^d$ denote the space of symmetric matrices in $\mathbb{R}^{d \times d}$. We let $\mathbb{B}_r(x)$ denote the Euclidean ball of radius r centered at $x \in \mathbb{R}^d$ (with dimension clear from context). We adopt non-asymptotic big-O notation, where $f = O(g)$ for $f, g : \mathcal{X} \to \mathbb{R}_+$ if $f(x) \le C g(x)$ for some constant $C > 0$.

2. Setup

We study the problem of finding ϵ-stationary and (ϵ, γ)-second order stationary points in the standard oracle complexity framework [29], which we briefly review here.

Function classes. We consider p-times differentiable functions satisfying standard regularity conditions, and define

$$\mathcal{F}_p(\Delta, L_{1:p}) = \left\{ F : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \begin{array}{l} F \in \mathcal{C}^p, \quad F(0) - \inf_x F(x) \le \Delta, \\ \|\nabla^q F(x) - \nabla^q F(y)\|_{\mathrm{op}} \le L_q \|x - y\| \text{ for all } x, y \in \mathbb{R}^d,\ q \in [p] \end{array} \right\},$$

so that $L_{1:p} := (L_1, \ldots, L_p)$ specifies the Lipschitz constants of the qth order derivatives $\nabla^q F$ with respect to the operator norm. We make no restriction on the ambient dimension d.

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

$$\mathsf{O}^p_F(x, z) := \left(\widehat{F}(x, z), \widehat{\nabla F}(x, z), \widehat{\nabla^2 F}(x, z), \ldots, \widehat{\nabla^p F}(x, z) \right), \tag{6}$$

where $\{\widehat{\nabla^q F}(\cdot, z)\}_{q=0}^p$ are unbiased estimators of the respective derivatives. That is, for all x, $\mathbb{E}_{z \sim P_z}[\widehat{F}(x, z)] = F(x)$ and $\mathbb{E}_{z \sim P_z}[\widehat{\nabla^q F}(x, z)] = \nabla^q F(x)$ for all $q \in [p]$.[4]

Given variance parameters $\sigma_{1:p} = (\sigma_1, \ldots, \sigma_p)$, we define the *oracle class* $\mathcal{O}_p(F, \sigma_{1:p})$ to be the set of all stochastic pth-order oracles for which the variance of the derivative estimators satisfies

$$\mathbb{E}_{z \sim P_z} \left\| \widehat{\nabla^q F}(x, z) - \nabla^q F(x) \right\|^2_{\mathrm{op}} \le \sigma_q^2, \quad q \in [p]. \tag{7}$$

The upper bounds in this paper hold even when $\sigma_0^2 := \max_{x \in \mathbb{R}^d} \mathrm{Var}(\widehat{F}(x, z))$ is infinite, while our lower bounds hold when $\sigma_0 = 0$, so to reduce notation, we leave dependence on this parameter tacit.

Optimization protocol. We consider stochastic pth-order optimization algorithms that access an unknown function $F \in \mathcal{F}_p(\Delta, L_{1:p})$ through multiple rounds of queries to a stochastic pth-order oracle $(\mathsf{O}^p_F, P_z) \in \mathcal{O}_p(F, \sigma_{1:p})$. When queried at $x^{(t)}$ in round t, the oracle performs an independent draw of $z^{(t)} \sim P_z$ and answers with $\mathsf{O}^p_F(x^{(t)}, z^{(t)})$. Algorithm queries depend on F only through the oracle answers; see e.g. Arjevani et al. [8, Section 2] for a more formal treatment.

4. For $p \ge 2$ we assume without loss of generality that $\widehat{\nabla^p F}(x, z)$ is a symmetric tensor.

3. Complexity of finding first-order stationary points

In this section we focus on the task of finding ϵ-approximate stationary points (satisfying $\|\nabla F(x)\| \le \epsilon$). As prior work observes [cf. 10, 2], stationary point search is a useful primitive for achieving the end goal of finding second-order stationary points (1). We begin with describing algorithmic upper bounds on the complexity of finding stationary points with stochastic second-order oracles, and then proceed to match their leading terms with general pth order lower bounds.

3.1. Upper bounds

Our algorithms rely on *recursive variance reduction* [31]: we sequentially estimate the gradient at the points $\{x^{(t)}\}_{t\ge 0}$ by accumulating cheap estimators of $\nabla F(x^{(\tau)}) - \nabla F(x^{(\tau-1)})$ for $\tau = t_0 + 1, \dots, t$, where at iteration t_0 we reset the gradient estimator by computing a high-accuracy approximation of $\nabla F(x^{(t_0)})$ with many oracle queries. Our implementation of recursive variance reduction, Algorithm 1, differs from previous approaches [17, 39, 37] in three aspects.

First, in Line 8 we estimate differences of the form $\nabla F(x^{(\tau)}) - \nabla F(x^{(\tau-1)})$ by averaging stochastic Hessian-vector products. This allows us to do away with multi-point queries and operate under weaker assumptions than prior work (see Appendix C), but it also introduces bias to our estimator, which makes its analysis more involved. This is the key novelty in our algorithm. Second, rather than resetting the gradient estimator every fixed number of steps, we reset with a user-defined probability b (Line 4); this makes the estimator stateless and greatly simplifies its analysis, especially in our algorithms for finding SOSPs, where we use a varying value of b. Finally, we dynamically select the batch size K for estimating gradient differences based on the distance between iterates (Line 2), while prior work uses a constant batch size. Our dynamic batch size scheme is crucial for controlling the bias in our estimator, while still allowing for large step sizes as in Wang et al. [37].

The core of our analysis is the following lemma, which bounds the gradient estimation error and expected oracle complexity. To state the lemma, we let $\{x^{(t)}\}_{t\ge 0}$ be sequence of queries to Algorithm 1, and let $g^{(t)} = \textsf{HVP-RVR-Gradient-Estimator}_{\epsilon,b}(x^{(t)}, x^{(t-1)}, g^{(t-1)})$ be the sequence of estimates it returns.

Lemma 1 *For any oracle in $\mathcal{O}_2(F, \sigma_{1:2})$ and $F \in \mathcal{F}_2(\Delta, L_{1:2})$, Algorithm 1 guarantees that*

$$\mathbb{E}\,\|g^{(t)} - \nabla F(x^{(t)})\|^2 \le \epsilon^2$$

for all $t \ge 1$. Furthermore, conditional on $x^{(t-1)}$, $x^{(t)}$ and $g^{(t-1)}$, the t^{th} execution of Algorithm 1 with reset probability b uses at most

$$O\Big(1 + b\frac{\sigma_1^2}{\epsilon^2} + \big\|x^{(t)} - x^{(t-1)}\big\|^2 \cdot \frac{\sigma_2^2 + \epsilon L_2}{b\epsilon^2}\Big)$$

stochastic gradient and Hessian-vector product queries in expectation.

We prove the lemma in Appendix D by bounding the per-step variance using the HVP oracle's variance bound (7), and by bounding the per-step bias relative to $\nabla F(x^{(t)}) - \nabla F(x^{(t-1)})$ using the Lipschitz continuity of the Hessian.

Our first algorithm for finding ϵ-stationary points, Algorithm 2, is simply stochastic gradient descent using the HVP-RVR gradient estimator (Algorithm 1); we bound its complexity by $O(\epsilon^{-3})$. Before stating the result formally, we briefly sketch the analysis here (see Appendix F.1 for details).

Algorithm 1 Recursive variance reduction with stochastic Hessian-vector products (HVP-RVR)

// **Gradient estimator for** $F \in \mathcal{F}_2(\Delta, L_{1:2})$ **given stochastic oracle in** $\mathcal{O}_2(F, \sigma_{1:2})$.

function HVP-RVR-GRADIENT-ESTIMATOR$_{\epsilon,b}(x, x_{\text{prev}}, g_{\text{prev}})$:

  Set $K = \left\lceil \frac{5(\sigma_2^2 + L_2\epsilon)}{b\epsilon^2} \cdot \|x - x_{\text{prev}}\|^2 \right\rceil$ and $n = \left\lceil \frac{5\sigma_1^2}{\epsilon^2} \right\rceil$.

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

  if C is 1 or g_{prev} is $\perp$ **then**

    Query the oracle n times at x and set $g \leftarrow \frac{1}{n}\sum_{j=1}^{n} \widehat{\nabla F}(x, z^{(j)})$, where $z^{(j)} \overset{\text{i.i.d.}}{\sim} P_z$.

  else

    Define $x^{(k)} := \frac{k}{K}x + \left(1 - \frac{k}{K}\right)x_{\text{prev}}$ for $k \in \{0, \dots, K\}$.

    Query the oracle at the set of points $\left(x^{(k)}\right)_{k=0}^{K-1}$ to compute

    $$g \leftarrow g_{\text{prev}} + \sum_{k=1}^{K} \widehat{\nabla^2 F}(x^{(k-1)}, z^{(k)})\left(x^{(k)} - x^{(k-1)}\right), \quad \text{where } z^{(k)} \overset{\text{i.i.d.}}{\sim} P_z.$$

  return g.

Algorithm 2 Stochastic gradient descent with HVP-RVR

Input: Oracle $(\mathsf{O}_F^2, P_z) \in \mathcal{O}_2(F, \sigma_{1:2})$ for $F \in \mathcal{F}_2(\Delta, L_1, L_2)$. Precision parameter ϵ.

Set $\eta = \frac{1}{2\sqrt{L_1^2 + \sigma_2^2 + \epsilon L_2}}$, $T = \left\lceil \frac{2\Delta}{\eta\epsilon^2} \right\rceil$, $n = \left\lceil \frac{4\sigma_1^2}{\epsilon^2} \right\rceil$, $b = \min\left\{1, \frac{\eta\epsilon\sqrt{\sigma_2^2 + \epsilon L_2}}{\sigma_1}\right\}$.

Initialize $x^{(0)}, x^{(1)} \leftarrow 0$, $g^{(0)} \leftarrow \perp$.

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

  $g^{(t)} \leftarrow$ HVP-RVR-Gradient-Estimator$_{\epsilon,b}(x^{(t)}, x^{(t-1)}, g^{(t-1)})$.

  $x^{(t+1)} \leftarrow x^{(t)} - \eta g^{(t)}$.

return $\widehat{x}$ chosen uniformly at random from $\left\{x^{(t)}\right\}_{t=1}^{T}$.

Standard analysis of SGD with step size $\eta \le \frac{1}{2L_1}$ shows that its iterates satisfy $\mathbb{E}\|\nabla F(x^{(t)})\|^2 \le \frac{1}{\eta}\mathbb{E}[F(x^{(t+1)}) - F(x^{(t)})] + O(1) \cdot \mathbb{E}\,\|g^{(t)} - \nabla F(x^{(t)})\|^2$. Telescoping over T steps, using Lemma 1 and substituting in the initial suboptimality bound Δ, this implies that

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla F(x^{(t)})\|^2 \le \frac{\Delta}{\eta T} + O(\epsilon^2). \tag{8}$$

Taking $T = \Omega(\frac{\Delta}{\eta\epsilon^2})$, we are guaranteed that a uniformly selected iterate has expected norm $O(\epsilon)$.

To account for oracle complexity, we observe from Lemma 1 that T calls to Algorithm 1 require at most $T(\frac{\sigma_1^2 b}{\epsilon^2} + 1) + \sum_{t=1}^{T} \mathbb{E}\,\|x^{(t)} - x^{(t-1)}\|^2 \cdot \left(\frac{\sigma_2^2 + L_2\epsilon}{b\epsilon^2}\right)$ oracle queries in expectation. Using $x^{(t)} - x^{(t-1)} = \eta g^{(t-1)}$, Lemma 1 and (8) imply that $\sum_{t=1}^{T} \mathbb{E}\,\|x^{(t)} - x^{(t-1)}\|^2 \le O(T\epsilon^2)$. We then choose b to out the terms $T\left(\frac{\sigma_1^2 b}{\epsilon^2}\right)$ and $T\left(\frac{\sigma_2^2 + L_2\epsilon}{b}\right)$. This gives the following guarantee.

Theorem 2 *For any function $F \in \mathcal{F}_2(\Delta, L_1, L_2)$, stochastic second-order oracle in $\mathcal{O}_2(F, \sigma_1, \sigma_2)$, and $\epsilon < \min\left\{\sigma_1, \sqrt{\Delta L_1}\right\}$, with probability at least $\frac{3}{4}$, Algorithm 2 returns a point $\widehat{x}$ such that $\|\nabla F(\widehat{x})\| \le \epsilon$ and performs at most*

$$O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} + \frac{\Delta L_2^{0.5}\sigma_1}{\epsilon^{2.5}} + \frac{\Delta L_1}{\epsilon^2}\right)$$

stochastic gradient and Hessian-vector product queries.

The oracle complexity of Algorithm 2 depends on the Lipschitz parameters of F only through lower-order terms in ϵ, with the leading term scaling only with the variance of the gradient and Hessian estimators. In the low noise regime where $\sigma_1 < \epsilon$ and $\sigma_2 < \max\{L_1, \sqrt{L_2\epsilon}\}$, the complexity becomes $O(\Delta L_1 \epsilon^{-2} + \Delta L_2^{0.5}\epsilon^{-1.5})$ which is simply the maximum of the noiseless guarantees for gradient descent and Newton's method. We remark, however, that in the noiseless regime $\sigma_1 = \sigma_2 = 0$, a slightly better guarantee $O(\Delta L_1^{0.5} L_2^{0.25} \epsilon^{-1.75} + \Delta L_2^{0.5}\epsilon^{-1.5})$ is achievable [10].

In the noiseless setting, any algorithm that uses only first-order and Hessian-vector product queries must have complexity scaling with L_1, but full Hessian access can remove this dependence [13]. We show that the same holds true in the stochastic setting: Algorithm 3, a subsampled cubic regularized trust-region method using Algorithm 1 for gradient estimation, enjoys a complexity bound independent of L_1. We defer the full description and analysis to Appendix F.2 and state the guarantee here.

Theorem 3 *For any function $F \in \mathcal{F}_2(\Delta, \infty, L_2)$, stochastic second order oracle in $\mathcal{O}_2(F, \sigma_1, \sigma_2)$, and $\epsilon < \sigma_1$, with probability at least $\frac{3}{4}$, Algorithm 3 returns a point $\widehat{x}$ such that $\|\nabla F(\widehat{x})\| \le \epsilon$ and performs at most*

$$O\Big(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} \cdot \log^{0.5} d + \frac{\Delta L_2^{0.5}\sigma_1}{\epsilon^{2.5}}\Big)$$

stochastic gradient and Hessian queries.

The guarantee of Theorem 3 constitutes an improvement in query complexity over Theorem 2 in the regime $L_1 \gtrsim (1 + \frac{\sigma_1}{\epsilon})(\sigma_2 + \sqrt{L_2\epsilon})$. However, depending on the problem, full stochastic Hessians can be up to d times more expensive to compute than stochastic Hessian-vector products.

3.2. Lower bounds

Having presented stochastic second-order methods with $O(\epsilon^{-3})$-complexity bound for finding ϵ-stationary points, our we next show that this rates cannot be improved. In fact, we show that this rate is optimal even when one is given access to stochastic higher derivatives of *any* order. We prove our lower bounds for the class of *zero-respecting* algorithms, which subsumes the majority of existing optimization methods; see Appendix H.1 for a formal definition. We believe that existing techniques [12, 8] can strengthen our lower bounds to apply to general randomized algorithms; for brevity, we do not pursue it here.

The lower bounds in this section closely follow a recent construction by Arjevani et al. [8, Section 3], who prove lower bounds for stochastic first-order methods. To establish complexity bounds for pth-order methods, we extend the 'probabilistic zero-chain' gradient estimator introduced in Arjevani et al. [8] to high-order derivative estimators. The most technically demanding part of our proof is a careful scaling of the basic construction to simultaneously meet multiple Lipschitz continuity and variance constraints. Deferring the proof details to Appendix H.1, our lower bound is as follows.

Theorem 4 *For all $p \in \mathbb{N}$, $\Delta, L_{1:p}, \sigma_{1:p} > 0$ and $\epsilon \le O(\sigma_1)$, there exists $F \in \mathcal{F}_p(\Delta, L_{1:p})$ and $(\mathsf{O}_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p})$, such that for any pth-order zero-respecting algorithm, the number of queries required to obtain an ϵ-stationary point with constant probability is bounded from below by*

$$\Omega(1) \cdot \frac{\Delta\sigma_1^2}{\epsilon^3} \min\left\{ \min_{q\in\{2,\dots,p\}} \left(\frac{\sigma_q}{\sigma_1}\right)^{\frac{1}{q-1}}, \min_{q'\in\{1,\dots,p\}} \left(\frac{L_{q'}}{\epsilon}\right)^{1/q'} \right\}. \tag{9}$$

A construction of dimension $\Theta\left(\frac{\Delta}{\epsilon}\min\left\{\min_{q\in\{2,\dots,p\}}\left(\frac{\sigma_q}{\sigma_1}\right)^{\frac{1}{q-1}}, \min_{q'\in\{1,\dots,p\}}\left(\frac{L_{q'}}{\epsilon}\right)^{1/q'}\right\}\right)$ *realizes this lower bound.*

For second-order methods ($p = 2$), Theorem 4 specializes to the complexity lower bound

$$\Omega(1)\cdot\min\left\{\frac{\Delta\sigma_1\sigma_2}{\epsilon^3}, \frac{\Delta L_2^{0.5}\sigma_1}{\epsilon^{3.5}}, \frac{\Delta L_1\sigma_1^2}{\epsilon^4}\right\}, \tag{10}$$

which is tight in that it matches (up to numerical constants) the convergence rate of Algorithm 2 in the regime where $\Delta\sigma_1\sigma_2\epsilon^{-3}$ dominates both the upper bound in Theorem 2 and expression (10). The lower bound (10) is also tight when the second-order information is not available or reliable (σ_2 is infinite or very large, respectively): Standard SGD matches the ϵ^{-4} term [22], while more sophisticated variants based on restarting [18] and normalized updates with momentum [15] match the $\epsilon^{-3.5}$ term (the former up to logarithmic factors)—neither of these algorithms requires stochastic second derivative estimation.

Theorem 4 implies that while higher-order methods (with $p > 2$) might achieve better dependence on the variance parameters than the upper bounds for Algorithm 2 or Algorithm 3, they cannot improve the ϵ^{-3} scaling. This highlights a fundamental limitation for higher-order methods in stochastic non-convex optimization which does not exist in the noiseless case. Indeed, without noise the optimal rate for finding ϵ-stationary point with a pth order method is $\Theta(\epsilon^{-1+\frac{1}{p}})$ [12]; we illustrate this contrast in Figure 1.

Altogether, the results presented in this section fully characterize (with respect to dependence on ϵ) the complexity of finding ϵ-stationary points with stochastic second-order methods and beyond in the single-point query model. We briefly remark that lower bound in (9) immediately extends to multi-point queries, which shows that even second-order methods offer little benefit once two or more simultaneous queries are allowed.

4. Complexity of finding second-order stationary points

Having established rates of convergence for finding ϵ-stationary points, we now turn our attention to (ϵ, γ)-second order stationary points, which have the additional requirement that $\lambda_{\min}(\nabla^2 F(x)) \geq -\gamma$, i.e. that F is γ-weakly convex around x. This section follows the general organization of the prequel: we first design and analyze an algorithm with improved upper bounds, and then develop nearly-matching lower bounds that apply to a broad class of algorithms.

4.1. Upper bounds

Our first contribution for this section is an algorithm that enjoys improved complexity for finding (ϵ, γ)-second-order stationary points, and that achieves this using only stochastic gradient and Hessian-vector product queries. To guarantee second-order stationarity, we follow the established technique of interleaving an algorithm for finding a first-order stationary point with negative curvature descent [10, 5]. However, we employ a randomized variant of this approach. Specifically, at every iteration we flip a biased coin to determine whether to perform a stochastic gradient step or a stochastic negative curvature descent step.

Our algorithm estimates stochastic gradients using the HVP-RVR scheme (Algorithm 1), where the value of the restart probability b depends on the type of the previous step (gradient or negative

curvature). To implement negative curvature descent, we apply Oja's method [32, 4] which detects directions of negative curvature using only stochastic Hessian-vector product queries. For technical reasons pertaining to the analysis of Oja's method, we require the stochastic Hessians to be bounded almost surely, i.e., $\|\widehat{\nabla^2 F}(x,z) - \nabla^2 F(x)\|_{\mathrm{op}} \le \bar{\sigma}_2$ a.s.; we let $\overline{\mathcal{O}}_2(F, \sigma_1, \bar{\sigma}_2)$ denote the class of such bounded noise oracles. Under this assumption, Algorithm 4—whose description is deferred to the Appendix G—enjoys the following convergence guarantee.[5]

Theorem 5 *For any function $F \in \mathcal{F}_2(\Delta, L_1, L_2)$, stochastic Hessian-vector product oracle in $\overline{\mathcal{O}}_2(F, \sigma_1, \bar{\sigma}_2)$, $\epsilon \le \min\{\sigma_1, \sqrt{\Delta L_1}\}$, and $\gamma \le \min\{\bar{\sigma}_2, L_1, \sqrt{\epsilon L_2}\}$, with probability at least $\frac{5}{8}$ Algorithm 4 returns a point $\widehat{x}$ such that*

$$\|\nabla F(\widehat{x})\| \le \epsilon \quad \text{and} \quad \lambda_{\min}\left(\nabla^2 F(\widehat{x})\right) \ge -\gamma,$$

and performs at most

$$\widetilde{O}\left(\frac{\Delta\sigma_1\bar{\sigma}_2}{\epsilon^3} + \frac{\Delta L_2\sigma_1\bar{\sigma}_2}{\gamma^2\epsilon^2} + \frac{\Delta L_2^2(\bar{\sigma}_2 + L_1)^2}{\gamma^5} + \frac{\Delta L_1}{\epsilon^2}\right)$$

stochastic gradient and Hessian-vector product queries.

Similar to the case for finding ϵ-stationary points (see discussion preceding Theorem 3), using full stochastic Hessian information allows us to design an algorithm (Algorithm 5) which removes the dependence on L_1 from the theorem above. Moreover, estimating negative curvature directly from empirical Hessian estimates saves us the need to use Oja's method, which means that we do not need the additional boundedness assumption on the stochastic Hessian used by Algorithm 4. We defer the complete description, complexity guarantee, and for analysis for Algorithm 5 to Appendix G.1.

4.2. Lower bounds

We now develop lower complexity bounds for the task of finding (ϵ, γ)-stationary points. To do so, we prove new lower bounds for the simpler sub-problem of finding a γ-*weakly convex* point, i.e., a point x such that $\lambda_{\min}(\nabla^2 F(x)) \ge -\gamma$ (with no restriction on $\|\nabla F(x)\|$). Lower bounds for finding (ϵ, γ)-SOSPs follow as the maximum (or, equivalently, the sum) of lower bounds we develop here and the lower bounds for finding ϵ-stationary points given in Theorem 6. To see why this is so, let F_ϵ and F_γ be hard instances for finding ϵ-stationary and γ-weakly-convex points respectively, and consider the "direct sum" $F_{\epsilon,\gamma}(x) := \frac{1}{2}F_\epsilon(x_1, \ldots, x_d) + \frac{1}{2}F_\gamma(x_{d+1}, \ldots, x_{2d})$; this is a hard instance for finding (ϵ, γ)-SOSPs that inherits all the regularity properties of its constituent functions.

The basic construction we use here is a modification of the zero-chain introduced in Carmon et al. [12] (see (74) in Appendix H) in which large $\lambda_{\min}(\nabla^2 F(x))$ is possible only when essentially none of the entries of x is zero. Given $T > 0$, we define the hard function

$$G_T(x) := \Psi(1)\Lambda(x_1) + \sum_{i=2}^{T} \left[\Psi(-x_{i-1})\Lambda(-x_i) + \Psi(x_{i-1})\Lambda(x_i)\right], \tag{11}$$

5. The notation $\widetilde{O}(\cdot)$ hides lower-order terms and logarithmic dependence on the dimension d. See the proof in Appendix G for the complete description of the algorithm and the full complexity bound, including lower order terms.

where $\Psi(x) := \exp(1 - \frac{1}{(2x-1)^2})\mathbf{1}\{x > \frac{1}{2}\}$ (as in Carmon et al. [12]) and $\Lambda(x) := 8(e^{\frac{-x^2}{2}} - 1)$.

Our design for the function Λ guarantees that any query whose last coordinate is zero has significant negative curvature, while maintaining the original chain structure which guarantees that zero-respecting algorithms require many queries before "discovering" the last coordinate. We complete the construction by specifying a collection of stochastic derivative estimators similar to those in Section 4.2, except for that we choose the stochastic gradient estimator $\widehat{\nabla G_T}$ to be exactly equal to ∇G_T, so that the lower bound holds even for $\sigma_1 = 0$; Appropriately scaling G_T allows us to tune the Lipschitz constants of its derivatives and the variance of the estimators, thereby establishing the following complexity bounds (see Appendix H.2 for a full derivation).

Theorem 6 *Let $p \geq 2$ and $\Delta, L_{1:p}, \sigma_{1:p} > 0$ be fixed. If $\gamma \leq O(\min\{\sigma_2, L_1\})$, then there exists $F \in \mathcal{F}_p(\Delta, L_{1:p})$ and $(\mathsf{O}_F^p, P_z) \in \mathcal{O}_p(F, \sigma_{1:p})$ such that for any stochastic pth-order zero-respecting algorithm, the number of queries to O_F^p required to obtain a γ-weakly convex point with constant probability is at least*

$$\Omega(1) \cdot \begin{cases} \frac{\Delta\sigma_2^2 L_2^2}{\gamma^5}, & p = 2, \\ \frac{\Delta\sigma_2^2}{\gamma^3} \min\left\{ \min_{q\in\{3,\dots,p\}} \left(\frac{\sigma_q}{\sigma_2}\right)^{\frac{2}{q-2}}, \min_{q'\in\{2,\dots,p\}} \left(\frac{L_{q'}}{\gamma}\right)^{\frac{2}{q'-1}} \right\}, & p > 2. \end{cases} \tag{12}$$

Theorem 6 is new even in the noiseless case (in which $\sigma_1 = \cdots = \sigma_p = 0$), where it specializes to $\frac{\Delta}{\gamma} \min_{q\in\{2,\dots,p\}} \left(\frac{L_q}{\gamma}\right)^{\frac{2}{q-1}}$. For the class $\mathcal{F}_p(\Delta, L_p)$, this further simplifies to $\Delta L_p^{\frac{2}{p-1}} \gamma^{-\frac{p+1}{p-1}}$, which is attained by the pth-order regularization method given in Cartis et al. [14, Theorem 3.6].

Together, these results characterize the deterministic complexity of finding γ-weakly convex points with noiseless pth-order methods.

Returning to the stochastic setting, the bound in Theorem 6, when combined with Theorem 4, implies the following oracle complexity lower bound bound for finding (ϵ, γ)-SOSPs with zero-respecting stochastic second-order methods ($p = 2$):

$$\Omega(1) \cdot \left(\min\left\{ \frac{\Delta\sigma_1\sigma_2}{\epsilon^3}, \frac{\Delta L_2^{0.5}\sigma_1}{\epsilon^{3.5}}, \frac{\Delta L_1\sigma_1^2}{\epsilon^4} \right\} + \frac{\Delta\sigma_2^2 L_2^2}{\gamma^5} \right). \tag{13}$$

Our lower bound matches the $\epsilon^{-3} + \gamma^{-5}$ terms in the upper bound given by Theorem 5, but does not match the mixed term $\epsilon^{-2}\gamma^{-2}$ appearing in the upper bound.[6] Overall, the rates match whenever $\gamma = \Omega(\epsilon^{0.5})$ or $\gamma = O(\epsilon^{2/3})$.

Theorem 6 is suggestive of another "elbow" phenomenon: In the stochastic regime, the rate does not improve beyond γ^{-3} for $p \geq 3$, while the optimal rate in the noiseless regime, $\gamma^{-\frac{p+1}{p-1}}$, continues improving for all p.[7] However, we are not yet aware of an algorithm using stochastic third-order information or higher that can achieve the γ^{-3} complexity bound.

Discussion

Due to space constraints, we defer conclusions and discussion to Appendix A.

6. Young's inequality only gives $\epsilon^{-3} + \gamma^{-5} \geq \Omega(\epsilon^{-9/5}\gamma^{-2})$.

7. Indeed, when high-order noise moments are assumed finite, the term $\min_{q\in\{3,\dots,p\}}(\sigma_q/\sigma_2)^{\frac{2}{q-2}}$ can longer be disregarded. This, in turn, implies that for sufficiently small γ, one cannot improve over γ^{-3}-scaling, as seen by (12).

Acknowledgements

We thank Blake Woodworth and Nati Srebo for helpful discussions. YA acknowledges partial support from the Sloan Foundation and Samsung Research. JCD acknowledges support from the NSF CAREER award CCF-1553086, ONR YIP N00014-19-2288, Sloan Foundation, NSF HDR 1934578 (Stanford Data Science Collaboratory), and the DAWN Consortium. DF acknowledges the support of TRIPODS award 1740751. KS acknowledges support from NSF CAREER Award 1750575 and a Sloan Research Fellowship.

References

[1] Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In *Conference On Learning Theory*, pages 774–792, 2018.

[2] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems*, pages 1165–1175, 2018.

[3] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems*, pages 2675–2686, 2018.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: Faster algorithms for matrix multiplicative weight updates. *International Conference on Machine Learning*, 2017.

[5] Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. *International Conference on Machine Learning*, 2018.

[6] Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *International Conference on Machine Learning*, pages 908–916, 2016.

[7] Yossi Arjevani and Ohad Shamir. Oracle complexity of second-order methods for finite-sum problems. In *Proceedings of the 34th International Conference on Machine Learning*, pages 205–213, 2017.

[8] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[9] Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1-2):327–360, 2019.

[10] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 654–663, 2017.

[11] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

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

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

[14] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *arXiv preprint arXiv:1708.04044*, 2017.

[15] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. *International Conference on Machine Learning*, 2020.

[16] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, 2019.

[17] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[18] Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 1192–1234, 2019.

[19] Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1319–1345, 2019.

[20] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[21] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

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

[23] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732, 2017.

[24] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

[25] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *Foundations of Computational Mathematics*, 2019.

[26] 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*, 42 (3):906–945, 2014.

[27] Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal*

on Optimization, 23(2):1092–1125, 2013.

[28] Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[29] Arkadi Nemirovski and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[30] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[31] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2613–2621, 2017.

[32] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[33] Barak A Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1): 147–160, 1994.

[34] Max Simchowitz. On the randomized complexity of minimizing a convex quadratic function. *arXiv preprint arXiv:1807.09386*, 2018.

[35] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[36] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2899–2908, 2018.

[37] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster stochastic variance reduction algorithms. In *Advances in Neural Information Processing Systems*, 2019.

[38] Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5530–5540, 2018.

[39] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3925–3936, 2018.

Contents of Appendix

Appendix A. Further discussion

This paper provides a fairly complete picture of the worst-case oracle complexity of finding stationary points with a stochastic second-order oracle: for ϵ-stationary points we characterize the leading term in ϵ^{-1} exactly and for (ϵ, γ)-SOSPs we characterize the leading term in γ^{-1} for a wide range of parameters. Nevertheless, our results point to a number of open questions.

Benefits of higher-order information for γ-weakly convex points. Our upper and lower bounds (in Theorem 20 and Theorem 6) resolve the optimal rate to find an (ϵ, γ)-stationary point for $p = 2$, i.e., when F is second-order smooth and the algorithm can query stochastic gradient and Hessian information. Furthermore, Theorem 4 shows that higher order information ($p \geq 3$) cannot improve the dependence of the rate on the first-order stationarity parameter ϵ. However, our lower bound for dependence on γ scales as γ^{-5} for $p = 2$, but scales as γ^{-3} for $p \geq 3$. The weaker lower bound for $p \geq 3$ leaves open the possibility of a stronger upper bound using third-order information or higher.

Global methods. For statistical learning and sample average approximation problems, it is natural to consider problem instances of the form $F(x) = \mathbb{E}\big[\widehat{F}(x, z)\big]$. For this setting, a more powerful oracle model is the *global oracle*, in which samples $z^{(1)}, \ldots, z^{(n)}$ are drawn i.i.d. and the learner observes the entire function $\widehat{F}(\cdot, z^{(t)})$ for each $t \in [n]$. Global oracles are more powerful than stochastic pth order oracles for every p, and lead to improved rates in the convex setting [19]. Is it possible to beat the ϵ^{-3} elbow for such oracles, or do our lower bounds extend to this setting?

Adaptivity and instance-dependent complexity. Our lower bounds show that stochastic higher-order methods cannot improve the ϵ^{-3} oracle complexity attained with stochastic gradients and Hessian-vector products. Furthermore, in the multi-point query model, stochastic second-order information does not even lead to improved rates over stochastic first-order information. However, these conclusions could be artifacts of our worst-case point of view—are there natural families of problem instances for which higher-order methods can adapt to additional problem structure and obtain stronger instance-dependent convergence guarantees? Developing a theory of instance-dependent complexity that can distinguish adaptive algorithms stands out as an exciting research prospect.

Appendix B. Detailed comparison with existing rates

Table 2 provides a detailed comparison between our upper bounds on the complexity of finding ϵ-stationary points and those of prior work.

Appendix C. Comparison: multi-point queries and mean-squared smoothness

Stochastic first-order methods that utilize variance reduction [24, 17, 39] employ the following *mean-squared smoothness* (MSS) assumption on the stochastic gradient estimator:

$$\mathbb{E}\,\|\widehat{\nabla F}(x, z) - \widehat{\nabla F}(y, z)\|^2 \leq \bar{L}^2\|x - y\|^2 \;\text{ for all }\; x, y \in \mathbb{R}^d.$$

Since $\mathbb{E}[\widehat{\nabla F}(x, z)] = \nabla F(x)$, this is equivalent to assuming

$$\mathbb{E}\,\|\widehat{\nabla F}(x, z) - \widehat{\nabla F}(y, z) - (\nabla F(x) - \nabla F(y))\|^2 \leq \sigma_{\mathrm{mss}}^2\|x - y\|^2 \;\text{ for all }\; x, y \in \mathbb{R}^d, \qquad (14)$$

Method	Uses $\widehat{\nabla^2 F}$?	Complexity bound	Additional assumptions
SGD [22]	No	$O(\Delta L_1 \sigma_1^2 \epsilon^{-4})$	
Restarted SGD [18]	No	$O(\Delta L_2^{0.5} \sigma_1^2 \epsilon^{-3.5})^\dagger$	$\widehat{\nabla F}$ Lipschitz almost surely
Normalized SGD [15]	No	$O(\Delta L_2^{0.5} \sigma_1^2 \epsilon^{-3.5})^\dagger$	
Subsampled regularized Newton [36]	Yes*	$O(\Delta L_2^{0.5} \sigma_1^2 \epsilon^{-3.5})^\dagger$	
Recursive variance reduction [e.g., 17]	No	$O(\Delta \sigma_1 \sigma_{\text{mss}} \epsilon^{-3} + \Delta L_1 \epsilon^{-2})$	Mean-squared smoothness $\sigma_{\text{mss}} \le \sigma_2$, simultaneous queries (Appendix C)
SGD with HVP-RVR (Algorithm 2)	Yes*	$O(\Delta \sigma_1 \sigma_2 \epsilon^{-3} + \Delta L_2^{0.5} \sigma_1 \epsilon^{-2.5} + \Delta L_1 \epsilon^{-2})$	
Subsampled Newton with HVP-RVR (Algorithm 3)	Yes	$O(\Delta \sigma_1 \sigma_2 \epsilon^{-3} + \Delta L_2^{0.5} \sigma_1 \epsilon^{-2.5} + \Delta \sigma_2 \epsilon^{-2})$	

Table 2: Detailed comparison of guarantees for finding ϵ-stationary points (satisfying $\mathbb{E}\|\nabla F(x)\| \le \epsilon$) for a function F with L_1-Lipschitz gradients and L_2-Lipschitz Hessian. Here Δ is the initial optimality gap, and σ_p is the variance of $\widehat{\nabla^p F}$. Algorithms marked with * require only stochastic Hessian-vector products. Complexity bounds marked with † only show leading order term in ϵ.

for some $\sigma_{\text{mss}} < \bar{L}$. In fact, while it always holds that $\bar{L}^2 \le L_1^2 + \sigma_{\text{mss}}^2$, inspection of the results of Fang et al. [17], Wang et al. [37] shows one can replace $\bar{L}$ with σ_{mss} in the leading terms of their complexity bounds without any change to the algorithms.

Algorithms that take advantage of the MSS structure rely on the following additional *simultaneous query* assumption (which is a special case of (4) for $n = 2$):

$$\textit{We may query } x, y \in \mathbb{R}^d \textit{ and observe } \mathsf{O}_F^1(x, z) \textit{ and } \mathsf{O}_F^1(y, z) \textit{ for the same draw of } z \sim P_z. \quad (15)$$

In empirical risk minimization problems, z represents the datapoint index and possibly data augmentation parameters, and the value of z is typically part of the query, which means that assumption (15) indeed holds. In certain online learning settings, however, the assumption can fail. For example, the variable z could represent the instantaneous power demands in an electric grid, and testing two grid configurations for the same grid state might be impractical.

We observe that assuming access to both an MSS gradient estimator and simultaneous two-point queries is stronger than assuming a bounded variance stochastic Hessian-vector product estimator. This holds because the former allows us to simulate the latter with finite differencing. Formally, we have the following.

Observation 1 *Let F have L_2-Lipschitz Hessian, let $\widehat{\nabla F}$ satisfy (14), and assume we have access to a two-point query oracle as in (15). Then, for any $\delta > 0$ and every unit-norm vector u, the Hessian-vector product estimator*

$$\widehat{\nabla^2 F}_\delta(x,z)u := \frac{1}{\delta}\left[\widehat{\nabla F}(x+\delta\cdot u,z) - \widehat{\nabla F}(x,z)\right] \tag{16}$$

satisfies

$$\left\|\mathbb{E}[\widehat{\nabla^2 F}_\delta(x,z)u] - \nabla^2 F(x)u\right\| \le \frac{L_2\delta}{2} \;\; \text{and} \;\; \mathbb{E}\left\|\widehat{\nabla^2 F}_\delta(x,z)u - \nabla^2 F(x)u\right\|^2 \le \sigma_{\mathrm{mss}}^2 + \frac{L_2^2\delta^2}{4}.$$

Proof. We have $\mathbb{E}[\widehat{\nabla^2 F}_\delta(x,z)u] = \frac{1}{\delta}[\nabla F(x+\delta\cdot u) - \nabla F(x)]$, and by Lipschitz continuity of $\nabla^2 F$,

$$\|\nabla F(x+\delta\cdot u) - \nabla F(x) - \nabla^2 F(x)[\delta u]\| \le \frac{L_2}{2}\delta^2\|u\|^2 = \frac{L_2}{2}\delta^2,$$

which implies the bound on the bias. To bound the variance, we note that

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\nabla^2 F}_\delta(x,z)u - \mathbb{E}[\widehat{\nabla^2 F}_\delta(x,z)u]\right\|^2 \\
&\quad\le \frac{1}{\delta^2}\mathbb{E}\left\|\widehat{\nabla F}(x+\delta u,z) - \widehat{\nabla F}(x,z) - [\nabla F(x+\delta u) - \nabla F(x)]\right\|^2 \le \frac{1}{\delta^2}\cdot\sigma_2^2\|\delta u\|^2 = \sigma_{\mathrm{mss}}^2,
\end{aligned}$$

by the MSS property (14). ∎

We conclude from Observation 1 that Algorithm 2, which only requires stochastic Hessian-vector products, attains $O(\epsilon^{-3})$ complexity under assumptions no stronger than previous algorithms. In fact, we show now that our assumptions are in fact strictly weaker than prior work. That is, while an MSS gradient estimator implies a bounded variance Hessian estimator, the opposite is not true in general. This is simply due to the fact that in our oracle model, $\widehat{\nabla F}$ and $\widehat{\nabla^2 F}$ can be completely unrelated. Consider for example the case where P_z is uniform on $\{-1,1\}$ and

$$\widehat{\nabla F}(x,z) = \begin{cases} \nabla F(x) + \frac{x}{\|x\|}z & x \neq 0 \\ \nabla F(x) & x = 0, \end{cases} \quad \text{while} \;\; \widehat{\nabla^2 F}(x,z) = \nabla^2 F(x).$$

Clearly $\widehat{\nabla F}$ is not MSS, even though $\widehat{\nabla^2 F}$ has zero variance.

There is, however, an important setting where bounded variance for $\widehat{\nabla^2 F}$ *does* imply that $\widehat{\nabla F}$ is MSS. Suppose that the derivative of $\widehat{\nabla F}(x,z)$ exists, and has the form

$$\nabla[\widehat{\nabla F}(x,z)] = \widehat{\nabla^2 F}(x,z). \tag{17}$$

That is, the Hessian estimator is the Jacobian of the gradient estimator. In this case, bounded variance for the Hessian estimator implies mean-squared smoothness.

Observation 2 *Let F have gradient and Hessian estimators $\widehat{\nabla F}$ and $\widehat{\nabla^2 F}$ satisfying (3) and (17). Then $\widehat{\nabla F}$ has the MSS property (14) with $\sigma_{\mathrm{mss}} \le \sigma_2$.*

Proof. Under the property (17), we have

$$\widehat{\nabla F}(x,z) - \widehat{\nabla F}(y,z) - [\nabla F(x) - \nabla F(y)] \\ = \int_0^1 \Big(\widehat{\nabla^2 F}(xt + y(1-t), z) - \nabla^2 F(xt + y(1-t))\Big)(x-y)dt.$$

Taking the squared norm, applying Jensen's inequality, and substituting the variance bound (3) gives the MSS property (14). ■

The property (17) holds for empirical risk minimization, where we have the more general relation $\widehat{\nabla^p F}(x,z) = \nabla^p \widehat{F}(x,z)$ for any p; That is, all the stochastic derivative estimators are themselves the derivatives of a single stochastic function. Therefore, by Observation 1 and Observation 2, in empirical risk minimization settings, mean-square smoothness is essentially equivalent to bounded variance of the stochastic Hessian estimator.

Appendix D. Variance-reduced gradient estimator (HVP-RVR)

In this section we prove Lemma 1. First, we formally describe the protocol in which our optimization algorithms query the gradient estimator HVP-RVR-Gradient-Estimator described in Algorithm 1, and define some additional notation.

Given a function $F \in \mathcal{F}_2(\Delta, L_1, L_2)$ and a stochastic second-order oracle in $\mathcal{O}_2(F, \sigma_{1:2})$, the optimization algorithm interacts with HVP-RVR-Gradient-Estimator by sequentially querying points $\{x^{(t)}\}_{t=1}^{\infty}$ with reset probabilities $\{b^{(t)}\}_{t=1}^{\infty}$, to obtain estimates $g^{(t)}$ for $\nabla F(x^{(t)})$ for each time t; that is,

$$x^{(t)} = \mathsf{A}^{(t)}(g^{(0)}, g^{(1)}, \ldots, g^{(t-1)}; r^{(t-1)}),\ b^{(t)} = \mathsf{B}^{(t)}(r^{(t-1)}),\ \text{ and} \\ g^{(t)} = \textsf{HVP-RVR-Gradient-Estimator}_{\epsilon, b^{(t)}}(x^{(t)}, x^{(t-1)}, g^{(t-1)}), \tag{18}$$

where $\mathsf{A}^{(t)}, \mathsf{B}^{(t)}$ are measurable mappings modeling the optimization algorithm and $\{r^{(t)}\}$ is an independent sequence of random seeds.[8] That is, Lemma 1 holds for any sequence of queries where $x^{(t)}, b^{(t)}$ are adapted to the filtration

$$\mathcal{G}^{(t)} = \sigma\Big(\{g^{(j)}, r^{(j)}\}_{j<t}\Big).$$

but $b^{(t)}$ is independent of $\mathcal{G}^{(t-1)}$ and $g^{(t-1)}$.

Lemma 1 is an immediate consequence of Lemma 7 and Lemma 8, proven below, which respectively establish the estimator's error and complexity bounds.

Lemma 7 *Given a function $F \in \mathcal{F}_2(\Delta, \infty, L_2)$, a stochastic oracle in $\mathcal{O}_2(F, \sigma_{1:2})$, and initial points $x^{(0)}$ and $g^{(0)} = \perp$, let $\{g^{(t)}\}_{t\geq 0}$ denote the sequence of gradient estimates at $\{x^{(t)}\}_{t\geq 0}$ respectively, returned by HVP-RVR-Gradient-Estimator under the protocol (18). Then, for all $t \geq 1$,*

$$\mathbb{E}\left\|g^{(t)} - \nabla F(x^{(t)})\right\|^2 \leq \epsilon^2.$$

8. This level of formalism is not used within the proof, but we include it here for clarity.

Proof. We prove that

$$\mathbb{E}\left\|g^{(t)} - \nabla F(x^{(t)})\right\|^2 \leq \left(1 - \frac{\mathbb{E}[b^{(t)}]}{2}\right)\mathbb{E}\left\|g^{(t-1)} - \nabla F(x^{(t-1)})\right\|^2 + \frac{\mathbb{E}[b^{(t)}]}{2}\epsilon^2,$$

whence the result follows by a simple induction whose basis is

$$\mathbb{E}\left\|g^{(1)} - \nabla F(x^{(1)})\right\|^2 \leq \frac{\sigma_1^2}{n} \leq \epsilon^2.$$

Let $C^{(t)}$ denote the value of the coin toss in the t^{th} call to Algorithm 1 (Line 3), recalling that $C^{(t)} \sim \text{Bernoulli}(b^{(t)})$. Writing $\mathfrak{e}^{(t)} = g^{(t)} - \nabla F(x^{(t)})$ for brevity, we have

$$\mathbb{E}\left[\left\|\mathfrak{e}^{(t)}\right\|^2 \,\middle|\, b^{(t)}\right] = b^{(t)}\,\mathbb{E}\left[\left\|\mathfrak{e}^{(t)}\right\|^2 \,\middle|\, C^{(t)} = 1\right] + (1 - b^{(t)})\,\mathbb{E}\left[\left\|\mathfrak{e}^{(t)}\right\|^2 \,\middle|\, C^{(t)} = 0\right]. \tag{19}$$

Clearly,

$$\mathbb{E}\left[\left\|\mathfrak{e}^{(t)}\right\|^2 \,\middle|\, C^{(t)} = 1\right] \leq \frac{\sigma_1^2}{n} = \frac{\epsilon^2}{5}. \tag{20}$$

Moreover, conditional on $C^{(t)} = 0$, we have from the definition of the gradient estimator that

$$\mathfrak{e}^{(t)} = \mathfrak{e}^{(t-1)} + \psi^{(t)},$$

where

$$\psi^{(t)} := \sum_{k=1}^{K^{(t)}} \widehat{\nabla^2 F}(x^{(t,k-1)}, z^{(t,k)})\left(x^{(t,k)} - x^{(t,k-1)}\right) - \nabla F(x^{(t)}) + \nabla F(x^{(t-1)}),$$

and

$$K^{(t)} = \left\lceil \frac{5\left(\sigma_2^2 + L_2\epsilon\right)}{b^{(t)}\epsilon^2} \cdot \|x^{(t)} - x^{(t-1)}\|^2 \right\rceil, \tag{21}$$

where $x^{(t,k)}$ and $x^{(t,k)}$ respectively denote the values of $x^{(k)}$ and $z^{(k)}$ (defined on Line 8) during the t^{th} call to Algorithm 1.

We may therefore decompose the error conditional on $C^{(t)} = 0$ as

$$\begin{aligned}\mathbb{E}\left[\left\|\mathfrak{e}^{(t)}\right\|^2 \,\middle|\, C^{(t)} = 0\right] &\overset{(i)}{=} \mathbb{E}\left\|\mathfrak{e}^{(t-1)} + \mathbb{E}\left[\psi^{(t)} \mid \mathcal{G}^{(t)}\right]\right\|^2 + \mathbb{E}\left\|\psi^{(t)} - \mathbb{E}\left[\psi^{(t)} \mid \mathcal{G}^{(t)}\right]\right\|^2 \\ &\overset{(ii)}{\leq} \mathbb{E}\left[\left(1 + \frac{b^{(t)}}{2}\right)\left\|\mathfrak{e}^{(t-1)}\right\|^2\right] + \mathbb{E}\left[\left(1 + \frac{2}{b^{(t)}}\right)\left\|\mathbb{E}\left[\psi^{(t)} \mid \mathcal{G}^{(t)}\right]\right\|^2\right] + \mathbb{E}\left\|\psi^{(t)} - \mathbb{E}\left[\psi^{(t)} \mid \mathcal{G}^{(t)}\right]\right\|^2,\end{aligned} \tag{22}$$

where (i) is due to $\mathfrak{e}^{(t-1)} \in \mathcal{G}^{(t)}$ and (ii) is due to Young's inequality.

The facts that $z^{(t,k)}$ is independent from $\mathcal{G}^{(t)}$, that $\nabla F(x^{(t)}) - \nabla F(x^{(t-1)}) \in \mathcal{G}^{(t)}$, and that $\widehat{\nabla^2 F}(\cdot)$ is unbiased give

$$\mathbb{E}\left[\psi^{(t)} \,\middle|\, \mathcal{G}^{(t)}\right] = \sum_{k=1}^{K^{(t)}} \nabla^2 F(x^{(t,k-1)})\left(x^{(t,k)} - x^{(t,k-1)}\right) - \nabla F(x^{(t)}) + \nabla F(x^{(t-1)})$$

for every t. Consequently, the scaling (21) and Hessian estimator variance bound imply

$$\begin{aligned}
&\mathbb{E}\left[\left\|\psi^{(t)} - \mathbb{E}\left[\psi^{(t)} \mid \mathcal{G}^{(t)}\right]\right\|^2 \mid \mathcal{G}^{(t)}\right] \\
&\quad\overset{(\star)}{=} \frac{1}{(K^{(t)})^2}\sum_{k=1}^{K^{(t)}} \mathbb{E}\left[\left\|(\widehat{\nabla^2 F}(x^{(t,k-1)}, z^{(t,k)}) - \nabla^2 F(x^{(t,k-1)}))(x^{(t)} - x^{(t-1)})\right\|^2 \mid \mathcal{G}^{(t)}\right] \\
&\quad\le \frac{1}{(K^{(t)})^2}\sum_{k=1}^{K^{(t)}} \mathbb{E}\left[\left\|\widehat{\nabla^2 F}(x^{(t,k-1)}, z^{(t,k)}) - \nabla^2 F(x^{(t,k-1)})\right\|_{\mathrm{op}}^2 \mid \mathcal{G}^{(t)}\right]\left\|x^{(t)} - x^{(t-1)}\right\|^2 \\
&\quad\le \sigma_2^2 \cdot \frac{\|x^{(t)} - x^{(t-1)}\|^2}{K^{(t)}} \le b^{(t)} \cdot \frac{\epsilon^2}{5},
\end{aligned} \tag{23}$$

where the equality $(\star)$ above is due to the fact that $z^{(t,1)}, \ldots, z^{(t,K^{(t)})}$ are i.i.d., as well as $x^{(t,k)} - x^{(t,k-1)} = \frac{1}{K^{(t)}}(x^{(t)} - x^{(t-1)})$.

Next, we observe that Taylor's theorem and fact that F has L_2-Lipschitz Hessian implies that $\|\nabla F(x') - \nabla F(x) - \nabla^2(x)F(x' - x)\| \le \frac{L_2}{2}\|x' - x\|^2$ for all $x, x' \in \mathbb{R}^d$. Therefore,

$$\begin{aligned}
\left\|\mathbb{E}\left[\psi^{(t)} \;\middle|\; \mathcal{G}^{(t)}\right]\right\| &= \left\|\sum_{k=1}^{K^{(t)}} \nabla F(x^{(t,k)}) - \nabla F(x^{(t,k-1)}) - \nabla^2 F(x^{(t,k-1)})\left(x^{(t,k)} - x^{(t,k-1)}\right)\right\| \\
&\le \sum_{k=1}^{K^{(t)}} \left\|\nabla F(x^{(t,k)}) - \nabla F(x^{(t,k-1)}) - \nabla^2 F(x^{(t,k-1)})\left(x^{(t,k)} - x^{(t,k-1)}\right)\right\| \\
&\le K^{(t)} \cdot \frac{L_2}{2} \cdot \left(\frac{\|x^{(t)} - x^{(t-1)}\|}{K^{(t)}}\right)^2 \le b^{(t)} \cdot \frac{\epsilon}{50},
\end{aligned} \tag{24}$$

where we used (21) again.

Substituting back through equations (24), (23), (22), (20) and (19), we have

$$\begin{aligned}
\mathbb{E}\left\|\mathfrak{e}^{(t)}\right\|^2 &\le \mathbb{E}\left[b^{(t)} \cdot \tfrac{\epsilon^2}{5} + (1 - b^{(t)})\left((1 + \tfrac{b^{(t)}}{2})\left\|\mathfrak{e}^{(t-1)}\right\|^2 + (1 + \tfrac{2}{b^{(t)}})(\tfrac{b^{(t)}\epsilon}{50})^2 + b^{(t)} \cdot \tfrac{\epsilon^2}{5}\right)\right] \\
&\le \left(1 - \tfrac{\mathbb{E}[b^{(t)}]}{2}\right)\mathbb{E}\left\|g^{(t-1)} - \nabla F(x^{(t-1)})\right\|^2 + \tfrac{\mathbb{E}[b^{(t)}]}{2}\epsilon^2 \le \epsilon^2,
\end{aligned}$$

as required; the second inequality follows from algebraic manipulation and the fact that $\mathfrak{e}^{(t-1)}$ is independent of $b^{(t)}$ by assumption. ■

The following lemma bounds the number of oracle queries made per call to the gradient estimator.

Lemma 8 *The expected number of stochastic oracle queries made by HVP-RVR-Gradient-Estimator when called a single time with arguments $(x, x_{\mathrm{prev}}, g_{\mathrm{prev}})$ and parameters (ϵ, b) is at most*

$$6\left(1 + \frac{b\sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \|x - x_{\mathrm{prev}}\|^2}{b\epsilon^2}\right).$$

Proof. Let m denote the number of oracle calls made by the gradient estimator when invoked with arguments $(x, x_{\text{prev}}, g_{\text{prev}})$. For any call to the estimator, there are two cases, either (a) $C = 1$, or (b) $C = 0$. In the first case, the gradient estimator queries the oracle n times at the point x and returns the empirical average of the returned stochastic estimates (see Line 5 in Algorithm 1). Thus, $m = n$ for this case. In the second case, the estimator queries the oracle once for each point in the set $\left(x^{(k-1)}\right)_{k=1}^{K}$, and updates the gradient using a stochastic path integral as in Line 8. Thus, $m = K$ for this case.

Combining the two cases, using $C \sim \text{Bernoulli}(b)$ and substituting in the values of n and K, we get

$$\begin{aligned}
\mathbb{E}[m] &= \Pr(C = 1)\,\mathbb{E}[m \mid C = 1] + \Pr(C = 0)\,\mathbb{E}[m \mid C = 0] \\
&= \mathbb{E}[b \cdot n + (1 - b) \cdot K] \\
&= \left\lceil \frac{5b\sigma_1^2}{\epsilon^2} \right\rceil + \left\lceil \frac{5(\sigma_2^2 + L_2\epsilon) \cdot \|x - x_{\text{prev}}\|^2}{b\epsilon^2} \right\rceil \\
&\leq 6\left(\frac{b\sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \|x - x_{\text{prev}}\|^2}{b\epsilon^2} + 1 \right),
\end{aligned}$$

where the final inequality follows from $\lceil x \rceil \leq x + 1$. ■

Appendix E. Supporting technical results

E.1. Error bound for empirical Hessian

In order to find the negative curvature direction at a given point or to build a cubic regularized sub-model, Algorithm 3 estimates the Hessian by computing an empirical average of the stochastic Hessian queries to the oracle. The following lemma is a standard result which bounds the expected error for the empirical Hessian.

Lemma 9 *Given a function $F \in \mathcal{F}_2(\Delta, \infty, L_2)$, a stochastic oracle in $\mathcal{O}_2(F, \sigma_{1:2})$ and a point x, let $H := \frac{1}{m}\sum_{i=1}^{m} \widehat{\nabla^2 F}(x, z^{(i)})$ denote the empirical Hessian at the point x estimated using m stochastic queries at x, where $z^{(i)} \overset{\text{i.i.d.}}{\sim} P_z$. Then*

$$\mathbb{E}\left[\left\| H - \nabla^2 F(x) \right\|_{\text{op}}^2 \right] \leq \frac{22\sigma_2^2 \log(d)}{m}.$$

Proof. This is an immediate consequence of Lemma 10 below, using $A_i := \widehat{\nabla^2 F}(x, z^{(i)})$ and $B := \nabla^2 F(x)$.

■

Lemma 10 *Let $(A_i)_{i=1}^n$ be a collection of i.i.d. matrices in $\mathbb{S}^d$, with $\mathbb{E}[A_i] = B$ and $\mathbb{E}\|A_i - B\|_{\text{op}}^2 \leq \sigma^2$. Then it holds that*

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} A_i - B \right\|_{\text{op}}^2 \leq \frac{22\sigma^2 \log d}{n}.$$

Proof. We drop the normalization by n throughout this proof. We first symmetrize. Observe that by Jensen's inequality we have

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^n A_i - B\right\|_{\mathrm{op}}^2 &\le \mathbb{E}_A \mathbb{E}_{A'}\left\|\sum_{i=1}^n A_i - A_i'\right\|_{\mathrm{op}}^2 \\ &= \mathbb{E}_A \mathbb{E}_{A'}\left\|\sum_{i=1}^n (A_i - B) - (A_i' - B)\right\|_{\mathrm{op}}^2 \\ &= \mathbb{E}_A \mathbb{E}_{A'} \mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i((A_i - B) - (A_i' - B))\right\|_{\mathrm{op}}^2 \le 4\,\mathbb{E}_A \mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i(A_i - B)\right\|_{\mathrm{op}}^2,\end{aligned}$$

where $(A')_{i=1}^n$ is a sequence of independent copies of $(A_i)_{i=1}^n$ and $(\epsilon_i)_{i=1}^n$ are Rademacher random variables. Henceforth we condition on A. Let $p = \log d$, and let $\|\cdot\|_{S_p}$ denote the Schatten p-norm. In what follows, we will use that for any matrix X, $\|X\|_{\mathrm{op}} \le \|X\|_{S_{2p}} \le e^{1/2}\|X\|_{\mathrm{op}}$. To begin, we have

$$\mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i(A_i - B)\right\|_{\mathrm{op}}^2 \le \mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i(A_i - B)\right\|_{S_{2p}}^2 \le \left(\mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i(A_i - B)\right\|_{S_{2p}}^{2p}\right)^{1/p},$$

where the second inequality follows by Jensen. We now apply the matrix Khintchine inequality [26, Corollary 7.4], which implies that

$$\begin{aligned}\left(\mathbb{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i(A_i - B)\right\|_{S_{2p}}^{2p}\right)^{1/p} &\le (2p-1)\left\|\sum_{i=1}^n (A_i - B)^2\right\|_{S_{2p}} \le (2p-1)\sum_{i=1}^n \|(A_i - B)\|_{S_{2p}}^2 \\ &\le e(2p-1)\sum_{i=1}^n \|(A_i - B)\|_{\mathrm{op}}^2.\end{aligned}$$

Putting all the developments so far together and taking expectation with respect to A, we have

$$\mathbb{E}\left\|\sum_{i=1}^n A_i - B\right\|_{\mathrm{op}}^2 \le 4e(2p-1)\sum_{i=1}^n \mathbb{E}_{A_i}\|(A_i - B)\|_{\mathrm{op}}^2 \le 4e(2p-1)n\sigma^2.$$

To obtain the final result we normalize by n^2. ∎

E.2. Descent lemma for stochastic gradient descent

The following lemma characterizes the effect of gradient descent update step used by Algorithm 2 and Algorithm 4.

Lemma 11 *Given a function $F \in \mathcal{F}_2(\Delta, L_1, \infty)$, a point x, and gradient estimator g at x, define*

$$y := x - \eta g.$$

Then, for any $\eta \le \frac{1}{2L_1}$, the point y satisfies

$$F(x) - F(y) \ge \frac{\eta}{8}\|\nabla F(x)\|^2 - \frac{3\eta}{4}\|\nabla F(x) - g\|^2.$$

Proof. Since, the gradient of F is L_1-Lipschitz, we have

$$\begin{aligned}
F(y) &\le F(x) + \langle \nabla F(x), y - x\rangle + \frac{L_1}{2}\|y-x\|^2 \\
&\overset{(i)}{=} F(x) - \eta\langle \nabla F(x), g\rangle + \frac{L_1\eta^2}{2}\|g\|^2 \\
&= F(x) - \eta\langle \nabla F(x) - g, g\rangle - \eta\|g\|^2 + \frac{L_1\eta^2}{2}\|g\|^2 \\
&\overset{(ii)}{\le} F(x) + \eta\|\nabla F(x) - g\|\|g\| - \eta\left(1 - \frac{L_1\eta}{2}\right)\|g\|^2 \\
&\overset{(iii)}{\le} F(x) + \frac{\eta}{2}\|\nabla F(x) - g\|^2 - \eta\left(\frac{1}{2} - \frac{L_1\eta}{2}\right)\|g\|^2 \\
&\overset{(iv)}{\le} F(x) + \frac{\eta}{2}\|\nabla F(x) - g\|^2 - \frac{\eta}{4}\|g\|^2 \\
&\overset{(v)}{\le} F(x) + \frac{3\eta}{4}\|\nabla F(x) - g\|^2 - \frac{\eta}{8}\|\nabla F(x)\|^2, \qquad (25)
\end{aligned}$$

where (i) uses that $y - x = \eta g$, (ii) is due to the Cauchy-Schwarz inequality, (iii) is given by an application of the AM-GM inequality and (iv) holds because $\eta \le \frac{1}{2L_1}$. Finally, (v) follows by invoking Jensen's inequality for the function $\|\cdot\|^2$ to upper bound $\|\nabla F(x)\|^2 \le 2\Big(\|\nabla F(x-g)\|^2 + \|g\|^2\Big)$. Rearranging the terms in (25), we get,

$$F(x) - F(y) \ge \frac{\eta}{8}\|\nabla F(x)\|^2 - \frac{3\eta}{4}\|\nabla F(x) - g\|^2.$$

■

E.3. Descent lemma for cubic-regularized trust-region method

The following lemmas establish properties for the updates step involving constrained minimization of the cubic regularized model in used in Algorithm 3.

Lemma 12 *Given a function $F \in \mathcal{F}_2(\Delta, \infty, L_2)$, gradient estimator $g \in \mathbb{R}^d$ and hessian estimator $H \in \mathbb{S}^d$, define*

$$m_x(y) = F(x) + \langle g, y - x\rangle + \frac{H}{2}[y-x, y-x] + \frac{M}{6}\|y-x\|^3,$$

and let $y \in \arg\min_{z \in \mathbb{B}_\eta(x)} m_x(z)$. Then, for any $M \ge 4L_2$ and $\eta \ge 0$, the point y satisfies

$$F(x) - F(y) \ge \frac{M}{12}\|y-x\|^3 - \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{4\eta^{\frac{3}{2}}}{\sqrt{M}}\big\|\nabla^2 F(x) - H\big\|^{\frac{3}{2}}.$$

Proof. Since $\nabla^2 F$ is L_2-Lipschitz, we have

$$F(y) - F(x) \le F(x) + \langle \nabla F(x), y - x\rangle + \frac{1}{2}\nabla^2 F(x)[y-x, y-x] + \frac{L_2}{6}\|y-x\|^3 - F(x)$$

$$\begin{aligned}
&\overset{(i)}{=} m_x(y) + \frac{L_2 - M}{6}\|y-x\|^3 + \langle \nabla F(x) - g, y - x\rangle + \frac{1}{2}\nabla^2 F(x)[y-x, y-x] \\
&\quad - \frac{1}{2}H[y-x, y-x] - m_x(x) \\
&\overset{(ii)}{\leq} -\frac{M}{8}\|y-x\|^3 + \|\nabla F(x) - g\|\|y-x\| + \frac{1}{2}\left\|\nabla^2 F(x)[y-x,\cdot] - H[y-x,\cdot]\right\|\|y-x\|,
\end{aligned} \tag{26}$$

where (i) follows from the definition of $m_x(\cdot)$ and (ii) follows by the fact that $y \in \arg\min_{y' \mathbb{B}_\eta(x)} m_x(y')$, along with an application of the Cauchy-Schwarz inequality for remainder of the terms, and because $M \geq 4L_2$. Additionally, using Young's inequality, we have

$$\|\nabla F(x) - g\|\|y-x\| \leq \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{M}{64}\|y-x\|^3,$$

and,

$$\left\|\nabla^2 F(x)[y-x,\cdot] - H[y-x,\cdot]\right\|\|y-x\| \leq \frac{8}{\sqrt{M}}\left\|\nabla^2 F(x)[y-x,\cdot] - H[y-x,\cdot]\right\|^{\frac{3}{2}} + \frac{M}{64}\|y-x\|^3.$$

Plugging these bounds into (26), we have

$$\begin{aligned}
F(y) - F(x) &\leq -\frac{M}{12}\|y-x\|^3 + \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{4}{\sqrt{M}}\left\|\nabla^2 F(x)[y-x,\cdot] - H[y-x,\cdot]\right\|^{\frac{3}{2}} \\
&\overset{(i)}{\leq} -\frac{M}{12}\|y-x\|^3 + \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{4}{\sqrt{M}}\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}^{\frac{3}{2}}\|y-x\|^{\frac{3}{2}} \\
&\overset{(ii)}{\leq} -\frac{M}{12}\|y-x\|^3 + \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{4}{\sqrt{M}}\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}^{\frac{3}{2}} \cdot \eta^{\frac{3}{2}},
\end{aligned}$$

where (i) follows by the definition of the operator norm and (ii) follows by observing that $\|y-x\| \leq \eta$. Rearranging the terms, we have

$$F(x) - F(y) \geq \frac{M}{12}\|y-x\|^3 - \frac{8}{\sqrt{M}}\|\nabla F(x) - g\|^{\frac{3}{2}} + \frac{4\eta^{\frac{3}{2}}}{\sqrt{M}}\left\|\nabla^2 F(x) - H\right\|^{\frac{3}{2}}.$$

■

Lemma 13 *Under the same setting as Lemma 12, the point y satisfies*

$$\mathbf{1}\left\{\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right\} \leq \frac{2}{\eta^2}\|y-x\|^2 + \frac{2}{M\eta^2}\left(\|\nabla F(x) - g\| + \eta\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}\right).$$

Proof. There are two scenarios: (i) either y lies on the boundary of $\mathbb{B}_\eta(x)$, or (ii) y is in the interior of $\mathbb{B}_\eta(x)$. In the first case, $\|y-x\| = \eta$. In the second case,

$$\begin{aligned}
\|\nabla F(y)\| &\overset{(i)}{\leq} \left\|\nabla F(y) - \nabla F(x) - \nabla^2 F(x)[y-x,\cdot]\right\| + \left\|\nabla F(x) + \nabla^2 F(x)[y-x,\cdot]\right\| \\
&\overset{(ii)}{\leq} \frac{L_2}{2}\|y-x\|^2 + \left\|\nabla F(x) + \nabla^2 F(x)[y-x,\cdot]\right\|
\end{aligned}$$

$$\overset{(iii)}{\leq} \frac{L_2}{2}\|y-x\|^2 + \|\nabla F(x) - g\| + \left\|\nabla^2 F(x)[y-x,\cdot] - H[y-x,\cdot]\right\| + \|g + H[y-x,\cdot]\|$$
$$\overset{(iv)}{\leq} \frac{L_2}{2}\|y-x\|^2 + \|\nabla F(x) - g\| + \left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}} \cdot \eta + \|g + H[y-x,\cdot]\|$$
$$\overset{(v)}{\leq} \frac{L_2 + M}{2}\|y-x\|^2 + \|\nabla F(x) - g\| + \left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}} \cdot \eta, \tag{27}$$

where (i) follows by triangle inequality, (ii) follows by Taylor expansion of $\nabla F(y)$ at x and observing that F is L_2-hessian Lipschitz, (iii) follows by another application of the triangle inequality, (iv) follows from Cauchy-Schwarz inequality and observing that $\|y-x\| \leq \eta$, and (v) follows by using first order optimization conditions for $y \in \arg\min_{\mathbb{B}_\eta(x)} m_x(y)$, i.e.,

$$\|\nabla \widehat{m}_x(y)\| = 0, \quad \text{or,} \quad g + H[y-x,\cdot] + \frac{M}{2}\|y-x\|(y-x) = \mathbf{0}.$$

Rearranging the terms in (27), we get,

$$\|y-x\|^2 \geq \frac{2}{L_2+M}\Big(\|\nabla F(y)\| - \|\nabla F(x) - g\| - \left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}} \cdot \eta\Big).$$

Since one of the two cases ($\|y-x\| < \eta$ or $\|y-x\| = \eta$) must hold, we have,

$$\|y-x\|^2 \geq \min\left\{\eta^2, \frac{2}{L_2+M}\Big(\|\nabla F(y)\| - \|\nabla F(x) - g\| - \eta \cdot \left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}^2\Big)\right\}$$
$$\geq \min\left\{\eta^2, \frac{2}{L_2+M}\|\nabla F(y)\|\right\} - \frac{2}{L_2+M}\|\nabla F(x) - g\| - \frac{2\eta}{L_2+M}\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}.$$

Rearranging the terms, and using the fact that $M \geq 2L_2$, we have

$$\min\left\{\frac{M\eta^2}{2}, \|\nabla F(y)\|\right\} \leq M\|y-x\|^2 + \|\nabla F(x) - g\| + \eta\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}.$$

Finally, using the fact that for any $a, b \geq 0$, $\min\{a,b\} \leq a\mathbf{1}\{b \geq a\}$, we have

$$\frac{M\eta^2}{2}\mathbf{1}\left\{\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right\} \leq M\|y-x\|^2 + \|\nabla F(x) - g\| + \eta\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}},$$

or, equivalently,

$$\mathbf{1}\left\{\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right\} \leq \frac{2}{\eta^2}\|y-x\|^2 + \frac{2}{M\eta^2}\Big(\|\nabla F(x) - g\| + \eta\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}\Big).$$

■

Lemma 14 *Consider the same setting as Lemma 12, but let $H \in \mathbb{S}^d$ and $g \in \mathbb{R}^d$ be random variables. Then the random variable y satisfies*

$$\mathbb{E}[F(x) - F(y)] \geq \frac{M\eta^3}{60}\Pr\left(\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right) - \frac{9}{\sqrt{M}} \cdot \mathbb{E}\Big[\|\nabla F(x) - g\|^2\Big]^{\frac{3}{4}}$$
$$- \frac{5\eta^{\frac{3}{2}}}{\sqrt{M}} \cdot \mathbb{E}\Big[\left\|\nabla^2 F(x) - H\right\|_{\mathrm{op}}^2\Big]^{\frac{3}{4}},$$

where $\Pr(\cdot)$ and $\mathbb{E}[\cdot]$ are taken with respect to the randomness over H and g.

Proof. For the ease of notation, let χ and ζ denote the error in the gradient estimator g and the hessian estimator H at x respectively, i.e.

$$\chi := \|\nabla F(x) - g\| \quad \text{and} \quad \zeta := \left\|\nabla^2 F(x) - H\right\|_{\text{op}}.$$

We prove the desired statement by combining the following two results.

- First, plugging $x = x$, and $z = y$ in to Lemma 12, we have

$$F(x) - F(y) \geq \frac{M}{12}\|y - x\|^3 - \frac{8}{\sqrt{M}}\chi_t^{\frac{3}{2}} - \frac{4}{\sqrt{M}}(\eta\zeta)^{\frac{3}{2}}.$$

  Taking expectations on both the sides, we get,

$$\begin{aligned}\mathbb{E}[F(x) - F(y)] &\geq \frac{M}{12}\,\mathbb{E}\Big[\|y - x\|^3\Big] - \frac{8}{\sqrt{M}}\,\mathbb{E}\Big[(\chi)^{\frac{3}{2}}\Big] - \frac{4}{\sqrt{M}}\,\mathbb{E}\Big[(\eta\zeta)^{\frac{3}{2}}\Big] \\ &\geq \frac{M}{12}\,\mathbb{E}\Big[\|y - x\|^3\Big] - \frac{8}{\sqrt{M}}\big(\mathbb{E}\big[\chi_t^2\big]\big)^{\frac{3}{4}} - \frac{4}{\sqrt{M}}\big(\eta^2\,\mathbb{E}\big[\zeta_t^2\big]\big)^{\frac{3}{4}}, \end{aligned} \tag{28}$$

  where the last inequality follows from an application of Jensen's inequality.

- Similarly, plugging $x = x$, $z = y$ in Lemma 13, we get

$$\mathbf{1}\left\{\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right\} \leq \frac{2}{\eta^2}\|y - x\|^2 + \frac{2}{M\eta^2}(\chi + \eta\zeta).$$

  Raising both the sides with the exponent of $\frac{3}{2}$, we get

$$\begin{aligned}\mathbf{1}\left\{\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right\} &\leq \left(\frac{2}{\eta^2}\|y - x\|^2 + \frac{2}{M\eta^2}(\chi + \eta\zeta)\right)^{\frac{3}{2}} \\ &\leq \frac{5}{\eta^3}\|y - x\|^3 + \frac{5}{M^{\frac{3}{2}}\eta^3}\Big(\chi^{\frac{3}{2}} + (\eta\zeta)^{\frac{3}{2}}\Big).\end{aligned}$$

  Taking expectations on both the sides and rearranging the terms implies that

$$\begin{aligned}\mathbb{E}\Big[\|x^{(t+1)} - x\|^3\Big] &\geq \frac{\eta^3}{5}\Pr\left(\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right) - \frac{1}{M^{\frac{3}{2}}}\,\mathbb{E}\Big[\chi^{\frac{3}{2}} + (\eta\zeta)^{\frac{3}{2}}\Big] \\ &\geq \frac{\eta^3}{5}\Pr\left(\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right) - \frac{1}{M^{\frac{3}{2}}}\Big(\big(\mathbb{E}\big[\chi_t^2\big]\big)^{\frac{3}{4}} + \big(\eta^2\,\mathbb{E}\big[\zeta_t^2\big]\big)^{\frac{3}{4}}\Big),\end{aligned} \tag{29}$$

  where the last inequality follows from an application of the Jensen's inequality.

Plugging (29) into (28), we get

$$\mathbb{E}[F(x) - F(y)] \geq \frac{M\eta^3}{60}\Pr\left(\|\nabla F(y)\| \geq \frac{M\eta^2}{2}\right) - \frac{9}{\sqrt{M}}\big(\mathbb{E}\big[\chi_t^2\big]\big)^{\frac{3}{4}} - \frac{5\eta^{\frac{3}{2}}}{\sqrt{M}}\big(\mathbb{E}\big[\zeta_t^2\big]\big)^{\frac{3}{4}}.$$

The final statement follows from the above inequality by using the definition of χ and ζ. ■

E.4. Stochastic negative curvature search

The following lemma establishes properties of the negative curvature search step used in Algorithm 4 and Algorithm 5.

Lemma 15 *Let $\gamma > 0$, and $F \in \mathcal{F}_2(\Delta, \infty, L_2)$ be given. Let $x \in \mathbb{R}^d$ be given, and let $H \in \mathbb{S}^d$ be a random variable (representing a stochastic estimator for the Hessian at x). Define y via*

$$y := \begin{cases} x + \frac{r\gamma}{L_2} \cdot u, & \text{if} \quad \lambda_{\min}(H) \leq -4\gamma, \\ x, & \text{otherwise.} \end{cases},$$

where r is an independent Rademacher random variable and u is an arbitrary unit vector such that $H[u,u] \leq -2\gamma$. Then, the point y satisfies

$$\mathbb{E}[F(x) - F(y)] \geq \frac{5\gamma^3}{6L_2^2} \Pr(\lambda_{\min}(H) \leq -4\gamma) - \frac{\gamma^2}{2L_2^2} \mathbb{E}\Big[\big\|\nabla^2 F(x) - H\big\|_{\mathrm{op}}\Big],$$

where $\Pr(\cdot)$ and $\mathbb{E}[\cdot]$ are taken with respect to the randomness in H and r.

Proof. There are two cases: either (a) $\lambda_{\min}(H) > -4\gamma$, or, (b) $\lambda_{\min}(H) \leq -4\gamma$. In the first case, $y = x$, and thus,

$$F(y) - F(x) = 0 \leq \frac{\gamma^2}{2L_2^2}\big\|H - \nabla^2 F(x)\big\|_{\mathrm{op}} \tag{30}$$

In the second case, Taylor expansion for $F(y)$ at $F(x)$ implies that

$$F(y) \leq F(x) + \langle \nabla F(x), \tilde{u}\rangle + \frac{1}{2}\nabla^2 F(x)[\tilde{u},\tilde{u}] + \frac{L_2}{6}\|\tilde{u}\|^3,$$

where $\tilde{u} := \frac{r\gamma}{L_2} \cdot u$. Taking expectations on both the sides with respect to r, we get

$$\begin{aligned}
\mathbb{E}_r[F(y)] &\overset{(i)}{=} F(x) + \frac{\gamma^2}{2L_2^2}\nabla^2 F(x)[u,u] + \frac{\gamma^3}{6L_2^2}\|u\|^3 \\
&\leq F(x) + \frac{\gamma^2}{2L_2^2}\big(H[u,u] + \nabla^2 F(x)[u,u] - H[u,u]\big) + \frac{\gamma^3}{6L_2^2}\|u\|^3 \\
&\overset{(ii)}{=} F(x) + \frac{\gamma^2}{2L_2^2}\Big(-2\gamma + \big\|\nabla^2 F(x) - H\big\|_{\mathrm{op}}\Big) + \frac{\gamma^3}{6L_2^2} \\
&\leq F(x) - \frac{5\gamma^3}{6L_2^2} + \frac{\gamma^2}{2L_2^2}\big\|\nabla^2 F(x) - H\big\|_{\mathrm{op}},
\end{aligned} \tag{31}$$

where (i) is given by the fact that $\mathbb{E}_r[\langle \nabla F(x), ru\rangle] = 0$, and (ii) follows from the fact that u is chosen such that $\mathbb{E}\big[\nabla^2 F(x)[u,u]\big] \leq -2\gamma$ and $\|u\| = 1$, and the fact that for any matrix A and vector b, $\|Ab\| \leq \|A\|_{\mathrm{op}}\|b\|$.

Since, one of the two cases ($\lambda_{\min}(H) > -4\gamma$ or $\lambda_{\min}(H) \leq -4\gamma$) must hold, combining (30) and (31), we have

$$\mathbb{E}_r[F(y)] \leq F(x) - \frac{5\gamma^3}{6L_2^2}\mathbf{1}\{\lambda_{\min}(H) \leq -4\gamma\} + \frac{\gamma^2}{2L_2^2}\big\|\nabla^2 F(x) - H\big\|_{\mathrm{op}}.$$

Taking expectation on both the sides gives the desired statement:

$$\mathbb{E}[F(x) - F(y)] \geq \frac{5\gamma^3}{6L_2^2} \Pr(\lambda_{\min}(H) \leq -4\gamma) - \frac{\gamma^2}{2L_2^2} \mathbb{E}\Big[\big\|\nabla^2 F(x) - H\big\|_{\mathrm{op}}\Big].$$

■

The following lemma establishes properties of Oja's method (Oja), as used in Algorithm 4.

Lemma 16 (Allen-Zhu [3], Lemma 5.3) *The procedure* Oja *takes as input a point $x \in \mathbb{R}^d$, a stochastic Hessian-vector product oracle $\mathsf{O}_F^2 \in \overline{\mathcal{O}}_2(F, \sigma_1, \bar{\sigma}_2)$ for some function $F \in \mathcal{F}_2(\Delta, L_1, \infty)$, a precision parameter $\gamma > 0$ and a failure probability $\delta \in (0, 1)$, and runs outputs $u \in \mathbb{R}^d \cup \{\perp\}$ such that with probability at least $1 - \delta$, either*[9]

a) $u = \perp$, *and* $\nabla^2 F(x) \succeq -2\gamma I$.

b) if $u \neq \perp$, *then* $\|u\| = 1$ *and* $\langle u, \nabla^2 F(x) u\rangle \leq -\gamma$.

Moreover, when invoked as above, the procedure uses at most

$$O\left(\frac{(\bar{\sigma}_2 + L_1)^2}{4\gamma^2} \log^2\left(\frac{d}{\delta}\right)\right)$$

queries to the stochastic Hessian-vector product oracle.

Appendix F. Upper bounds for finding ϵ-stationary points

F.1. Proof of Theorem 2

Proof of Theorem 2. In the following, we first show that Algorithm 2 returns a point $\widehat{x}$ such that, $\mathbb{E}[\|\nabla F(\widehat{x})\|] \leq 32\epsilon$. We then bound the expected number of oracle queries used throughout the execution.[10]

Since, $\eta = \frac{1}{2\sqrt{L_1^2 + \bar{\sigma}_2^2 + \bar{\epsilon} L_2}} \leq \frac{1}{2L_1}$ and F has L_1-Lipschitz gradient, Lemma 11 implies that the point $x^{(t+1)}$ computed using the update rule $x^{(t+1)} \leftarrow x^{(t)} - \eta g^{(t)}$ satisfies

$$\frac{\eta}{8}\left\|\nabla F(x^{(t)})\right\|^2 \leq F(x^{(t)}) - F(x^{(t+1)}) + \frac{3\eta}{4}\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2.$$

Telescoping the above from t from 1 to T, this implies

$$\begin{aligned}\frac{\eta}{8}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2 &\leq F(x^{(0)}) - F(x^{(T+1)}) + \frac{3\eta}{4}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2 \\ &\leq \Delta + \frac{3\eta}{4}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2,\end{aligned}$$

9. Note that if this event fails, the algorithm still returns either $\perp$ or a unit vector u.

10. In the proof, we show convergence to a 32ϵ-stationary point. A simple change of variable, i.e. running Algorithm 2 with $\epsilon \leftarrow \frac{\epsilon}{32}$, returns a point $\widehat{x}$ that enjoys the guarantee that $\|\nabla F(\hat{x})\| \leq \epsilon$.

where the last inequality follows from the fact that $F(x^{(0)}) - F(x^{(T+1)}) \leq \Delta$. Next, taking expectation on both the sides (with respect to the stochasticity of the oracle and the algorithm's internal randomization), we get

$$\frac{\eta}{8}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2\right] \leq \Delta + \frac{3\eta}{4}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2\right].$$

Using Lemma 7, we have $\mathbb{E}\left[\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2\right] \leq \epsilon^2$ for all $t \geq 1$. Dividing both the sides by $\frac{\eta T}{8}$, and plugging in the value of the parameters T and η, we get,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2\right] \leq \frac{8\Delta}{\eta T} + \frac{6}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2\right] \leq \frac{8\Delta}{\eta T} + 6\epsilon^2 \leq 14\epsilon^2. \tag{32}$$

Thus, for $\widehat{x}$ chosen uniformly at random from the set $\left(x^{(t)}\right)_{t=1}^{T}$, we have

$$\mathbb{E}\|\nabla F(\widehat{x})\| = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla F(x^{(t)})\right\| \leq \sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2\right]} \leq 4\epsilon.$$

Finally, Markov's inequality implies that with probability at least $\frac{7}{8}$,

$$\|\nabla F(\widehat{x})\| \leq 32\epsilon. \tag{33}$$

Bound on the number of oracle queries. Algorithm 2 queries the stochastic oracle in only when it invokes HVP-RVR in Line 4 to compute the gradient estimate $g^{(t)}$ at time t. Let M denote the total number of oracle calls made up until time T. Invoking Lemma 8 to bound the expected number of stochastic oracle calls for each $t \geq 1$, and ignoring all the mutiplicative constants, we get

$$\begin{aligned}
\mathbb{E}[M] &\leq 5\sum_{t=1}^{T}\mathbb{E}\left[\frac{b\sigma_1^2}{\epsilon^2} + \frac{\left\|x^{(t+1)} - x^{(t)}\right\|^2 \cdot \left(\sigma_2^2 + \epsilon L_2\right)}{b\epsilon^2} + 1\right] \\
&\overset{(i)}{\leq} O\left(\sum_{t=1}^{T}\mathbb{E}\left[\frac{b\sigma_1^2}{\epsilon^2} + \frac{\left\|\eta g^{(t)}\right\|^2 \cdot \left(\sigma_2^2 + \epsilon L_2\right)}{b\epsilon^2} + 1\right]\right) \\
&\overset{(ii)}{\leq} O\left(\frac{\Delta}{\eta\epsilon^2} \cdot \left(\frac{b\sigma_1^2}{\epsilon^2} + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|g^{(t)}\right\|^2\right] \cdot \frac{\eta^2\left(\sigma_2^2 + \epsilon L_2\right)}{b\epsilon^2} + 1\right)\right) \\
&\overset{(iii)}{=} O\left(\frac{\Delta}{\eta\epsilon^2} \cdot \left(\frac{b\sigma_1^2}{\epsilon^2} + \frac{\eta^2\left(\sigma_2^2 + \epsilon L_2\right)}{b} + 1\right)\right),
\end{aligned} \tag{34}$$

where (i) is given by plugging in the update rule from Line 5 and by dropping multiplicative constants, (ii) is given by rearranging the terms, plugging in the value of T and using that $T \geq 1$ (to simplify the ceiling operator) under the assumption $\epsilon \leq \sqrt{\Delta L_1}$, and (iii) follows by observing that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|g^{(t)}\right\|^2\right] \leq 2\,\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|g^{(t)} - \nabla F(x^{(t)})\right\|^2 + \frac{1}{T}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2\right] \leq 30\epsilon^2,$$

Algorithm 3 Subsampled cubic-regularized trust-region method with HVP-RVR

Input: Oracle $(\mathsf{O}_F^2, P_z) \in \mathcal{O}_2(F, \sigma_{1:2})$ for $F \in \mathcal{F}_2(\Delta, \infty, L_2)$. Precision parameter ϵ.

Set $M = 5\max\left\{L_2, \frac{\epsilon\sigma_2^2\log(d)}{\sigma_1^2}\right\}$, $\eta = 25\sqrt{\frac{\epsilon}{M}}$, $T = \left\lceil \frac{5\Delta}{3\eta\epsilon} \right\rceil$ and $n_H = \left\lceil \frac{22\sigma_2^2\eta^2\log(d)}{\epsilon^2} \right\rceil$.

Set $b = \min\left\{1, \frac{\eta\sqrt{\sigma_2^2+\epsilon L_2}}{25\sigma_1}\right\}$.

Initialize $x^{(0)}, x^{(1)} \leftarrow 0$, $g^{(0)} \leftarrow \perp$.

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

  Query the oracle n_H times at $x^{(t)}$ and compute

$$H^{(t)} \leftarrow \frac{1}{n_H}\sum_{j=1}^{n_H} \widehat{\nabla^2 F}(x^{(t)}, z^{(t,j)}), \quad \text{where} \quad z^{(t,j)} \overset{\text{i.i.d.}}{\sim} P_z.$$

  $g^{(t)} \leftarrow \text{HVP-RVR-Gradient-Estimator}_{\epsilon,b}\left(x^{(t)}, x^{(t-1)}, g^{(t-1)}\right)$.

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

$$x^{(t+1)} \leftarrow \operatorname*{arg\,min}_{y:\|y-x^{(t)}\|\le\eta} \left\langle g^{(t)}, y - x^{(t)}\right\rangle + \frac{1}{2}\left\langle y - x^{(t)}, H^{(t)}(y - x^{(t)})\right\rangle + \frac{M}{6}\left\|y - x^{(t)}\right\|^3.$$

return $\widehat{x}$ chosen uniformly at random from $\left\{x^{(t)}\right\}_{t=2}^{T+1}$.

as a consequence of Lemma 7 and the bound in (32). Next, note that since we assume $\epsilon < \sigma_1$, and since we have $\eta \le \frac{1}{2\sqrt{\sigma_2^2+\epsilon L_2}}$, the parameter b is equal to $\frac{\eta\epsilon\sqrt{\sigma_2^2+\epsilon L_2}}{\sigma_1}$ (as this is smaller than 1). Thus, plugging the value of b and η in the bound (34), we get,

$$\begin{aligned}\mathbb{E}[m(T)] &= O\left(\frac{\Delta\sigma_1\sqrt{\sigma_2^2+\epsilon L_2}}{\epsilon^3} + \frac{\Delta\sqrt{L_1^2+\sigma_2^2+\epsilon L_2}}{\epsilon^2}\right)\\ &= O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} + \frac{\Delta\sigma_1\sqrt{L_2}}{\epsilon^{2.5}} + \frac{\Delta\sigma_2}{\epsilon^2} + \frac{\Delta L_1}{\epsilon^2} + \frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right).\end{aligned}$$

Using Markov's inequality, we have that with probability at least $\frac{7}{8}$,

$$M \le O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} + \frac{\Delta\sigma_1\sqrt{L_2}}{\epsilon^{2.5}} + \frac{\Delta\sigma_2}{\epsilon^2} + \frac{\Delta L_1}{\epsilon^2} + \frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right). \tag{35}$$

The final statement follows by taking a union bound with failure probabilities for (33) and (35). ■

F.2. Full statement and proof for Algorithm 3

Proof of Theorem 3. In the following, we first show that Algorithm 3 returns a point $\hat{x}$, such that with probability at least $\frac{7}{8}$, $\|\nabla F(\hat{x})\| \le 350\epsilon$. We then bound, with probability at least $\frac{7}{8}$, the total number of oracle queries made up until time T.

Note that, using Lemma 7 and Lemma 9, we have for all $t \ge 0$,

$$\mathbb{E}\left[\left\|\nabla F\left(x^{(t)}\right) - g^{(t)}\right\|\right] \le \epsilon^2, \quad \text{and} \quad \mathbb{E}\left[\left\|\nabla^2 F\left(x^{(t)}\right) - H^{(t)}\right\|_{\text{op}}\right] \le \frac{\epsilon^2}{\eta^2}. \tag{36}$$

Thus, for each $t \geq 1$, invoking Lemma 14 and plugging in the bounds from (36), and using the value of η, we get

$$\begin{aligned}\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big] &\geq \frac{M\eta^3}{60}\Pr\left(\left\|\nabla F(x^{(t+1)})\right\| \geq \frac{M\eta^2}{2}\right) - \frac{14\epsilon^{\frac{3}{2}}}{\sqrt{M}} \\ &\geq \frac{240\epsilon^{\frac{3}{2}}}{\sqrt{M}}\left(\Pr\left(\left\|\nabla F(x^{(t+1)})\right\| \geq 350\epsilon\right) - \frac{1}{16}\right).\end{aligned}$$

Telescoping this inequality from $t = 1$ to T, we have that

$$\begin{aligned}\mathbb{E}\Big[F(x^{(1)}) - F(x^{(T+1)})\Big] &\geq \frac{240\epsilon^{\frac{3}{2}}}{\sqrt{M}} \cdot T \cdot \left(\frac{1}{T}\sum_{t=1}^{T}\Pr\left(\left\|\nabla F(x^{(t+1)})\right\| \geq 350\epsilon\right) - \frac{1}{16}\right) \\ &= \frac{240\epsilon^{\frac{3}{2}}}{\sqrt{M}} \cdot T \cdot \left(\Pr(\|\nabla F(\widehat{x})\| \geq 350\epsilon) - \frac{1}{16}\right),\end{aligned}$$

where the equality follows because $\widehat{x}$ is sampled uniformly at random from the set $\left\{x^{(t)}\right\}_{t=2}^{T+1}$. Next, using the fact that, $F(x^{(t)}) - F\left(x^{(T+1)}\right) \leq \Delta$, rearranging the terms, and plugging in the value of T, we get

$$\Pr(\|\nabla F(\widehat{x})\| \geq 350\epsilon) \leq \frac{\Delta\sqrt{M}}{240\epsilon^{\frac{3}{2}}T} + \frac{1}{16} \leq \frac{1}{8}.$$

Thus, with probability at least $\frac{7}{8}$,

$$\|\nabla F(\widehat{x})\| \leq 350\epsilon. \tag{37}$$

Bound on the number of oracle queries. Algorithm 3 queries the stochastic oracle in Line 5 and Line 6 only to compute the respective Hessian and gradient estimates. Let M_h and M_g denote the total number of stochastic oracle queries made by Line 5 and Line 6 till time T respectively. Further, Let $M = M_h + M_g$ denote the total number of oracle queries made till time T.

In what follows, we first bound $\mathbb{E}[M_h]$ and $\mathbb{E}[M_g]$. Then, we invoke Markov's inequality to deduce that the desired bound on M holds with probability at least $\frac{7}{8}$.

1. **Bound on $\mathbb{E}[M_h]$.** Since the algorithm queries the stochastic Hessian oracle n_H times per iteration, $M_h = T \cdot n_H$. Plugging the values of T, n_H and M as specified in Algorithm 3, and ignoring multiplicative constant, we get,

$$\begin{aligned}\mathbb{E}[M_h] &= \left\lceil\frac{5\Delta\sqrt{M}}{3\epsilon^{1.5}}\right\rceil \cdot \left\lceil\frac{22\sigma_2^2\eta^2\log(d)}{\epsilon^2}\right\rceil \\ &\leq O\left(\frac{\Delta\sqrt{M}}{\epsilon^{1.5}} + \frac{\Delta\sigma_2^2\log(d)}{\epsilon^{2.5}\sqrt{M}}\right) \\ &\leq O\left(\frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}} + \frac{\Delta\sigma_2}{\epsilon^2} + \frac{\Delta\sigma_1\sigma_2\sqrt{\log(d)}}{\epsilon^3}\right),\end{aligned} \tag{38}$$

where the first inequality above follows from the fact that $\frac{\Delta\sqrt{M}}{\epsilon^{1.5}} \geq 1$ under the natural choice for the precision parameter $\epsilon \leq \Delta^{\frac{2}{3}}M^{\frac{1}{3}}$ and using the identity $\lceil x\rceil \leq x + 1$ for $x \geq 0$.

2. **Bound on $\mathbb{E}[M_g]$.** Invoking Lemma 8 for each $t \geq 1$, we get

$$\begin{aligned}\mathbb{E}[M_g] &= 6\sum_{t=1}^{T}\mathbb{E}\left[\frac{b\sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon)\cdot\left\|x^{(t)} - x^{(t-1)}\right\|^2}{b\epsilon^2} + 1\right] \\ &\overset{(i)}{=} O\left(T\cdot\left(\frac{b\sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon)\cdot\eta^2}{b\epsilon^2} + 1\right)\right) \\ &\overset{(ii)}{=} O\left(\frac{\Delta}{\eta\epsilon}\cdot\left(\frac{b\sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon)\cdot\eta^2}{b\epsilon^2} + 1\right)\right) \end{aligned} \tag{39}$$

where (i) follows by observing $\left\|x^{(t)} - x^{(t-1)}\right\| \leq \eta$ due to the update rule in Line 7 and (ii) is given by plugging in the value of $T \leq O(\frac{\Delta}{\eta\epsilon})$ for the natural choice of parameter $\epsilon = O(\Delta^{\frac{2}{3}}M^{\frac{1}{3}})$. Next, note that since $M > L_2$, and since we assume $\epsilon < \sigma_1$, the parameter b is equal to $\frac{\eta\sqrt{\sigma_2^2+\epsilon L_2}}{25\sigma_1}$ (which is smaller than 1). Thus, plugging the value of b and η in the bound (39), we get,

$$\begin{aligned}\mathbb{E}[M_g] &= O\left(\frac{\Delta\sigma_1\sqrt{\sigma_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta\sqrt{M}}{\epsilon^{1.5}}\right) \\ &= O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} + \frac{\Delta\sigma_1\sqrt{L_2}}{\epsilon^{2.5}} + \frac{\Delta\sigma_2}{\epsilon^2}\sqrt{\log(d)} + \frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right), \end{aligned} \tag{40}$$

where the second equality follows by using that $\epsilon \leq \sigma_1$ to simplify the term $\frac{\Delta\sqrt{M}}{\epsilon^{1.5}}$.

Adding (40) and (38), the total number of oracle queries made by Algorithm 3 till time T is bounded, in expectation, by

$$\mathbb{E}[M] = \mathbb{E}[M_g + M_h] = O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3}\sqrt{\log(d)} + \frac{\Delta\sigma_1\sqrt{L_2}}{\epsilon^{2.5}} + \frac{\Delta\sigma_2}{\epsilon^2}\sqrt{\log(d)} + \frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right).$$

Using Markov's inequality, we get that, with probability at least $\frac{7}{8}$,

$$M \leq O\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3}\sqrt{\log(d)} + \frac{\Delta\sigma_1\sqrt{L_2}}{\epsilon^{2.5}} + \frac{\Delta\sigma_2}{\epsilon^2}\sqrt{\log(d)} + \frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right). \tag{41}$$

The final statement follows by taking a union bound for the failure probability of (37) and (41). ■

Appendix G. Upper bounds for finding (ϵ, γ)-second-order-stationary points

G.1. Full statement and proof for Algorithm 4

Algorithm 4 Stochastic gradient descent with negative curvature search and HVP-RVR

Input: Oracle $(\mathsf{O}_F^2, P_z) \in \overline{\mathcal{O}}_2(F, \sigma_1, \bar{\sigma}_2)$ for $F \in \mathcal{F}_2(\Delta, L_1, L_2)$. Precision parameters ϵ, γ.

1: Set $\eta = \min\left\{\frac{\gamma}{\epsilon L_2}, \frac{1}{2\sqrt{L_1^2+\bar{\sigma}_2^2+\epsilon L_2}}\right\}$, $T = \left\lceil \frac{20\Delta L_2^2}{\gamma^3} + \frac{2\Delta}{\eta\epsilon^2}\right\rceil$, $p = \frac{\gamma^3}{\gamma^3+10\Delta L_2^2\eta\epsilon^2}$, $\delta = \frac{\gamma}{40^2 L_2}$.

2: Set $b_g = \min\{1, \frac{\eta\epsilon\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\sigma_1}\}$ and $b_H = \min\{1, \frac{\gamma\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\sigma_1 L_2}\}$.

3: Initialize $x^{(0)}$, $x^{(1)} \leftarrow 0$, $g^{(1)} \leftarrow \mathsf{HVP\text{-}RVR\text{-}Gradient\text{-}Estimator}_{\epsilon,b_g}(x^{(1)}, x^{(0)}, \perp)$.

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

5: Sample $Q_t \sim \text{Bernoulli}(p)$.

6: **if** $Q_t = 1$ **then**

7: $x^{(t+1)} \leftarrow x^{(t)} - \eta \cdot g^{(t)}$.

8: $g^{(t+1)} \leftarrow \mathsf{HVP\text{-}RVR\text{-}Gradient\text{-}Estimator}_{\epsilon,b_g}(x^{(t+1)}, x^{(t)}, g^{(t)})$.

9: **else**

10: $u^{(t)} \leftarrow \mathsf{Oja}\big(x^{(t)}, \mathsf{O}_F^2, 2\gamma, \delta\big)$. `// Oja's algorithm (Lemma 16).`

11: **if** $u^{(t)} \equiv \perp$ **then**

12: $x^{(t+1)} \leftarrow x^{(t)}$.

13: $g^{(t+1)} \leftarrow g^{(t)}$.

14: **else**

15: Sample $r^{(t)} \sim \text{Uniform}(\{-1, 1\})$.

16: $x^{(t+1)} \leftarrow x^{(t)} + \frac{\gamma}{L_2} \cdot r^{(t)} \cdot u^{(t)}$.

17: $g^{(t+1)} \leftarrow \mathsf{HVP\text{-}RVR\text{-}Gradient\text{-}Estimator}_{\epsilon,b_H}(x^{(t+1)}, x^{(t)}, g^{(t)})$.

18: **return** $\widehat{x}$ chosen uniformly at random from $\big(x^{(t)}\big)_{t=1}^{T}$.

Proof of Theorem 5. We first show that Algorithm 4 returns a point $\widehat{x}$ such that, $\mathbb{E}[\|\nabla F(\widehat{x})\|] \le 8\epsilon$ and $\lambda_{\min}\big(\nabla^2 F(\widehat{x})\big) \ge -4\gamma$. We then bound the expected number of oracle queries used throughout the execution.

To begin, note that, for any $t \ge 1$, there are two scenarios: (a) either $Q_t = 1$ and $x^{(t+1)}$ is set using the update rule in Line 7, or, (b) $Q_t = 0$ and we set $x^{(t+1)}$ using Line 10, respectively. We analyze the two cases separately below.

Case 1: $Q_t = 1$. Since, $\eta \le \frac{1}{2\sqrt{L_1^2+\bar{\sigma}_2^2+\epsilon L_2}} \le \frac{1}{2L_1}$ and F has L_1-Lipschitz gradient, using Lemma 11, we have

$$F(x^{(t)}) - F(x^{(t+1)}) \ge \frac{\eta}{8}\left\|\nabla F(x^{(t)})\right\|^2 - \frac{3\eta}{4}\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2.$$

Taking expectation on both the sides, while conditioning on the event that $Q_t = 1$, we get

$$\begin{aligned}\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 1\Big] &\ge \frac{\eta}{8}\,\mathbb{E}\left[\left\|\nabla F(x^{(t)})\right\|^2\right] - \frac{3\eta}{4}\,\mathbb{E}\left[\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2\right] \\ &\ge \frac{\eta}{8}\,\mathbb{E}\left[\left\|\nabla F(x^{(t)})\right\|^2\right] - \frac{3\eta\epsilon^2}{4}, \end{aligned} \tag{42}$$

where the last inequality follows using Lemma 7.

Case 2: $Q_t = 0$**.** Let $\mathsf{E}^{\mathsf{Oja}}(t)$ denote the event that Oja succeeds at time t, in the sense that the event in Lemma 16 holds: (i) if $u^{(t)} = \perp$ then $\nabla^2 F(x^{(t)}) \succeq -2\gamma I$, and (ii) otherwise, $u^{(t)}$ satisfies $\langle u^{(t)}, \nabla^2 F(x^{(t)}) u^{(t)} \rangle \leq -\gamma$. Then using Lemma 17, we are guaranteed that

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0\Big] \geq \frac{5\gamma^3}{6L_2^2}\bigg(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\Pr\Big(\neg\mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0\Big)\bigg).$$

In particular, we are guaranteed by Lemma 16 that

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0\Big] \geq \frac{5\gamma^3}{6L_2^2}\bigg(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\delta\bigg). \tag{43}$$

Combining the two cases ($Q_t = 0$ and $Q_t = 1$) from (42) and (43) above, we get

$$\begin{aligned}
&\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big] &(44)\\
&= \sum_{q\in\{0,1\}} \Pr(Q_t = q)\,\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = q\Big]\\
&\geq \frac{5(1-p)\gamma^3}{6L_2^2}\bigg(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\delta\bigg) + p\bigg(\frac{\eta}{8}\,\mathbb{E}\Big[\Big\|\nabla F(x^{(t)})\Big\|^2\Big] - \frac{3\eta\epsilon^2}{4}\bigg). &(45)
\end{aligned}$$

Using that $\mathbb{E}\Big[\Big\|\nabla F(x^{(t)})\Big\|^2\Big] \geq (8\epsilon)^2 \cdot \Pr\Big(\Big\|\nabla F(x^{(t)})\Big\| \geq 8\epsilon\Big)$ and that $\delta \leq \frac{\gamma}{1600 L_1}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big]\\
&\geq \frac{5(1-p)\gamma^3}{6L_2^2}\bigg(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{1}{800}\bigg) + 8p\eta\epsilon^2\bigg(\Pr\Big(\Big\|\nabla F(x^{(t)})\Big\| \geq 8\epsilon\Big) - \frac{3}{32}\bigg).
\end{aligned}$$

Telescoping this inequality for t from 1 to T and using the bound $\mathbb{E}\big[F(x^{(1)}) - F(x^{(T+1)})\big] \leq \Delta$, we get

$$\begin{aligned}
\Delta &\geq \mathbb{E}\Big[F(x^{(1)}) - F(x^{(T+1)})\Big]\\
&\geq \frac{5T(1-p)\gamma^3}{6L_2^2}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{1}{800}\Big)\\
&\quad + 8Tp\eta\epsilon^2\Big(\frac{1}{T}\sum_{t=0}^{T-1}\Pr\Big(\Big\|\nabla F(x^{(t)})\Big\| \geq 8\epsilon\Big) - \frac{3}{32}\Big)\\
&\overset{(i)}{\geq} \frac{5T(1-p)\gamma^3}{6L_2^2}\bigg(\Pr\big(\lambda_{\min}(\nabla^2 F(\widehat{x})) \leq -4\gamma\big) - \frac{1}{800}\bigg) + 8Tp\eta\epsilon^2\Big(\Pr(\|\nabla F(\widehat{x})\| \geq 8\epsilon) - \frac{3}{32}\Big)\\
&\overset{(ii)}{\geq} 16\Delta\bigg(\Pr\big(\lambda_{\min}(\nabla^2 F(\widehat{x})) \leq -4\gamma\big) + \Pr(\|\nabla F(\widehat{x})\| \geq 8\epsilon) - \frac{1}{4}\bigg), &(46)
\end{aligned}$$

where (i) follows because $\widehat{x}$ is sampled uniformly at random from $\left(x^{(t)}\right)_{t=1}^{T}$ and (ii) follows from Lemma 19. Rearranging the terms, we get

$$\Pr\left(\lambda_{\min}(\nabla^2 F(\widehat{x})) \leq -4\gamma\right) + \Pr(\|\nabla F(\widehat{x})\| \geq 8\epsilon) \leq \frac{5}{16},$$

which further implies that

$$\Pr\left(\lambda_{\min}(\nabla^2 F(\widehat{x})) \geq -4\gamma \wedge \|\nabla F(\widehat{x})\| \leq 8\epsilon\right) \geq \frac{11}{16}. \tag{47}$$

Bound on the number of oracle queries. At every iteration, Algorithm 4 queries the stochastic oracle in either Line 8 or Line 17 (to compute the stochastic gradient estimator and to execute Oja's algorithm, respectively), and possibly Line 10 (to update the gradient estimator after a negative curvature step). Let $m_g(t)$ denote the total number of stochastic oracle queries made by Line 8 or Line 17 at time t, and let $M_{\mathrm{g}} = \sum_{t=1}^{T} m_g(t)$. Further, let M_{nc} denote the total number of oracle calls made by Line 10, and further let $M = M_{\mathrm{g}} + M_{\mathrm{nc}}$ be the total number of oracle queries made up until time T.

In what follows, we first bound $\mathbb{E}[M_{\mathrm{g}}]$ and $\mathbb{E}[M_{\mathrm{nc}}]$. Then, we invoke Markov's inequality to bound M with probability at least $\frac{19}{20}$.

Bound on M_{g}. For any $t > 0$, there are two scenarios, either (a) $Q_t = 1$ and we go through Line 7, or (b) $Q_t = 0$ and Line 17 is executed. Thus,

$$\mathbb{E}[M_{\mathrm{g}}] = \sum_{t=1}^{T} \Pr(Q_t = 0)\, \mathbb{E}[m_g(t) \mid Q_t = 0] + \sum_{t=1}^{T} \Pr(Q_t = 1)\, \mathbb{E}[m_g(t) \mid Q_t = 1] \tag{48}$$

We denote the two terms on the right hand side above by $(\mathbf{A})$ and $(\mathbf{B})$, respectively. We bound them separately as follows.

- **Bound on $(\mathbf{A})$.** Using Lemma 8 with the fact that $\Pr(Q_t = 0) = 1 - p$, we get

$$\begin{aligned}
(\mathbf{A}) &= O(1) \sum_{t=1}^{T} (1-p) \cdot \mathbb{E}\left[b_H \frac{\sigma_1^2}{\epsilon^2} + \left\|x^{(t+1)} - x^{(t)}\right\|^2 \cdot \frac{\bar{\sigma}_2^2 + \epsilon L_2}{b_H \epsilon^2} + 1 \;\middle|\; Q_t = 0 \right] \\
&\overset{(i)}{=} O\left(T \cdot (1-p) \cdot \left(\frac{\gamma \sigma_1 \sqrt{\bar{\sigma}_2^2 + \epsilon L_2}}{L_2 \epsilon^2} + \frac{\gamma^2}{\epsilon^2} \cdot \frac{\bar{\sigma}^2 + \epsilon L_2}{L_2^2} + 1 \right) \right) \\
&\overset{(ii)}{\leq} O\left(\frac{\Delta L_2 \sigma_1 \sqrt{\bar{\sigma}_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta(\bar{\sigma}_2^2 + \epsilon L_2)}{\gamma \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3} \right),
\end{aligned} \tag{49}$$

  where (i) is given by plugging in $\|x^{(t)} - x^{(t-1)}\| = \gamma / L_2$. The inequality (ii) follows by using the bound on $T \cdot (1-p)$ from Lemma 19.

- **Bound on $(\mathbf{B})$.** Using Lemma 8 with the fact that $\Pr(Q_t = 1) = p$, we get

$$(\mathbf{B}) = O(1) \sum_{t=1}^{T} p \cdot \mathbb{E}\left[b_g \frac{\sigma_1^2}{\epsilon^2} + \left\|x^{(t+1)} - x^{(t)}\right\|^2 \cdot \frac{\bar{\sigma}_2^2 + \epsilon L_2}{b_g \epsilon^2} + 1 \;\middle|\; Q_t = 1 \right]$$

$$\begin{aligned}
&\overset{(i)}{=} O(1)\sum_{t=1}^{T} p\cdot\mathbb{E}\left[b_g\frac{\sigma_1^2}{\epsilon^2}+\left\|\eta g^{(t)}\right\|^2\cdot\frac{\bar{\sigma}_2^2+\epsilon L_2}{b_g\epsilon^2}+1\;\middle|\; Q_t=1\right]\\
&\overset{(ii)}{\leq} O\left(\frac{\Delta}{\eta\epsilon^2}\cdot\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|g^{(t)}\right\|^2\right]\cdot\frac{\eta^2(\bar{\sigma}_2^2+\epsilon L_2)}{b_g\epsilon^2}+b_g\frac{\sigma_1^2}{\epsilon^2}+1\right)\right)\\
&\overset{(iii)}{=} O\Big(\frac{\Delta\sigma_1\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\epsilon^3}+\frac{\Delta}{\eta\epsilon^2}\Big),
\end{aligned}\tag{50}$$

where (i) follows by plugging in the update rule from Line 7 (when $Q_t=1$), (ii) follows by rearranging the terms and using the bound on $T\cdot p$ from Lemma 19, and (iii) is follows from the choices of b_g (in particular, our assumption that $\epsilon\leq\sigma_1$ implies that $b_g=\frac{\eta\epsilon\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\sigma_1}$) and η, as well as the following bound for $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|g^{(t)}\right\|^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|g^{(t)}\|^2\right]&\leq\mathbb{E}\left[\frac{2}{T}\sum_{t=1}^{T}\left\|g^{(t)}-\nabla F(x^{(t)})\right\|^2+\frac{2}{T}\sum_{t=1}^{T}\left\|\nabla F(x^{(t)})\right\|^2\right]\\
&\leq O\big(\epsilon^2+\|\nabla F(\widehat{x})\|^2\big)\leq O(\epsilon^2),
\end{aligned}$$

where the last inequality is uses Lemma 7 and Lemma 18.

Combining the bounds from (49) and (50) in (48), we have

$$\mathbb{E}[M_{\mathrm{g}}]\leq O\left(\frac{\Delta L_2\sigma_1\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\gamma^2\epsilon^2}+\frac{\Delta(\bar{\sigma}_2^2+\epsilon L_2)}{\gamma\epsilon^2}+\frac{\Delta L_2^2}{\gamma^3}+\frac{\Delta\sigma_1\sqrt{\bar{\sigma}_2^2+\epsilon L_2}}{\epsilon^3}+\frac{\Delta}{\eta\epsilon^2}\right).\tag{51}$$

Bound on M_{nc}. Using the law of total probability with the observation that Algorithm 4 enters Line 10 only if $Q_t=0$, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}m_{\mathrm{nc}}(t)\right]&=\sum_{t=1}^{T}\sum_{q\in\{0,1\}}\Pr(Q_t=q)\,\mathbb{E}[m_{\mathrm{nc}}(t)\mid Q_t=q]\\
&=\sum_{t=1}^{T}\Pr(Q_t=0)\,\mathbb{E}[m_{\mathrm{nc}}(t)\mid Q_t=0]\\
&=T\cdot(1-p)\cdot n_H\leq O\left(\frac{\Delta L_2^2}{\gamma^3}\cdot n_H\right),
\end{aligned}\tag{52}$$

where n_H denotes the number of oracle queries made by Oja, the last inequality follows by bounding $T\cdot(1-p)$ as in (46). Note that Lemma 16 implies that for $\delta=\frac{\gamma}{1600L_1}$,

$$n_H\leq O\left(\frac{(\bar{\sigma}_2+L_1)^2}{\gamma^2}\log^2\left(\frac{L_1}{\gamma}d\right)\right).\tag{53}$$

Combining the above bounds for M_{g} and M_{nc} (in (51) and (52) respectively), we get

$$\mathbb{E}[M]\leq 20\,\mathbb{E}[M_{\mathrm{g}}+M_{\mathrm{nc}}]$$

$$= O\left(\frac{\Delta L_2 \sigma_1 \sqrt{\bar{\sigma}_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta(\bar{\sigma}_2^2 + \epsilon L_2)}{\gamma \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3} + \frac{\Delta \sigma_1 \sqrt{\bar{\sigma}_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta}{\eta \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3} \cdot n_H\right).$$

Plugging in the value of η from Algorithm 4 and n_H from (53), and using Markov's inequality, we get that, with probability at least $\frac{15}{16}$,

$$M = O\left(\frac{\Delta \sigma_1 \sqrt{\bar{\sigma}_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta L_2\left(\sigma_1 \bar{\sigma}_2 + \sqrt{\epsilon L_2} + \gamma \bar{\sigma}_2^2 / L_2 + \gamma \epsilon\right)}{\gamma^2 \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3}\left(\frac{\left(\bar{\sigma}_2 + L_1\right)^2}{\gamma^2} \log^2\left(\frac{L_1}{\gamma} d\right)\right)\right.$$
$$\left. + O\left(\frac{\Delta L_2^2}{\gamma^3} + \frac{\Delta \sqrt{L_1^2 + \bar{\sigma}_2^2 + \epsilon L_2}}{\epsilon^2}\right). \tag{54}$$

Ignoring the lower-order terms, we have

$$M = \widetilde{O}\left(\frac{\Delta \sigma_1 \bar{\sigma}_2}{\epsilon^3} + \frac{\Delta L_2 \sigma_1 \bar{\sigma}_2}{\gamma^2 \epsilon^2} + \frac{\Delta L_2^2 (\bar{\sigma}_2 + L_1)^2}{\gamma^5}\right).$$

The final statement follows by taking a union bound for the failure probability of the claims in (47) and (54). ■

Lemma 17 *Under the setting of Theorem 5, we are guaranteed that*

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0\Big]$$
$$\geq \frac{5\gamma^3}{6L_2^2}\left(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\Pr\Big(\neg \mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0\Big)\right).$$

Proof. Recall that Algorithm 4 calls Oja with the precision parameter 2γ. To begin, suppose that $\mathsf{E}^{\mathsf{Oja}}(t)$ holds. Then if Oja returns $\perp$, then $\lambda_{\min}\big(\nabla^2 F(x^{(t)})\big) \geq -4\gamma$, otherwise Oja returns a unit vector $u^{(t)}$ such that $\nabla^2 F(x^{(t)})[u^{(t)}, u^{(t)}] \leq -2\gamma$. Thus, using Lemma 15 with $H = \nabla^2 F(x^{(t)})$ and $u^{(t)}$, we conclude that—conditioned on the history up to time t, and on $Q_t = 0$—we have

$$\mathbf{1}\{\mathsf{E}^{\mathsf{Oja}}(t)\}(F(x^{(t)}) - F(x^{(t+1)})) \geq \frac{5\gamma^3}{6L_2^2}\mathbf{1}\{\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma \wedge \mathsf{E}^{\mathsf{Oja}}(t)\}.$$

In particular, this implies that

$$F(x^{(t)}) - F(x^{(t+1)})$$
$$\geq \frac{5\gamma^3}{6L_2^2}\Big(\mathbf{1}\{\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\} - \mathbf{1}\{\neg \mathsf{E}^{\mathsf{Oja}}(t)\}\Big) - \mathbf{1}\{\neg \mathsf{E}^{\mathsf{Oja}}(t)\}(F(x^{(t)}) - F(x^{(t+1)})).$$

Taking conditional expectations, this further implies that

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0\Big] \geq \frac{5\gamma^3}{6L_2^2}\Big(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \Pr\Big(\neg \mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0\Big)\Big)$$
$$- \mathbb{E}\Big[\mathbf{1}\{\neg \mathsf{E}^{\mathsf{Oja}}(t)\}(F(x^{(t)}) - F(x^{(t+1)})) \mid Q_t = 0\Big].$$

Now, consider the term

$$\begin{aligned}&\mathbb{E}\Big[\mathbf{1}\{\neg\mathsf{E}^{\mathsf{Oja}}(t)\}(F(x^{(t)}) - F(x^{(t+1)})) \mid Q_t = 0\Big]\\ &= \Pr(\neg\mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0) \cdot \mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0, \neg\mathsf{E}^{\mathsf{Oja}}(t)\Big].\end{aligned}$$

Given that Oja fails, there are two cases two consider: The first case is where it returns $\perp$ (even though we may not have $\lambda_{\min}\big(\nabla^2 F(x^{(t)})\big) \geq -4\gamma$), which we denote by $P_t = 0$, and the second case is that it returns some vector $u^{(t)}$ (which may not actually satisfy $\nabla^2 F(x^{(t)})[u^{(t)}, u^{(t)}] \leq -2\gamma$), which we denote $P_t = 1$. If $P_t = 0$, we have $x^{(t+1)} - x^{(t)}$, so

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0, \neg\mathsf{E}^{\mathsf{Oja}}(t), P_t = 0\Big] = 0.$$

Otherwise, using a third-order Taylor expansion, and following the same reasoning as the proof of Lemma 15, we have

$$\begin{aligned}&\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0, \neg\mathsf{E}^{\mathsf{Oja}}(t), P_t = 1\Big]\\ &\leq \mathbb{E}\left[\frac{\gamma^2}{2L_2^2}\nabla^2 F(x)[u^{(t)}, u^{(t)}] + \frac{\gamma^3}{6L_2^2}\|u^{(t)}\|^3 \mid Q_t = 0, \neg\mathsf{E}^{\mathsf{Oja}}(t), P_t = 1\right]\\ &\leq \frac{\gamma^2}{2L_2^2}L_1 + \frac{\gamma^3}{6L_2^2} \leq \frac{2}{3}\frac{\gamma^2 L_1}{L_2^2}.\end{aligned}$$

Combining this bound with the earlier inequalities (and being rather loose with constants), we conclude that

$$\begin{aligned}&\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = 0\Big]\\ &\geq \frac{5\gamma^3}{6L_2^2}\left(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \left(1 + \frac{L_1}{\gamma}\right)\Pr\Big(\neg\mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0\Big)\right)\\ &\geq \frac{5\gamma^3}{6L_2^2}\left(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\Pr\Big(\neg\mathsf{E}^{\mathsf{Oja}}(t) \mid Q_t = 0\Big)\right).\end{aligned}$$

■

Lemma 18 *Under the same setting as Theorem 5, the point $\widehat{x}$ returned by Algorithm 4 satisfies*

$$\mathbb{E}\Big[\|\nabla F(x^{(t)})\|^2\Big] \leq 17\epsilon^2.$$

Proof. Starting from (45) in the proof of Theorem 5, we have

$$\begin{aligned}&\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big]\\ &\geq \frac{5(1-p)\gamma^3}{6L_2^2}\left(\Pr\Big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\Big) - \frac{2L_1}{\gamma}\delta\right) + p\left(\frac{\eta}{8}\mathbb{E}\Big[\|\nabla F(x^{(t)})\|^2\Big] - \frac{3\eta\epsilon^2}{4}\right).\end{aligned}$$

Ignoring the positive term $\Pr\big(\lambda_{\min}(\nabla^2 F(x^{(t)})) \le -4\gamma\big)$ on the right hand side in the above, we get

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big] \ge \frac{p\eta}{8}\Big(\mathbb{E}\Big[\|\nabla F(x^{(t)})\|^2\Big] - 6\epsilon^2\Big) - \frac{5(1-p)\gamma^3}{3L_2^2}\frac{L_1}{\gamma}\delta.$$

Telescoping this inequality for t from 1 to T and using that $F(x^{(1)}) - F(x^{(T+1)}) \le \Delta$, we get

$$\Delta \ge \frac{Tp\eta}{8}\big(\mathbb{E}\big[\|\nabla F(\widehat{x})\|^2\big] - 6\epsilon^2\big) - T\frac{5(1-p)\gamma^3}{3L_2^2}\frac{L_1}{\gamma}\delta \ge \frac{\Delta}{4\epsilon^2}\big(\mathbb{E}\big[\|\nabla F(\widehat{x})\|^2\big] - 12\epsilon^2\big) - 70\Delta\frac{L_1}{\gamma}\delta,$$

where the last inequality follows from Lemma 19. Rearranging the terms, we get

$$\mathbb{E}\big[\|\nabla F(\widehat{x})\|^2\big] \le 16\epsilon^2 + 280\epsilon^2 \cdot \frac{L_1}{\gamma}\delta \le 17\epsilon^2,$$

where the last inequality uses that $\delta \le \frac{\gamma}{1600L_1}$. ■

Lemma 19 *For the values of the parameters T and p specified in Algorithm 4,*

$$\frac{2\Delta}{\eta\epsilon^2} \le Tp \le \frac{4\Delta}{\eta\epsilon^2}, \quad \textit{and,} \quad \frac{20\Delta L_2^2}{\gamma^3} \le T(1-p) \le \frac{40\Delta L_2^2}{\gamma^3}.$$

Proof. Since, $\eta \le \frac{1}{2\sqrt{L_1^2+\bar{\sigma}_2^2+\epsilon L_2}} \le \frac{1}{2L_1}$ and $\epsilon \le \sqrt{\Delta L_1}$, we have that

$$T \ge \frac{2\Delta}{\eta\epsilon^2} \ge \frac{4\Delta L_1}{\epsilon^2} \ge 4.$$

Thus, using the fact that $x \le \lceil x \rceil \le 2x$ for all $x \ge 1$, we get

$$\frac{20\Delta L_2^2}{\gamma^3} + \frac{2\Delta}{\eta\epsilon^2} \le T \le \frac{40\Delta L_2^2}{\gamma^3} + \frac{4\Delta}{\eta\epsilon^2}. \tag{55}$$

Consequently, by plugging in the values of T and p, we have

$$\begin{aligned} T(1-p) &= \left\lceil \frac{20\Delta L_2^2}{\gamma^3} + \frac{2\Delta}{\eta\epsilon^2} \right\rceil \cdot \left(1 - \frac{\gamma^3}{\gamma^3 + 10\Delta L_2^2\eta\epsilon^2}\right) \\ &\le \left(\frac{40\Delta L_2^2}{\gamma^3} + \frac{4\Delta}{\eta\epsilon^2}\right) \cdot \left(\frac{10\Delta L_2^2\eta\epsilon^2}{\gamma^3 + 10\Delta L_2^2\eta\epsilon^2}\right) = \frac{40\Delta L_2^2}{\gamma^3}, \end{aligned}$$

where the first inequality is due to (55). Similarly, we have that

$$T(1-p) \ge \left(\frac{20\Delta L_2^2}{\gamma^3} + \frac{2\Delta}{\eta\epsilon^2}\right) \cdot \left(\frac{10\Delta L_2^2\eta\epsilon^2}{\gamma^3 + 10\Delta L_2^2\eta\epsilon^2}\right) = \frac{20\Delta L_2^2}{\gamma^3}.$$

Together, the above two bounds imply that

$$\frac{20\Delta L_2^2}{\gamma^3} \le T(1-p) \le \frac{40\Delta L_2^2}{\gamma^3}.$$

The bound on $T \cdot p$ follows similarly. ■

G.2. Full statement and proof for Algorithm 5

Algorithm 5 Subsampled cubic-regularized trust-region method with HVP-RVR

Input:

Stochastic second-order oracle $(\mathsf{O}_F^2, P_z) \in \mathcal{O}_2(F, \sigma_{1:2})$, where $F \in \mathcal{F}_2(\Delta, \infty, L_2)$.

Precision parameter ϵ.

Set $M = 4\max\left\{L_2, \frac{\sigma_2^2 \epsilon \log(d)}{\sigma_1^2}\right\}$, $\eta = 30\sqrt{\frac{\epsilon}{M}}$, $T = \left\lceil \frac{18\Delta L_2^2}{\gamma^3} + \frac{\Delta\sqrt{M}}{30\epsilon^{3/2}} \right\rceil$, $p = \frac{\sqrt{M}\gamma^{3/2}}{\sqrt{M}\gamma^{3/2} + 540 L_2^2 \epsilon^{3/2}}$.

Set $m_1 = \left\lceil \frac{2\cdot 10^4 \cdot \sigma_2^2 \log(d)}{\epsilon M} \right\rceil$, $m_2 = \left\lceil \frac{440\sigma_2^2 \log(d)}{\gamma^2} \right\rceil$.

Set $b_g = \min\left\{1, \frac{\eta\sqrt{\sigma_2^2 + \epsilon L_2}}{30\sigma_1}\right\}$ and $b_H = \min\left\{1, \frac{\gamma\sqrt{\sigma_2^2 + \epsilon L_2}}{\sigma_1 L_2}\right\}$.

Initialize $x^{(0)}, x^{(1)} \leftarrow 0$, $g^{(1)} \leftarrow \text{HVP-RVR-Gradient-Estimator}_{\epsilon, b_g}\left(x^{(1)}, x^{(0)}, \perp\right)$.

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

  Sample $Q_t \sim$ Bernoulli(p) with bias p.

  if $Q_t = 1$ **then**

    Query the oracle m_1 times at $x^{(t)}$ and compute

$$H_1^{(t)} \leftarrow \frac{1}{m_1} \sum_{j=1}^{m_1} \widehat{\nabla^2 F}(x^{(t)}, z^{(t,j)}), \quad \text{where} \quad z^{(t,j)} \overset{\text{i.i.d.}}{\sim} P_z.$$

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

$$x^{(t+1)} \leftarrow \operatorname*{arg\,min}_{\|y - x^{(t)}\| \le \eta} \left\langle g^{(t)}, y - x^{(t)} \right\rangle + \frac{1}{2}\left\langle y - x^{(t)}, H_1^{(t)}(y - x^{(t)}) \right\rangle + \frac{M}{6}\|y - x^{(t)}\|^3.$$

    $g^{(t+1)} \leftarrow \text{HVP-RVR-Gradient-Estimator}_{\epsilon, b_g}\left(x^{(t+1)}, x^{(t)}, g^{(t)}\right)$.

  else

    Query the oracle m_2 times at $x^{(t)}$ and compute

$$H_2^{(t)} \leftarrow \frac{1}{m_2} \sum_{j=1}^{m_2} \widehat{\nabla^2 F}(x^{(t)}, z^{(t,j)}), \quad \text{where} \quad z^{(t,j)} \overset{\text{i.i.d.}}{\sim} P_z.$$

    if $\lambda_{\min}\left(H_2^{(t)}\right) \le -4\gamma$ **then**

      Find a unit vector $u^{(t)}$ such that $H_2^{(t)}\left[u^{(t)}, u^{(t)}\right] \le -2\gamma$.

      $x^{(t+1)} \leftarrow x^{(t)} + \frac{\gamma}{L_2} \cdot r^{(t)} \cdot u^{(t)}$, where $r^{(t)} \sim \text{Uniform}(\{-1, 1\})$.

      $g^{(t+1)} \leftarrow \text{HVP-RVR-Gradient-Estimator}_{\epsilon, b_H}\left(x^{(t+1)}, x^{(t)}, g^{(t)}\right)$.

    else

      $x^{(t+1)} \leftarrow x^{(t)}$.

      $g^{(t+1)} \leftarrow g^{(t)}$.

return $\widehat{x}$ chosen uniformly at random from $\left\{x^{(t)}\right\}_{t=1}^{T-1}$.

Theorem 20 *For any function $F \in \mathcal{F}_2(\Delta, \infty, L_2)$, stochastic second order oracle in $\mathcal{O}_2(F, \sigma_1, \sigma_2)$, $\epsilon \leq \sigma_1$, and $\gamma \leq \min\{\sigma_2, \sqrt{\epsilon L_2}, \Delta^{\frac{1}{3}} L_2^{\frac{2}{3}}\}$, with probability at least $\frac{3}{5}$, Algorithm 5 returns a point $\widehat{x}$ such that*

$$\|\nabla F(\widehat{x})\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}\left(\nabla^2 F(\widehat{x})\right) \geq -\gamma,$$

and performs at most

$$\widetilde{O}\left(\frac{\Delta\sigma_1\sigma_2}{\epsilon^3} + \frac{\Delta L_2\sigma_1\sigma_2}{\gamma^2\epsilon^2} + \frac{\Delta L_2^2\sigma_2^2}{\gamma^5}\right)$$

stochastic gradient and Hessian queries.

Proof. We first show that Algorithm 5 returns a point $\widehat{x}$ such that, $\|\nabla F(\widehat{x})\| \leq 450\epsilon$ and $\lambda_{\min}\left(\nabla^2 F(\widehat{x})\right) \geq -4\gamma$. We then bound the expected number of oracle queries used throughout the execution.
Before we delve into the proof, first note that using Lemma 7, we have for all $t \geq 1$,

$$\mathbb{E}\left[\left\|\nabla F(x^{(t)}) - g^{(t)}\right\|^2\right] \leq \epsilon^2.$$

Further, using Lemma 9 with our choice of m_1 and m_2, we have, for all $t \geq 1$,

$$\mathbb{E}\left[\left\|\nabla^2 F(x^{(t)}) - H_1^{(t)}\right\|_{\text{op}}^2\right] \leq \frac{\epsilon M}{900}, \quad \text{and,} \quad \mathbb{E}\left[\left\|\nabla^2 F(x^{(t)}) - H_2^{(t)}\right\|_{\text{op}}^2\right] \leq \frac{\gamma^2}{20}. \tag{56}$$

To begin the proof, we observe that for any $t \geq 0$, there are two scenarios: (a) either $Q_t = 1$ and the algorithm goes through Line 9, or, (b) $Q_t = 0$ and the algorithm goes through Line 15. We analyze the two cases separately below.

(a) **Case 1:** $Q_t = 1$**.** In this case, we set $x^{(t+1)}$ using the update rule in Line 9. Invoking Lemma 14 with the bound in (56) and $\eta = 30\sqrt{\frac{\epsilon}{M}}$, we get

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \;\Big|\; Q_t = 1\Big] \geq \frac{450\epsilon^{3/2}}{\sqrt{M}}\left(\Pr\left(\left\|\nabla F(x^{(t+1)})\right\| \geq 450\epsilon\right) - \frac{1}{32}\right). \tag{57}$$

(b) **Case 2:** $Q_t = 0$**.** In this case, either $\lambda_{\min}\left(H_2^{(t)}\right) > -4\gamma$, in which case we set $x^{(t+1)} = x^{(t)}$, or we compute $x^{(t+1)}$ using the update rule in Line 15 in Algorithm 5. Thus, using Lemma 15 with (56), we get

$$\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \;\Big|\; Q_t = 0\Big] \geq \frac{5\gamma^3}{6L_2^2}\left(\Pr\left(\lambda_{\min}\left(H_2^{(t)}\right) \leq \gamma\right) - \frac{1}{32}\right). \tag{58}$$

Combining the two cases ($Q_t = 0$ or $Q_t = 1$) from (57) and (58) above, we get

$$\begin{aligned}\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)})\Big] &= \sum_{q\in\{0,1\}} \Pr(Q_t = q)\,\mathbb{E}\Big[F(x^{(t)}) - F(x^{(t+1)}) \mid Q_t = q\Big] \\ &\geq (1-p)\cdot\frac{5\gamma^3}{6L_2^2}\left(\Pr\left(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma\right) - \frac{1}{32}\right)\end{aligned}$$

$$+ p \cdot \frac{450\epsilon^{3/2}}{\sqrt{M}} \left(\Pr\left(\left\| \nabla F(x^{(t+1)}) \right\| \geq 450\epsilon \right) - \frac{1}{32} \right).$$

Telescoping the inequality above for t from 0 to $T-1$, and using the bound $\mathbb{E}\left[F(x^{(0)}) - F(x^{(T)})\right] \leq \Delta$, we get

$$\begin{aligned}
\Delta &\geq \mathbb{E}\left[F(x^{(0)}) - F(x^{(T)})\right] \\
&\geq \frac{5T(1-p)\gamma^3}{6L_2^2} \left(\frac{1}{T} \sum_{t=0}^{T-1} \Pr\left(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma \right) - \frac{1}{32} \right) \\
&\qquad + \frac{450Tp\epsilon^{3/2}}{\sqrt{M}} \left(\frac{1}{T} \sum_{t=1}^{T} \Pr\left(\left\| \nabla F(x^{(t)}) \right\| \geq 450\epsilon \right) - \frac{1}{32} \right) \\
&\overset{(i)}{\geq} 15\Delta \left(\frac{1}{T} \sum_{t=0}^{T-1} \Pr\left(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma \right) + \frac{1}{T} \sum_{t=1}^{T} \Pr\left(\left\| \nabla F(x^{(t)}) \right\| \geq 450\epsilon \right) - \frac{1}{8} \right) \\
&\overset{(ii)}{\geq} 15\Delta \left(\frac{5}{6(T-1)} \sum_{t=1}^{T-1} \left(\Pr\left(\lambda_{\min}(\nabla^2 F(x^{(t)})) \leq -4\gamma \right) + \Pr\left(\left\| \nabla F(x^{(t)}) \right\| \geq 450\epsilon \right) \right) - \frac{1}{8} \right) \\
&\overset{(iii)}{\geq} 15\Delta \left(\frac{5}{6} \left(\Pr\left(\lambda_{\min}(\nabla^2 F(\widehat{x})) \leq -4\gamma \right) + \Pr(\|\nabla F(\widehat{x})\| \geq 450\epsilon) \right) - \frac{1}{8} \right),
\end{aligned} \tag{59}$$

where the inequality in (i) follows from Lemma 21. The inequality in (ii) is given by ignoring the (non-negative) terms $\Pr\left(\nabla^2 F(x^{(0)}) \leq -4\gamma\right)$ and $\Pr\left(\left\|\nabla F(x^{(T)})\right\| \geq 450\epsilon\right)$ on the right-hand side and using the fact that $T \geq 6$. Finally, (iii) follows by recalling the definition of $\widehat{x}$ as samples uniformly at random from the set $(x^{(t)})_{t=1}^{T-1}$. Rearranging the terms, we get

$$\Pr\left(\lambda_{\min}(\nabla^2 F(\widehat{x})) \leq -4\gamma\right) + \Pr(\|\nabla F(\widehat{x})\| \geq 450\epsilon) \leq \frac{1}{4},$$

which further implies that the returned point $\widehat{x}$ satisfies

$$\Pr\left(\lambda_{\min}(\nabla^2 F(\widehat{x})) \geq -\gamma \wedge \|\nabla F(\widehat{x})\| \leq 450\epsilon\right) \geq \frac{3}{4}. \tag{60}$$

Bound on the number of oracle queries. Let us first introduce some notation to count the number of oracle calls made in each iteration of the algorithm.

- On Line 10 and Line 16, Algorithm 5 queries the stochastic oracle through the subroutine HVP-RVR-Gradient-Estimator. Let $m_g(t)$ denote the total number of oracle queries resulting from either line at iteration t.
- Let $m_{h,1}(t)$ and $m_{h,2}(t)$ denote the total number of oracle calls made by Line 8 and Line 12 at iteration t to compute $H_1^{(t)}$ and $H_2^{(t)}$ respectively.

Define M_g, $M_{h,1}$ and $M_{h,2}$ by $\sum_{t=1}^{T} m_g(t)$, $\sum_{t=1}^{T} m_{h,1}(t)$ and $\sum_{t=1}^{T} m_{h,2}(t)$ respectively. In what follows, we give separate bounds for $\mathbb{E}[M_g]$, $\mathbb{E}[M_{h,1}]$ and $\mathbb{E}[M_{h,2}]$. The final statement on the total number of oracle calls follows by an application of Markov's inequality.

Bound on $\mathbb{E}[M_g]$. For any $t > 0$, there are two scenarios, either (a) $Q_t = 1$ and we update $x^{(t+1)}$ through Line 9, or (b) $Q_t = 0$ and we update $x^{(t+1)}$ through Line 15 orLine 18. Thus, using the law of total expectation

$$\mathbb{E}[M_g] = \sum_{t=0}^{T-1} \Pr(Q_t = 0)\, \mathbb{E}[m_g(t) \mid Q_t = 0] + \sum_{t=0}^{T-1} \Pr(Q_t = 1)\, \mathbb{E}[m_g(t) \mid Q_t = 1]. \tag{61}$$

We denote the two terms on the right hand side above by $(\mathbf{A})$ and $(\mathbf{B})$, respectively. We bound them separately in as follows.

(a) **Bound on $(\mathbf{A})$.** Using Lemma 8 with the fact that $\Pr(Q_t = 0) = 1 - p$, we get

$$\begin{aligned}
(\mathbf{A}) &= 6 \sum_{t=1}^{T} (1-p) \cdot \mathbb{E}\left[\frac{b_H \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \left\| x^{(t+1)} - x^{(t)} \right\|^2}{b_H \epsilon^2} + 1 \,\middle|\, Q_t = 0 \right] \\
&\overset{(i)}{=} 6T(1-p) \cdot \left(\frac{b_H \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \gamma^2}{b_H \epsilon^2 L_2^2} + 1 \right) \\
&\overset{(ii)}{=} O\left(\frac{\Delta L_2^2}{\gamma^3} \cdot \left(\frac{b_H \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \gamma^2}{b_H \epsilon^2 L_2^2} + 1 \right) \right) \\
&\overset{(iii)}{=} O\left(\frac{\Delta L_2 \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta(\sigma_2^2 + \epsilon L_2)}{\gamma \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3} \right),
\end{aligned} \tag{62}$$

where (i) holds because when $Q_t = 0$, we either have $\left\| x^{(t)} - x^{(t-1)} \right\| \leq \frac{\gamma}{L_2}$ (if we follow the update rule in Line 15) or $\left\| x^{(t)} - x^{(t-1)} \right\| = 0$ (if we follow Line 18). The inequality (ii) uses the bound on $T \cdot (1-p)$ from Lemma 21 and (iii) follows from plugging in the value of b_H.

(b) **Bound on $(\mathbf{B})$.** Using Lemma 8 with the definition $\Pr(Q_t = 1) = p$, we get

$$\begin{aligned}
(\mathbf{B}) &= 6 \sum_{t=1}^{T} p \cdot \mathbb{E}\left[\frac{b_g \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \left\| x^{(t+1)} - x^{(t)} \right\|^2}{b_g \epsilon^2} + 1 \,\middle|\, Q_t = 1 \right] \\
&\overset{(i)}{=} 6Tp \cdot \left(\frac{b_g \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \eta^2}{b_g \epsilon^2} + 1 \right) \\
&\overset{(ii)}{=} O\left(\frac{\Delta \sqrt{M}}{\epsilon^{1.5}} \cdot \left(\frac{b_g \sigma_1^2}{\epsilon^2} + \frac{(\sigma_2^2 + L_2\epsilon) \cdot \eta^2}{b_g \epsilon^2} + 1 \right) \right) \\
&\overset{(iii)}{=} O\left(\frac{\Delta \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta \sqrt{M}}{\epsilon^{1.5}} \right) \\
&\overset{(iv)}{=} O\left(\frac{\Delta \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta \sqrt{L_2}}{\epsilon^{1.5}} + \frac{\Delta \sigma_2 \sqrt{\log(d)}}{\epsilon^2} \right),
\end{aligned} \tag{63}$$

where (i) is given by the update rule from Line 9 and the fact that HVP-RVR-Gradient-Estimator uses parameter b_g in this case, and (ii) follows by using the bound on $T \cdot p$ from Lemma 21. The inequality (iii) follows because for the choice of parameters η and M and the assumed range of ϵ in the theorem statement, $b_g = \frac{\eta \sqrt{\sigma_2^2 + \epsilon L_2}}{\sigma_1} < 1$. Finally, the inequality (iv) is given by plugging in the value of M and using that $\epsilon \leq \sigma_1$.

Plugging the bound in (62) and (63) back in (61), we get

$$\mathbb{E}[M_g] = O\left(\frac{\Delta L_2 \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta\left(\sigma_2^2 + \epsilon L_2\right)}{\gamma \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3}\right) + O\left(\frac{\Delta \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\epsilon^3} + \frac{\Delta \sqrt{L_2}}{\epsilon^{1.5}} + \frac{\Delta \sigma_2 \sqrt{\log(d)}}{\epsilon^2}\right). \tag{64}$$

Bound on $\mathbb{E}[M_{H,1}]$. For each $t \geq 0$, Algorithm 5 samples an independent Bernoulli Q_t with bias $\mathbb{E}[Q_t] = p$ and executes Line 8 if $Q_t = 1$. For every such pass through Line 8, the algorithm queries the stochastic Hessian oracle m_1 times. Thus,

$$\mathbb{E}[M_H] = \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{1}\{Q_t = 1\} \cdot m_1\right] = T \cdot p \cdot m_1 \overset{(i)}{=} O\left(\frac{\Delta\sqrt{M}}{\epsilon^{1.5}} \cdot \left\lceil \frac{900 \sigma_2^2 \log(d)}{\epsilon M} \right\rceil\right) = O\left(\frac{\Delta \sqrt{L_2}}{\epsilon^{1.5}} + \frac{\Delta \sigma_1 \sigma_2 \sqrt{\log(d)}}{\epsilon^3}\right), \tag{65}$$

where (i) follows by plugging in the values of m_1 and M as specified in Algorithm 5 (using that $\epsilon \leq \sigma_1$ to simplify), and using the bound on $T \cdot p$ from Lemma 21 .

Bound on $\mathbb{E}[M_{H,2}]$. The algorithm executes Line 12 only if $Q_t = 0$, which happens with probability $1 - p$. For every such pass through Line 12, the algorithm queries the stochastic Hessian oracle m_2 times. Consequently,

$$\mathbb{E}[M_H] = \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{1}\{Q_t = 0\} \cdot m_1\right] = T \cdot (1-p) \cdot m_1 \overset{(i)}{=} O\left(\frac{\Delta L_2^2}{\gamma^3} \cdot \left\lceil \frac{20 \sigma_2^2 \log(d)}{\gamma^2} \right\rceil\right) = O\left(\frac{\Delta L_2^2 \sigma_2^2 \log(d)}{\gamma^5} + \frac{\Delta L_2^2}{\gamma^3}\right), \tag{66}$$

where (i) follows by plugging in the values of m_1 as specified in Algorithm 5, and using the bound on $T \cdot p$ from Lemma 21.

Adding together all the bounds above (from (64), (65), and (66)), we have that the total number of oracle queries by Algorithm 5 till time T is bounded in expectation by

$$\begin{aligned}\mathbb{E}[M] &= \mathbb{E}[M_g + M_{H,1} + M_{H,2}] \\ &= O\left(\frac{\Delta L_2^2 \sigma_2^2 \log(d)}{\gamma^5} + \frac{\Delta L_2 \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta \sigma_1 \sigma_2 \sqrt{\log(d)}}{\epsilon^3} + \frac{\Delta \sigma_1 \sqrt{L_2}}{\epsilon^{2.5}}\right) \\ &\quad + O\left(\frac{\Delta\left(\sigma_2^2 + \epsilon L_2\right)}{\gamma \epsilon^2} + \frac{\Delta L_2^2}{\gamma^3} + \frac{\Delta \sigma_2 \sqrt{\log(d)}}{\epsilon^2} + \frac{\Delta \sqrt{L_2}}{\epsilon^{1.5}}\right).\end{aligned}$$

Using Markov's inequality, this implies that with probability at least $\frac{7}{8}$,

$$M = O\left(\frac{\Delta L_2^2 \sigma_2^2 \log(d)}{\gamma^5} + \frac{\Delta L_2 \sigma_1 \sqrt{\sigma_2^2 + \epsilon L_2}}{\gamma^2 \epsilon^2} + \frac{\Delta \sigma_1 \sigma_2 \sqrt{\log(d)}}{\epsilon^3} + \frac{\Delta \sigma_1 \sqrt{L_2}}{\epsilon^{2.5}}\right)$$

$$+ O\left(\frac{\Delta\left(\sigma_2^2+\epsilon L_2\right)}{\gamma\epsilon^2}+\frac{\Delta L_2^2}{\gamma^3}+\frac{\Delta\sigma_2\sqrt{\log(d)}}{\epsilon^2}+\frac{\Delta\sqrt{L_2}}{\epsilon^{1.5}}\right).$$

Ignoring the lower order terms, we have

$$M=\widetilde{O}\left(\frac{\Delta L_2^2\sigma_2^2}{\gamma^5}+\frac{\Delta L_2\sigma_1\sigma_2}{\gamma^2\epsilon^2}+\frac{\Delta\sigma_1\sigma_2}{\epsilon^3}\right). \tag{67}$$

The final statement follows by union bound, using the failure probabilities for (60) and (67). ■

Lemma 21 *For the values of the parameters T and p specified in Algorithm 5, we have*

$$\frac{\Delta\sqrt{M}}{30\epsilon^{\frac{3}{2}}}\le Tp\le\frac{2\Delta\sqrt{M}}{30\epsilon^{\frac{3}{2}}}\quad\text{and}\quad\frac{18\Delta L_2^2}{\gamma^3}\le T(1-p)\le\frac{36\Delta L_2^2}{\gamma^3}.$$

Proof. Under the assumption that $\gamma\le\Delta^{\frac{1}{3}}L_2^{\frac{2}{3}}$, we have that

$$T\ge\frac{18\Delta L_2^2}{\gamma^3}\ge 18.$$

Thus, using the fact that $x\le\lceil x\rceil\le 2x$ for any $x\ge 1$, we get

$$\frac{18\Delta L_2^2}{\gamma^3}+\frac{\Delta\sqrt{M}}{30\epsilon^{\frac{3}{2}}}\le T\le\frac{36\Delta L_2^2}{\gamma^3}+\frac{\Delta\sqrt{M}}{15\epsilon^{\frac{3}{2}}}. \tag{68}$$

Thus, plugging in the value of T and p, we get

$$\begin{aligned}T(1-p)&=\left\lceil\frac{18\Delta L_2^2}{\gamma^3}+\frac{\Delta\sqrt{M}}{30\epsilon^{\frac{3}{2}}}\right\rceil\cdot\left(1-\frac{\sqrt{M}\gamma^{\frac{3}{2}}}{\sqrt{M}\gamma^{\frac{3}{2}}+540L_2^2\epsilon^{\frac{3}{2}}}\right)\\&\le\left(\frac{36\Delta L_2^2}{\gamma^3}+\frac{\Delta\sqrt{M}}{15\epsilon^{\frac{3}{2}}}\right)\cdot\frac{540L_2^2\epsilon^{\frac{3}{2}}}{\sqrt{M}\gamma^{\frac{3}{2}}+540L_2^2\epsilon^{\frac{3}{2}}}=\frac{36\Delta L_2^2}{\gamma^3},\end{aligned}$$

where the first inequality is due to (68). Similarly, we have that

$$T(1-p)\ge\left\lceil\frac{18\Delta L_2^2}{\gamma^3}+\frac{\Delta\sqrt{M}}{30\epsilon^{\frac{3}{2}}}\right\rceil\cdot\frac{540L_2^2\epsilon^{\frac{3}{2}}}{\sqrt{M}\gamma^{\frac{3}{2}}+540L_2^2\epsilon^{\frac{3}{2}}}=\frac{18\Delta L_2^2}{\gamma^3}.$$

Together, the above two bounds imply that

$$\frac{18\Delta L_2^2}{\gamma^3}\le T(1-p)\le\frac{36\Delta L_2^2}{\gamma^3}$$

The bound on $T\cdot p$ follows similarly. ■

Appendix H. Lower bounds

H.1. Proof of Theorem 4

In this section, we prove Theorem 4. We begin by generalizing the lower bound framework of [8]—which centers around the notion of zero-respecting algorithms and stochastic gradient estimators called *probabilistic zero-chains*—to higher-order derivatives. Given a qth-order tensor $T \in \mathbb{R}^{\otimes^q d}$, we define $\operatorname{support}\{T\} \coloneqq \{i \in [d] \mid T_i \neq 0\}$, where T_i is the $(q-1)$-order subtensor defined by $[T_i]_{j_1,\ldots,j_{q-1}} = T_{i,j_1,\ldots,j_{q-1}}$. Given a tuple of tensors $\mathcal{T} = \big(T^{(1)}, T^{(2)}, \ldots\big)$, we let $\operatorname{support}\{\mathcal{T}\} \coloneqq \bigcup_i \operatorname{support}\{T^{(i)}\}$ be the union of the supports of $T^{(i)}$. Lastly, given an algorithm A and a an oracle O_F^p, we let $x^{(t)}_{\mathsf{A}[\mathsf{O}_F^p]}$ denote the (possibly randomized) tth query point generated by A when fed by information from O (i.e., $x^{(t)}_{\mathsf{A}[\mathsf{O}_F^p]}$ is a measurable function of $\big\{\mathsf{O}_F^p(x^{(i)}, z^{(i)})\big\}_{i=1}^{t-1}$, and possibly a random seed $r^{(t)}$).

Definition 22 *A stochastic pth-order algorithm* A *is* zero-respecting *if for any function F and any pth-order oracle O_F^p, the iterates $\{x^{(t)}\}_{t\in\mathbb{N}}$ produced by A by querying O_F^p satisfy*

$$\operatorname{support}\big(x^{(t)}\big) \subseteq \bigcup_{i<t} \operatorname{support}\big(\mathsf{O}_F^p(x^{(i)}, z^{(i)})\big), \quad \text{for all } t \in \mathbb{N}, \tag{69}$$

with probability one with respect the randomness of the algorithm and the realizations of $\{z^{(t)}\}_{t\in\mathbb{N}}$.

Given $x \in \mathbb{R}^d$, we define

$$\operatorname{prog}_\alpha(x) \coloneqq \max\{i \geq 0 \mid |x_i| > \alpha\} \ \text{ (where we set } x_0 \coloneqq 1), \tag{70}$$

which represents the highest index of x whose entry is α-far from zero, for some threshold $\alpha \in [0, 1)$. To lighten notation, we further let $\operatorname{prog} \coloneqq \operatorname{prog}_0$. For a tensor T, we let $\operatorname{prog}(T) \coloneqq \max\{\operatorname{support}\{T\}\}$ denote the highest index in $\operatorname{support}\{T\}$ (where $\operatorname{prog}(T) \coloneqq 0$ if $\operatorname{support}\{T\} = \emptyset$), and let $\operatorname{prog}(\mathcal{T}) \coloneqq \max_i \operatorname{prog}(T^{(i)})$ be the overall maximal index of $\operatorname{prog}(T^{(i)})$ for a tuple of tensors $\mathcal{T} = \big(T^{(1)}, T^{(2)}, \ldots\big)$.

Definition 23 *A collection of derivative estimators $\widehat{\nabla^1 F}(x,z), \ldots, \widehat{\nabla^p F}(x,z)$ for a function F forms a probability-ρ zero-chain if*

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

and

$$\Pr\Big(\exists x \ \mid \ \operatorname{prog}(\widehat{\nabla^1 F}(x,z), \ldots, \widehat{\nabla^p F}(x,z)) = \operatorname{prog}_{\frac{1}{4}}(x) + i\Big) = 0, \ i > 1.$$

No constraint is imposed for $i \leq \operatorname{prog}_{\frac{1}{4}}(x)$.

We note that the constant $1/4$ is used here for compatibility with the analysis in Arjevani et al. [8, Section 3]. Any non-negative constant less than $1/2$ would suffice in its place. The next lemma formalizes the idea that any zero-respecting algorithm interacting with a probabilistic zero-chain must wait many rounds to activate all the coordinates.

Lemma 24 *Let $\widehat{\nabla^1 F}(x,z),\dots,\widehat{\nabla^p F}(x,z)$ be a collection of probability-ρ zero-chain derivative estimators for $F:\mathbb{R}^T\to\mathbb{R}$, and let O_F^p be an oracle with $\mathsf{O}_F^p(x,z)=(\widehat{\nabla^q F}(x,z))_{q\in[p]}$. Let $\{x^{(t)}_{\mathsf{A}[\mathsf{O}_F]}\}$ be a sequence of queries produced by $\mathsf{A}\in\mathcal{A}_{\mathsf{zr}}(K)$ interacting with O_F^p. Then, with probability at least $1-\delta$,*

$$\operatorname{prog}\left(x^{(t)}_{\mathsf{A}[\mathsf{O}_F^p]}\right) < T, \quad \text{for all } t \le \frac{T-\log(1/\delta)}{2\rho}.$$

The proof of Lemma 24 is a simple adaptation of the proof of Lemma 1 of [8] to high-order zero-respecting methods—we provide it here for completeness. The proof idea is that any zero-respecting algorithm must activate coordinates in sequence, and must wait on average at least $\Omega(1/\rho)$ rounds between activations, leading to a total wait time of $\Omega(T/\rho)$ rounds.

Proof. Let $\{\widehat{\nabla^q F}(x^{(i)},z^{(i)})\}_{q\in[p]}$ denote the oracle responses for the ith query made at the point $x^{(i)}$, and let $\mathcal{G}^{(i)}$ be the natural filtration for the algorithm's iterates, the oracle randomness, and the oracle answers up to time i. We measure the progress of the algorithm through two quantities:

$$\begin{aligned}
\pi^{(t)} &:= \max_{i\le t}\operatorname{prog}\left(x^{(i)}\right) = \max\left\{j\le d \mid x_j^{(i)}\neq 0 \text{ for some } i\le t\right\},\\
\delta^{(t)} &:= \max_{i\le t}\operatorname{prog}\left(\nabla^q F(x^{(i)},z^{(i)})\right)\\
&= \max\left\{j\le d \mid \nabla^q f(x^{(i)},z^{(i)})_j\neq 0 \text{ for some } i\le t \text{ and } q\in[p]\right\}.
\end{aligned}$$

Note that $\pi^{(t)}$ is the largest non-zero coordinate in $\operatorname{support}\{(x^{(i)})_{i\le t}\}$, and that $\pi^{(0)}=0$ and $\delta^{(0)}=0$. Thus, for any zero-respecting algorithm

$$\pi^{(t)}\le\delta^{(t-1)}, \tag{71}$$

for all t. Moreover, observe that with probability one,

$$\operatorname{prog}\left(\nabla^q F(x^{(t)},z^{(t)})\right)\le 1+\operatorname{prog}_{\frac14}(x^{(t)})\le 1+\operatorname{prog}(x^{(t)})\le 1+\pi^{(t)}\le 1+\delta^{(t-1)}, \tag{72}$$

where the first inequality follows by the zero-chain property. Further, using the ρ-zero chain property, it follows that conditioned on $\mathcal{G}^{(i)}$, with probability at least $1-\rho$,

$$\operatorname{prog}\left(\nabla^q F(x^{(t)},z^{(t)})\right)\le \operatorname{prog}_{\frac14}(x^{(t)})\le\operatorname{prog}(x^{(t)})\le\pi^{(t)}\le\delta^{(t-1)}. \tag{73}$$

Combining (72) and (73), we have that conditioned on $\mathcal{G}^{(i-1)}$,

$$\delta^{(t-1)}\le\delta^{(t)}\le\delta^{(t-1)}+1 \qquad\text{and}\qquad \Pr\left[\delta^{(t)}=\delta^{(t-1)}+1\right]\le\rho.$$

Thus, denoting the increments $\iota^{(t)}:=\delta^{(t)}-\delta^{(t-1)}$, we have via the Chernoff method,

$$\begin{aligned}
\Pr\left[\delta^{(t)}\ge T\right] = \Pr\left[\sum_{j=1}^t \iota^{(j)}\ge T\right] &\le \frac{\mathbb{E}\left[\exp\left(\sum_{j=1}^t\iota^{(j)}\right)\right]}{\exp(T)} = e^{-T}\,\mathbb{E}\left[\prod_{i=1}^t \mathbb{E}\left[\exp\left(\iota^{(i)}\right)\mid\mathcal{G}^{(i-1)}\right]\right]\\
&\le e^{-T}(1-\rho+\rho\cdot e)^t \le e^{2\rho t-T}.
\end{aligned}$$

Thus, $\Pr\left[\delta^{(t)} \geq T\right] \leq \delta$ for all $t \leq \frac{T - \log(1/\delta)}{2\rho}$; combined with (71), this yields the desired result. ■

In light of Lemma 24, our lower bound strategy is as follows. We construct a function $F \in \mathcal{F}_p(\Delta, L_p)$ that both admits probability-ρ zero-chain derivative estimators and has large gradients for all $x \in \mathbb{R}^T$ with $\operatorname{prog}\left(x^{(i)}\right) < T$. Together with Lemma 24, this ensures that any zero-respecting algorithm interacting with a pth-order oracle must perform $\Omega(T/\rho)$ steps to make the gradient of F small. We make this approach concrete by adopting the construction used in [8], and adjusting it so as to be consistent with the additional high-order Lipschitz and variance parameters. For each $T \in \mathbb{N}$, we define

$$F_T(x) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^{T} [\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)], \tag{74}$$

where the component functions Ψ and Φ are

$$\Psi(x) = \begin{cases} 0, & x \leq 1/2, \\ \exp\left(1 - \frac{1}{(2x-1)^2}\right), & x > 1/2 \end{cases} \qquad \text{and} \qquad \Phi(x) = \sqrt{e} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt. \tag{75}$$

We start by collecting some relevant properties of F_T.

Lemma 25 (Carmon et al. [12]) *The function F_T satisfies:*

1. $F_T(0) - \inf_x F_T(x) \leq \Delta_0 \cdot T$, *where* $\Delta_0 = 12$.
2. *For $p \geq 1$, the pth order derivatives of F_T are ℓ_p-Lipschitz continuous, where $\ell_p \leq e^{\frac{5}{2}p\log p + cp}$ for a numerical constant $c < \infty$.*
3. *For all $x \in \mathbb{R}^T$, $p \in \mathbb{N}$ and $i \in [T]$, we have $\left\|\nabla_i^p F_T(x)\right\|_{\text{op}} \leq \ell_{p-1}$.*
4. *For all $x \in \mathbb{R}^T$ and $p \in \mathbb{N}$, $\operatorname{prog}(\nabla^p F_T(x)) \leq \operatorname{prog}_{\frac{1}{2}}(x) + 1$.*
5. *For all $x \in \mathbb{R}^T$, if $\operatorname{prog}_1(x) < T$ then $\|\nabla F_T(x)\| \geq |\nabla_{\operatorname{prog}_1(x)+1} F_T(x)| > 1$.*

Proof. Parts 1 and 2 follow from Lemma 3 in [12] and its proof; Part 3 is proven in Section H.1.1; Part 4 follows from Observation 3 in [12] and Part 5 is the same as Lemma 2 in [12]. ■

The derivative estimators we use are defined as

$$\left[\widehat{\nabla^q F_T}(x, z)\right]_i := \left(1 + \mathbf{1}\left\{i > \operatorname{prog}_{\frac{1}{4}}(x)\right\}\left(\frac{z}{\rho} - 1\right)\right) \cdot \nabla_i^q F_T(x), \tag{76}$$

where $z \sim \text{Bernoulli}(\rho)$.

Lemma 26 *The estimators $\widehat{\nabla^q F_T}$ form a probability-ρ zero-chain, are unbiased for $\nabla^q F_T$, and satisfy*

$$\mathbb{E}\left\|\widehat{\nabla^q F_T}(x, z) - \nabla^q F_T(x)\right\|^2 \leq \frac{\ell_{q-1}^2(1-\rho)}{\rho}, \quad \text{for all } x \in \mathbb{R}^T. \tag{77}$$

Proof. First, we observe that $\mathbb{E}\left[\widehat{\nabla^q F_T}(x,z)\right] = \nabla^q F_T(x)$ for all $x \in \mathbb{R}^T$, as $\mathbb{E}[z/\rho] = 1$. Second, we argue that the probability-ρ zero-chain property holds. Recall that $\mathrm{prog}_\alpha(x)$ is non-increasing in α (in particular, $\mathrm{prog}_{\frac{1}{4}}(x) \geq \mathrm{prog}_{\frac{1}{2}}(x)$). Therefore, by Lemma 25.4, $[\widehat{\nabla^q F_T}(x,z)]_i = \nabla_i F_T(x) = 0$ for all $i > \mathrm{prog}_{\frac{1}{4}}(x) + 1$, all $x \in \mathbb{R}^T$ and all $z \in \{0,1\}$. In addition, since $z \sim \mathrm{Bernoulli}(\rho)$, we have $\Pr\left(\exists x \;\middle|\; \mathrm{prog}(\widehat{\nabla^1 F_T}(x,z), \ldots, \widehat{\nabla^p F_T}(x,z)) = \mathrm{prog}_{\frac{1}{4}}(x) + 1\right) \leq \rho$, establishing that the oracle is a probability-ρ zero-chain.

To bound the variance of the derivative estimators, we observe that $\widehat{\nabla^q F_T}(x,z) - \nabla^q F_T(x)$ has at most one nonzero $(q-1)$-subtensor in the coordinate $i_x = \mathrm{prog}_{\frac{1}{4}}(x) + 1$. Therefore,

$$\mathbb{E}\|\widehat{\nabla^q F_T}(x,z) - \nabla^q F_T(x)\|^2 = \left\|\nabla^q_{i_x} F_T(x)\right\|^2 \mathbb{E}\left(\frac{z}{\rho} - 1\right)^2 = \left\|\nabla^q_{i_x} F_T(x)\right\|^2 \frac{1-\rho}{\rho} \leq \frac{(1-\rho)\ell_{q-1}^2}{\rho},$$

where the final inequality is due to Lemma 25.3, establishing the variance bound in (77). ∎

Proof of Theorem 4. We now prove the Theorem 4 by scaling the construction F_T appropriately. Let Δ_0 and ℓ_2 be the numerical constants in Lemma 25. Let the accuracy parameter ϵ, initial suboptimality Δ, derivative order $p \in \mathbb{N}$, smoothness parameters $L_1, \ldots, L_p$, and variance parameters $\sigma_1, \ldots, \sigma_p$ be fixed. We set

$$F_T^\star(x) = \alpha F_T\left(\beta x\right),$$

for some scalars α and β to be determined. The relevant properties of $F_T^\star$ scale as follows

$$F_T^\star(0) - \inf_x F_T^\star(x) = \alpha\big(F_T\left(0\right) - \inf_x F_T\left(\alpha x\right)\big) \leq \alpha\Delta_0 T, \tag{78}$$

$$\left\|\nabla^{q+1} F_T^\star(x)\right\| = \alpha\beta^{q+1}\left\|\nabla^{q+1} F_T\left(\beta x\right)\right\| \leq \alpha\beta^{q+1}\ell_q, \tag{79}$$

$$\|\nabla F_T^\star(x)\| \geq \alpha\beta\|\nabla F_T(x)\| \geq \alpha\beta, \;\forall x \text{ s.t., } \mathrm{prog}_1(x) < T. \tag{80}$$

The corresponding scaled derivative estimators $\widehat{\nabla^q F_T^\star}(x,z) = \alpha\beta^q\widehat{\nabla^q F_T}(\beta x, z)$ clearly form a probability-ρ zero-chain. Therefore, by Lemma 24, we have that for every zero respecting algorithm A interacting with $\mathsf{O}^p_{F_T^\star}$, with probability at least $1/2$, $\mathrm{prog}\left(x^{(t)}_{\mathsf{A}[\mathsf{O}^p_{F_T^\star}]}\right) < T$ for all $t \leq (T-1)/2\rho$. Hence, since $\mathrm{prog}_1(x) \leq \mathrm{prog}(x)$ for any $x \in \mathbb{R}^T$, we have by Lemma 25,

$$\mathbb{E}\|\nabla F_T^\star\big(x^{(t)}_{\mathsf{A}[\mathsf{O}^p_{F_T^\star}]}\big)\| = \alpha\beta\mathbb{E}\|\nabla F_T\big(\beta x^{(t)}_{\mathsf{A}[\mathsf{O}^p_{F_T^\star}]}\big)\| \geq \frac{\alpha\beta}{2}, \quad \forall t \leq (T-1)/2\rho. \tag{81}$$

We bound the variance of the scaled derivative estimators as

$$\mathbb{E}\|\widehat{\nabla^q F_T^\star}(x,z) - \nabla^q F_T^\star(x)\|^2 = \alpha^2\beta^{2q}\mathbb{E}\left\|\widehat{\nabla^q F_T}\left(\beta x, z\right) - \nabla^q F_T\left(\beta x\right)\right\|^2 \leq \frac{\alpha^2\beta^{2q}\ell_{q-1}^2(1-\rho)}{\rho},$$

where the last inequality follows by Lemma 26. Our goal now is to meet the following set of constraints:

- Δ-constraint: $\quad \alpha\Delta_0 T \leq \Delta$

- L_q-constraint: $\alpha\beta^{q+1}\ell_q \le L_q$, for $q \in [p]$
- ϵ-constraint: $\frac{\alpha\beta}{2} \ge \epsilon$
- σ_q-constraint: $\frac{\alpha^2\beta^{2q}\ell_{q-1}^2(1-\rho)}{\rho} \le \sigma_q^2$, for $q \in [p]$

Generically, since there are more inequalities to satisfy than the number of degrees of freedom (α, β, T and ρ) in our construction, not all inequalities can be activated (that is, met by equality) simultaneously. Different compromises will yield different rates.

First, to have a tight dependence in terms of ϵ, we activate the ϵ-constraint by setting $\alpha = 2\epsilon/\beta$. Next, we activate the σ_1-constraint, by setting $\rho = \min\{(\alpha\beta\ell_0/\sigma_1)^2, 1\} = \min\{(2\epsilon\ell_0/\sigma_1)^2, 1\}$. The bound on the variance of the qth-order derivative now reads

$$\frac{\alpha^2\beta^{2q}\ell_{q-1}^2(1-\rho)}{\rho} \le \frac{\sigma_1^2\alpha^2\beta^{2q}\ell_{q-1}^2}{(\alpha\beta\ell_0)^2} = \frac{\ell_{q-1}^2\beta^{2(q-1)}\sigma_1^2}{\ell_0^2}, \quad q = 2, \dots, p.$$

Since β is the only degree of freedom which can be tuned to meet though (not necessarily activate) the σ_q-constraint for $q = 2, \dots, p$ and the L_q-constraints for $q = 1, \dots, p$, we are forced to set

$$\beta = \min_{\substack{q=2,\dots,p \\ q'=1,\dots,p}} \min\left\{ \left(\frac{\ell_0\sigma_q}{\ell_{q-1}\sigma_1}\right)^{\frac{1}{q-1}}, \left(\frac{L_{q'}}{2\epsilon\ell_{q'}}\right)^{1/q'} \right\}. \tag{82}$$

Lastly, we activate the Δ-constraint by setting

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

Assuming $(2\epsilon\ell_0/\sigma_1)^2 \le 1$ and $T \ge 3$, we have by (81) that the number of oracle queries required to obtain an ϵ-stationary point for $G_T^\star$ is bounded from below by

$$\begin{aligned}
\frac{T-1}{2\rho} &= \frac{1}{2\rho}\left(\left\lfloor \frac{\Delta\beta}{2\Delta_0\epsilon} \right\rfloor - 1 \right) \\
&\overset{(\star)}{\ge} \frac{1}{2\rho} \cdot \frac{\Delta\beta}{4\Delta_0\epsilon} \\
&\ge \frac{\sigma_1^2}{2(2\ell_0\epsilon)^2} \cdot \frac{\Delta}{4\Delta_0\epsilon} \cdot \min_{\substack{q=2,\dots,p \\ q'=1,\dots,p}} \min\left\{ \left(\frac{\ell_0\sigma_q}{\ell_{q-1}\sigma_1}\right)^{\frac{1}{q-1}}, \left(\frac{L_{q'}}{2\epsilon\ell_{q'}}\right)^{1/q'} \right\} \\
&\ge \frac{\Delta\sigma_1^2}{2^5\Delta_0\ell_0^2\epsilon^3} \cdot \min_{\substack{q=2,\dots,p \\ q'=1,\dots,p}} \min\left\{ \left(\frac{\ell_0\sigma_q}{\ell_{q-1}\sigma_1}\right)^{\frac{1}{q-1}}, \left(\frac{L_{q'}}{2\epsilon\ell_{q'}}\right)^{1/q'} \right\},
\end{aligned} \tag{83}$$

where $(\star)$ uses $\lfloor \xi \rfloor - 1 \ge \xi/2$ whenever $\xi \ge 3$, implying the desired bound. Lastly, we note that one can obtain tight lower complexity bounds for deterministic oracles by setting $\rho = 1$. Following the same chain of inequalities as in (83), in this case we get a lower oracle-complexity bound of

$$\frac{\Delta}{8\Delta_0\epsilon} \min_{q=1,\dots,p} \left(\frac{L_q}{2\epsilon\ell_q}\right)^{1/q}. \tag{84}$$

■

H.1.1. BOUNDING THE OPERATOR NORM OF $\nabla_i^p F_T$

In this subsection we complete the proof of Lemma 25 by proving Part 3. Our proof follows along the lines of the proof of Lemma 3 of [12]. Let $x \in \mathbb{R}^T$ and $i_1, \dots, i_p \in [T]$, and note that by the chain-like structure of F_T, $\partial_{i_1} \cdots \partial_{i_p} F_T(x)$ is non-zero if and only if $|i_j - i_k| \le 1$ for any $j, k \in [p]$. A straightforward calculation yields

$$|\partial_{i_1} \cdots \partial_{i_p} F_T(x)| \le \max_{i \in [T]} \max_{\delta \in \{0,1\}^{p-1} \cup \{0,-1\}^{p-1}} |\partial_{i+\delta_1} \cdots \partial_{i+\delta_{p-1}} \partial_i F_T(x)| \tag{85}$$
$$\le \max_{k \in [p]} \left\{ 2 \sup_{\xi \in \mathbb{R}} \left| \Psi^k(\xi) \right| \sup_{\xi' \in \mathbb{R}} \left| \Phi^{p-k}(\xi') \right| \right\} \le \exp(2.5p \log p + 4p + 9) \le \frac{\ell_{p-1}}{2^{p+1}},$$

where the penultimate inequality is due to Lemma 1 of [12]. Therefore, for a fixed $i \in [T]$, we have

$$\begin{aligned}
\|\nabla_i^p F_T(x)\|_{\mathrm{op}} &\overset{(a)}{=} \sup_{\|v\|=1} |\langle \nabla_i^p F_T(x), v \rangle| \\
&= \sup_{\|v\|=1} \left| \sum_{i_1, \dots, i_{p-1} \in [T]} \partial_{i_1} \cdots \partial_{i_{p-1}} \partial_i F_T(x) v_{i_1} \cdots v_{i_{p-1}} \right| \\
&\overset{(b)}{\le} \sum_{\delta \in \{0,1\}^{p-1} \cup \{0,-1\}^{p-1}} |\partial_{i+\delta_1} \cdots \partial_{i+\delta_{p-1}} \partial_i F_T(x)| \\
&\overset{(c)}{\le} (2^p - 1) \frac{\ell_{p-1}}{2^{p+1}} \le \ell_{p-1},
\end{aligned}$$

where (a) follows from the definition of the operator norm, (b) follows by the chain-like structure of F_T, and (c) follows from (85), concluding the proof.

H.2. Proof of Theorem 6

In this section we prove Theorem 6 following the schema outlined in Section 4.2. We start by collecting all the relevant properties of Ψ and Λ from the construction in (11).

Lemma 27 *The functions Ψ and Λ satisfy the following properties:*

1. *For all $x \le 1/2$ and for all $k \in \mathbb{N} \cup \{0\}$, $\Psi^{(k)}(x) = 0$.*
2. *The function Ψ is non-negative and its first- and second-order derivatives are bounded by*
$$0 \le \Psi \le e, \qquad 0 \le \Psi' \le \sqrt{54/e}, \qquad -40 \le \Psi'' \le 40.$$
3. *The function Λ and its first- and second-order derivatives are bounded by*
$$-8 \le \Lambda \le 0, \qquad -6 \le \Lambda' \le 6, \qquad -8 \le \Lambda'' \le 4.$$
4. *Both Ψ and Λ are infinitely differentiable, and for all $k \in \mathbb{N}$, we have*
$$\sup_x \left| \Psi^{(k)}(x) \right| \le \exp\left(\frac{5k}{2} \log(4k) \right) \quad \text{and} \quad \sup_x \left| \Lambda^{(k)}(x) \right| \le \frac{8}{\sqrt{e}} \cdot \exp\left(\frac{3(k+1)}{2} \log\left(\frac{3(k+1)}{2} \right) \right).$$

Proof. Parts 1-4 are immediate. Part 5 follows from Lemma 1 of [12] and by noting that

$$\sup_x \left|\Lambda^{(k)}(x)\right| = \frac{8}{\sqrt{e}} \sup_x \left|\Phi^{(k+1)}(x)\right| \leq \frac{8}{\sqrt{e}} \cdot \exp\left(\frac{3(k+1)}{2} \log\left(\frac{3(k+1)}{2}\right)\right).$$

■

Using these basic properties of Ψ and Λ, we establish the following properties of the construction G_T (analogous to Lemma 25).

Lemma 28 *The function G_T satisfies the following properties:*

1. $G_T(0) - \inf_x(G_T(x)) \leq \overline{\Delta}_0 T$, *with* $\overline{\Delta}_0 = 40$.
2. *For $p \geq 1$, the pth order derivatives of G_T are $\tilde{\ell}_p$-Lipschitz continuous, where $\tilde{\ell}_p \leq e^{cp \log p + c'p}$ for a numerical constant $c, c' < \infty$.*
3. *For all $x \in \mathbb{R}^T$, and $i \in [T]$, we have $f \left\|\nabla_i^p G_T(x)\right\|_{\text{op}} \leq \tilde{\ell}_p$.*
4. *For all $x \in \mathbb{R}^T$ and $q \in [p]$, $\text{prog}\big(\nabla^{(q)} G_T(x)\big) \leq \text{prog}_{\frac{1}{2}}(x) + 1$.*
5. *For all $x \in \mathbb{R}^T$, if $\text{prog}_{\frac{9}{10}}(x) < T-1$ then $\lambda_{\min}\big(\nabla^2 G_T(x)\big) \leq -0.5$, and $\lambda_{\min}\big(\nabla^2 G_T(x)\big) \leq 700$ otherwise.*

Proof. We prove the individual parts of the lemma one by one:

1. Since $\Psi(0) = \Lambda(0) = 0$, we have

$$G_T(0) = \Psi(1)\Lambda(0) + \sum_{i=2}^{T} \left[\Psi(0)\Lambda(0) + \Psi(0)\Lambda(0))\right] = -\Psi(1)\Lambda(0) = 0.$$

   On the other hand,

$$\begin{aligned} G_T(x) &= \Psi(1)\Lambda(x_1) + \sum_{i=2}^{T} \left[\Psi(-x_{i-1})\Lambda(-x_i) + \Psi(x_{i-1})\Lambda(x_i))\right] \\ &\geq -8eT \qquad \text{(by Lemma 27.2. and Lemma 27.3)} \\ &\geq -40T. \end{aligned}$$

2. The proof follows along the same lines of Lemma 3 of [12] together with the derivative bounds stated in Lemma 27.4.
3. The claim follows using the same calculation as in Section H.1.1, with the derivative bounds replaced by those in Lemma 27.4, mutatis mutandis.
4. The claim follows Observation 3 in [12], mutatis mutandis.

5. We have

$$\frac{\partial G_T}{\partial x_j} = -\Psi(-x_{j-1})\Lambda'(-x_j) + \Psi(x_{j-1})\Lambda'(x_j) - \Psi'(-x_j)\Lambda(-x_{j+1}) + \Psi'(x_j)\Lambda(x_{j+1}). \tag{86}$$

Therefore, for any $x \in \mathbb{R}^d$, $\nabla^2 G_T(x)$ is a tridiagonal matrix specified as follows.

$$\nabla^2 G_T(x)_{i,j} = \begin{cases} \Psi(-x_{i-1})\Lambda''(-x_i) + \Psi(x_{i-1})\Lambda''(x_i) & \\ \qquad +\Psi''(-x_i)\Lambda(-x_{i+1}) + \Psi''(x_i)\Lambda(x_{i+1}) & \text{if } i = j, \\ \Psi'(-x_j)\Lambda'(-x_i) + \Psi'(x_j)\Lambda'(x_i) & \text{if } j = i-1, \\ \Psi'(-x_i)\Lambda'(-x_j) + \Psi'(x_i)\Lambda'(x_j) & \text{if } j = i+1, \\ 0 & \text{otherwise.} \end{cases}$$

The following facts can be verified by a straightforward calculation:

(i) $\Psi(x) \geq 0.5$ for all $x \geq 9/10$.

(ii) $\Psi''(x) \geq 0$ for all $|x| < 9/10$.

(iii) $\Lambda''(x) \leq -1$ for all $|x| < 9/10$.

Next, assuming $k \coloneqq \operatorname{prog}_{\frac{9}{10}}(x) + 1 < T$, we have, by definition, that $|x_{k+1}|, |x_k| < \frac{9}{10} \leq |x_{k-1}|$, implying,

$$\begin{aligned} \lambda_{\min}(\nabla^2 G_T(x)) &= \min_{y \in \mathbb{R}^n} \frac{y^T \nabla^2 G_T(x) y}{y^T y} && \text{(Rayleigh quotient)} \\ &\leq \frac{e_k^T \nabla^2 G_T(x) e_k}{e_k^T e_k} \\ &= \nabla^2 G_T(x)_{k,k} \\ &= \Psi(-x_{k-1})\Lambda''(-x_k) + \Psi(x_{k-1})\Lambda''(x_k) \\ &+ \Psi''(-x_k)\Lambda(-x_{k+1}) + \Psi''(x_k)\Lambda(x_{k+1}) \\ &\leq \Psi(-x_{k-1})\Lambda''(-x_k) + \Psi(x_{k-1})\Lambda''(x_k) && \text{((ii) and } \Lambda \leq 0) \\ &= \Psi(|x_{k-1}|)\Lambda''(\operatorname{sign}\{x_{k-1}\}x_k) && (\Psi(x) = 0,\ \forall x < 0) \\ &\leq -1 \cdot 0.5 = -0.5. && \text{((i) and (iii))} \end{aligned}$$

Otherwise, if nothing is assumed on x, then the same chain of inequalities, using $k = 2$, can be used to bound the minimal value of $\nabla^2 G_T(x)$.

$$\begin{aligned} \lambda_{\min}(\nabla^2 G_T(x)) &= \min_{y \in \mathbb{R}^n} \frac{y^T \nabla^2 G_T(x) y}{y^T y} && \text{(Rayleigh quotient)} \\ &\leq \frac{e_k^T \nabla^2 G_T(x) e_k}{e_k^T e_k} \\ &= \nabla^2 G_T(x)_{k,k} \\ &= \Psi(-x_{k-1})\Lambda''(-x_k) + \Psi(x_{k-1})\Lambda''(x_k) \end{aligned}$$

$$
\begin{aligned}
&+ \Psi''(-x_k)\Lambda(-x_{k+1}) + \Psi''(x_k)\Lambda(x_{k+1}) \\
&\leq 2(4e + 320) \leq 700,
\end{aligned}
$$

thus giving the desired bound.

■

We employ similar derivative estimators to the proof of Theorem 4, only this time we provide a noiseless estimate for the gradient. Formally, we set

$$
\left[\widehat{\nabla^q G_T}(x, z)\right]_i := \begin{cases} \nabla_i G_T(x) & q = 1, \\ \left(1 + \mathbf{1}\left\{i > \mathrm{prog}_{\frac{1}{4}}(x)\right\}\left(\frac{z}{p} - 1\right)\right) \cdot \nabla_i^q G_T(x) & q \geq 2, \end{cases} \tag{87}
$$

where $z \sim \mathrm{Bernoulli}(\rho)$. The dynamics of zero-respecting methods can be now characterized in an analogous way to the proof of Theorem 4. The only difference is that here, since $\Lambda'(0) = \Psi'(0) = 0$, it follows that $\mathrm{prog}_0(\nabla G_T(x)) = \mathrm{prog}_0(x)$. Therefore, the collection of estimators defined above is a ρ-probability zero-chain—with respect to prog_0 (rather than $\mathrm{prog}_{\frac{1}{4}}$ as in Definition 23)[11]—in which the variance of the gradient estimator is 0; a key property that shall be used soon. Following the proof of Lemma 24, mutatis mutandis, gives us the same bound on the number of non-zero entries acquired over time. That is, we have that with probability at least $1 - \delta$,

$$
\mathrm{prog}\left(x^{(t)}_{\mathsf{A}[\mathsf{O}^p_F]}\right) < T, \quad \text{for all } t \leq \frac{T - \log(1/\delta)}{2\rho}, \tag{88}
$$

where we employ the same notation as in Lemma 24. The proof now proceeds along the same lines of the proof of Theorem 4. The estimators have variance bounded as

$$
\mathbb{E}\left\|\widehat{\nabla^q G_T}(x, z) - \nabla^q G_T(x)\right\|^2 \leq \begin{cases} 0 & q = 1, \\ \frac{\tilde{\ell}^2_{q-1}(1-\rho)}{\rho}, \quad \text{for all } x \in \mathbb{R}^T & q \geq 2, \end{cases} \tag{89}
$$

which can established the same fashion as Lemma 26 by invoking Lemma 28.3 and Lemma 28.4.

Proof of Theorem 6. We now complete the proof of Theorem 6 for $p \geq 2$ by scaling G_T appropriately. Let Δ_0 and $\tilde{\ell}_p$ be the numerical constants in Lemma 28. Let the accuracy parameter γ, initial suboptimality Δ, derivative order $p \in \mathbb{N}$, smoothness parameter $L_1, \ldots, L_p$, and variance parameter $\sigma_1, \sigma_2, \ldots, \sigma_p$ be fixed. We let

$$
G^\star_T(x) := \alpha G_T\left(\beta x\right),
$$

for scalars α and β to be determined. The relevant properties of $G^\star_T$ are as follows:

$$
G^\star_T(0) - \inf_x G^\star_T(x) = \alpha\big(G_T\left(0\right) - \inf_x G_T\left(\alpha x\right)\big) \leq \alpha\tilde{\Delta}_0 T, \tag{90}
$$

11. Using prog_0, rather than $\mathrm{prog}_{\frac{1}{4}}$, carries one major disadvantage: our bounds for finding γ-weakly convex points cannot be directly extended to arbitrary randomized algorithm using the technique presented in Section 3.4 of [12] as is (at least, not without the degrading the dependence on problem parameters). We defer such an extension to future work.

$$\left\|\nabla^{q+1} G_T^\star(x)\right\| = \alpha\beta^{q+1}\left\|\nabla^{q+1} G_T\left(\beta x\right)\right\| \le \alpha\beta^{q+1}\tilde{\ell}_q, \tag{91}$$

$$\lambda_{\min}(\nabla^2 G_T^\star(x)) = \alpha\beta^2\lambda_{\min}(\nabla^2 G_T(x)) \le -\frac{\alpha\beta^2}{2}, \qquad \forall x \text{ s.t., } \mathrm{prog}_{9/10}(x) < T. \tag{92}$$

The corresponding scaled derivative estimators $\widehat{\nabla^q G_T^\star}(x,z) = \alpha\beta^q\widehat{\nabla^q G_T}(\beta x, z)$ clearly form a probability-ρ zero-chain, thus by (88), we have that for every zero respecting algorithm A interacting with $\mathsf{O}^p_{G_T^\star}$, with probability at least $1 - 1/(4\cdot 700)$, $\mathrm{prog}\left(x^{(t)}_{\mathsf{A}[\mathsf{O}^p_F]}\right) < T-1$ for all $t \le (T-2)/2\rho$. Therefore, since $\mathrm{prog}_{9/10}(x) \le \mathrm{prog}(x)$ for any $x \in \mathbb{R}^T$, we have by Lemma 28.5,

$$\begin{aligned}
\mathbb{E}\left[\lambda_{\min}(\nabla^2 G_T^\star(x^{(t)}_{\mathsf{A}[\mathsf{O}^p_F]}))\right] &= \alpha\beta^2\lambda_{\min}(\nabla^2 G_T(\beta x^{(t)}_{\mathsf{A}[\mathsf{O}^p_F]})) \\
&\le \alpha\beta^2\left(-0.5\cdot(1-\frac{1}{4\cdot 700}) + 700\cdot\frac{1}{4\cdot 700}\right) \\
&\le \frac{-\alpha\beta^2}{5},
\end{aligned} \tag{93}$$

for any $t \le (T-2)/2\rho$. The variance of the scaled derivative estimators can be bounded as

$$\mathbb{E}\|\widehat{\nabla^q G_T^\star}(x,z) - \nabla^q G_T^\star(x)\|^2 = \alpha^2\beta^{2q}\mathbb{E}\left\|\widehat{\nabla^q G_T}\left(\beta x, z\right) - \nabla^q G_T\left(\beta x\right)\right\|^2 \le \frac{\alpha^2\beta^{2q}\tilde{\ell}_{q-1}^2(1-\rho)}{\rho},$$

where the last inequality is by (89). Our goal now is to meet the following set of constraints:

- Δ-constraint: $\alpha\tilde{\Delta}_0 T \le \Delta$.
- L_q-constraint: $\alpha\beta^{q+1}\tilde{\ell}_q \le L_q$ for $q = 1,\ldots,p$.
- γ-constraint: $-\frac{\alpha\beta^2}{5} \le -\gamma$.
- σ_q-constraint: $\frac{\alpha^2\beta^{2q}\tilde{\ell}_{q-1}^2(1-\rho)}{\rho} \le \sigma_q^2$ for $q = 1,\ldots,p$.

As there are more inequalities to satisfy than the four degrees of freedom (α, β, T and ρ) in our construction, generically, not all inequalities can be activated (that is, met by equality) simultaneously. Different compromises may yield different bounds. First, to have a tight dependence in terms of γ, we activate the γ-constraint by setting $\alpha = 5\gamma/\beta^2$. Next, we activate the σ_2-constraint, by setting $\rho = \min\{(\alpha\beta^2\tilde{\ell}_1/\sigma_2)^2, 1\} = \min\{(5\tilde{\ell}_1\gamma/\sigma_2)^2, 1\}$. The bound on the variance of the qth derivative for $q = 3,\ldots,p$, now reads

$$\frac{\alpha^2\beta^{2q}\tilde{\ell}_{q-1}^2(1-\rho)}{\rho} \le \frac{\sigma_2^2\alpha^2\beta^{2q}\tilde{\ell}_{q-1}^2}{(\alpha\beta^2\tilde{\ell}_1)^2} = \frac{\tilde{\ell}_{q-1}^2\beta^{2(q-2)}\sigma_2^2}{\tilde{\ell}_1^2}, \quad q = 3,\ldots,p.$$

Since β is the only degree of freedom which can be tuned to meet (though not necessarily activate) the σ_q-constraints for $q = 3,\ldots,p$, and the $L_{q'}$-constraint for $q' = 2,\ldots,p$, we are forced to have

$$\beta = \min_{\substack{q=3,\ldots,p \\ q'=2,\ldots,p}} \min\left\{\left(\frac{\tilde{\ell}_1\sigma_q}{\tilde{\ell}_{q-1}\sigma_2}\right)^{\frac{1}{q-2}}, \left(\frac{L_{q'}}{5\tilde{\ell}_{q'}\gamma}\right)^{\frac{1}{q'-1}}\right\}. \tag{94}$$

Note that, by definition, the σ_1-constraint always holds (as the variance of the gradient estimator is zero, see (89)). To satisfy the L_1-constraint, i.e., $\alpha\beta^2\tilde{\ell}_1 \leq L_1$, we must have

$$\gamma \leq L_1/5\tilde{\ell}_1. \tag{95}$$

This constraint holds w.l.o.g. as L_1 also bounds the absolute value of the Hessian eigenvalues (in other words, any point x is trivially $O(L_1)$-weakly convex). Lastly, we activate the Δ-constraint, by setting

$$T = \left\lfloor \frac{\Delta}{\alpha\tilde{\Delta}_0} \right\rfloor = \left\lfloor \frac{\Delta\beta^2}{5\tilde{\Delta}_0\gamma} \right\rfloor.$$

Assuming $(5\tilde{\ell}_1\gamma/\sigma_2)^2 \leq 1$ (i.e., $\gamma = O(\sigma_2)$) and $T \geq 3$, we have by (93) that the number of oracle queries required to obtain a point x such that $\lambda_{\min}(\nabla^2 G_T^\star(x)) \leq \lambda$, is bounded from below by

$$\begin{aligned}
\frac{T-2}{2\rho} &= \frac{1}{2\rho}\left(\left\lfloor \frac{\Delta\beta^2}{5\tilde{\Delta}_0\gamma} \right\rfloor - 2\right) \\
&\overset{(\star)}{\geq} \frac{1}{2\rho}\frac{\Delta\beta^2}{5^2\tilde{\Delta}_0\gamma} \\
&\geq \frac{\sigma_2^2}{(5\tilde{\ell}_1\gamma)^2}\cdot\frac{\Delta\beta^2}{5^2\tilde{\Delta}_0\gamma} \\
&= \frac{\sigma_2^2}{(5\tilde{\ell}_1\gamma)^2}\cdot\frac{\Delta}{5^2\tilde{\Delta}_0\gamma}\min_{\substack{q=3,\ldots,p \\ q'=2,\ldots,p}} \min\left\{\left(\frac{\tilde{\ell}_1\sigma_q}{\tilde{\ell}_{q-1}\sigma_2}\right)^{\frac{2}{q-2}}, \left(\frac{L_{q'}}{5\tilde{\ell}_{q'}\gamma}\right)^{\frac{2}{q'-1}}\right\} \\
&= \frac{1}{5^4\tilde{\ell}_1^2\tilde{\Delta}_0}\cdot\frac{\Delta\sigma_2^2}{\gamma^3}\min_{\substack{q=3,\ldots,p \\ q'=2,\ldots,p}} \min\left\{\left(\frac{\tilde{\ell}_1\sigma_q}{\tilde{\ell}_{q-1}\sigma_2}\right)^{\frac{2}{q-2}}, \left(\frac{L_{q'}}{5\tilde{\ell}_{q'}\gamma}\right)^{\frac{2}{q'-1}}\right\},
\end{aligned} \tag{96}$$

where $(\star)$ uses that $\lfloor\xi\rfloor - 2 \geq \xi/5$ whenever $\xi \geq 3$, implying the desired result (note that this bound does not depend on L_1 and σ_1.).

If $\sigma_1 = \cdots = \sigma_p = 0$, we obtain the following lower complexity bound for noiseless oracles (where ρ is effectively set to one), assuming $\gamma = O(L_1)$ (this holds without loss of generality, as we discuss above). As before, we set $\alpha = 5\gamma/\beta^2$. The L_1-constraint is satisfied under the same condition stated in (95). Thus, letting

$$\beta = \min_{q=2,\ldots,p}\left\{\left(\frac{L_q}{5\tilde{\ell}_q\gamma}\right)^{\frac{1}{q-1}}\right\},$$

it follows that our construction is L_q-Lipschitz for any $q = 1,\ldots,p$. Following the same chain of inequalities as in (96) yields an oracle complexity lower bound of

$$\frac{\Delta\beta^2}{5^3\tilde{\Delta}_0\gamma} = \frac{\Delta}{5^3\tilde{\Delta}_0\gamma}\min_{q=2,\ldots,p}\left\{\left(\frac{L_q}{5\tilde{\ell}_q\gamma}\right)^{\frac{2}{q-1}}\right\}.$$

Note that this bound does not depend on L_1.

■