Escaping Saddle-Point Faster under Interpolation-like Conditions

Abhiskek Roy
Department of Statistics
University of California, Davis
abroy@ucdavis.edu

Krishnakumar Balasubramanian
Department of Statistics
University of California, Davis
kbala@ucdavis.edu

Saeed Ghadimi
Department of Management Sciences
University of Waterloo
sghadimi@uwaterloo.ca

Prasant Mohapatra
Department of Computer Science
University of California, Davis
pmohapatra@ucdavis.edu

Abstract

In this paper, we show that under over-parametrization several standard stochastic optimization algorithms escape saddle-points and converge to local-minimizers much faster. One of the fundamental aspects of over-parametrized models is that they are capable of interpolating the training data. We show that, under interpolation-like assumptions satisfied by the stochastic gradients in an over-parametrization setting, the first-order oracle complexity of Perturbed Stochastic Gradient Descent (PSGD) algorithm to reach an ϵ-local-minimizer, matches the corresponding deterministic rate of $\tilde{\mathcal{O}}(1/\epsilon^2)$. We next analyze Stochastic Cubic-Regularized Newton (SCRN) algorithm under interpolation-like conditions, and show that the oracle complexity to reach an ϵ-local-minimizer under interpolation-like conditions, is $\tilde{\mathcal{O}}(1/\epsilon^{2.5})$. While this obtained complexity is better than the corresponding complexity of either PSGD, or SCRN without interpolation-like assumptions, it does not match the rate of $\tilde{\mathcal{O}}(1/\epsilon^{1.5})$ corresponding to deterministic Cubic-Regularized Newton method. It seems further Hessian-based interpolation-like assumptions are necessary to bridge this gap. We also discuss the corresponding improved complexities in the zeroth-order settings.

1 Introduction

Over-parametrized models, for which the training stage involves solving nonconvex optimization problems, are common in modern machine learning. A canonical example of such a model is deep neural networks. Such over-parametrized models have several interesting statistical and computational properties. On the statistical side, such over-parametrized models are highly expressive and are capable of nearly perfectly interpolating the training data. Furthermore, despite the highly nonconvex training landscape, most local minimizers have good generalization properties under regularity conditions; see for example [39, 27, 22, 21] for empirical and theoretical details. We emphasize here that over-parametrization plays an important role for both phenomenon to occur. Furthermore, it is to be noted that not all critical points exhibit nice generalization properties. Hence, from a computational perspective, designing algorithms that do not get trapped in saddle-points, and converge to local minimizers during the training process, becomes extremely important [12].

Indeed, recently there has been extensive research in the machine learning and optimization communities on designing algorithms that escape saddle-points and converge to local minimizers. The

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

authors of [30] proved the folklore result that in the deterministic setting for sufficiently regular functions, vanilla gradient descent algorithms converges almost surely to local minimizers, even when initialized randomly; see also [29]. However, [30, 29] only provide asymptotic results, that have limited consequence for practice. Understandably, it has been shown by the authors of [15], that gradient descent might take exponential-time to escape saddle points in several cases. In this context, injecting artificial noise in each step of the gradient descent algorithm has been empirically observed to help escape saddle points. Several works, for example, [24, 26], showed that such *perturbed* gradient descent algorithms escape saddles faster in a non-asymptotic sense. Such algorithms are routinely used in training highly over-parametrized deep neural network and other over-parameterized nonconvex machine learning models. However, existing theoretical analysis of such algorithms fail to take advantage of the interpolation-like properties enjoyed by over-parametrized machine learning models. Hence, such theoretical results are conservative. Specifically, there is a gap between the assumptions used in the theoretical analysis of algorithms that escape saddle-points and the assumptions commonly satisfied by over-parametrized models which are trained by those algorithms.

In this work, we consider nonconvex stochastic optimization problems of the following form:

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} \{f(x) := \mathbf{E}_\xi[F(x,\xi)]\}\,. \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is nonconvex function satisfying certain regularity properties described next, and ξ is a random variable characterizing the stochasticity in the problem. We assume that the function f has a lower bound f^* throughout this work. We analyze two standard algorithms that escape saddle-points, namely the perturbed stochastic gradient descent (PSGD) and stochastic cubic-regularized Newton's method (SCRN) for problems of the form in (1). We show that under interpolation-like assumptions (see Section 2 for exact definitions) on the stochastic gradient, it could be proved that both PSGD and SCRN escape saddle-points and converge to local minimizers much faster. In particular, we show that in order for PSGD algorithm to escape saddle-points and find an ϵ-local-minimizer, the number of calls to the stochastic first-order oracle is of the order $\tilde{\mathcal{O}}(1/\epsilon^2)$[1] which matches number of calls when the objective being optimized is a deterministic objective (for which exact gradient could be obtained in each step of the algorithm)[2]. As a point of comparison, [19, 25] showed that without the interpolation-like conditions that we make, PSGD requires $\tilde{\mathcal{O}}(1/\epsilon^4)$ calls to the stochastic gradient oracle. Furthermore, [17] analyzed a version of PSGD with averaging and improved the oracle complexity to $\tilde{\mathcal{O}}(1/\epsilon^{3.5})$. It is also worth noting that, with a mean-square Lipschitz gradient assumption on the objective function being optimized, and using complicated variance reduction techniques, the authors of [16] showed that it is possible for a double-loop version of PSGD to converge to ϵ-local minimizers with $\tilde{\mathcal{O}}(1/\epsilon^3)$ number of calls to the stochastic first-order oracle. However, recent empirical investigations seem to suggest that variance reduction techniques are inefficient for the nonconvex deep learning problems [13, 43]. Our results, on the other hand exploit the naturally available structure present in over-parametrized models and obtains the best-known oracle complexity for escaping saddle-points using only the vanilla versions of PSGD algorithm (which is oftentimes the version of PSGD used in practice). We also analyze the corresponding Zeroth-Order version of the PSGD algorithm. In this setting, we are able to observe only potentially noisy evaluations of the function being optimized. In this setting, we show that PSGD algorithm requires $\tilde{\mathcal{O}}(d^{1.5}/\epsilon^{4.5})$ calls to the stochastic zeroth-order oracle. In this context, we are not aware of a result to compare with. The recent works of [4, 18] provided results for bounded functions in the zeroth-order deterministic setting, where one obtains exact function values; such a setting though is highly unrealistic in practice.

Next, we consider the question of whether using second-order methods helps reduce the number of calls. Indeed, in the deterministic setting, it is well-known that second-order information helps escape saddle point at a much faster rate. For example, [37] proposed that Cubic-regularized Newton's method and showed that the method requires only $\tilde{\mathcal{O}}(1/\epsilon^{1.5})$ calls to the gradient and Hessian oracle; see also [8, 11] for related results. Correspondingly, in the stochastic setting [47] showed that SCRN method requires $\tilde{\mathcal{O}}(1/\epsilon^{3.5})$ calls, which is better than that of PSGD (without further assumptions). In this work, we show that under interpolation-like assumptions on (only) the stochastic gradient, SCRN method requires only $\tilde{\mathcal{O}}(1/\epsilon^{2.5})$ calls. In contrast to the PSGD setting, SCRN requires more calls than its corresponding deterministic counterpart. However, it should be noted that the complexity of SCRN

[1] Here, $\tilde{\mathcal{O}}$ hides $\log$ factors.

[2] It is possible to obtain $\tilde{\mathcal{O}}(1/\epsilon^{11.75})$ complexity using accelerated method in deterministic setting; see [26].

Algorithm	**With SGC (This paper)**		Without SGC		Deterministic
	ZO	HO	ZO	HO	HO
Perturbed GD	$\tilde{\boldsymbol{\mathcal{O}}}\left(\boldsymbol{d^{1.5}\epsilon^{-4.5}}\right)$ Theorem 3.1	$\tilde{\boldsymbol{\mathcal{O}}}\left(\boldsymbol{\epsilon^{-2}}\right)$ Theorem 3.1	$\tilde{\boldsymbol{\mathcal{O}}}\left(\boldsymbol{d^{1.5}\epsilon^{-5.5}}\right)$ Theorem 3.2	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ Theorem 17 [25]	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$ Theorem 3 [24]
Cubic Newton	$\tilde{\boldsymbol{\mathcal{O}}}\left(\boldsymbol{d^{4}\epsilon^{-2.5}}\right)$ Theorem 4.1	$\tilde{\boldsymbol{\mathcal{O}}}\left(\boldsymbol{\epsilon^{-2.5}}\right)$ Theorem 4.1	$\tilde{\mathcal{O}}\left(d^4\epsilon^{-2.5}\right) + \mathcal{O}\left(d\epsilon^{-3.5}\right)$ Theorem 4.1 [5]	$\tilde{\mathcal{O}}\left(\epsilon^{-3.5}\right)$ Theorem 1 [47]	$\mathcal{O}\left(\epsilon^{-1.5}\right)$ Theorem 3 [37]

Table 1: Oracle complexities of perturbed stochastic gradient descent (PSGD) and stochastic cubic-regularized Newton's method (SCRN). ZO corresponds to number of calls to zeroth-order oracle and HO corresponds to number of calls to first or second-order oracles. The result for PSGD and SCRN are given respectively in high-probability and in expectation. The results in the deterministic case corresponds to projected gradient descent and cubic-Regularized Newton's method (without stochastic gradients).

is still better than that of the PSGD, with or without interpolation-like assumptions. We belive that without further interpolation-like assumptions also on the stochastic Hessians, the oracle complexity of SCRN cannot be improved, in particular to match the deterministic rate of $\tilde{\mathcal{O}}(1/\epsilon^{1.5})$ (see also Remark 6). We also provide similar improved results for a zeroth-order version of SCRN method, thereby improving upon the results of [5]. All of our results, along with comparison to existing results in the literature and the corresponding assumption required, are summarized in Table 1. We conclude this section with a other related works.

More Related Works. In the interpolation regime, [33] recently showed that mini-batch stochastic gradient descent (SGD) algorithm enjoys exponential rates of convergence for unconstrained strongly-convex optimization problems; see also [45, 36] for related earlier work. For the non-convex setting, [6] analyze SGD for non-convex functions satisfying the Polyak-Lojasiewicz (PL) inequality [41] under the interpolation condition and show that SGD can achieve a linear convergence rate. Recently, [48] introduced a more practical form of interpolation condition, and prove that the constant step-size SGD can obtain the optimal convergence rate for strongly-convex and smooth convex functions. They also show the first results in the non-convex setting that the constant step-size SGD can obtain the deterministic rate in the interpolation regime for converging to first-order stationary solution. Subsequently, [34] investigate the regularized subsampled Newton method (R-SSN) and the stochastic BFGS algorithm under the interpolation-like conditions. We emphasize that all the above works consider the case of convex objective function predominantly; the only exception is [48] that consider the nonconvex case but only study convergence to first-order stationary solution. There has been several works on obtaining oracle complexity of escaping saddle-points in the finite-sum setting; we refer the interested reader to [1, 51, 50, 49] and references therein for such results. We emphasize that a majority of the above works are based on complicated variance reduction techniques that increase the implementation complexity of such methods and make them less appealing in practice. There exist only few works on escaping saddle-points for constrained optimization problems; see [32, 31, 40, 35] for more details.

We also briefly discuss the consequences of our results to deep neural network training and related works. Roughly speaking, there are now two potential explanations for the success of optimization methods for training deep neural networks [46]. The first explanation is based on landscape analysis. This involves two steps: Showing the optimization landscape has favorable geometry [28] (i.e., all local minima are (approximate) global minima under suitable regularity conditions), and hence constructing optimization algorithms that can efficiently escape saddle-points. The second explanation is based on the NTK viewpoint; see, for example [23, 9, 10, 3, 14, 52], for a partial overview. However, a majority of the results based on NTK viewpoint are for polynomially (in depth and sample-size) large-width networks (indeed, [3] mention that their polynomial degrees are impractical). Our results in this paper are geared towards the former program.

2 Preliminaries

We now present the assumptions and definitions used throughout the paper. Section-specific additional details are in the respective sections. In this paper we use $\|\cdot\|$, and $\|\cdot\|_*$ to denote a norm and the corresponding dual norm on $\mathbb{R}^d$. We now describe some regularity conditions made on the objective function in (1) assumptions in this work.

Assumption 2.1 (Lipschitz Function) *The function F is L-Lipschitz, almost surely for any ξ, i.e., $|F(x,\xi) - F(y,\xi)| \leq L\|x-y\|$. Here we assume $\|\cdot\| = \|\cdot\|_2$, unless specified explicitly.*

Assumption 2.2 (Lipschitz Gradient) *The function F has Lipschitz continuous gradient, almost surely for any ξ, i.e., $\|\nabla F(x,\xi) - \nabla F(y,\xi)\| \leq L_G\|x-y\|_*$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. This also implies $|F(y,\xi) - F(x,\xi) - \nabla F(x,\xi)^\top (y-x)| \leq \frac{L_G}{2}\|y-x\|^2$.*

Assumption 2.3 (Lipschitz Hessian) *The function F has Lipschitz continuous Hessian, almost surely for any ξ, i.e., $\left\|\nabla^2 F(x,\xi) - \nabla^2 F(y,\xi)\right\| \leq L_H\|x-y\|$.*

Note that if Assumptions 2.1–2.3 are true for F, then they also hold for $f(\cdot) = \mathbf{E}\left[F(\cdot,\xi)\right]$; but the other way around is not true. For our higher-order results, we make the above assumptions only on $f(\cdot)$, which is a weaker assumption. In the interpolation regime, the stochastic gradients become small when the true gradient is small. The following condition, known as Strong Growth Condition (SGC) [48], captures how fast the stochastic gradient goes to 0 with respect to the true gradient.

Assumption 2.4 (SGC [48]) *For any point $x \in \mathbb{R}^d$, the stochastic gradient satisfies $\mathbf{E}_\xi\|\nabla F(x,\xi)\|^2 \leq \rho\|\nabla f(x)\|^2$, for $\rho > 1$ (as $\rho = 1$, corresponds to the deterministic setting).*

SGC controls the variance of the obtained stochastic gradient in the above mentioned way. Note in particular that in the case when $\|\nabla f(x)\|^2 = 0$, under SGC, we have almost surely $\|\nabla F(x,\xi)\|^2 = 0$. This means that when the point x is a stationary point of the function f, then it is also a stationary point of the function F almost surely. In the context of deep neural networks, the function F corresponds to the risk based on training sample ξ and the function f corresponds to the risk. Hence, the strong growth condition states that that deep neural network is capable of interpolating the training data almost surely. Such a phenomenon is observed in practice with deep neural networks, which provides a strong motivation for using this assumption for analyzing the performance of PSGD and SCRN for escaping saddle-points.

In this work, we study the algorithms under two oracles settings: Stochastic zeroth-order oracle, where one obtains noisy unbiased function evaluations, and the stochastic higher-order oracle, where one obtains noisy unbiased estimators of the gradients, and hessians. We now define them formally.

Assumption 2.5 (Zeroth-order oracle) *For any $x \in \mathbb{R}^d$, the zeroth order oracle outputs an estimator $F(x,\xi)$ of $f(x)$ such that $\mathbf{E}\left[F(x,\xi)\right] = f(x)$, $\mathbf{E}\left[\nabla F(x,\xi)\right] = \nabla f(x)$, $\mathbf{E}\left[\nabla^2 F(x,\xi)\right] = \nabla^2 f(x)$, and $\mathbf{E}\left[\|\nabla^2 F(x,\xi) - \nabla^2 f(x)\|_F^4\right] \leq \sigma_2^4$, where $\|\cdot\|_F$ is the Frobenius norm.*

Assumption 2.6 (Higher-order oracles) *For any $x \in \mathbb{R}^d$, (i) the first-order oracle outputs an estimate $\nabla F(x,\xi)$ of $\nabla f(x)$ such that $\mathbf{E}\left[\nabla F(x,\xi)\right] = \nabla f(x)$ and (ii) the second-order oracle, in addition outputs an estimate $\nabla^2 F(x,\xi)$ of $\nabla^2 f(x)$ such that, $\mathbf{E}\left[\nabla^2 F(x,\xi)\right] = \nabla^2 f(x)$, and $\mathbf{E}\left[\|\nabla^2 F(x,\xi) - \nabla^2 f(x)\|_F^4\right] \leq \sigma_2^4$.*

Such assumptions on the zeroth-order and higher-order oracles are standard in the literature; see for example [20, 38**?**]. Our goal in this paper is to reach an approximate local minimizer (also called as a second-order stationary point) of a non-convex function, which is defined as follows:

Definition 2.1 (ϵ-Local Minimizer) *Let Assumption 2.3 hold for a function f. Then a point $\bar{x}$ is called a ϵ-second-order stationary point if,*

$$\max\left(\sqrt{\|\nabla f(\bar{x})\|}, -\frac{\lambda_{min}\left(\nabla^2 f(\bar{x})\right)}{L_H}\right) \leq \sqrt{\epsilon} \tag{2}$$

where $\lambda_{min}\left(\nabla^2 f(\bar{x})\right)$ is the minimum eigenvalue of $\nabla^2 f(\bar{x})$.

Algorithm 1 Perturbed Stochastic Gradient Descent Algorithm

Input: $x_0 \in \mathbb{R}^d$, η, r.
for $t = 0$ to T **do**
Set $g_t = \frac{1}{n_1}\sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(x_t, \xi_{t,i}\right) \qquad \text{(First-order)}$$

$$g_{t,i} = \frac{F(x_t + \nu u_{t,i}, \xi_{t,i}) - F(x_t, \xi_{t,i})}{\nu} u_i \qquad \text{(Zeroth-order)}$$

and $u_{t,i} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right) \forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1$
Sample $\theta_t \in \mathcal{N}\left(\mathbf{0}, r^2 \boldsymbol{I}_d\right)$
Update $x_{t+1} = x_t - \eta\left(g_t + \theta_t\right)$
end for

Note that for stochastic optimization problems, the quantity on the left hand side of (2), is a random variable. In this paper we prove a high-probability bound, and an expectation bound for the above quantity for PSGD, and SCRN respectively.

For a point x_t, we will use ∇_t, ∇_t^2, h_t, and $\lambda_{1,t}$ to denote ∇_t, $\nabla^2 f(x_t)$, $(x_{t+1} - x_t)$,and $\lambda_{min}\left(\nabla^2 f\left(x_t\right)\right)$ respectively. The zeroth-order minibatch gradient [38], and Hessian estimator [**?**] g_t, and H_t are defined as:

$$g_t = \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{F(x_t + \nu u_{t,i}, \xi_{t,i}) - F(x_t, \xi_{t,i})}{\nu} u_i, \qquad H_t = \frac{1}{n_2}\sum_{i=1}^{n_2} \mathfrak{H}_{t,i}\left(u_{t,i}u_{t,i}^\top - I\right), \quad (3)$$

where $\mathfrak{H}_{t,i} = \frac{F(x_t+\nu u_{t,i},\xi_{t,i}) + F(x_t - \nu u_{t,i}, \xi_{t,i}) - 2F(x_t,\xi_{t,i})}{2\nu^2}$, and $u_{t,i} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}_d\right)$ $\forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1$. We will use $\zeta_t = g_t - \nabla_t = \frac{1}{n_1}\sum_{i=1}^{n_1} g_{t,i} - \nabla_t$, and $\tilde{\zeta}_t = \zeta_t + \theta_t$. In the following lemma we show that under SGC, the variance of $\nabla F(x_t, \xi)$ is of the order of the gradient norm squared.

Lemma 2.1 *Let Assumption 2.4 hold for a function f. Then, for both zeroth-order, and first-order oracle, we have,*

$$\mathbf{E}\left[\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\nabla F(x_t, \xi_i) - \nabla_t\right\|^2\right] \leq \frac{\rho - 1}{n_1}\|\nabla_t\|^2. \quad (4)$$

Proof Let $g_t = \frac{1}{n_1}\sum_{i=1}^{n_1}\nabla F(x_t, \xi_i)$. Then we have

$$\begin{aligned}\mathbf{E}\left[\|g_t - \nabla_t\|^2\right] =& \mathbf{E}\left[\|g_t\|^2 + \|\nabla_t\|^2 - 2g_t^\top \nabla_t\right] = \frac{1}{n_1^2}\mathbf{E}\left[\left\|\sum_{i=1}^{n_1}\nabla F(x_t, \xi_i)\right\|^2\right] - \|\nabla_t\|^2 \\ \leq& \frac{1}{n_1^2}\left(\rho n_1 \|\nabla_t\|^2 + n_1(n_1 - 1)\|\nabla_t\|^2\right) - \|\nabla_t\|^2 = \frac{\rho - 1}{n_1}\|\nabla_t\|^2,\end{aligned}$$

which completes the proof. ∎

Remark 1 *The above simple results actually turns out to have far-reaching consequences for obtaining improved complexity bounds for both PSGD and SCRN algorithms. It implies that when the true gradient is small, the variance of the stochastic gradient is also small. Typically, in the analysis of PSGD and SCRN, it is assumed that the stochastic gradients are assumed to have a constant variance. But for over-parametrized models, we will use Lemma 2.1 to prove deterministic rate for PSGD and improved rates for SCRN.*

3 Perturbed Stochastic Gradient Descent

In this section we show that under SGC, PSGD attain deterministic rate in the first-order setting and obtains much better rate than previously known rates in the zeroth-order setting. An intuitive

explanation of this phenomenon is as follows: in the general stochastic setting, at time t where $\|\nabla_t\| \geq \epsilon$, PSGD does not descend as much as in the deterministic setting due to noisy gradient. So it takes more iterations to average out the noise. While escaping a saddle point, due to noisy gradient, the iterates follow the direction of the most negative curvature with more difficulty leading to higher complexity. Under SGC, when $\|\nabla_t\| \geq \epsilon$, the noise variance is of the order of $\|\nabla_t\|^2$ as shown in Lemma 2.1. So the algorithm still manages to descent. While escaping a saddle point under SGC, as $\|\nabla_t\| \leq \epsilon$, and the gradient noise is also small leading to deterministic rates.

The outline of the proof of the bounds for PSGD in the first-order setting is similar to [25] except that we analyze PSGD under interpolation regime. At a high level the proof has two stages: firstly, we show that when $\|\nabla_t\| \geq \epsilon$, the function descends as fast as the deterministic case; Secondly, when $\|\nabla f(x_t)\| \leq \epsilon$, and $\lambda_{min}(\nabla^2 f(x_t)) \leq -\sqrt{L_H \epsilon}$, i.e., x_t is a saddle point, by a coupling argument it is shown that either the function descends or the sequence of iterates are stuck around the saddle point. But then it is shown that the stuck region is narrow enough so that the iterates escape the saddle points with high probability. We now require a condition on the tail of the stochastic gradient.

Assumption 3.1 *For any $x \in \mathbb{R}^d$, $\mathbb{P}\left(\|\nabla F(x,\xi) - \nabla f(x)\| \geq \tau\right) \leq 2e^{-\frac{\tau^2}{2\mathbf{E}\left[\|\nabla F(x,\xi)-\nabla f(x)\|^2\right]}}$.*

Such light-tail conditions are common in the stochastic optimization literature to obtain high-probability bounds; see for example [20, 25]. Note that under Assumption 2.4, Assumption 3.1 is equivalent to

$$\mathbb{P}\left(\|\nabla F(x,\xi) - \nabla f(x)\| \geq \tau\right) \leq 2e^{-\tau^2/(2(\rho-1)\|\nabla f(x)\|^2)} \tag{5}$$

We now present our main result on PSGD.

Theorem 3.1 *a) Under Assumptions 2.2, 2.3 on the function $f(\cdot)$, and Assumptions 2.4, and 3.1, choosing,*

$$\eta = \log\left(\frac{1}{\epsilon}\right)^{-2} \Big/ a_0 \log\left(\frac{f(x_0)-f^*}{\delta\epsilon}\right),\ r = \epsilon^{1.5}\log(\epsilon^{-1})^{-3}, n_1 = 512c(\rho-1)\log(\epsilon^{-1}), \tag{6}$$

with probability at least $1-\delta$, half of the iterations of Algorithm 1 will be ϵ-local minimizers after T iterations where,

$$T = a_1 \max\left\{\frac{(f(x_0)-f^*)\mathcal{T}_1}{\mathcal{F}_1}, \frac{(f(x_0)-f^*)}{\eta\epsilon^2}\right\} = \tilde{\mathcal{O}}\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right), \tag{7}$$

where $a_0,\ a_1$ are constants, and $\mathcal{T}_1 = 0.5\log\left(\frac{1}{\epsilon}\right)^3/\sqrt{\epsilon}$, and $\mathcal{F}_1 = \epsilon^{1.5}/\log\left(\frac{1}{\epsilon}\right)^7$.

b) Under Assumptions 2.1, 2.2, 2.3, 2.4, and 3.1 in the zeroth order-setting, choosing,

$$\eta = \frac{\kappa_0}{\log\left(\frac{f(x_0)-f^*}{\delta\epsilon}\right)} \quad r = \kappa_1\epsilon \quad \nu = \frac{\kappa_4\epsilon}{d\log\left(\frac{1}{\epsilon}\right)} \quad n_1 = \frac{\kappa_5\log\left(\frac{1}{\epsilon}\right)^5 d^{1.5}\sqrt{\rho-1}}{\epsilon^{2.5}} \tag{8}$$

with probability at least $1-\delta$, half of the iterations of Algorithm 1 will be ϵ-local minimzers, after T iterations, where,

$$T = \kappa_9 \max\left\{\frac{(f(x_0)-f^*)\mathcal{T}_0}{\mathcal{F}_0}, \frac{(f(x_0)-f^*)}{\eta\epsilon^2}\right\} = \tilde{\mathcal{O}}\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right). \tag{9}$$

Here, $\kappa_i, i = 1, 2, \cdots, 9$ are absolute constants, and $\mathcal{T}_0 = \kappa_3\frac{\log\left(\frac{1}{\epsilon}\right)^2\log(d)^2}{\sqrt{\epsilon}}$, and $\mathcal{F}_0 = \kappa_8\epsilon^{1.5}$. Hence, the total number of zeroth-order oracle calls is $Tn_1 = \tilde{\mathcal{O}}\left(\frac{d^{1.5}\sqrt{\rho-1}}{\epsilon^{4.5}}\right)$.

Remark 2 *Note that the complexity result in* (7) *for the PSGD in the first-order setting matches corresponding complexity of perturbed gradient descent on deterministic optimization problems.*

Remark 3 *We briefly highlight on the difficulty associated with proving the result in* (9)*. First note that in the first-order proof, and also in [25], it is assumed that the noise ξ is sub-gaussian. But for the zeroth-order gradient g_t as defined in* (3)*, $\|g_t - \nabla_t\|$ no longer has sub-Gaussian tails. Also note*

that we have from [38], $\mathbf{E}_{u_{t,i}}[g_{t,i}] = \nabla F_\nu(x_t, \xi_{t,i}) = \nabla \mathbf{E}_{u_{t,i}}[F(x + \nu u_{t,i}, \xi_{t,i})]$*. So* g_t *is not an unbiased estimator of* ∇_t*. But as shown in [38],* $\nabla F_\nu(x_t, \xi_{t,i})$ *is close to* $\nabla F(x_t, \xi_{t,i})$*. So we first need to establish concentration properties for* g_t *in the zeroth-order setting. Towards this, we show that* g_t *is* α*-sub-exponential with* $\alpha = 2/3$*, even if* ξ *is sub-gaussian, i.e., the noise in the gradient estimates has heavier tail (Lemma A.1). This leads to the obtained complexity bounds in* (9)*.*

Remark 4 *Note that* $\mathcal{T}_1$ *and* $\mathcal{T}_0$ *are the number of iterations required to descend by* $\mathcal{F}_1$ *and* $\mathcal{F}_0$ *respectively in the first and zeroth-order setting, after the algorithm hits a saddle point. As shown in [25], without SGC,* $\mathcal{T}_1 = \tilde{O}(\epsilon^{-2.5})$*. In this paper we show that, under SGC,* $\mathcal{T}_1 = \mathcal{T}_0 = \tilde{O}(\epsilon^{-0.5})$*. This shows under SGC, it is indeed possible to escape saddle point faster.*

We highlight here that [4, 18] recently considered escaping saddle points in the zeroth-order setting. However they assume that the function being optimized is deterministic (which means exact gradients could be obtained) and is bounded (which means sub-Gaussian tails are possible for the zeroth-order gradient estimator). These two assumptions are however highly impractical and are not satisfied by several situations in practice where zeroth-order optimization techniques are utilized. To the best of our knowledge, there is no known bound on the number of times zeroth-order oracle should accessed for (9) to hold, when SGC does not hold and only the following standard variance assumption on the unseen stochastic gradient holds (see, e.g., [20]) for some $\sigma > 0$,

$$\mathbf{E}\left[\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}\nabla F(x_t, \xi_i) - \nabla_t\right\|^2\right] \leq \frac{\sigma^2}{n_1}. \tag{10}$$

For completeness we present the corresponding result below, which serves as a reference to compare our results with the SGC assumption to what one could obtain without it.

Theorem 3.2 *Under Assumptions 2.1, 2.2, 2.3, 2.4, and 3.1, we have the following: In the zeroth order-setting, choosing,*

$$\eta = \frac{\kappa_0}{\log\left(\frac{f(x_0)-f^*}{\delta\epsilon}\right)} \quad r = \kappa_1\epsilon \quad \nu = \frac{\kappa_4\epsilon}{d\log\left(\frac{1}{\epsilon}\right)} \quad n_1 = \frac{\kappa_5 \log\left(\frac{1}{\epsilon}\right)^5 d^{1.5}\sigma}{\epsilon^{3.5}} \tag{11}$$

with probability at least $1-\delta$*, half of the iterations of Algorithm 1 will be* ϵ*-local minimzers, after* T *iterations, where,*

$$T = \kappa_9 \max\left\{\frac{(f(x_0)-f^*)\mathcal{T}_0}{\mathcal{F}_0}, \frac{(f(x_0)-f^*)}{\eta\epsilon^2}\right\} = \tilde{\mathcal{O}}\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right). \tag{12}$$

Here, $\kappa_i, i = 1, 2, \cdots, 9$ *are absolute constants, and* $\mathcal{T}_0 = \kappa_3 \frac{\log\left(\frac{1}{\epsilon}\right)^2 \log(d)^2}{\sqrt{\epsilon}}$ *and* $\mathcal{F}_0 = \kappa_8\epsilon^{1.5}$*. Hence, the total number of zeroth-order oracle calls is* $Tn_1 = \tilde{\mathcal{O}}\left(\frac{d^{1.5}\sigma}{\epsilon^{5.5}}\right)$*.*

Remark 5 *A generic reduction was proposed in [2] for using any algorithm that converges to a first-order stationary points at a particular rate, to converge to a local minimizer at the same rate. The results in [2] are not directly applicable to the zeroth-order setting due to their assumptions. However, assuming that their assumption could be relaxed to get it work in the zeroth-order setting, it is interesting to examine if the results in [20] for converging to first-order stationary solution could be combined with the reduction proposed in [2] to establish a result similar to Theorem 3.2. To make the result of [20] hold with the same probability as in Theorem 3.2, we would require* $O(d\,\epsilon^{-6})$ *calls to the stochastic zeroth-order oracle. Hence, in certain regimes it is plausible we obtain improved results. It is interesting future work to examine this further rigorously.*

4 Stochastic Cubic-Regularized Newton's Method

In this section we analyze Cubic-Regularized (CR) Newton method under interpolation regime. In non-interpolation like stochastic setting, CR Newton achieves a rate of $\mathcal{O}(\epsilon^{-3.5})$ as compared to $\mathcal{O}(\epsilon^{-4})$ attained by PSGD. Here we show that CR Newton achieves a rate of $\mathcal{O}(\epsilon^{-2.5})$ under

Algorithm 2 Cubic-Regularized Newton Algorithm

Input: $x_1 \in \mathbb{R}^d$, T, M, n_1, n_2
for $t = 1$ to T **do**
Set $g_t = \frac{1}{n_1}\sum_{i=1}^{n_1} g_{t,i}$ **where**

$$g_{t,i} = \nabla F\left(x_t, \xi_{t,i}^G\right) \quad \text{(Higher-order)}$$

$$g_{t,i} = \frac{F(x_t + \nu u_{t,i}^G, \xi_{t,i}^G) - F(x_t, \xi_{t,i}^G)}{\nu} u_i^G \quad \text{(Zeroth-order)}$$

Set $H_t = \frac{1}{n_2}\sum_{i=1}^{n_2} H_{t,i}$ **where**

$$H_{t,i} = \nabla^2 F\left(x_t, \xi_{t,i}^H\right) \quad \text{(Higher-order)}$$

$$H_{t,i} = \frac{F(x_t + \nu u_{t,i}^H, \xi_{t,i}^H) + F(x_t - \nu u_{t,i}^H, \xi_{t,i}^H) - 2F(x_t, \xi_{t,i}^H)}{2\nu^2}\left(u_{t,i}^H {u_{t,i}^H}^\top - I\right) \quad \text{(Zeroth-order)}$$

where $u_{t,i}^{G[H]} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)\ \forall t = 1, 2, \cdots, T, i = 1, 2, \cdots, n_1[n_2]$
Update

$$x_{t+1} = \operatorname*{argmin}_y m_t\left(x_t, y, g_t, H_t, M\right), \tag{13}$$

where

$$m_t(y) = f(x_t) + (y - x_t)^\top g_t + \frac{1}{2}(y - x_t)^\top H_t (y - x_t) + \frac{M}{6}\|y - x_t\|^3 \tag{14}$$

end for

SGC. Even though this rate is better than non-interpolation like stochastic setting, quite interestingly, CR Newton method fails to achieve deterministic rate of $\mathcal{O}(\epsilon^{-1.5})$ unlike PSGD. We believe that without stronger assumption on the Hessian estimator noise as well, CR Newton will perform worse than PSGD. In this section let $\mathcal{F}_t$ be the filtration generated until time t, i.e., in the higher-order setting $\mathcal{F}_t = \sigma(\{\xi_{i,j}^G\}_{i,j=1}^{t,n_1}, \{\xi_{i,j}^H\}_{i,j=1}^{t,n_2})$, and in the zeroth-order setting $\mathcal{F}_t = \sigma(\{\xi_{i,j}^G\}_{i,j=1}^{t,n_1}, \{u_{i,j}^G\}_{i,j=1}^{t,n_1}, \{\xi_{i,j}^H\}_{i,j=1}^{t,n_2}, \{u_{i,j}^H\}_{i,j=1}^{t,n_2})$. We now present our main result.

Theorem 4.1 *Let f be a function for which Assumptions 2.2, and 2.3 are true. Then under SGC, i.e., under Assumption 2.4, for Algorithm 2, we have:*

a) In the higher-order setting, choosing

$$T = \frac{144\left(f(x_1) - f^*\right)}{M\epsilon^{\frac{3}{2}}}\, n_1 = \frac{\mu_0(\rho - 1)}{\epsilon}\, n_2 = \epsilon^{-1}, M = \max\left(L_H, \frac{1}{4}, \left(0.004 L_G \epsilon^{\frac{1}{4}} + \sigma_2 \epsilon^{\frac{1}{4}}\right), 40\sigma_2\right) \tag{15}$$

we get, $\max\left(\sqrt{\frac{\mathbf{E}[\|\nabla f(x_R)\|]}{144M}}, -\frac{\mathbf{E}[\lambda_{1,R}]}{9M}\right) \leq \sqrt{\epsilon}$*, where μ_0 is a constant independent of ϵ and d, and R is an integer random variable uniformly distributed over the support $\{1, 2, \cdots, T\}$. The total number of first-order and second-order oracle calls are hence $\mathcal{O}\left(\epsilon^{-\frac{5}{2}}\right)$.*

b) In the zeroth-order setting, choosing

$$T = \frac{\mu_0\left(f(x_1) - f^*\right)}{M\epsilon^{\frac{3}{2}}}, n_1 = \frac{\mu_1(d+5)}{\epsilon}, M = \mu_4, \nu = \frac{\mu_3\epsilon}{(d+16)^{\frac{5}{2}}}, n_2 = \frac{\mu_2(1 + 2\log 2d)(d+16)^4}{\epsilon} \tag{16}$$

we get, $\max\left(\sqrt{\mathbf{E}\left[\|\nabla f\left(x_R\right)\|\right]}, -\mathbf{E}\left[\lambda_{1,R}\right]\right) \leq \mathcal{O}\left(\sqrt{\epsilon}\right)$*, where $\mu_0, \mu_1, \mu_2, \mu_3, \mu_4$ are constants independent of ϵ, and d, and R is an integer random variable uniformly distributed over the support $\{1, 2, \cdots, T\}$.. The total number of first-order oracle calls is $\mathcal{O}\left(d/\epsilon^{\frac{5}{2}}\right)$, and the number of second-order oracle calls is $\mathcal{O}\left(d^4 \log d/\epsilon^{\frac{5}{2}}\right)$.*

Remark 6 *The above results only require Assumption 2.4, which is a gradient-level property of interpolation condition. As SCRN is a second-order algorithm, an assumption like "if the min eigenvalue of true Hessian at a point is non-negative, then min eigenvalue of stochastic Hessian is almost surely also non-negative" might be required to capture second-order properties of interpolation. Such an assumption could then be used to obtain a result similar to Lemma 2.1 for stochastic Hessians, to improve the rates in Theorem 4.1. Formalizing this intuition is an extremely interesting future work.*

Remark 7 *In comparison to the PSGD algorithm, we obtain the results for the SCRN algorithm in expectation. We highlight that it is straightforward to obtain to obtain a high-probability result in the higher-order setting. However, it is technically challenging to do so for the zeroth-order setting. This is due to the difficulty associated with obtaining sharper concentration results for the zeroth-order Hessian estimator, which we leave as future work. In Theorem 4.1, we presented the results in expectation for both settings to maintain uniformity of presentation. In Algorithm 2 we assume that the exact solution to* (13) *is available. We remark that it is possible to relax this assumption following the approach of [47] which in turn leveraged the results in [7] showing that the subproblem in* (13) *can be solved with high probability using gradient descent.*

5 Summary

In this work, we analyze the oracle complexity of two standard algorithms –the perturbed stochastic gradient descent algorithm and the stochastic cubic-regularized Newton's method–for escaping saddle-points in nonconvex stochastic optimization. We show that under interpolation-like conditions satisfied in modern over-parametrized machine learning problems, PSGD and SCRN obtain improved rates for escaping saddle-points. In particular the above stated improvements are obtained for the vanilla versions of PSGD and SCRN algorithms and are not based on any complicated variance reduction techniques. For future work, it is extremely interesting to bridge the gap between SCRN and its deterministic counterpart. The key to this is come up with a Hessian-based interpolation-like assumption, which is both practically meaningful and theoretically sound.

Broader Impact

We focus in this work on establishing theoretical justification for a practically observed phenomenon: Stochastic gradient method and its relatives perform well for training deep neural networks with complicated nonconvex landscape. The result presented will benefit researchers and practitioners who are interested in understanding the theoretical underpinnings of stochastic optimization for deep learning. Although our work in this draft is theoretical, it might have a positive impact for various practical applications of neural networks.

Funding Disclosure

The AR was supported in parts by the NSF Grant CCF-1934568. The research of KB was supported in parts by UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program. The research of PM was sponsored in part by the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pages 3716–3726, 2018.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[4] Qinbo Bai, Mridul Agarwal, and Vaneet Aggarwal. Escaping saddle points for zeroth-order non-convex optimization using estimated gradient descent. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2020.

[5] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points. *arXiv preprint arXiv:1809.06474*, 2018.

[6] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[7] Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2016.

[8] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part II: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011.

[9] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

[10] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[11] Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $\epsilon^{-3/2}$ for nonconvex optimization. *Mathematical Programming*, 162(1-2):1–32, 2017.

[12] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[13] Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pages 1753–1763, 2019.

[14] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

[15] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.

[16] 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.

[17] Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234, 2019.

[18] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *arXiv preprint arXiv:1910.13021*, 2019.

[19] 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.

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

[21] Benjamin Haeffele, Eric Young, and Rene Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International conference on machine learning*, pages 2007–2015, 2014.

[22] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[23] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[24] 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. JMLR. org, 2017.

[25] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019.

[26] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085, 2018.

[27] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

[28] Kenji Kawaguchi and Leslie Kaelbling. Elimination of all bad local minima in deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 853–863, 2020.

[29] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[30] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[31] Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. *arXiv preprint arXiv:1907.04450*, 2019.

[32] Songtao Lu, Ziping Zhao, Kejun Huang, and Mingyi Hong. Perturbed projected gradient descent converges to approximate second-order points for bound constrained nonconvex problems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5356–5360. IEEE, 2019.

[33] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334, 2018.

[34] Si Yi Meng, Sharan Vaswani, Issam Laradji, Mark Schmidt, and Simon Lacoste-Julien. Fast and furious convergence: Stochastic second order methods under interpolation. *arXiv preprint arXiv:1910.04920*, 2020.

[35] Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie. Escaping saddle points in constrained optimization. In *Advances in Neural Information Processing Systems*, pages 3629–3639, 2018.

[36] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

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

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

[39] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[40] Maher Nouiehed and Meisam Razaviyayn. A trust region method for finding second-order stationarity in linearly constrained non-convex optimization. *arXiv preprint arXiv:1904.06784*, 2019.

[41] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[42] Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Multi-point bandit algorithms for nonstationary online nonconvex optimization. *arXiv preprint arXiv:1907.13616*, 2019.

[43] Mark Schmidt. Faster algorithms for deep learning? (presentation in vector institute: https://www.cs.ubc.ca/ schmidtm/documents/2020_vector_smallresidual.pdf), 2020.

[44] Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Non-asymptotic results for langevin monte carlo: Coordinate-wise and black-box sampling. *arXiv preprint arXiv:1902.01373*, 2019.

[45] Thomas Strohmer and Roman Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[46] Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

[47] 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.

[48] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

[49] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. *arXiv preprint arXiv:1802.07372*, 2018.

[50] Junyu Zhang, Lin Xiao, and Shuzhong Zhang. Adaptive stochastic variance reduction for subsampled newton method with cubic regularization. *arXiv preprint arXiv:1811.11637*, 2018.

[51] Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. *arXiv preprint arXiv:1901.11518*, 2019.

[52] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

A Proof of Theorem 3.1.

Preliminaries I: We first present preliminary results regarding the zeroth-order setting.

Lemma A.1 *Let Assumption 2.2, and 3.1 be true for F. Then, in the zeroth-order setting, $\|\zeta_t\|$ is a $2/3$-sub-exponential variable. i.e.,*

$$\mathbb{P}\left(\|\zeta_t\| \geq \tau\right) \leq 4d\exp\left(-K_1 \min\left[\left(\frac{\sqrt{n_1}\tau'}{\Upsilon_t\sqrt{d}}\right)^2, \left(\frac{n_1\tau'}{\Upsilon_t\sqrt{d}}\right)^{2/3}\right]\right), \tag{17}$$

where $\tau' = \tau - \frac{\nu}{2}L_G(d+3)^{\frac{3}{2}}$, and $\Upsilon_t = \frac{\nu L_G(d+2)}{2} + c_0\sqrt{(\rho-1)(d+1)}\|\nabla_t\|$.

We will choose n_1 such that we have $(n_1\tau'/(\Upsilon_t\sqrt{d}))^{2/3} \leq (\sqrt{n_1}\tau'/(\Upsilon_t\sqrt{d}))^2$. So from now on we will only consider the heavier subexponential tail.

Lemma A.2 *Let Assumption 2.2, and 3.1 be true for F. Then, in the zeroth-order setting*

$$\mathbf{E}\left[\exp(s(c_t\|\zeta_t\|)^{\frac{1}{3}})\right] \leq 9d\exp(s^2/b_{1,t}),$$

where $s > 0$, $b_{1,t} = b_{0,t}/c_t^{2/3}$, and $b_{0,t} = K_1 n_1^{2/3}/(\Upsilon_t\sqrt{d})^{2/3}$.

Lemma A.3 *Let Assumption 2.2, and 3.1 be true for F. Then, in the zeroth-order setting for $l > 0$, with probability at least $1 - e^{-l}$ we have*

$$\eta\sum_{i=0}^{t-1}\nabla_i^\top\zeta_i \leq \frac{8\eta\sqrt{dt}}{K_1^{\frac{3}{2}}n_1}(t\log 9d + l)^{\frac{3}{2}}\sum_{i=0}^{t-1}\left(\frac{\nu L_G(d+2)}{2}\|\nabla_i\| + C_0\sqrt{(\rho-1)(d+1)}\|\nabla_i\|^2\right).$$

Lemma A.4 *Let Assumption 2.2, and 3.1 be true for F. Then, for $l > 0$, with probability at least $1 - e^{-l}$ we have*

$$\sum_{i=0}^{t-1}\|\zeta_i\|^2 \leq \frac{128dt^2(t\log 9d + l)^3}{K_1^3 n_1^2}\sum_{i=0}^{t-1}\left(\left(\frac{\nu L_G(d+2)}{2}\right)^2 + C_0^2(\rho-1)(d+1)\|\nabla_i\|^2\right).$$

Preliminaries II: We next present preliminary results regarding the iterates of PSGD. First, we show that the effect of PSGD updates comprises of two parts - the first term on the RHS of (19), and (22) represent the decrease in the function values, and the rest of the terms on the RHS represent possible increase in function value due to noise in the gradient estimator and introduced perturbation.

Lemma A.5 *Under Assumption 2.1, 2.2, 2.3, 2.4 and 3.1, for any fixed $\mathcal{T}_0, \mathcal{T}_1, l > \log 4$, with probability at least $1 - 4e^{-l}$, for Algorithm 1 we get*

a) for the first-order setting, choosing

$$n_1 \geq 512lc(\rho-1) \quad \eta \leq \frac{32lc}{3L_G(l+c)} \tag{18}$$

we have

$$f(x_{\mathcal{T}_1}) - f(x_0) \leq -\frac{\eta}{16}\sum_{i=0}^{\mathcal{T}_1}\|\nabla_i\|^2 + 3c\eta^2r^2(\mathcal{T}_1+l)L_G + 32cl\eta r^2 \tag{19}$$

b) for the zeroth-order case, selecting parameters such that

$$\frac{384L_GC_0^2d(\rho-1)(d+1){\mathcal{T}_0}^2(\mathcal{T}_0\log 9d + l)^3}{K_1^3n_1^2} \leq \frac{1}{16} \tag{20}$$

$$\frac{8C_0\sqrt{(\rho-1)d(d+1)\mathcal{T}_0}}{K_1^{\frac{3}{2}}n_1}(\mathcal{T}_0\log 9d + l)^{\frac{3}{2}} \leq \frac{1}{16} \tag{21}$$

we have

$$f(x_{\mathcal{T}_0}) - f(x_0) \leq -\frac{\eta}{16}\sum_{i=0}^{\mathcal{T}_0-1}\|\nabla_i\|^2 + \wp(r,l,\nu,\eta,d,\mathcal{T}_0) \tag{22}$$

where

$$\wp(r,l,\nu,\eta,d,\mathcal{T}_0) = 16cl\eta r^2 + 3cL_G\eta^2 r^2(\mathcal{T}_0+l)$$
$$+\frac{8\nu\eta LL_G(d+2)\sqrt{d}{\mathcal{T}_0}^{\frac{3}{2}}}{2K_1^{\frac{3}{2}}n_1}(\mathcal{T}_0\log 9d + l)^{\frac{3}{2}} + \frac{96L_G^3\nu^2\eta^2 d(d+2)^2{\mathcal{T}_0}^3(\mathcal{T}_0\log 9d+l)^3}{K_1^3 n_1^2}$$

In the following Lemma we show that when the function descent is small the iterates move only in a small region.

Lemma A.6 *Under conditions of Lemma A.5, Algorithm 1 satisfies*

a) for first-order setting, with probability at least $1-8d\mathcal{T}_1e^{-l}$*, for all* $\tau \leq \mathcal{T}_1$

$$\|x_\tau - x_0\|^2 \leq 32\eta\left(\mathcal{T}_1 + 2cl\frac{\rho-1}{n_1}\right)\left(f(x_0) - f(x_{\mathcal{T}_1}) + 3c\eta^2r^2(\mathcal{T}_1+l)L_G + 32cl\eta r^2\right) + 4cl\mathcal{T}_1\eta^2r^2 \tag{23}$$

b) for zeroth-order setting, with probability at least $1-3d\mathcal{T}_0e^{-l}$*, for all* $\tau \leq \mathcal{T}_0$

$$\|x_\tau - x_0\|^2 \leq \eta\mathcal{T}_0\left(32 + \frac{16}{3L_G}\right)(f(x_0) - f(x_{\mathcal{T}_0}) + \wp(r,l,\nu,\eta,d,\mathcal{T}_0)) + 4cl\mathcal{T}_0\eta^2r^2$$
$$+\frac{L_G\eta^2{\mathcal{T}_0}^2\nu^2(d+2)^2}{48C_0^2(\rho-1)(d+1)} \tag{24}$$

We also require the following definition from [25], to proceed.

Definition A.1 *[25] Let* e_1 *be the eigen-vector corresponding to the minimum eigen-value of* $\mathcal{H} = \nabla^2 f(x_0)$*, and* $\gamma := \lambda_{min}(\nabla^2 f(x_0))$*. Also let* $\mathcal{P}_{-1}$ *be the projection on to the complement subspace of* e_1*. Consider sequences* x_t*, and* x_t' *that are obtained as separate versions of Algorithm 1, both starting from* x_0*. They are coupled in the first-order (zero-order) setting if both sequences are generated by the same* $\mathcal{P}_{-1}\theta_\tau$*, and* ξ_τ $(\{\xi_\tau, \{u_{\tau,i}\}_{i=1}^{n_1}\})$*, while in* e_1 *direction we have* $e_1^\top\theta_\tau = -e_1^\top\theta_\tau'$*.*

We next state some intermediate results in Lemma A.7–A.10, to prove in Lemma A.11 that starting from a saddle-point PSGD should either descend or the iterates will be stuck around the saddle point. Then in Lemma A.12 we will show that the stuck region is narrow enough so that the iterates will escape and consequently the function will have sufficient descent.

Lemma A.7 *[25] Consider the coupling sequences* x_τ *and* x_τ' *as in Definition A.1 and let* $\hat{x}_\tau = x_\tau - x_\tau'$*. Then* $\hat{x}_t = -q_h(t) - q_{sg}(t) - q_p(t)$*, where:*

$$q_h(t) := \eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}\Delta_\tau\hat{x}_\tau, \quad q_{sg}(t) := \eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}\hat{\zeta}_\tau, \quad q_p(t) := \eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}\hat{\theta}_\tau$$

where $\Delta_t := \int_0^1(\nabla^2 f(\phi x_t + (1-\phi)x_t')d\phi - \mathcal{H}$*, and* $\hat{\zeta}_\tau := \zeta_\tau - \zeta_\tau'$*,* $\hat{\theta}_\tau = \theta_\tau - \theta_\tau'$*.*

Lemma A.8 *[25] Denote* $\alpha(t) := \left[\sum_{i=0}^{t-1}(1+\eta\gamma)^{2(t-1-\tau)}\right]^{\frac{1}{2}}$*, and* $\beta(t) = (1+\eta\gamma)^t/\sqrt{2\eta\gamma}$*. If* $\eta\gamma \in [0,1]$*, then (1)*$\alpha(t)\beta(t)$ *for any* $t \in \mathbb{N}$*; and (2)* $\alpha(t) \geq \beta(t)/\sqrt{3}$ *for* $t \geq \ln(2)/(\eta\gamma)$*.*

Lemma A.9 *[25] Under the notation of Lemma A.7, and A.8, we have* $\forall t > 0$*:*

$$\mathbb{P}\left(\|q_p(t)\| \leq \frac{c\beta(t)\eta r}{\sqrt{d}}\sqrt{l}\right) \geq 1 - 2e^{-l}$$
$$\mathbb{P}\left(\|q_p(\mathcal{T}_{1[0]})\| \geq \frac{\beta(\mathcal{T}_{1[0]})\eta r}{10\sqrt{d}}\right) \geq \frac{2}{3}$$

We use $1[0]$ *to denote that the inequality holds for both subscripts* 1 *and* 0*.*

Lemma A.10 *Under the notation of Lemma A.7 and A.8, if*

$$\eta \mathcal{S} \mathcal{T}_{1[0]} \max(L_H, L_G) \leq \frac{1}{l} \quad c \leq \sqrt{l}/40 \tag{25}$$

a) [25]then in the first-order case, we have

$$\begin{aligned} &\mathcal{P}\left(\min\{f(x_{\mathcal{T}_1}) - f(x_0), f(x'_{\mathcal{T}_1}) - f(x_0)\} \leq -\mathcal{F}_1, \text{ or } \forall t \leq \mathcal{T}_1 : \|q_h(t) + q_{sg}(t)\| \leq \frac{\beta(t)\eta t}{20\sqrt{d}}\right) \\ \geq\ & 1 - 10d{\mathcal{T}_1}^2 \log\left(\frac{\mathcal{S}_1\sqrt{d}}{\eta r}\right) e^{-l} \end{aligned}$$

b) in the zeroth-order case, we have

$$\begin{aligned} &\mathcal{P}\left(\min\{f(x_{\mathcal{T}_0}) - f(x_0), f(x'_{\mathcal{T}_0}) - f(x_0)\} \leq -\mathcal{F}_0, \text{ or } \forall t \leq \mathcal{T}_0 : \|q_h(t) + q_{sg}(t)\| \leq \frac{\beta(t)\eta t}{20\sqrt{d}}\right) \\ \geq\ & 1 - 3{\mathcal{T}_0}^2 e^{-l} \end{aligned}$$

Lemma A.11 *a) [25] Under the setting of Lemma A.5, for the first-order setting, we have*

$$\begin{aligned} &\mathbb{P}\left(\min\{f(x_{\mathcal{T}_1}) - f(x_0), f(x'_{\mathcal{T}_1}) - f(x_0)\} \leq -\mathcal{F}_1, \text{ or } \forall t \leq \mathcal{T}_1 : \max\{\|x_t - x_0\|^2, \|x'_t - x_0\|^2\} \leq \mathcal{S}_1^2\right) \\ \geq\ & 1 - 16d\mathcal{T}_1 e^{-l} \end{aligned}$$

b) for the zeroth-order setting, we have

$$\begin{aligned} &\mathbb{P}\left(\min\{f(x_{\mathcal{T}_0}) - f(x_0), f(x'_{\mathcal{T}_0}) - f(x_0)\} \leq -\mathcal{F}_0, \text{ or } \forall t \leq \mathcal{T}_0 : \max\{\|x_t - x_0\|^2, \|x'_t - x_0\|^2\} \leq \mathcal{S}_0^2\right) \\ \geq\ & 1 - 4d\mathcal{T}_0 e^{-l} \end{aligned}$$

In the following Lemma we show that while escaping from a saddle point, the PSGD descends more than it ascends with high probability.

Lemma A.12 *Let Under Assumption 2.1, 2.2, 2.3, 2.4, and 3.1 are true. Under condition* (25)*, for any fixed $t_0 > 0$, let x_0 satisfies*

$$\|\nabla_0\| \leq \epsilon \quad \lambda_{min}(\nabla^2 f(x_0)) \leq -\sqrt{L_H \epsilon}.$$

Then

a) if η, r, n_1 are chosen as in (6)*, $\mathcal{T}_1 = 0.5\log\left(\frac{1}{\epsilon}\right)^3/\sqrt{\epsilon}$, $\mathcal{F}_1 = \epsilon^{1.5}/\log\left(\frac{1}{\epsilon}\right)^7$, $\mathcal{S}_1 = \frac{\sqrt{\epsilon}}{\log\left(\frac{1}{\epsilon}\right)^2}$, $l = a_0 \log\left(\frac{f(x_0) - f^*}{\delta\epsilon}\right)$, then the sequence generated by Algorithm 1 in the first-order case satisfies*

$$\mathbb{P}\left(f(x_{t_0+\mathcal{T}_1}) - f(x_{t_0}) \leq 0.1\mathcal{F}_1\right) \geq 1 - 4e^{-l} \qquad \text{and} \tag{26}$$

$$\mathbb{P}\left(f(x_{t_0+\mathcal{T}_1}) - f(x_{t_0}) \leq -\mathcal{F}_1\right) \geq \frac{1}{3} - 9d{\mathcal{T}_1}^2 \log\left(\frac{\mathcal{S}_1\sqrt{d}}{\eta r}\right) e^{-l} \tag{27}$$

b) if η, r, n_1 are chosen as in (8)*, $\mathcal{T}_0 = \kappa_3 \frac{\log\left(\frac{1}{\epsilon}\right)^2 \log(d)^2}{\sqrt{\epsilon}}$, $\mathcal{F}_0 = \kappa_8 \epsilon^{1.5}$, $\mathcal{S}_0 = \frac{\kappa_7\sqrt{\epsilon}}{\log\left(\frac{1}{\epsilon}\right)^2}$ and $l = \kappa_6 \log\left(\frac{d(f(x_0) - f^*)}{\delta\epsilon}\right)$, then the sequence generated by Algorithm 1 in the zeroth-order case satisfies*

$$\mathbb{P}\left(f(x_{t_0+\mathcal{T}_0}) - f(x_{t_0}) \leq 0.1\mathcal{F}_0\right) \geq 1 - 4e^{-l} \qquad \text{and} \tag{28}$$

$$\mathbb{P}\left(f(x_{t_0+\mathcal{T}_0}) - f(x_{t_0}) \leq -\mathcal{F}_0\right) \geq \frac{1}{3} - \frac{3}{2}{\mathcal{T}_0}^2 e^{-l} \tag{29}$$

Finishing the proof: By combining the above results, we prove Theorem 3.1. The proof is divided in two parts – in the first part we show that the function descends enough when the gradient is large and in the second part we show that the iterates do escape from the saddle points and then function has sufficient descent.

Choice of parameters for Zeroth-order case. As the expressions involved in the analysis of the zeroth order case are little complicated, we show explicitly here how to choose the parameters. First define,

$$\Xi := \frac{32\sqrt{d(\mathcal{T}_0+1)}\eta\beta(\mathcal{T}_0+1)((\mathcal{T}_0+1)\log 9d+\log 2+l)^{\frac{3}{2}}}{K_1^{3/2}n_1} \tag{30}$$

The choice of the parameters should be such that the following equations are satisfied:

$$\begin{aligned}
&\frac{384L_GC_0^2d(\rho-1)(d+1){\mathcal{T}_0}^2(\mathcal{T}_0\log 9d+l)^3}{K_1^3n_1^2} \leq \frac{1}{16},\\
&\frac{8C_0\sqrt{(\rho-1)d(d+1)\mathcal{T}_0}}{K_1^{\frac{3}{2}}n_1}(\mathcal{T}_0\log 9d+l)^{\frac{3}{2}} \leq \frac{1}{16},\\
&\eta\mathcal{S}_0\mathcal{T}_0\max(L_H,L_G) \leq \frac{1}{l}, \quad c \leq \sqrt{l}/40,\\
&\Xi\cdot\sum_{i=0}^{\mathcal{T}_0}\left(\frac{\nu L_G(d+2)}{2}+C_0\sqrt{(\rho-1)(d+1)}L\right) \leq \frac{\beta(\mathcal{T}_0)r}{40\sqrt{d}},\\
&\frac{(1+\eta\gamma)^{\mathcal{T}_0}\sqrt{\eta}r}{40\sqrt{2\gamma d}} > \mathcal{S}_0, \quad \wp(r,l,\nu,\eta,d,\mathcal{T}_0) \leq 0.1\mathcal{F}_0.
\end{aligned}$$

Furthermore, we need to ensure the RHS of (24) is of the same order of $\mathcal{S}_0^2$.

Proof [Proof of Theorem 3.1]

a) 1. First we look at the time instants where $\|\nabla_t\| \geq \epsilon$. If there are more than $\frac{T}{4}$ such time steps, then using Lemma A.5 we have, with probability at least $1-4e^{-l}$

$$\begin{aligned}
f(x_T)-f(x_0) &\leq -\frac{T\epsilon^2}{64\log\left(\frac{1}{\epsilon}\right)^2}+3cL_G\frac{\epsilon^3}{\log\left(\frac{1}{\epsilon}\right)^{10}}\left(\frac{0.5\log\left(\frac{1}{\epsilon}\right)^3}{\sqrt{\epsilon}}+\log\left(\frac{1}{\epsilon}\right)\right)+32c\frac{\epsilon^3}{\log\left(\frac{1}{\epsilon}\right)^7}\\
&\leq -\frac{T\epsilon^2}{128\log\left(\frac{1}{\epsilon}\right)^2}
\end{aligned}$$

Letting T as in (7), we get $f(x_T) \leq f(x_0) - T\epsilon^2/128\log\left(\frac{1}{\epsilon}\right)^2 < f^*$ which is impossible.

2. As follows from Claim 2 in the proof of Theorem 16 of [25], we have, with probability at least $1-10d{\mathcal{T}_0}^2T^2\log(\mathcal{S}_1\sqrt{d}/(\eta r))e^{-l}$

$$f(x_T)-f(x_0) \leq -0.1\frac{T\mathcal{F}_1}{\mathcal{T}_1}$$

which implies $f(x_T) \leq f(x_0) - 0.1T\mathcal{F}_1/\mathcal{T}_1 < f^*$ which is impossible.

b) 1. First we look at the time instants where $\|\nabla_t\| \geq \epsilon$. If the parameters are chosen as in (8), $\mathcal{T}_0 = \kappa_3\frac{\log\left(\frac{1}{\epsilon}\right)^2\log(d)^2}{\sqrt{\epsilon}}$, and $l = \kappa_6\log\left(\frac{d(f(x_0)-f^*)}{\delta\epsilon}\right)$ then we have,

$$\wp(r,l,\nu,\eta,d,\mathcal{T}_0) = \mathcal{O}\left(\epsilon^{1.5}\right)$$

If there are more than $\frac{T}{4}$ such time steps, then using Lemma A.5 we have, with probability at least $1-4e^{-l}$

$$f(x_T)-f(x_0) \leq -\frac{\kappa_0T\epsilon^{2.5}}{64\log\left(\frac{1}{\epsilon}\right)}+\mathcal{O}\left(\epsilon^{1.5}\right) \leq -\frac{\kappa_0T\epsilon^{2.5}}{128\log\left(\frac{1}{\epsilon}\right)}$$

Letting T as in (9), $\kappa_9 \geq 128$, and $\kappa_0\kappa_3/\kappa_8 \geq 128$ we get $f(x_T) \leq f(x_0) - \frac{\kappa_0T\epsilon^{2.5}}{128\log\left(\frac{1}{\epsilon}\right)} < f^*$ which is impossible.

2. As follows from Claim 2 in the proof of Theorem 16 of [25], we have, with probability at least $1 - 3{\mathcal{T}_0}^2 T^2 e^{-l}$

$$f(x_T) - f(x_0) \leq -0.1\frac{T\mathcal{F}_0}{\mathcal{T}_0}$$

which implies $f(x_T) \leq f(x_0) - 0.1T\mathcal{F}_0/\mathcal{T}_0 < f^*$ when $\kappa_9 \geq 128$ and T is as in (8), which is impossible.

■

Proof [Proof of Theorem 3.2] The proof of Theorem 3.2 is same as Theorem 3.1 except for the concentration properties of $\|\zeta_t\|$. In this case we have $\|\zeta_t\|$ to be α-sub-exponential with coefficient $(\Upsilon_t\sqrt{d}/n_1)^{2/3}$ where

$$\Upsilon_t = \frac{\nu L_G(d+2)}{2} + C_0(\sigma + \|\nabla f(x_t)\|)\sqrt{d+1}.$$

So there is an extra term $C_0\sigma\sqrt{d+1}$ which can neither be made smaller using ν nor is of the same order as $\nabla f(x_t)$ so that it can be subsumed in other terms involving $\nabla f(x_t)$. Hence, the only way to make the coefficient smaller, which is essential in the proof, is to increase n_1. This is main reason why the rate deteriorates in the absence if SGC. For the sake of completeness, we provide below the set of conditions that need to be satisfied to pick the parameters in this setting, below.

Choice of parameters for Zeroth-order case when SGC does not hold. When SGC does not hold in the zeroth-order setting the conditions to be satisfied are:

$$\frac{384L_G C_0^2 d(\rho-1)(d+1){\mathcal{T}_0}^2(\mathcal{T}_0\log 9d + l)^3}{K_1^3 n_1^2} \leq \frac{\epsilon^2}{16},$$

$$\frac{8C_0\sqrt{(\rho-1)d(d+1)\mathcal{T}_0}}{K_1^{\frac{3}{2}} n_1}(\mathcal{T}_0\log 9d + l)^{\frac{3}{2}} \leq \frac{\epsilon}{16},$$

$$\eta\mathcal{S}_0\mathcal{T}_0\max(L_H, L_G) \leq \frac{1}{l}, \quad c \leq \sqrt{l}/40,$$

$$\Xi\cdot\sum_{i=0}^{\mathcal{T}_0}\left(\frac{\nu L_G(d+2)}{2} + C_0\sqrt{(\rho-1)(d+1)}L\right) \leq \frac{\beta(\mathcal{T}_0)r}{40\sqrt{d}},$$

$$\frac{(1+\eta\gamma)^{\mathcal{T}_0}\sqrt{\eta}r}{40\sqrt{2\gamma d}} > \mathcal{S}_0, \quad \wp(r, l, \nu, \eta, d, \mathcal{T}_0) \leq 0.1\mathcal{F}_0.$$

Furthermore, we need to ensure the RHS of (24) is of the same order of $\mathcal{S}_0^2$. ■

A.1 Proofs of Lemmas related to Perturbed Stochastic Gradient Descent

Assumption A.1 *[25] Consider random vectors $X_1, X_2, \cdots, X_n \in \mathbb{R}^d$, and the corresponding filtrations $\mathcal{F}_i = \sigma(X_1, X_2, \cdots, X_i)$ for $i = 1, 2, \cdots, n$, such that $X_i|\mathcal{F}_{i-1}$ is zero-mean nSG(σ_i) with $\sigma_i \in \mathcal{F}_{i-1}$. That is,*

$$\mathbf{E}\left[X_i|\mathcal{F}_{i-1}\right] = 0, \quad P(\|X_i\| \geq t|\mathcal{F}_{i-1}) \leq e^{-\frac{t^2}{2\sigma_i^2}}, \quad \forall t \in \mathbb{R}, \forall i = 1, 2, \cdots, n.$$

Lemma A.13 *[25] Let $X_1, X_2, \cdots, X_n \in \mathbb{R}^d$ satisfy Assumption 3.1. $u_i \in \mathcal{F}_{i-1}$ be a random vector for $i = 1, 2, \cdots, n$. Then for any $l > 0,\ \lambda > 0$, there exists absolute constant c such that, with probability at least $1 - e^{-l}$:*

$$\sum_i u_i^\top X_i \leq c\lambda\sum_i \|u_i\|^2\sigma_i^2 + \frac{l}{\lambda}$$

Lemma A.14 *[25] Let $X_1, X_2, \cdots, X_n \in \mathbb{R}^d$ satisfy Assumption 3.1 with $\sigma_1 = \sigma_2 = \cdots = \sigma_n = \sigma$. Then for any $l > 0,\ \lambda > 0$, there exists absolute constant c such that, with probability at least $1 - e^{-l}$:*

$$\sum_i \|X_i\|^2 \leq c\sigma^2(n + l)$$

Lemma A.15 *[25] Let $X_1, X_2, \cdots, X_n \in \mathbb{R}^d$ satisfy Assumption 3.1 with fixed $\{\sigma_i\}$ then for any $l > 0$, there exists an aboslute coonstant c such that, with probability at least $1 - 2de^{-l}$:*

$$\|\sum_{i=1}^{n} X_i\| \leq c\sqrt{\sum_{i=1}^{n} \sigma_i^2 l}$$

Let $F_\nu(x,\xi) = \mathbf{E}_u\left[F(x+\nu u, \xi)\right]$, and $g_{t,i}^j$, and $\nabla F_\nu(x_t,\xi_i)^j$ denote the j-th coordinate of the vector $g_{t,i} = \frac{F(x_t+\nu u_i,\xi_i)-F(x_t,\xi_i)}{\nu}u_i$, and $\nabla F_\nu(x_t,\xi_i)$ respectively.

Lemma A.16 *[38] Let Assumption 2.2 be true for F. Then*

$$\|\nabla F_\nu(x,\xi) - \nabla F(x,\xi)\| \leq \frac{\nu}{2} L_G (d+3)^{\frac{3}{2}}$$

Lemma A.17 *[38] For a Gaussian random vector $u \sim N(0, I_d)$, we have*

$$\mathbf{E}\left[\|u\|^k\right] \leq (d+k)^{\frac{k}{2}}$$

Lemma A.18 *[44] Let (X_i, Y_i), $i = 1, 2, \cdots, n$ be n independent copies of random variables X and Y. Let X be a sub-Gaussian random variable with sub-gaussian norm $\|X\|_{\psi_2} \leq \Upsilon_1$, and Y be a sub-exponential random variable with sub-exponential norm $\|Y\|_{\psi_1} \leq \Upsilon_2$ for some constants Υ_1 and Υ_2. Then for any $t \geq K \max\left(\Upsilon_1, \Upsilon_1^3\right)\Upsilon_2$ we have*

$$\mathbb{P}\left(|\sum_{i=1}^{n} X_i Y_i - \mathbf{E}\left[XY\right]| \geq t\right) \leq 4\exp\left(-K_1 \min\left[\left(\frac{t}{\sqrt{n}\Upsilon_1\Upsilon_2}\right)^2, \left(\frac{t}{\Upsilon_1\Upsilon_2}\right)^{2/3}\right]\right)$$

where K and K_1 are absolute constants.

Proof [Proof of Lemma A.1] Let us write $g_{t,i} = \phi(\nu, u_i, \xi_i)u_i$ where $\phi(\nu, u_i, \xi_i) = \frac{F(x_t+\nu u_i,\xi_i)-F(x_t,\xi_i)}{\nu}$. We will show that $\phi(\nu, u_i, \xi_i)$ is a sub-exponential random variable by showing that its sub-exponential norm or ψ_1-norm, defined as $\|.\|_{\psi_1} = \sup_{p\geq 1} p^{-1}\mathbf{E}\left[|.|^p\right]^{p^{-1}}$, is finite.

$$\|\phi(\nu,u_i,\xi_i)\|_{\psi_1} = \sup_{p\geq 1}\frac{1}{p}\mathbf{E}\left[|\phi(\nu,u_i,\xi_i)|^p\right]^{\frac{1}{p}} = \sup_{p\geq 1}\frac{1}{p}\mathbf{E}_{\xi_i}\left[\mathbf{E}_{u_i}\left[|\phi(\nu,u_i,\xi_i)|^p\right]\right]^{\frac{1}{p}} \tag{31}$$

We first concentrate on the term $\mathbf{E}_{u_i}\left[|\phi(\nu,u_i,\xi_i)|^p\right]$.

$$\mathbf{E}_{u_i}\left[|\phi(\nu,u_i,\xi_i)|^p\right] = \mathbf{E}_{u_i}\left[\left|\frac{F(x_t+\nu u_i,\xi_i) - F(x_t,\xi_i) - \nu\nabla F(x_t,\xi_i)^\top u_i}{\nu} + \nabla F(x_t,\xi_i)^\top u_i\right|^p\right]$$

By Minkowski's inequality,

$$\begin{aligned}
&\mathbf{E}_{u_i}\left[|\phi(\nu,u_i,\xi_i)|^p\right] \\
&\leq \left[\mathbf{E}_{u_i}\left[\left|\frac{F(x_t+\nu u_i,\xi_i) - F(x_t,\xi_i) - \nu\nabla F(x_t,\xi_i)^\top u_i}{\nu}\right|^p\right]^{\frac{1}{p}} + \mathbf{E}_{u_i}\left[\left|\nabla F(x_t,\xi_i)^\top u_i\right|^p\right]^{\frac{1}{p}}\right]^p \\
&\leq \left[\frac{\nu L_G}{2}\mathbf{E}_{u_i}\left[\|u_i\|^{2p}\right]^{\frac{1}{p}} + \|\nabla F(x_t,\xi_i)\|\mathbf{E}_{u_i}\left[\|u_i\|^p\right]^{\frac{1}{p}}\right]^p
\end{aligned}$$

Using Lemma A.17,

$$\mathbf{E}_{u_i}\left[|\phi(\nu,u_i,\xi_i)|^p\right] \leq \left[\frac{\nu L_G(d+2p)}{2} + \sqrt{d+p}\|\nabla F(x_t,\xi_i)\|\right]^p$$

Now from (31), using Minkowski's inequality, we get

$$\|\phi(\nu,u_i,\xi_i)\|_{\psi_1} \leq \sup_{p\geq 1}\frac{1}{p}\mathbf{E}_{\xi_i}\left[\left(\frac{\nu L_G(d+2p)}{2}\right)^p\right]^{\frac{1}{p}} + \sup_{p\geq 1}\frac{1}{p}\mathbf{E}_{\xi_i}\left[\left(\sqrt{d+p}\|\nabla F(x_t,\xi_i)\|\right)^p\right]^{\frac{1}{p}}$$

$$\leq \frac{\nu L_G(d+2)}{2} + \sup_{p\geq 1}\sqrt{\frac{d+p}{p}}\frac{1}{\sqrt{p}}\mathbf{E}_{\xi_i}\left[\|\nabla F(x_t,\xi_i)\|^p\right]^{\frac{1}{p}}$$
$$\leq \frac{\nu L_G(d+2)}{2} + \sup_{p\geq 1}\left(\sqrt{\frac{d+p}{p}}\sup_{p\geq 1}\frac{1}{\sqrt{p}}\mathbf{E}_{\xi_i}\left[\|\nabla F(x_t,\xi_i)\|^p\right]^{\frac{1}{p}}\right)$$

Now,

$$\begin{aligned}
&\mathbf{E}_{\xi_i}\left[\|\nabla F(x_t,\xi_i)\|^p\right]^{p^{-1}}\\
\leq&\mathbf{E}_{\xi_i}\left[(\|\nabla F(x_t,\xi_i)-\nabla f(x_t)+\nabla f(x_t)\|)^p\right]^{p^{-1}}\\
\leq&\mathbf{E}_{\xi_i}\left[2^{p-1}\|\nabla F(x_t,\xi_i)-\nabla f(x_t)\|^p+2^{p-1}\|\nabla f(x_t)\|^p\right]^{p^{-1}}\\
\leq&2\mathbf{E}_{\xi_i}\left[\|\nabla F(x_t,\xi_i)-\nabla f(x_t)\|^p\right]^{p^{-1}}+2\|\nabla f(x_t)\|
\end{aligned}$$

From (5) we have, $\sup_{p\geq 1} p^{-1/2}\mathbf{E}_{\xi_i}\left[(\|\nabla F(x_t,\xi_i)-\nabla f(x_t)\|)^p\right]^{p^{-1}} \leq c_0'\sqrt{\rho-1}\|\nabla_t\|$ where c_0 is a constant. Then,

$$\|\phi(\nu,u_i,\xi_i)\|_{\psi_1} \leq \frac{\nu L_G(d+2)}{2} + (2+c_0'\sqrt{(\rho-1)})\sqrt{d+1}\|\nabla_t\|$$

We also have, $\|u_i^j\|_{\psi_2} \leq 1$, and $\mathbf{E}\left[g_{t,i}\right] = \nabla f_\nu(x_t)$. Then using Lemma A.18, we have $\forall\ j = 1,2,\cdots,d$

$$\mathbb{P}\left(\frac{1}{n_1}\left|\sum_{i=1}^{n_1}\left(g_{t,i}^j-\nabla f_\nu(x_t)^j\right)\right|\geq\tau\right)\leq 4\text{exp}\left(-K_1\min\left[\left(\frac{\sqrt{n_1}\tau}{\Upsilon_t}\right)^2,\left(\frac{n_1\tau}{\Upsilon_t}\right)^{2/3}\right]\right)$$

where $\Upsilon_t = \frac{\nu L_G(d+2)}{2} + c_0\sqrt{(\rho-1)(d+1)}\|\nabla_t\|$. Using union bound,

$$\mathbb{P}\left(\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}g_{t,i}-\nabla f_\nu(x_t)\right\|\geq\tau\right)\leq\mathbb{P}\left(\exists\, j\in\{1,2,\cdots,d\}s.t.\ \left|\frac{1}{n_1}\sum_{i=1}^{n_1}g_{t,i}^j-\nabla f_\nu(x_t)^j\right|\geq\tau/\sqrt{d}\right)$$
$$\leq\sum_{j=1}^{d}\mathbb{P}\left(\left|\frac{1}{n_1}\sum_{i=1}^{n_1}g_{t,i}^j-\nabla f_\nu(x_t)^j\right|\geq\tau/\sqrt{d}\right)\leq 4d\text{exp}\left(-K_1\min\left[\left(\frac{\sqrt{n_1}\tau}{\Upsilon_t\sqrt{d}}\right)^2,\left(\frac{n_1\tau}{\Upsilon_t\sqrt{d}}\right)^{2/3}\right]\right)$$

Using Lemma A.16 we have

$$\mathbb{P}\left(\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}g_{t,i}-\nabla f(x_t)\right\|\geq\tau\right)\leq\mathbb{P}\left(\left\|\frac{1}{n_1}\sum_{i=1}^{n_1}g_{t,i}-\nabla_\nu f(x_t)\right\|\geq\tau-\frac{\nu L_G(d+3)^{\frac{3}{2}}}{2}\right)$$

■

Proof [Proof of Lemma A.2]

$$\mathbf{E}\left[(c_t\|\zeta_t\|)^{\frac{k}{3}}\right] = \int_0^\infty\mathbb{P}\left((c_t\|\zeta_t\|)^{\frac{k}{3}}>\tau\right)d\tau = \int_0^\infty\mathbb{P}\left(\|\zeta_t\|>\tau^{\frac{3}{k}}/c_t\right)d\tau$$
$$\leq\int_0^\infty 4d\text{exp}(-b_{1,t}\tau'^{2/k})d\tau\leq\int_{-\frac{\nu L_G(d+3)^{\frac{3}{2}}}{2}}^\infty 4d\text{exp}(-b_{1,t}\tau^{2/k})d\tau\leq\int_0^\infty 8d\text{exp}(-b_{1,t}\tau^{2/k})d\tau$$

Substituting, $u = b_{1,t}\tau^{2/k}$ we have,

$$\mathbf{E}\left[(c_t\|\zeta_t\|)^{\frac{k}{3}}\right]\leq\int_0^\infty 4dkb_{1,t}^{-k/2}e^{-u}u^{k/2-1}du = 4dkb_{1,t}^{-k/2}\Gamma\left(k/2\right)$$

Using $2(k!)^2\leq(2k)!$, and $\Gamma(k+1/2) = (2k)!\sqrt{\pi}/(4^k k!)$, we have

$$\mathbf{E}\left[e^{s(c_t\|\zeta_t\|)^{\frac{1}{3}}}\right] = 1+\sum_{k=1}^\infty\mathbf{E}\left[\frac{s^k(c_t\|\zeta_t\|)^{\frac{k}{3}}}{k!}\right]\leq 1+\sum_{k=1}^\infty\frac{s^k}{k!}4dkb_{1,t}^{-k/2}\Gamma\left(k/2\right)$$
$$\leq 1+4d\left[\sum_{k=1}^\infty\frac{2ks^{2k}b_{1,t}^{-k}}{(2k)!}\Gamma(k)+\sum_{k=0}^\infty\frac{(2k+1)s^{2k+1}b_{1,t}^{-k-1/2}}{(2k+1)!}\Gamma(k+1/2)\right]$$

$$\leq 1 + 4d\left[\sum_{k=1}^{\infty}\frac{s^{2k}b_{1,t}^{-k}}{k!} + \sqrt{\frac{\pi s^2}{b_{1,t}}}\sum_{k=0}^{\infty}\frac{s^{2k}b_{1,t}^{-k}}{4^k k!}\right]$$
$$\leq 1 + 4d\left[e^{\frac{s^2}{b_{1,t}}} + \sqrt{\frac{\pi s^2}{b_{1,t}}}e^{\frac{s^2}{4b_{1,t}}}\right] \leq 1 + 8de^{\frac{s^2}{b_{1,t}}} \leq 9de^{\frac{s^2}{b_{1,t}}}$$

■

Proof [Proof of Lemma A.3] Setting $c = \eta\|\nabla_i\|$, using Lemma A.2 we have

$$\mathbf{E}\left[e^{s(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}}}\right] \leq 9de^{\frac{s^2}{b_{1,i}}}.$$

Hence, we have the following:

$$\begin{aligned}
&\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-1}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\right]\\
=&\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-2}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\mathbf{E}\left[\exp\left(s(\eta\|\nabla_{t-1}\|\|\zeta_{t-1}\|)^{\frac{1}{3}}\right)|\mathcal{F}_{t-2}\right]\right]\\
=&9d\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-2}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)e^{\frac{s^2}{b_{1,t-1}}}\right]\\
=&9d\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-2}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-2}\frac{s^2}{b_{1,i}}\right)\right].
\end{aligned}$$

Continuing like above we get,

$$\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-1}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\right] \leq (9d)^t. \tag{32}$$

Now, we attempt the main result. Note that, we have

$$\begin{aligned}
\mathbb{P}\left(\eta\sum_{i=0}^{t}\nabla_i^\top\zeta_i \geq \tau\right) &\leq \mathbb{P}\left(\eta\sum_{i=0}^{t}\|\nabla_i\|\|\zeta_i\| \geq \tau\right)\\
&\leq \mathbb{P}\left(\sum_{i=0}^{t}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} \geq \tau^{\frac{1}{3}}\right)\\
&= \mathbb{P}\left(s\sum_{i=0}^{t}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}} \geq s\tau^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\\
&= \mathbb{P}\left(\exp\left(s\sum_{i=0}^{t}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right) \geq \exp\left(s\tau^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\right)\\
&\leq \frac{\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-1}(\eta\|\nabla_i\|\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)\right]}{\exp\left(s\tau^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right)}\\
&\leq \exp\left(t\log 9d - s\tau^{\frac{1}{3}} + \sum_{i=0}^{t-1}\frac{s^2}{b_{1,i}}\right).
\end{aligned}$$

The RHS is minimized at $s = \frac{\tau^{1/3}}{2\sum_{i=0}^{t-1}\frac{1}{b_{1,i}}}$. Substituting for s this value, for some $l > 0$ we have:

$$t\log 9d - \tau^{\frac{2}{3}}/\left(4\sum_{i=0}^{t-1}\frac{1}{b_{1,i}}\right) = -l.$$

Hence, we have

$$\tau = \left(4\sum_{i=0}^{t-1}\frac{1}{b_{1,i}}(t\log 9d + l)\right)^{3/2}.$$

Finally, to prove the statement of the Lemma, note that

$$\left(\sum_{i=0}^{t-1}\frac{1}{b_{1,i}}\right)^{\frac{3}{2}} = \frac{\sqrt{d}}{K_1^{\frac{3}{2}}n_1}\left(\sum_{i=0}^{t-1}(c_i\Upsilon_i)^{\frac{2}{3}}\right)^{\frac{3}{2}} \leq \frac{\eta\sqrt{dt}}{K_1^{\frac{3}{2}}n_1}\sum_{i=0}^{t-1}\left(\frac{\nu L_G(d+2)}{2}\|\nabla_i\| + C_0\sqrt{(\rho-1)(d+1)}\|\nabla_i\|^2\right)$$

■

Proof [Proof of Lemma A.4] From (32) we have,

$$\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-1}(\|\zeta_i\|)^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)\right] \leq (9d)^t$$

where $b_{0,i}$ is as defined in Lemma A.2.

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=0}^{t-1}\|\zeta_i\|^2 \geq \tau\right) \leq \mathbb{P}\left(s\sum_{i=0}^{t-1}\|\zeta_i\|^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}} \geq s\tau^{\frac{1}{6}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)\\
=&\mathbb{P}\left(\exp\left(s\sum_{i=0}^{t-1}\|\zeta_i\|^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right) \geq \exp\left(s\tau^{\frac{1}{6}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)\right)\\
\leq&\frac{\mathbf{E}\left[\exp\left(s\sum_{i=0}^{t-1}\|\zeta_i\|^{\frac{1}{3}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)\right]}{\exp\left(s\tau^{\frac{1}{6}} - \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)} \leq \exp\left(t\log 9d - s\tau^{\frac{1}{6}} + \sum_{i=0}^{t-1}\frac{s^2}{b_{0,i}}\right)
\end{aligned}$$

Following steps as in Lemma A.3 we have, $\tau = \left(4\sum_{i=0}^{t-1}\frac{1}{b_{0,i}}(t\log 9d + l)\right)^3$.

$$\left(\sum_{i=0}^{t-1}\frac{1}{b_{0,i}}\right)^3 = \frac{d}{K_1^3 n_1^2}\left(\sum_{i=0}^{t-1}\Upsilon_i^{\frac{2}{3}}\right)^3 \leq \frac{2dt^2}{K_1^3 n_1^2}\sum_{i=0}^{t-1}\left(\left(\frac{\nu L_G(d+2)}{2}\right)^2 + C_0^2(\rho-1)(d+1)\|\nabla_i\|^2\right)$$

■

Proof [Proof of Lemma A.5]

a)

$$\begin{aligned}
f(x_{t+1}) \leq& f(x_t) + \nabla_t^\top(x_{t+1}-x_t) + \frac{L_G}{2}\|x_{t+1}-x_t\|^2\\
\leq& f(x_t) - \eta\nabla_t^\top(\nabla_t + \tilde{\zeta}_t) + \frac{\eta^2 L_G}{2}\left(\frac{3}{2}\|\nabla_t\|^2 + 3\|\tilde{\zeta}_t\|^2\right)\\
\leq& f(x_t) - \frac{\eta}{4}\|\nabla_t\|^2 - \eta\nabla_t^\top\tilde{\zeta}_t + \frac{3\eta^2 L_G}{2}\|\tilde{\zeta}_t\|^2
\end{aligned}$$

The last inequality holds as we will choose $\eta \leq 1/L_G$. Summing both sides,

$$f(x_t) - f(x_0) \leq -\frac{\eta}{4}\sum_{i=0}^{t-1}\|\nabla_t\|^2 - \eta\sum_{i=0}^{t-1}\nabla_i^\top\tilde{\zeta}_i + \frac{3\eta^2 L_G}{2}\sum_{i=0}^{t-1}\|\tilde{\zeta}_i\|^2 \tag{33}$$

Observe that, by Assumption 2.4,

$$\mathbb{P}(\nabla_t^\top\zeta_t \geq \tau|\mathcal{F}_{t-1}) \leq \mathbb{P}(\|\nabla_t\|\|\zeta_t\| \geq \tau|\mathcal{F}_{t-1}) \leq 2\exp(-\tau^2/(\frac{2(\rho-1)}{n_1}\|\nabla_t\|^4)) \tag{34}$$

So $\nabla_t^\top\zeta_t|\mathcal{F}_{t-1}$ is $c\sqrt{\frac{\rho-1}{n_1}}\|\nabla_t\|^2$-subGaussian. Using Lemma A.13, we have, with probability at least $1-e^{-l}$,

$$-\eta\sum_{i=0}^{t-1}\nabla_i^\top\zeta_i \leq \lambda\eta c\frac{\rho-1}{n_1}\sum_{i=0}^{t-1}\|\nabla_i\|^4 + \eta\frac{l}{\lambda} \leq \lambda\eta c\frac{\rho-1}{n_1}\left(\sum_{i=0}^{t-1}\|\nabla_i\|^2\right)^2 + \eta\frac{l}{\lambda}$$

Plugging $\lambda = \frac{32l}{\sum_{i=0}^{t-1}\|\nabla_i\|^2}$, we have,

$$-\eta\sum_{i=0}^{t-1}\nabla_i^\top\zeta_i \le \eta\left(32cl\frac{\rho-1}{n_1}+\frac{1}{32}\right)\sum_{i=0}^{t-1}\|\nabla_i\|^2 \tag{35}$$

Using Lemma A.13, with probability at least $1-e^{-l}$ we have,

$$-\eta\sum_{i=0}^{t-1}\nabla_i^\top\theta_i \le \frac{\eta}{32}\sum_{i=0}^{t-1}\|\nabla_i\|^2 + 32cl\eta r^2 \tag{36}$$

Using Lemma A.14, we have with probability at least $1-e^{-l}$,

$$\sum_{i=0}^{t-1}\|\theta_i\|^2 \le cr^2(t+l) \tag{37}$$

Note that by Assumption 2.4, $\mathbf{E}\left[\|\zeta_t\|^2|\mathcal{F}_{t-1}\right] \le \frac{\rho-1}{n_1}\|\nabla_t\|^2$, and $\|\zeta_t\|^2|\mathcal{F}_{t-1}$ is $c\frac{\rho-1}{n_1}\|\nabla_t\|^2$-subExponential. So we have, with probability at least $1-e^{-l}$,

$$\sum_{i=0}^{t-1}\|\zeta_i\|^2 \le (c+l)\frac{\rho-1}{n_1}\sum_{i=0}^{t-1}\|\nabla_i\|^2 \tag{38}$$

Combining (33), (35), (36), (37) and (38), using $\|\tilde{\zeta}_t\|^2 \le 2(\|\zeta_t\|^2+\|\theta\|^2)$, and using union bound, we have with probability at least $1-4e^{-l}$,

$$f(x_t)-f(x_0) \le \left(-\frac{\eta}{4}+\eta\left(32lc\frac{\rho-1}{n_1}+\frac{1}{32}\right)+\frac{\eta}{32}+3\eta^2L_G(c+l)\frac{\rho-1}{n_1}\right)\sum_{i=0}^{t-1}\|\nabla_i\|^2 + 3c\eta^2r^2(t+l)L_G + 32cl\eta r^2$$

We need to choose η such that $\left(-\frac{\eta}{4}+\eta\left(32lc\frac{\rho-1}{n_1}+\frac{1}{32}\right)+\frac{\eta}{32}+3\eta^2L_G(c+l)\frac{\rho-1}{n_1}\right) < -\frac{\eta}{16}$. Choosing n_1, and η as in (18), and setting $t=\mathcal{T}_1$, we get (19).

b) Using Lemma A.3, and Lemma A.4, and Assumption 2.1 we have, with probability at least $1-4e^{-l}$

$$f(x_t)-f(x_0) \le -\frac{\eta}{4}\sum_{i=0}^{t-1}\|\nabla_i\|^2+\frac{\eta}{16}\sum_{i=0}^{t-1}\|\nabla_i\|^2+16cl\eta r^2+3cL_G\eta^2r^2(t+l)$$
$$+\frac{8\eta\sqrt{dt}}{K_1^{\frac{3}{2}}n_1}(t\log 9d+l)^{\frac{3}{2}}\sum_{i=0}^{t-1}\left(\frac{\nu LL_G(d+2)}{2}+C_0\sqrt{(\rho-1)(d+1)}\|\nabla_i\|^2\right)$$
$$+\frac{384L_Gd\eta^2t^2(t\log 9d+l)^3}{K_1^3n_1^2}\sum_{i=0}^{t-1}\left(\left(\frac{\nu L_G(d+2)}{2}\right)^2+C_0^2(\rho-1)(d+1)\|\nabla_i\|^2\right)$$

We will choose $\mathcal{T}_0$, η, and n_1 such that, (20), and (21) are true. Then, with probability at least $1-4e^{-l}$, we get (22).

■

Proof [Proof of Lemma A.6]

a) For a fixed $\tau \le t$, we have

$$\|x_\tau-x_0\|^2 \le \eta^2\|\sum_{i=0}^{\tau-1}(\nabla_i+\tilde{\zeta}_i)\|^2 \le 2\eta^2t\sum_{i=0}^{t-1}\|\nabla_i\|^2+4\eta^2(\|\sum_{i=0}^{t-1}\zeta_i\|^2+\|\sum_{i=0}^{t-1}\theta_i\|^2)$$

Using Lemma A.15, we have with probability at least $1-4de^{-l}$,

$$\|\sum_{i=0}^{t-1}\zeta_i\|^2+\|\sum_{i=0}^{t-1}\theta_i\|^2 \le cl\left(\frac{\rho-1}{n_1}\sum_{i=0}^{t-1}\|\nabla_i\|^2+tr^2\right)$$

Combining this with Lemma A.5, with probability at least $1-4e^{-l}-4de^{-l}$, setting $t=\mathcal{T}_1$, and using union bound we have (23).

b)

$$\mathbb{P}\left(\|\sum_{i=0}^{t-1}\zeta_i\|^2 \geq \tau\right) \leq \mathbb{P}\left(\sum_{i=0}^{t-1}\|\zeta_i\|^2 \geq \tau/t\right)$$

So from Lemma A.4, we have with probability at least $1-e^{-l}$

$$\|\sum_{i=0}^{t-1}\zeta_i\|^2 \leq \frac{128dt^3(t\log 9d+l)^3}{K_1^3 n_1^2}\sum_{i=0}^{t-1}\left(\left(\frac{\nu L_G(d+2)}{2}\right)^2 + C_0^2(\rho-1)(d+1)\|\nabla_i\|^2\right)$$

Plugging $t=\mathcal{T}_0$, under condition (20), we have,

$$\|\sum_{i=0}^{\mathcal{T}_0-1}\zeta_i\|^2 \leq \frac{L_G\mathcal{T}_0^2\nu^2(d+2)^2}{192C_0^2(\rho-1)(d+1)} + \frac{\mathcal{T}_0}{12L_G}\sum_{i=0}^{\mathcal{T}_0-1}\|\nabla_i\|^2$$

From (22) we have, with probability at least $1-e^{-l}$

$$\sum_{i=0}^{\mathcal{T}_0-1}\|\nabla_i\|^2 \leq \frac{16}{\eta}(f(x_0)-f(x_{\mathcal{T}_0})+\wp(r,l,\nu,\eta,d,\mathcal{T}_0))$$

Then we have with probability with at least $1-3d\mathcal{T}_0e^{-l}$, we have (24).

■

Proof [Proof of Lemma A.10]

a) Proof for the first-order setting is as in [25].

b) Note that $q_h(t)$ is the same as in part (a). If we can ensure that for the zeroth-order case $\forall t \leq \mathcal{T}_0$ we have $\|q_{sg}(t+1)\| \leq \beta(t)r/(40\sqrt{d})$, then the rest of the proof follows from [25]. For a fixed t, using Cauchy–Schwarz inequality,

$$\mathbb{P}\left(\|q_{sg}(t+1)\| \geq \tau\right) = \mathbb{P}\left(\eta\left\|\sum_{i=0}^{t}(I-\eta\mathcal{H})^{t-i}\hat{\zeta}_i\right\| \geq \tau\right) \leq \mathbb{P}\left(\eta\sum_{i=0}^{t}\|(I-\eta\mathcal{H})\|^{t-i}\|\zeta_i-\zeta_i'\| \geq \tau\right)$$

$$\leq 2\mathbb{P}\left(\eta\sum_{i=0}^{t}\|(I-\eta\mathcal{H})\|^{t-i}\|\zeta_i\| \geq \tau/2\right)$$

$$\leq 2\mathbb{P}\left(\eta\sqrt{\sum_{i=0}^{t}\|(I-\eta\mathcal{H})\|^{2t-2i}}\sqrt{\sum_{i=0}^{t}\|\zeta_i\|^2} \geq \tau/2\right)$$

$$\leq 2\mathbb{P}\left(\sum_{i=0}^{t}\|\zeta_i\|^2 \geq \left(\frac{\tau}{2\eta\beta(t+1)}\right)^2\right)$$

From Lemma A.4, we have with probability at least $1-e^{-l}$

$$\|q_{sg}(t+1)\| \leq \frac{32\sqrt{d(t+1)}\eta\beta(t+1)((t+1)\log 9d+\log 2+l)^{\frac{3}{2}}}{K_1^{3/2}n_1}$$

$$\sum_{i=0}^{t}\left(\frac{\nu L_G(d+2)}{2} + C_0\sqrt{(\rho-1)(d+1)}\|\nabla_i\|\right)$$

Recalling the definition of Ξ from (30), and setting $t=\mathcal{T}_0$, we will choose $\mathcal{T}_0$, r, η, l, and ν such that

$$\Xi\cdot\sum_{i=0}^{\mathcal{T}_0}\left(\frac{\nu L_G(d+2)}{2} + C_0\sqrt{(\rho-1)(d+1)}\|\nabla_i\|\right) \leq \frac{\beta(\mathcal{T}_0)r}{40\sqrt{d}} \tag{39}$$

■

Proof [Proof of Lemma A.12]

a) For the first part, we have from Lemma A.5, with probability at least $1 - 4e^{-l}$,

$$f(x_{\mathcal{T}_1}) - f(x_0) \leq 3c\eta^2 r^2(\mathcal{T}_1 + l)L_G + 32cl\eta r^2 \leq 0.1\mathcal{F}_1$$

By similar methods in [25], we have, with probability at least $2/3 - 10d{\mathcal{T}_1}^2 \log\left(\frac{\mathcal{S}_1\sqrt{d}}{\eta r}\right) e^{-l}$, if $\min\{f(x_{\mathcal{T}_1}) - f(x_0), f(x'_{\mathcal{T}_1}) - f(x_0)\} \geq -\mathcal{F}_1$, then

$$\max\{\|x_{\mathcal{T}_1} - x_0\|, \|x'_{\mathcal{T}_1} - x_0\|\} \geq \frac{\beta(\mathcal{T}_1)\eta r}{40\sqrt{d}} = \frac{(1+\eta\gamma)^{\mathcal{T}_1}\sqrt{\eta}r}{40\sqrt{2\gamma d}} > \mathcal{S}_1 \tag{40}$$

This is in contradiction with Lemma A.11. Then we have with probability at least $2/3 - 10d{\mathcal{T}_1}^2 \log\left(\frac{\mathcal{S}_1\sqrt{d}}{\eta r}\right) e^{-l}$, $\min\{f(x_{\mathcal{T}_1}) - f(x_0), f(x'_{\mathcal{T}_1}) - f(x_0)\} \leq -\mathcal{F}_1$. As the marginal distributions of $x_{\mathcal{T}_1}$ and $x'_{\mathcal{T}_1}$ are same we have,

$$\mathbb{P}(f(x'_{\mathcal{T}_1}) - f(x_0)\} \leq -\mathcal{F}_1) \geq \frac{1}{2}\mathbb{P}(\min\{f(x_{\mathcal{T}_1}) - f(x_0), f(x'_{\mathcal{T}_1}) - f(x_0)\} \leq -\mathcal{F}_1)$$

$$\geq 1/3 - 9d{\mathcal{T}_1}^2 \log\left(\frac{\mathcal{S}_1\sqrt{d}}{\eta r}\right) e^{-l}$$

b) Note that the probability for the second statement being true is at least $1/3 - 1.5{\mathcal{T}_0}^2 e^{-l}$ which is different from [25] but the proof method is same. So we omit the proof here.

■

B Proof of Theorem 4.1

We first state the following optimality conditions for CR Newton method updates due to [37].

Lemma B.1 *[37]*

$$g_t + H_t h_t^* + \frac{M}{2}\|h_t^*\|h_t = 0 \tag{41a}$$

$$H_t + \frac{M}{2}\|h_t^*\|I \succcurlyeq 0 \tag{41b}$$

Intuitively, the proof follows through three stages. First, in Lemma B.2, we show that the descent at each time point is proportional to the cube of the step size.

Lemma B.2 *[47] Let m_t be as defined in* (14)*. Then for all t,*

$$m_t(x_t + h_t^*) - m_t(x_t) \leq -\frac{M}{12}\|h_t^*\|^3 \tag{42}$$

Then, in Lemma B.3 we show that the second-order staionarity of an iterate is upper bounded by the step size at that time point.

Lemma B.3 *Let Assumption 2.2, and 2.3 hold true for f. Then the following holds $\forall t$*

a) for the first-order update of a CR Newton method,

$$\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]} \geq \max\left(\left(A\mathbf{E}\left[\|\nabla f\left(x_t + h_t^*\right)\||\mathcal{F}_t\right] - B\right)^{\frac{1}{2}}, \right.$$

$$\left. \frac{2}{M + 2L_H}\left(-\sqrt{\frac{\sigma_2^2}{n_2}} - \mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right)\right) \tag{43}$$

where $A = \frac{1}{2(L_H+M)}\left(1 - \sqrt{\frac{\rho-1}{n_1}}\right)$, and $B = \frac{1}{4(L_H+M)^2}\left(\frac{\rho-1}{2n_1}L_G^2 + \frac{\sigma_2^2}{n_2}\right)$.

b) for the zeroth-order update of a CR Newton Method

$$\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]} \geq \max\left(\left(A'\mathbf{E}\left[\|\nabla f\left(x_t+h_t^*\right)\||\mathcal{F}_t\right]\ -B'\right)^{\frac{1}{2}},\right.$$
$$\left.\frac{2}{M+2L_H}\left(-\sqrt{\frac{128(1+2\log 2d)(d+16)^4L_G^2}{3n_2}}-\sqrt{3}\nu L_H(d+16)^{\frac{5}{2}}-\mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right)\right) \tag{44}$$

where $A'=\frac{1}{2(L_H+M)}\left(1-\sqrt{\frac{\rho'-1}{n_1}}\right)$, *and*
$B'=\frac{1}{4(L_H+M)^2}\left(\frac{\rho'-1}{n_1}L_G^2+\frac{128(1+2\log 2d)(d+16)^4L_G^2}{3n_2}+3L_H^2\nu^2(d+16)^5+\sqrt{6}\nu(L_H+M)L_G(d+3)^{\frac{3}{2}}\right)$.

Finally, in Lemma B.4, we prove that the expected step size becomes smaller with the horizon.

Lemma B.4 *Let f be a function for which Assumptions 2.2, and 2.3 are true. Then,*

a) for first-order updates generated by Algorithm 2 the following holds:

$$\left(\frac{M}{72}-\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\frac{8}{\sqrt{M}A^{\frac{3}{2}}}\right)\mathbf{E}\left[\|h_R^*\|^3|\mathcal{F}_t\right]$$
$$\leq\frac{f(x_1)-f^*}{T}+\frac{1152L_G^3}{M^2}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{2}}$$
$$+\frac{8}{\sqrt{M}}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\left(\frac{B}{A}\right)^{\frac{3}{2}}+\frac{324}{M^2}\frac{\sigma_2^3}{n_2^{3/2}} \tag{45}$$

where R is an integer random variable uniformly distributed over the support $\{1,2,\cdots,T\}$.

b) for zeroth-order updates generated by Algorithm 2 the following holds:

$$\left(\frac{M}{144}-\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\frac{6}{\sqrt{M}A'^{\frac{3}{2}}}\right)\mathbf{E}\left[\|h_R^*\|^3|\mathcal{F}_t\right]$$
$$\leq\frac{f(x_1)-f^*}{T}+\frac{864L_G^3}{M^2}\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{2}}+\frac{4}{M}(\nu L_G)^{\frac{3}{2}}(d+3)^{\frac{9}{4}}$$
$$+\frac{6}{\sqrt{M}}\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\left(\frac{B'}{A'}\right)^{\frac{3}{2}}+\frac{162}{M^2}\left(\frac{160\sqrt{1+2\log 2d}(d+16)^6L_G^3}{n_2^{\frac{3}{2}}}+21L_H^3(d+16)^{\frac{15}{2}}\nu^3\right) \tag{46}$$

where R is an integer random variable uniformly distributed over the support $\{1,2,\cdots,T\}$.

Combining the above three facts, we complete proof of Theorem 4.1.

Proof [Proof of Theorem 4.1]

a) From Lemma B.3 we have,

$$\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]}+\sqrt{B}+\frac{2}{(2L_H+M)}\sqrt{\frac{\sigma_2^2}{n_2}}\geq$$
$$\max\left(\sqrt{A\mathbf{E}\left[\|\nabla f\left(x_t+h_t^*\right)\||\mathcal{F}_t\right]},-\frac{2}{(2L_H+M)}\mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right) \tag{47}$$

From Lemma B.4, we have

$$\left(\left(\frac{M}{72}-\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\frac{8}{\sqrt{M}A^{\frac{3}{2}}}\right)\mathbf{E}\left[\|h_R^*\|^3|\mathcal{F}_t\right]\right)^{\frac{1}{3}}$$

$$\leq \left(\frac{f(x_1)-f^*}{T}\right)^{\frac{1}{3}} + \frac{11L_G}{M^{\frac{2}{3}}}\left(\frac{\rho-1}{n_1}\right)^{\frac{1}{2}}$$
$$+ \frac{2}{M^{\frac{1}{6}}}\left(\frac{\rho-1}{n_1}\right)^{\frac{1}{4}}\left(\frac{B}{A}\right)^{\frac{1}{2}} + \frac{7}{M^{\frac{2}{3}}}\frac{\sigma_2}{n_2^{1/2}} \tag{48}$$

Combining (47) with (48), using Jensens's inequality we have, and choosing n_1, n_2, T, and M as in (15), we have $\max\left(\sqrt{\frac{\mathbf{E}[\|\nabla f(x_R)\|]}{144M}}, -\frac{\mathbf{E}[\lambda_{1,R}]}{9M}\right) \leq \sqrt{\epsilon}$. Total number of first-order oracle calls, and second-order oracle calls are $Tn_1 = Tn_2 = \mathcal{O}\left(\frac{1}{\epsilon^{\frac{5}{2}}}\right)$.

b) From Lemma B.3 we have,

$$\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]} + \sqrt{B'} + \frac{2}{(2L_H+M)}\left(\sqrt{\frac{128(1+2\log 2d)(d+16)^4L_G^2}{3n_2}} + \sqrt{3}\nu L_H(d+16)^{\frac{5}{2}}\right) \geq$$
$$\max\left(\sqrt{A'\mathbf{E}\left[\|\nabla f\left(x_t+h_t^*\right)\||\mathcal{F}_t\right]}, -\frac{2}{(2L_H+M)}\mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right) \tag{49}$$

From Lemma B.4, we have

$$\left(\left(\frac{1}{144} - \left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\frac{6}{M^{\frac{3}{2}}A'^{\frac{3}{2}}}\right)\mathbf{E}\left[\|h_R^*\|^3|\mathcal{F}_t\right]\right)^{\frac{1}{3}}$$
$$\leq \left(\frac{f(x_1)-f^*}{MT}\right)^{\frac{1}{3}} + \frac{10L_G}{M}\left(\frac{\rho'-1}{n_1}\right)^{\frac{1}{2}} + \frac{2}{M^{\frac{2}{3}}}(\nu L_G)^{\frac{1}{2}}(d+3)^{\frac{3}{4}}$$
$$+ \frac{2}{M^{\frac{1}{2}}}\left(\frac{\rho'-1}{n_1}\right)^{\frac{1}{4}}\left(\frac{B'}{A'}\right)^{\frac{1}{2}} + \frac{6}{M}\left(\frac{6L_G(1+2\log 2d)^{\frac{1}{6}}(d+16)^2}{n_2^{\frac{1}{2}}} + 3L_H(d+16)^{\frac{5}{2}}\nu\right) \tag{50}$$

Combining (49) with (50), using Jensens's inequality we have, and choosing n_1, n_2, T, ν, and M as in (16), we have $\max\left(\sqrt{\mathbf{E}\left[\|\nabla f\left(x_R\right)\|\right]}, -\mathbf{E}\left[\lambda_{1,R}\right]\right) \leq \mathcal{O}\left(\sqrt{\epsilon}\right)$. Total number of first-order oracle calls is $Tn_1 = \mathcal{O}\left(\frac{d}{\epsilon^{\frac{5}{2}}}\right)$, and second-order oracle calls is $Tn_2 = \mathcal{O}\left(\frac{d^4\log d}{\epsilon^{\frac{5}{2}}}\right)$.

■

B.1 Proofs of Lemmas related to CR Newton method

Lemma B.5 *[42]*

$$\mathbf{E}\left[\left\|\nabla_t^2 - \frac{1}{n_2}\sum_{i=1}^{n_2}\nabla^2 F(x_t,\xi_i)\right\|^2\right] \leq \frac{\sigma_2^2}{n_2} \tag{51}$$

$$\mathbf{E}\left[\left\|\nabla_t^2 - \frac{1}{n_2}\sum_{i=1}^{n_2}\nabla^2 F(x_t,\xi_i)\right\|^3\right] \leq \frac{2\sigma_2^3}{n_2^{\frac{3}{2}}} \tag{52}$$

For the zeroth-order estimates of gradient and Hessian as defined in (3) we have the following concentration result.

Lemma B.6 *[5]*

$$\mathbf{E}\left[\|g_t - \nabla_t\|^2\right] \leq \frac{\rho'-1}{n_1}\|\nabla_t\|^2 + \frac{3\nu^2}{2}L_G^2(d+3)^3 \tag{53a}$$

$$\mathbf{E}\left[\|\nabla_t^2 - H_t\|^2\right] \leq \frac{128(1+2\log 2d)(d+16)^4 L_G^2}{3n_2} + 3L_H^2(d+16)^5\nu^2 \tag{53b}$$

$$\mathbf{E}\left[\|\nabla_t^2 - H_t\|^3\right] \leq \frac{160\sqrt{1+2\log 2d}(d+16)^6 L_G^3}{n_2^{\frac{3}{2}}} + 21L_H^3(d+16)^{\frac{15}{2}}\nu^3 \tag{53c}$$

where $\rho' = 1 + 4(d+5)\rho$

Proof [Proof of Lemma B.3]

a) Using (41a) we get,

$$\|g_t + H_t h_t^*\| = \frac{M}{2}\|h_t^*\|^2$$

Then, using Assumption 2.3, and Young's inequality we get,

$$\begin{aligned}
&\|\nabla f\left(x_t + h_t^*\right)\| \\
\leq &\|\nabla f\left(x_t + h_t^*\right) - \nabla_t - \nabla_t^2 h_t^*\| + \|\nabla_t + \nabla_t^2 h_t^*\| \\
\leq &\|\nabla f\left(x_t + h_t^*\right) - \nabla_t - \nabla_t^2 h_t^*\| + \|g_t + H_t h_t^*\| \\
+ &\|g_t - \nabla_t\| + \|\left(H_t - \nabla_t^2\right) h_t^*\| \\
\leq &\frac{M + L_H}{2}\|h_t^*\|^2 + \|g_t - \nabla_t\| + \|\left(H_t - \nabla_t^2\right) h_t^*\| \\
\leq &\left(M + L_H\right)\|h_t^*\|^2 + \|g_t - \nabla_t\| + \frac{1}{2(L_H + M)}\|H_t - \nabla_t^2\|^2
\end{aligned}$$

Taking expectation on both sides, and using Lemma 2.1, Lemma B.5, and Jensen's inequality we have

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla f\left(x_t + h_t^*\right)\||\mathcal{F}_t\right] \\
&\leq (L_H + M)\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right] + \sqrt{\frac{\rho-1}{n_1}}\|\nabla_t\| + \frac{\sigma_2^2}{2(L_H+M)n_2}
\end{aligned}$$

Using Assumption 2.2 we get

$$\begin{aligned}
&\left(1 - \sqrt{\frac{\rho-1}{n_1}}\right)\mathbf{E}\left[\|\nabla f\left(x_t + h_t^*\right)\||\mathcal{F}_t\right] \\
&\leq (L_H + M)\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right] + \sqrt{\frac{\rho-1}{n_1}}L_G\mathbf{E}\left[\|h_t^*\||\mathcal{F}_t\right] + \frac{\sigma_2^2}{2(L_H+M)n_2} \\
&\leq 2(L_H + M)\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right] + \frac{1}{2(L_H+M)}\left(\frac{\rho-1}{n_1}L_G^2 + \frac{\sigma_2^2}{n_2}\right)
\end{aligned}$$

Rearranging we have,

$$\begin{aligned}
&\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]} \\
&\geq \left(\frac{1}{2(L_H+M)}\left(1 - \sqrt{\frac{\rho-1}{n_1}}\right)\mathbf{E}\left[\|\nabla f\left(x_t + h_t^*\right)\||\mathcal{F}_t\right]\right. \\
&\left.-\frac{1}{4(L_H+M)^2}\left(\frac{\rho-1}{n_1}L_G^2 + \frac{\sigma_2^2}{n_2}\right)\right)^{\frac{1}{2}}
\end{aligned} \tag{54}$$

Now, using Assumption 2.3 we get

$$\begin{aligned}
&\mathbf{E}\left[\nabla^2 f\left(x_t + h_t^*\right)|\mathcal{F}_t\right] \succcurlyeq \mathbf{E}\left[\nabla_t^2 - L_H\|h_t^*\|I|\mathcal{F}_t\right] \\
&\succcurlyeq \mathbf{E}\left[H_t - L_H\|h_t^*\|I|\mathcal{F}_t\right] - \sqrt{\frac{\sigma_2^2}{n_2}}I \\
&\succcurlyeq -\sqrt{\frac{\sigma_2^2}{n_2}}I - \left(L_H + \frac{M}{2}\right)\mathbf{E}\left[\|h_t^*\||\mathcal{F}_t\right]I
\end{aligned}$$

$$\mathbf{E}\left[\|h_t^*\||\mathcal{F}_t\right] \geq \frac{2}{M+2L_H}\left(-\sqrt{\frac{\sigma_2^2}{n_2}} - \mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right) \tag{55}$$

Now using Jensen's inequality, and (54) we get (43).

b) Using Lemma B.6, and following the proof of part (a), (54) becomes

$$\sqrt{\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]} \geq \left(\frac{1}{2(L_H+M)}\left(1-\sqrt{\frac{\rho'-1}{n_1}}\right)\mathbf{E}\left[\|\nabla f\left(x_t+h_t^*\right)\||\mathcal{F}_t\right]\right.$$
$$\left.-\frac{1}{4(L_H+M)^2}\left(\frac{\rho'-1}{n_1}L_G^2+\frac{128(1+2\log 2d)(d+16)^4L_G^2}{3n_2}+3L_H^2\nu^2(d+16)^5+\sqrt{6}\nu(L_H+M)L_G(d+3)^{\frac{3}{2}}\right)\right. \tag{56}$$

Similarly, (55) becomes

$$\mathbf{E}\left[\|h_t^*\||\mathcal{F}_t\right] \geq \frac{2}{(2L_H+M)}\left(-\sqrt{\frac{128(1+2\log 2d)(d+16)^4L_G^2}{3n_2}} - \sqrt{3}\nu L_H(d+16)^{\frac{5}{2}} - \mathbf{E}\left[\lambda_{1,t+1}|\mathcal{F}_t\right]\right) \tag{57}$$

■

Proof [Proof of Lemma B.4]

a) Using Young's inequality, and (42), we get

$$\begin{aligned}
&f(x_t+h_t^*)-f(x_t) \leq m_t(x_t+h_t^*)-m_t(x_t)\\
&+(\nabla_t-g_t)^\top h_t^*+\frac{1}{2}h_t^{*\top}(\nabla_t^2-H_t)h_t^*\\
&\leq m_t(x_t+h_t^*)-m_t(x_t)\\
&+\frac{4}{\sqrt{3M}}\|\nabla_t-g_t\|^{\frac{3}{2}}+\frac{162}{M^2}\|\nabla_t^2-H_t\|^3+\frac{M}{18}\|h_t^*\|^3\\
&\leq -\frac{M}{36}\|h_t^*\|^3+\frac{4}{\sqrt{3M}}\|\nabla_t-g_t\|^{\frac{3}{2}}+\frac{162}{M^2}\|\nabla_t^2-H_t\|^3
\end{aligned}$$

Taking expectation on both sides, and using Lemma 2.1 with Jensen's inequality, and Lemma B.5, we get

$$\begin{aligned}
&\mathbf{E}\left[f(x_t+h_t^*)|\mathcal{F}_t\right]-f(x_t) \leq -\frac{M}{36}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right]\\
&+\frac{4}{\sqrt{3M}}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\|\nabla_t\|^{\frac{3}{2}}+\frac{162}{M^2}\frac{2\sigma_2^3}{n_2^{3/2}}
\end{aligned} \tag{58}$$

Now let us relate the gradient size $\|\nabla_t\|$ with $\|h_t^*\|$. Note that, as $x_{t+1}=x_t+h_t^*$ we will use ∇_{t+1} to denote $\nabla f(x_t+h_t^*)$ here. Using triangle inequality, the fact $(a+b)^{3/2} \leq \sqrt{2}(a^{3/2}+b^{3/2})$ for $a,b>0$, Assumption 2.2, and Jensen's inequality we get

$$\begin{aligned}
\|\nabla_t\|^{\frac{3}{2}} &= \|\nabla_t-\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]+\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]\|^{\frac{3}{2}}\\
&\leq (\|\nabla_t-\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]\|+\|\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]\|)^{\frac{3}{2}}\\
&\leq \sqrt{2}(\|\nabla_t-\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]\|^{\frac{3}{2}}+\|\mathbf{E}\left[\nabla_{t+1}|\mathcal{F}_t\right]\|^{\frac{3}{2}})\\
&\leq \sqrt{2}(L_G^{\frac{3}{2}}\mathbf{E}\left[\|h_t^*\|^{\frac{3}{2}}|\mathcal{F}_t\right]+\mathbf{E}\left[\|\nabla_{t+1}\||\mathcal{F}_t\right]^{\frac{3}{2}})
\end{aligned} \tag{59}$$

From Lemma B.3 we have,

$$\mathbf{E}\left[\|h_t^*\|^2|\mathcal{F}_t\right]+B \geq A\mathbf{E}\left[\|\nabla_{t+1}\||\mathcal{F}_t\right]$$

Again using the fact $(a+b)^{3/2} \leq \sqrt{2}(a^{3/2}+b^{3/2})$ for $a,b>0$, and Jensens's inequality we get

$$\sqrt{2}\left(\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right]+B^{\frac{3}{2}}\right) \geq \left(A\mathbf{E}\left[\|\nabla_{t+1}\||\mathcal{F}_t\right]\right)^{\frac{3}{2}} \tag{60}$$

Combining (59), and (60), we get

$$\|\nabla_t\|^{\frac{3}{2}} \leq \sqrt{2}L_G^{\frac{3}{2}}\mathbf{E}\left[\|h_t^*\|^{\frac{3}{2}}|\mathcal{F}_t\right] + \frac{2}{A^{\frac{3}{2}}}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + 2\left(\frac{B}{A}\right)^{\frac{3}{2}}$$

Now, using Young's inequality

$$\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\|\nabla_t\|^{\frac{3}{2}} \leq \frac{288L_G^3}{M^{\frac{3}{2}}}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{2}} + \frac{\sqrt{3}M^{\frac{3}{2}}}{288}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\frac{2}{A^{\frac{3}{2}}}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + 2\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\left(\frac{B}{A}\right)^{\frac{3}{2}} \tag{61}$$

Combining (58), and (61) we get

$$\mathbf{E}\left[f(x_t+h_t^*)|\mathcal{F}_t\right] - f(x_t) \leq -\frac{M}{72}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \frac{1152L_G^3}{M^2}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{2}} + \left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\frac{8}{\sqrt{M}A^{\frac{3}{2}}}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \frac{8}{\sqrt{M}}\left(\frac{\rho-1}{n_1}\right)^{\frac{3}{4}}\left(\frac{B}{A}\right)^{\frac{3}{2}} + \frac{324}{M^2}\frac{\sigma_2^3}{n_2^{3/2}} \tag{62}$$

Rearranging and summing from $t=1$ to T, and dividing both sides by T we get (45).

b) Using Lemma B.6, and following the proof of Lemma B.4 we have the following inequality corresponding to (58)

$$\mathbf{E}\left[f(x_t+h_t^*)|\mathcal{F}_t\right] - f(x_t) \leq -\frac{M}{36}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \frac{3}{\sqrt{M}}\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\|\nabla_t\|^{\frac{3}{2}} + \frac{4}{M}(\nu L_G)^{\frac{3}{2}}(d+3)^{\frac{9}{4}} + \frac{162}{M^2}\left(\frac{160\sqrt{1+2\log 2d}(d+16)^6L_G^3}{n_2^{\frac{3}{2}}} + 21L_H^3(d+16)^{\frac{15}{2}}\nu^3\right) \tag{63}$$

Eventually we get the following descent in the function value similar to (62)

$$\mathbf{E}\left[f(x_t+h_t^*)|\mathcal{F}_t\right] - f(x_t) \leq -\frac{M}{144}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \frac{864L_G^3}{M^2}\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{2}} + \left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\frac{6}{\sqrt{M}A'^{\frac{3}{2}}}\mathbf{E}\left[\|h_t^*\|^3|\mathcal{F}_t\right] + \frac{4}{M}(\nu L_G)^{\frac{3}{2}}(d+3)^{\frac{9}{4}} + \frac{6}{\sqrt{M}}\left(\frac{\rho'-1}{n_1}\right)^{\frac{3}{4}}\left(\frac{B'}{A'}\right)^{\frac{3}{2}} + \frac{162}{M^2}\left(\frac{160\sqrt{1+2\log 2d}(d+16)^6L_G^3}{n_2^{\frac{3}{2}}} + 21L_H^3(d+16)^{\frac{15}{2}}\nu^3\right) \tag{64}$$

Rearranging and summing from $t=1$ to T, and dividing both sides by T we get (46).

■