Journal of Machine Learning Research 23 (2022) 1-61 Submitted 1/21; Revised 5/22; Published 8/22

Simple and Optimal Stochastic Gradient Methods for Nonsmooth Nonconvex Optimization*

Zhize Li ZHIZELI@CMU.EDU
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Jian Li LIJIAN83@MAIL.TSINGHUA.EDU.CN
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing 100084, China

Editor: Ambuj Tewari

Abstract

We propose and analyze several stochastic gradient algorithms for finding stationary points or local minimum in nonconvex, possibly with nonsmooth regularizer, finite-sum and online optimization problems. First, we propose a simple proximal stochastic gradient algorithm based on variance reduction called ProxSVRG+. We provide a clean and tight analysis of ProxSVRG+, which shows that it outperforms the deterministic proximal gradient descent (ProxGD) for a wide range of mini-batch sizes, hence solves an open problem proposed in Reddi et al. (2016b). Also, ProxSVRG+ uses much less proximal oracle calls than ProxSVRG (Reddi et al., 2016b) and extends to the online setting by avoiding full gradient computations. Then, we further propose an optimal algorithm, called SSRGD, based on SARAH (Nguyen et al., 2017) and show that SSRGD further improves the gradient complexity of ProxSVRG+ and achieves the the optimal upper bound, matching the known lower bound of (Fang et al., 2018; Li et al., 2021). Moreover, we show that both ProxSVRG+ and SSRGD enjoy automatic adaptation with local structure of the objective function such as the Polyak-Łojasiewicz (PL) condition for nonconvex functions in the finite-sum case, i.e., we prove that both of them can automatically switch to faster global linear convergence without any restart performed in prior work ProxSVRG (Reddi et al., 2016b). Finally, we focus on the more challenging problem of finding an (ϵ, δ)-local minimum instead of just finding an ϵ-approximate (first-order) stationary point (which may be some bad unstable saddle points). We show that SSRGD can find an (ϵ, δ)-local minimum by simply adding some random perturbations. Our algorithm is almost as simple as its counterpart for finding stationary points, and achieves similar optimal rates.

Keywords: nonconvex optimization, optimal algorithm, proximal gradient descent, variance reduction, local minimum

*. Some preliminary results of this paper appear in two conference papers NeurIPS'18 (Li and Li, 2018) and NeurIPS'19 (Li, 2019). This paper further simplifies some of the proofs, improves the bounds and extends various results to more general settings. The detailed differences between the present paper and the preliminary conference papers are summarized in Section 2.4.

1. Introduction

Nonconvex optimization is ubiquitous in machine learning problems, especially in training deep neural networks. In this paper, we consider the nonsmooth (composite) nonconvex optimization problems of the form

$$\min_{x\in\mathbb{R}^d} \{\Phi(x) := f(x) + h(x)\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable and possibly nonconvex function, while $h : \mathbb{R}^d \to \mathbb{R}$ is nonsmooth but convex (e.g., ℓ_1 norm $\|x\|_1$ or indicator function $I_C(x)$ for some convex set C). In particular, we are interested in functions of f having the *finite-sum* form

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{2}$$

where functions f_is are also possibly nonconvex. The finite-sum form captures the standard empirical risk minimization problems and thus is fundamental to many machine learning problems, ranging from convex optimization (f_is are convex functions) such as logistic regression, SVM to highly nonconvex problem such as optimizing deep neural networks. Moreover, if the number of data samples n is very large or even infinite, e.g., in the online/streaming case, then function f usually is modeled via the *online* form

$$f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)]. \tag{3}$$

For notational convenience, we adopt the notation of the finite-sum form (2) in the descriptions and algorithms in the rest of this paper. However, our results apply to the online form (3) as well by letting $f_i(x) := F(x,\zeta_i)$ and treating n as a very large number or even infinite.

There is a large body of literature for solving the standard problem (1) with finite-sum form (2) or online form (3). The convex setting (i.e., f_is are convex) are well-understood (see e.g., Xiao and Zhang, 2014; Lin et al., 2015; Lan and Zhou, 2015; Woodworth and Srebro, 2016; Lan and Zhou, 2018; Allen-Zhu, 2017; Lan et al., 2019; Li, 2021). Due to the increasing popularity of deep learning, the nonconvex case has attracted significant attention in recent years. In search of the optimal algorithms, a large family of *variance-reduced* methods plays an important role. In particular, SVRG (Johnson and Zhang, 2013), SAGA (Schmidt et al., 2013; Defazio et al., 2014) and SARAH (Nguyen et al., 2017) are representative variance-reduced methods which were originally designed to solve convex optimization problems. They were extended to solve nonconvex problems in subsequent works, such as SCSG (Lei and Jordan, 2017; Lei et al., 2017), SVRG+ (Li and Li, 2018), L-SVRG (Kovalev et al., 2019), SNVRG (Zhou et al., 2018b), SPIDER (Fang et al., 2018), SpiderBoost (Wang et al., 2019), SSRGD (Li, 2019), PAGE (Li et al., 2021). Particularly, Li and Richtárik (2020) provided a unified analysis for a large family of stochastic gradient methods in nonconvex optimization such as SGD, SGD with arbitrary sampling, SGD with compressed gradient, variance-reduced methods such as SVRG and SAGA, and their distributed variants. There are also many advanced variants in the distributed/federated settings such as (Karimireddy et al., 2020; Li et al., 2020; Zhao et al., 2021b; Li and Richtárik, 2021a; Gorbunov et al., 2021; Richtárik et al., 2021; Fatkhullin et al., 2021; Li and Richtárik, 2021b; Zhao et al., 2021a; Richtárik et al., 2022; Li et al., 2022a; Zhao et al., 2022; Li et al., 2022b).

While much prior work focused on the smooth convex/nonconvex case (i.e., $h(x) \equiv 0$ in (1)), relatively less work studied the more general *nonsmooth* nonconvex case. Here we briefly survey

previous work that are directly related to our work. Ghadimi et al. (2016) analyzed the deterministic proximal gradient method (i.e., computing the full-gradient in every iteration) for this nonsmooth nonconvex setting. Here we denote it as ProxGD. Ghadimi et al. (2016) also considered the stochastic variant (here we denote it as ProxSGD). However, ProxSGD requires the minibatch size being a very large number (i.e., $b = O(\sigma^2/\epsilon^2)$) for showing the convergence. Later, Reddi et al. (2016b) provided two algorithms called ProxSVRG and ProxSAGA, which are based on SVRG (Johnson and Zhang, 2013) and SAGA (Defazio et al., 2014). However, their convergence results (using constant or moderate size minibatches) are still worse than the deterministic ProxGD in terms of *proximal oracle complexity* (see Definition 2 for the formal definition). Note that their algorithms (i.e., ProxSVRG/SAGA) outperform the ProxGD only if they use a quite large minibatch size $b = O(n^{2/3})$, where n is the number of training samples. Note that from the perspectives of both computational efficiency and statistical generalization, always computing full-gradient (GD or ProxGD) may not be desirable for large-scale machine learning problems. A reasonable minibatch size is desirable in practice, since the computation of minibatch stochastic gradients can be much cheaper and also implemented in parallel. In fact, practitioners typically use moderate minibatch sizes, often ranging from something like 16 or 32 to a few hundreds (sometimes to a few thousands). Hence, it is important to study the convergence in moderate and constant minibatch size regime. In light of this consideration, Reddi et al. (2016b) presented an important open problem of *developing stochastic methods with provably better performance than ProxGD with constant minibatch size*. In this paper, we provide algorithms for solving this open problem and also achieving optimal results. See Table 1 and 2 for more related works and their detailed convergence results.

Besides, we show that better convergence can be achieved if the objective/loss function satisfies the Polyak-Łojasiewicz (PL) condition (Assumption 4). Note that under the PL condition, one can obtain a faster linear convergence $O(\cdot \log \frac{1}{\epsilon})$ rather than the sublinear convergence $O(\cdot \frac{1}{\epsilon^2})$. In many cases, although the objective function is globally nonconvex, some local regions (e.g., large gradient regions) may satisfy the PL condition. Thus, we also prove that our algorithms can achieve faster linear convergence rates under the PL condition. In particular, the parameter settings of our algorithm remain the same for the finite-sum case, i.e., our algorithms can automatically switch to the faster linear convergence rate in these regions where the PL condition is satisfied.

For nonconvex problems, the point with zero gradient $\nabla f(x) = 0$ can be a local minimum, a local maximum or a saddle point. To avoid stucking in bad saddle points (or local maxima), we want to further find a local minimum, i.e., $\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$ (this is a sufficient condition for x being a local minimum). We note that although finding the global minimum for nonconvex problems is NP-hard in general, it is known that for some special nonconvex problems all local minima are also global minima, such as matrix sensing (Bhojanapalli et al., 2016), matrix completion (Ge et al., 2016), and some special neural networks (Ge et al., 2017). In our paper, we also consider the goal of finding an approximate (ϵ, δ)-local minimum (i.e., $\|\nabla f(x)\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \geq -\delta$) instead of just finding the ϵ-approximate first-order solution (i.e., $\|\nabla f(x)\| \leq \epsilon$). For this purpose, Xu et al. (2018) and Allen-Zhu and Li (2018) independently proposed generic reductions Neon/Neon2, that can be combined with algorithms that finds ϵ-approximate (first-order) solutions in order to find an (ϵ, δ)-local minimum. However, algorithms obtained via such reduction are quite complicated and rarely used in practice. In particular, the reduction needs to extract negative curvature directions from the Hessian to escape saddle points by using a negative curvature search subroutine: given a point x, find an approximate eigenvector corresponding to the smallest eigenvalue of $\nabla^2 f(x)$. This also makes the analysis more complicated. In practice, standard stochastic gradient algorithms can

often work well in nonconvex setting (they can escape bad saddle points) without a negative curvature search subroutine. Intuitively, the saddle points are not very stable, and some stochasticity can escape such saddle points. This raises the following natural theoretical question "Is there any simple modification to the standard first-order gradient method, that can achieve second-order optimality guarantee (local minimum)?". For gradient descent (GD), Jin et al. (2017) showed that a simple perturbation step (by injecting Gaussian noises) is enough to escape saddle points for finding an (ϵ, δ)-local minimum, and this is necessary (Du et al., 2017). Jin et al. (2018) showed that an accelerated GD version can achieve faster convergence rate. Note that both (Jin et al., 2017, 2018) require computing the full gradients. Ge et al. (2015), Daneshmand et al. (2018), Jin et al. (2019), and Fang et al. (2019) analyzed stochastic gradients can also find approximate local minimum if some Gaussian noises are injected.[1] Recently, Ge et al. (2019) showed that a simple perturbation step is also enough to find an (ϵ, δ)-local minimum for SVRG algorithm (Johnson and Zhang, 2013; Li and Li, 2018). Moreover, Ge et al. (2019) also developed a stabilized trick to further improve the dependency of Hessian Lipschitz parameter. See also Table 5 for the convergence rates of the aforementioned works.

In the next section, we present our contributions and provide more discussions and details of related work.

2. Our Contributions

In this section, we review previous related work and present our contributions. Concretely, in Section 2.1, we compare the convergence results of our ProxSVRG+ and SSRGD with previous work, and show that SSRGD further improves on ProxSVRG+ and achieves the optimal convergence results. In Section 2.2, we present the convergence results of ProxSVRG+ and SSRGD under the PL condition. In this PL setting, SSRGD achieves new state-of-the-art results. Note that both ProxSVRG+ and SSRGD can automatically switch to the faster global linear convergence in the finite-sum case under the PL condition. In Section 2.3, we further present the convergence results of SSRGD for finding an *approximate local minimum* which is a more challenging guarantee compared with just finding an approximate first-order stationary point.

2.1 Nonsmooth nonconvex optimization

We list the convergence results of ProxGD, ProxSGD, ProxSVRG/SAGA, ProxSVRG+ and SSRGD in Table 1. Our goal in this section is to find an ϵ-approximate solution of (1) (see Definition 1). The convergence results are stated in terms of the number of stochastic first-order oracle (SFO) calls and proximal oracle (PO) calls (see Definition 2). Although the algorithm of ProxSVRG+ is the same as in the conference version (Li and Li, 2018), our convergence analysis is notably different. In this paper, we further simplify our original proofs of ProxSVRG+ provided in Li and Li (2018), which allows for using larger step size and also leads to better constant in the convergence results.

The original version of SSRGD (Simple Stochastic Recursive Gradient Descent) in Li (2019) was designed to solve the smooth nonconvex problems (i.e., $h(x) \equiv 0$ in (1)). In this paper, we extend it to solve the *nonsmooth* nonconvex problems (1). Compared with the SFO complexity of ProxSVRG+, SSRGD improves the factor $\sqrt{b}$ to b, e.g., $O(\frac{n}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2})$ to $O(\frac{n}{\epsilon^2 b} + \frac{b}{\epsilon^2})$ in the finite-

1. Daneshmand et al. (2018) and Fang et al. (2019) also show that the plain SGD can find approximate local minimum under certain assumptions.

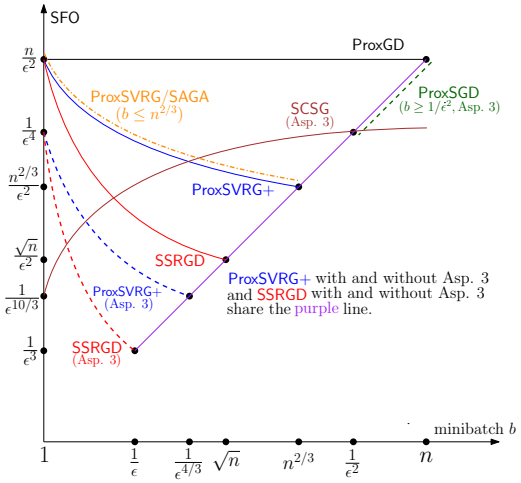

Figure 1: SFO complexity w.r.t. minibatch b. [2]

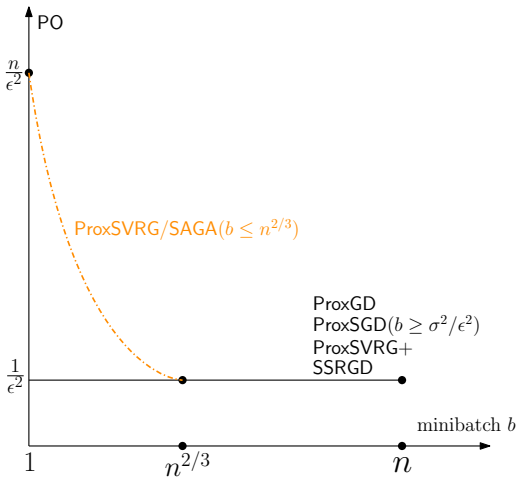

Figure 2: PO complexity w.r.t. minibatch b.

Table 1: SFO and PO complexity for finding an ϵ-approximate solution of problem (1)

Algorithms	Stochastic first-order oracle (SFO)	Proximal oracle (PO)	Assumptions
ProxGD (Ghadimi et al., 2016)	$O(\frac{n}{\epsilon^2})$	$O(\frac{1}{\epsilon^2})$	Asp 1 (finite-sum)
ProxSGD (Ghadimi et al., 2016)	$O(\frac{b}{\epsilon^2})$, where $b \ge \frac{\sigma^2}{\epsilon^2}$	$O(\frac{1}{\epsilon^2})$	Asp 2, 3 (finite-sum or online [3])
ProxSVRG/SAGA (Reddi et al., 2016b)	$O(\frac{n}{\epsilon^2\sqrt{b}} + n)$, where $b \le n^{2/3}$	$O(\frac{n}{b^{3/2}\epsilon^2})$	Asp 2 (finite-sum)
SCSG (Lei et al., 2017) (smooth case $h(x) \equiv 0$ in (1))	$O(\frac{b^{1/3}B^{2/3}}{\epsilon^2} + B)$ [4]	NA [5]	Asp 2, 3 (finite-sum or online [3])
ProxSVRG+ (this paper, Theorem 5)	$O(\frac{n}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2} + n)$	$O(\frac{1}{\epsilon^2})$	Asp 2 (finite-sum)
ProxSVRG+ (this paper, Theorem 5)	$O(\frac{B}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2} + B)$ [4]	$O(\frac{1}{\epsilon^2})$	Asp 2, 3 (finite-sum or online [3])
SSRGD (this paper, Theorem 6)	$O(\frac{n}{\epsilon^2 b} + \frac{b}{\epsilon^2} + n)$	$O(\frac{1}{\epsilon^2})$	Asp 2 (finite-sum)
SSRGD (this paper, Theorem 6)	$O(\frac{B}{\epsilon^2 b} + \frac{b}{\epsilon^2} + B)$ [4]	$O(\frac{1}{\epsilon^2})$	Asp 2, 3 (finite-sum or online [3])

sum case, where b is the minibatch size (See Table 1). Although SSRGD yields better convergence results than ProxSVRG+, in our opinion, the analysis of ProxSVRG+ is quite simple and clean, and useful for understanding the analysis of SSRGD. Hence, we choose to present the details of ProxSVRG+ as well.

2. In this figure, we assume that $\sigma^2/\epsilon^2 \le n$, i.e., $B := \min\{n, \sigma^2/\epsilon^2\} = \sigma^2/\epsilon^2$. Otherwise there is no difference from the finite-sum case if $B = n$. We also omit σ for simplicity of presentation.
3. Note that we refer to the finite-sum problem (2) with large or infinite n as the online problem (3), as discussed in Section 1. In the online problem, computing the full gradient may be very expensive or simply impossible (e.g., if n is infinite), so the bounded variance assumption of stochastic gradient (Assumption 3) is needed.
4. $B := \min\{n, \sigma^2/\epsilon^2\}$.
5. SCSG (Lei et al., 2017) only considered the smooth case, i.e., $h(x) \equiv 0$ in problem (1). The proximal oracle is required only for the nonsmooth setting.

Table 2: SFO and PO complexity of recent algorithms for solving problem (1) [6]

Algorithms	Stochastic first-order oracle (SFO)	Proximal oracle (PO)	Assumptions
SNVRG (Zhou et al., 2018b)	$\widetilde{O}(n+\frac{\sqrt{n}}{\epsilon^2})$ [7]	NA	Asp 2 (finite-sum)
SPIDER (Fang et al., 2018)	$O(n+\frac{\sqrt{n}}{\epsilon^2})$ [7]	NA	Asp 2 (finite-sum)
SpiderBoost (Wang et al., 2019)	$O(n+\frac{\sqrt{n}}{\epsilon^2})$ [7]	$O(\frac{1}{\epsilon^2})$	Asp 2 (finite-sum)
ProxSARAH (Pham et al., 2019)	$O(n+\frac{\sqrt{n}}{\epsilon^2})$	$O(\frac{\sqrt{n}}{b\epsilon^2})$, where $b\le\sqrt{n}$	Asp 2 (finite-sum)
PAGE (Li et al., 2021)	$O(n+\frac{\sqrt{n}+b}{\epsilon^2})$	NA	Asp 2 (finite-sum)
SSRGD (this paper, Theorem 6)	$O(n+\frac{\sqrt{n}}{\epsilon^2})$	$O(\frac{1}{\epsilon^2})$	Asp 2 (finite-sum)
Lower bound (Fang et al., 2018)	$\Omega(\frac{\sqrt{n}}{\epsilon^2})$, where $n\le O(\frac{1}{\epsilon^4})$	NA	Asp 2 (finite-sum)
Lower bound (Li et al., 2021)	$\Omega(n+\frac{\sqrt{n}}{\epsilon^2})$	NA	Asp 2 (finite-sum)

We highlight the following results yielded by ProxSVRG+ and SSRGD:

1) Reddi et al. (2016b) proposed the following open question: developing stochastic methods with provably better performance than ProxGD with constant minibatch size b. Note that #PO of ProxSVRG (Reddi et al., 2016b) is $n/b^{2/3}$ times larger than ProxGD. Our ProxSVRG+ is $\sqrt{b}$ (resp. $\sqrt{b}n/B$ in the online case) times faster than ProxGD in terms of #SFO when $b\le n^{2/3}$ (resp. $b\le B^{2/3}$), and n/b times faster than ProxGD when $b>n^{2/3}$ (resp. $b>B^{2/3}$), where $B:=\min\{n,\frac{\sigma^2}{\epsilon^2}\}$. SSRGD is b (resp. bn/B in the online case) times faster than ProxGD in terms of #SFO when $b\le n^{1/2}$ (resp. $b\le B^{1/2}$), and n/b times faster than ProxGD when $b>n^{1/2}$ (resp. $b>B^{1/2}$). Note that the number of proximal oracle (PO) calls for ProxGD, ProxSVRG+ and SSRGD are the same, i.e., #PO $=O(1/\epsilon^2)$. Hence, both results answers the open question posed by Reddi et al. (2016b). Also see Figure 1 and 2 for an overview.

2) For the online case (which needs an extra bounded variance Assumption 3 since the full gradient may not be available), ProxSVRG+ and SSRGD generalize and improve the result achieved by SCSG (Lei et al., 2017) for the smooth nonconvex case ($h(x)\equiv 0$ in form (1)) to this nonsmooth setting. We note that ProxSVRG+ is also more straightforward than SCSG and the proof is also simpler. Also note that SCSG (Lei et al., 2017) achieves its best convergence result with minibatch size $b=1$ (see Figure 1), while ProxSVRG+ and SSRGD achieves their best results with moderate minibatch sizes and thus can also enjoy speed up with parallelism/vectorization.

6. Similar results hold for these recent algorithms and our SSRGD by replacing n with $B:=\min\{n,\frac{\sigma^2}{\epsilon^2}\}$ in finite-sum or online setting (under Asp 2 and 3) similar to Table 1. Note that SSRGD also matches the lower bound $\Omega(B+\frac{\sqrt{B}}{\epsilon^2})$ (Li et al., 2021) in the online setting (i.e., it achieves optimal results in both finite-sum and online settings).
7. They only analyzed a fixed choice of minibatch size b.

Table 3: SFO and PO complexity of algorithms under PL condition for solving problem (1)

Algorithms	Stochastic first-order oracle (SFO)	Proximal oracle (PO)	Assumptions
ProxGD (Karimi et al., 2016)	$O(n\kappa \log \frac{1}{\epsilon})$	$O(\kappa \log \frac{1}{\epsilon})$	Asp 1, 4 (finite-sum)
ProxSVRG/SAGA (Reddi et al., 2016b)	$O\big(\frac{n\kappa}{\sqrt{b}} \log \frac{1}{\epsilon} + n \log \frac{1}{\epsilon}\big)$, where $b \le n^{2/3}$	$O\big(\frac{n\kappa}{b^{3/2}} \log \frac{1}{\epsilon}\big)$	Asp 2, 4 (finite-sum)
SCSG (Lei et al., 2017) (smooth case $h(x) \equiv 0$ in (1))	$O\big(b^{1/3}B^{2/3}\kappa \log \frac{1}{\epsilon}\big)$ [8]	NA [9]	Asp 2, 3, 4 (finite-sum or online)
ProxSVRG+ (this paper, Theorem 7)	$O\big(\frac{n\kappa}{\sqrt{b}} \log \frac{1}{\epsilon} + b\kappa \log \frac{1}{\epsilon}\big)$	$O(\kappa \log \frac{1}{\epsilon})$	Asp 2, 4 (finite-sum)
ProxSVRG+ (this paper, Theorem 7)	$O\big(\frac{B\kappa}{\sqrt{b}} \log \frac{1}{\epsilon} + b\kappa \log \frac{1}{\epsilon}\big)$ [8]	$O(\kappa \log \frac{1}{\epsilon})$	Asp 2, 3, 4 (finite-sum or online)
SSRGD (this paper, Theorem 8)	$O\big(\frac{n\kappa}{b} \log \frac{1}{\epsilon} + b\kappa \log \frac{1}{\epsilon}\big)$	$O(\kappa \log \frac{1}{\epsilon})$	Asp 2, 4 (finite-sum)
SSRGD (this paper, Theorem 8)	$O\big(\frac{B\kappa}{b} \log \frac{1}{\epsilon} + b\kappa \log \frac{1}{\epsilon}\big)$ [8]	$O(\kappa \log \frac{1}{\epsilon})$	Asp 2, 3, 4 (finite-sum or online)

3) By choosing the best minibatch size $b = \sqrt{n}$ ($b = \sqrt{B}$ in online case), SSRGD achieves the *optimal* results in both finite-sum and online settings, which match the lower bounds given in (Fang et al., 2018; Li et al., 2021). Note that the optimal SFO complexity $O(n+\frac{\sqrt{n}}{\epsilon^2})$ have already been achieved by several recent works such as SNVRG (Zhou et al., 2018a), SPIDER (Fang et al., 2018), SpiderBoost (Wang et al., 2019), ProxSARAH (Pham et al., 2019) and PAGE (Li et al., 2021). We highlight some differences between SSRGD and previous results. For SNVRG, SPIDER and PAGE, they only considered the smooth case ($h(x) \equiv 0$ in form (1)). SpiderBoost only analyzed a fixed choice of minibtach size b and ProxSARAH requires much more #PO calls if minibatch size b is small. SSRGD provides the results for all minibatch size $b \in [1, n]$ and the number of #PO calls is always the same as ProxGD. We also note that ProxSARAH with $\gamma_t \equiv 1$ (Algorithm 1 in (Pham et al., 2019)) is the same as SSRGD. However, the convergence analysis in Pham et al. (2019) (Theorem 6 in their paper) requires $\gamma_t \equiv \gamma = \frac{1}{L\sqrt{\omega m}}$ (where $\omega = \frac{3(n-b)}{2b(n-1)}$), hence does not cover the case that $\gamma_t \equiv 1$. Our main technical contribution is a simple and clean analysis (arguably simpler than that in the previous optimal algorithms) that is inspired by our analysis of ProxSVRG+. See Table 2 for these recent results. Note that these results were not stated in terms of minibatch size b, so we use a separate table for them.

2.2 PL setting

Note that under the PL condition (Assumption 4), one can obtain the faster linear convergence rates $O(\cdot \log \frac{1}{\epsilon})$ (see Theorem 7 and 8) rather than the sublinear convergence rates $O(\cdot \frac{1}{\epsilon^2})$ (see Theorem 5 and 6).

Now we summarize the convergence results of prior work, ProxSVRG+ and SSRGD under PL condition (Assumption 4) in Table 3. The convergence result of these algorithms are very similar

8. $B := \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$.
9. SCSG (Lei et al., 2017) also only considered the smooth case (i.e., $h(x) \equiv 0$ in problem (1)) in the PL setting. The proximal oracle is only required for the nonsmooth setting.

Table 4: SFO and PO complexity of recent algorithms under PL condition for solving problem (1)

Algorithms	Stochastic first-order oracle (SFO)	Proximal oracle (PO)	Assumptions
SNVRG (Zhou et al., 2018b)	$O\big(\big((n+\sqrt{n}\kappa)\log^3 n\big)\log\frac{1}{\epsilon}\big)$	NA [10]	Asp 2, 4 (finite-sum)
SNVRG (Zhou et al., 2018b)	$O\big(\big((B+\sqrt{B}\kappa)\log^3 B\big)\log\frac{1}{\epsilon}\big)$ [11]	NA [10]	Asp 2, 3, 4 (finite-sum or online)
Prox-SpiderBoost-PL (Wang et al., 2019)	$O\big((n+\kappa^2)\log\frac{1}{\epsilon}\big)$	$O(\kappa\log\frac{1}{\epsilon})$	Asp 2, 4 (finite-sum)
PAGE (Li et al., 2021)	$O\big((n+\sqrt{n}\kappa)\log\frac{1}{\epsilon}\big)$	NA [10]	Asp 2, 4 (finite-sum)
PAGE (Li et al., 2021)	$O\big((B+\sqrt{B}\kappa)\log\frac{1}{\epsilon}\big)$ [11]	NA [10]	Asp 2, 3, 4 (finite-sum or online)
SSRGD (this paper, Theorem 8)	$O\big((n+\sqrt{n}\kappa)\log\frac{1}{\epsilon}\big)$	$O(\kappa\log\frac{1}{\epsilon})$	Asp 2, 4 (finite-sum)
SSRGD (this paper, Theorem 8)	$O\big((B+\sqrt{B}\kappa)\log\frac{1}{\epsilon}\big)$ [11]	$O(\kappa\log\frac{1}{\epsilon})$	Asp 2, 3, 4 (finite-sum or online)

to Table 1 by replacing $\frac{1}{\epsilon^2}$ with $\kappa\log\frac{1}{\epsilon}$. Similarly, under PL condition, ProxSVRG+ also improves ProxSVRG by using less PO calls and extends the choice of minibatch size to all $b\in[1,n]$. SSRGD further improves ProxSVRG+ by a factor of $\sqrt{b}$, e.g., from $O(\frac{n\kappa}{\sqrt{b}}\log\frac{1}{\epsilon}+b\kappa\log\frac{1}{\epsilon})$ to $O(\frac{n\kappa}{b}\log\frac{1}{\epsilon}+b\kappa\log\frac{1}{\epsilon})$ in the finite-sum case (See Table 3). In particular, the best result for SSRGD is $\widetilde{O}(\sqrt{n}\kappa)$ while the best results for ProxSVRG and ProxSVRG+ are $\widetilde{O}(n^{2/3}\kappa)$. For the online case, the best result for SSRGD is $\widetilde{O}(\sqrt{B}\kappa)$ while the best results for SCSG and ProxSVRG+ are $\widetilde{O}(B^{2/3}\kappa)$, where $B:=\min\{n,\frac{\sigma^2}{\mu\epsilon}\}$. See Table 3 for more details.

By choosing the best minibatch size, SSRGD achieves new state-of-the-art results in the PL setting. See Table 4 for convergence results of SSRGD (with best minibatch b) and some prior results. Note that we are mainly interested in the case where the condition number $\kappa>\sqrt{n}$. Hence, one can see that SSRGD is better than Prox-SpiderBoost-PL (Wang et al., 2019) in term of the number of SFO calls. If the condition number $\kappa\leq\sqrt{n}$, the SFO complexity of both Prox-SpiderBoost-PL and SSRGD can be bounded by $O(n\log\frac{1}{\epsilon})$.

Note that we do not combine Table 3 and 4 since all prior results in Table 4 were not stated in terms of the minibatch size b. Hence, we use a separate table to list the best results they achieved. We emphasize that our analysis of SSRGD in this PL setting is new and its convergence result also improves over all prior results (see Table 3 and 4).

2.3 Finding local minimum

Now, we consider the problem of finding the approximate (ϵ,δ)-local minimum (i.e., $\|\nabla f(\widehat{x})\|\leq\epsilon$ and $\lambda_{\min}(\nabla^2 f(\widehat{x}))\geq-\delta$) in nonconvex optimization problems. We compare our solution with several other recent theoretical results on finding approximate local minimum. This includes those that adopt Neon/Neon2 (Xu et al., 2018; Allen-Zhu and Li, 2018) (which involve some negative

10. 'NA' in the PO column means that these algorithms only considered the smooth case (i.e., $h(x)\equiv 0$ in problem (1)) in the PL setting. The proximal oracle is only required for the nonsmooth setting.

11. $B:=\min\{n,\frac{\sigma^2}{\mu\epsilon}\}$.

Table 5: Gradient complexity of algorithms for nonconvex finite-sum problem (2) under Asp 5

Algorithms	Stochastic gradient complexity	Guarantee	NC [12]
PGD (Jin et al., 2017)	$\widetilde{O}(\frac{n}{\epsilon^2}+\frac{n}{\delta^4})$	(ϵ,δ)-local min	No
PAGD (Jin et al., 2018)	$\widetilde{O}(\frac{n}{\epsilon^{1.75}}+\frac{n}{\delta^{3.5}})$	(ϵ,δ)-local min	No
Neon2+FastCubic/CDHS (Agarwal et al., 2016; Carmon et al., 2016)	$\widetilde{O}(\frac{n}{\epsilon^{1.5}}+\frac{n^{3/4}}{\epsilon^{1.75}}+\frac{n^{3/4}}{\delta^{3.5}}+\frac{n}{\delta^3})$	(ϵ,δ)-local min	Needed
Neon2+SVRG (Allen-Zhu and Li, 2018)	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n^{3/4}}{\delta^{3.5}}+\frac{n}{\delta^3})$	(ϵ,δ)-local min	Needed
Neon2+SNVRG (Zhou et al., 2018a)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{3/4}}{\delta^{3.5}}+\frac{n}{\delta^3})$	(ϵ,δ)-local min	Needed
Neon2+SPIDER (Fang et al., 2018)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon\delta^2}+\frac{1}{\epsilon\delta^3}+\frac{1}{\delta^5})$	(ϵ,δ)-local min	Needed
Stabilized SVRG (Ge et al., 2019)	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n^{2/3}}{\delta^4}+\frac{n}{\delta^3})$	(ϵ,δ)-local min	No
SSRGD (this paper, Theorem 9)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\delta^4}+\frac{n}{\delta^3})$	(ϵ,δ)-local min	No

curvature searching procedure), such as (Agarwal et al., 2016; Carmon et al., 2016; Allen-Zhu and Li, 2018; Zhou et al., 2018a; Fang et al., 2018) and those by adding simple pertubations to fairly standard gradient methods, such as PGD (Jin et al., 2017), PAGD (Jin et al., 2018), CNC-SGD (Daneshmand et al., 2018) and Stabilized SVRG (Ge et al., 2019).

We show that our SSRGD can find an (ϵ,δ)-local minimum and further improve the convergence result of Stabilized SVRG (Ge et al., 2019) from $n^{2/3}/\epsilon^2$ to $n^{1/2}/\epsilon^2$ (see Table 5). Similar to Ge et al. (2019), SSRGD for finding a local minimum is as simple as its counterpart for finding a first-order stationary point. This is done by just adding a random perturbation in each superepoch, and it does not require a negative curvature (NC) search subroutine (such as Neon/Neon2) or computing Hessian-vector products (such as FastCubic/CDHS). Thus SSRGD (only uses stochastic gradients and random perturbations) can be easily applied in practice. We note that the convergence rate of SSRGD can be better than Neon2+SPIDER (Fang et al., 2018) if δ is very small (i.e., higher accuracy for second-order guarantee $\lambda_{\min}(\nabla^2 f(\widehat{x})) \geq -\delta$). Also Neon2+SPIDER (Fang et al., 2018) requires a negative curvature (NC) search subroutine (such as Neon/Neon2) and thus is more complicated than SSRGD. Our convergence analysis is also arguably simpler. The previous results and our new results are summarized in Table 5 (finite-sum case) and 6 (online case). Also note that the first term of the convergence result of SSRGD (i.e., $\frac{\sqrt{n}}{\epsilon^2}$ or $\frac{1}{\epsilon^3}$) matches the corresponding result for finding the first-order optimal solution (See previous Table 1 or Figure 1) and hence is optimal.

Finally, if we further assume that f has L_3-Lipschitz continuous third-order derivative (i.e., Assumption 7), we show that better convergence rate can be achieved, by replacing the super epoch part of SSRGD (Algorithm 3) by a negative-curvature search step (e.g., Neon2 (Allen-Zhu and Li, 2018))). Currently, the best known result under this setting is achieved in Zhou et al. (2018a), which also uses a negative-curvature search procedure. Our approach is similar to theirs and we obtain the same convergence rate (see Table 7).

12. Negative Curvature search subroutine.

Table 6: Gradient complexity of algorithms for nonconvex online problem (3) under Asp 5 and 6

Algorithms	Stochastic gradient complexity	Guarantee	NC [12]
Noisy SGD (Ge et al., 2015)	$\text{poly}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$	(ϵ, δ)-local min	No
CNC-SGD (Daneshmand et al., 2018)	$\widetilde{O}(\frac{1}{\epsilon^4} + \frac{1}{\delta^{10}})$	(ϵ, δ)-local min	No
Perturbed SGD (Jin et al., 2019)	$\widetilde{O}(\frac{1}{\epsilon^4} + \frac{1}{\delta^8})$	(ϵ, δ)-local min	No
SGD with averaging (Fang et al., 2019)	$\widetilde{O}(\frac{1}{\epsilon^{3.5}} + \frac{1}{\delta^7})$	(ϵ, δ)-local min	No
Neon2+SCSG (Allen-Zhu and Li, 2018)	$\widetilde{O}(\frac{1}{\epsilon^{10/3}} + \frac{1}{\epsilon^2\delta^3} + \frac{1}{\delta^5})$	(ϵ, δ)-local min	Needed
Neon2+Natasha2 (Allen-Zhu, 2018)	$\widetilde{O}(\frac{1}{\epsilon^{3.25}} + \frac{1}{\epsilon^3\delta} + \frac{1}{\delta^5})$	(ϵ, δ)-local min	Needed
Neon2+SNVRG (Zhou et al., 2018a)	$\widetilde{O}(\frac{1}{\epsilon^3} + \frac{1}{\epsilon^2\delta^3} + \frac{1}{\delta^5})$	(ϵ, δ)-local min	Needed
Neon2+SPIDER (Fang et al., 2018)	$\widetilde{O}(\frac{1}{\epsilon^3} + \frac{1}{\epsilon^2\delta^2} + \frac{1}{\delta^5})$	(ϵ, δ)-local min	Needed
SSRGD (this paper, Theorem 9)	$\widetilde{O}(\frac{1}{\epsilon^3} + \frac{1}{\epsilon^2\delta^3} + \frac{1}{\epsilon\delta^4})$	(ϵ, δ)-local min	No

Table 7: Gradient complexity for nonconvex online problem (3) under Asp 5, 6 and 7

Algorithms	Stochastic gradient complexity	Guarantee	NC [12]
FLASH (Yu et al., 2017)	$\widetilde{O}(\frac{1}{\epsilon^{10/3}} + \frac{1}{\epsilon^2\delta^2} + \frac{1}{\delta^4})$	(ϵ, δ)-local min	Needed
SNVRG (Zhou et al., 2018a)	$\widetilde{O}(\frac{1}{\epsilon^3} + \frac{1}{\epsilon^2\delta^2} + \frac{1}{\delta^4})$	(ϵ, δ)-local min	Needed
SSRGD (this paper, Theorem 10)	$\widetilde{O}(\frac{1}{\epsilon^3} + \frac{1}{\epsilon^2\delta^2} + \frac{1}{\delta^4})$	(ϵ, δ)-local min	Needed

2.4 Comparison with the preliminary conference papers

The present paper significantly extends the preliminary two conference papers (Li and Li, 2018; Li, 2019). The major differences between the present paper and the conference papers are summarized as follows. (1) We further simplify the proof of ProxSVRG+ in (Li and Li, 2018). See the proof of Theorem 5 in Appendix A. (2) We extend the original SSRGD in (Li, 2019), which can only handle smooth functions, to a proximal version that can handle nonsmooth functions as well. See Algorithm 2 and Theorem 6. (3) We show SSRGD can achieve linear convergence rate if PL condition is satisfied. Moreover, SSRGD obtains new state-of-the-art results in this classical PL setting. This part is not published elsewhere. See Theorem 8. (4) We provide more details and intuitions in the analysis of SSRGD for escaping saddle point. See the proof of Theorem 9 in Appendix D. (5) We briefly note that SSRGD, when combined with Neon2, can achieve better convergence rate under an additional third order smoothness assumption (Assumption 7). See Theorem 10. This result is not published elsewhere.

2.5 Organization

The remaining paper is organized as follows. Section 3 introduces the notations, standard assumptions and definitions in nonconvex optimization. Section 4 presents the ProxSVRG+ algorithm and its convergence results. Section 5 present the SSRGD algorithm and its convergence results. Then,

Section 6 present the results of ProxSVRG+ and SSRGD in the PL setting, where faster linear convergence can be obtained. Finally, we show how to find an approximate local minimum instead of first-order stationary point via SSRGD and present the corresponding convergence results in Section 7. All proofs are deferred to the appendix.

3. Preliminaries

Let $[n]$ denote the set $\{1, 2, \cdots, n\}$ and $\|\cdot\|$ the Euclidean norm for a vector or the spectral norm for a matrix. Let $\langle u, v\rangle$ denote the inner product of two vectors u and v. Let $\lambda_{\min}(A)$ denote the smallest eigenvalue of a symmetric matrix A. Let $\mathbb{B}_x(r)$ denote a Euclidean ball with center x and radius r. We use $O(\cdot)$ and $\Omega(\cdot)$ to hide the absolute constant, and $\widetilde{O}(\cdot)$ to hide the logarithmic factor.

We assume that the nonsmooth function $h(x)$ in problem (1) is well structured such that the following proximal operator on h can be computed efficiently:

$$\mathrm{prox}_{\eta h}(x) := \arg\min_{y\in\mathbb{R}^d} \Big(h(y) + \frac{1}{2\eta}\|y - x\|^2\Big). \tag{4}$$

For convex problems, one typically uses the optimality gap $\Phi(x) - \Phi(x^*)$ as the convergence criterion for problem (1) (see e.g., (Nesterov, 2004)). But for general nonconvex problems, one typically uses the gradient norm as the convergence criterion. E.g., for smooth nonconvex problems (i.e., $h(x) \equiv 0$), Ghadimi and Lan (2013), Reddi et al. (2016a) and Lei et al. (2017) used $\|\nabla f(x)\|$ to measure the convergence. In order to analyze the convergence results for *nonsmooth* nonconvex problems, following with Ghadimi et al. (2016); Reddi et al. (2016b), we use the *gradient mapping*:

$$\mathcal{G}_\eta(x) := \frac{1}{\eta}\Big(x - \mathrm{prox}_{\eta h}\big(x - \eta\nabla f(x)\big)\Big). \tag{5}$$

Note that if $h(x)$ is a constant function (in particular, zero), this gradient mapping reduces to the ordinary gradient: $\mathcal{G}_\eta(x) = \nabla\Phi(x) = \nabla f(x)$. Thus we use the norm of the gradient mapping $\mathcal{G}_\eta(x)$ as the convergence criterion for problem (1) in the same way as in Ghadimi et al. (2016); Reddi et al. (2016b).

Definition 1 *$\widehat{x}$ is called an ϵ-approximate solution for problem* (1) *if $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|] \le \epsilon$. In particular, if $h(x) \equiv 0$ in* (1)*, this is equivalent to $\mathbb{E}[\|\nabla f(\widehat{x})\|] \le \epsilon$.*

Note that $\mathcal{G}_\eta(x)$ has been already normalized by the step-size η, i.e., it is independent of different algorithms. Let $x^+ := \mathrm{prox}_{\eta h}\big(x - \eta\nabla f(x)\big)$. Then one can see that $\mathcal{G}_\eta(x) := \frac{1}{\eta}\big(x - x^+\big) = \nabla f(x) + \partial h(x^+)$. Moreover, if $\mathcal{G}_\eta(x^*) = 0$, then x^* indeed is a first-order stationary point for problem (1), i.e., $\partial\Phi(x^*) = 0$.

To measure the efficiency of a stochastic algorithm for solving problem (1), we use the following SFO and PO oracle complexities.

Definition 2
1. *Stochastic first-order oracle (SFO): given a point x, SFO outputs a stochastic gradient $\nabla f_i(x)$ (i.e., gradient of one component/data in* (2)*) such that $\mathbb{E}_i[\nabla f_i(x)] = \nabla f(x)$.*
2. *Proximal oracle (PO): given a point x, PO outputs the result of the proximal projection $\mathrm{prox}_{\eta h}(x)$ (see (4)).*

Moreover, in order to prove convergence results, we usually need the following standard smoothness assumptions. Besides, for stochastic/online problems (3), we also usually need the extra bounded variance assumption. These assumptions are very standard in the optimization literature (see e.g., Nesterov, 2004; Ghadimi et al., 2016; Lei et al., 2017; Li and Li, 2018; Allen-Zhu, 2018; Zhou et al., 2018b; Fang et al., 2018; Pham et al., 2019; Li and Richtárik, 2020).

Assumption 1 (L-smoothness) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if*

$$\exists L > 0, \text{ such that } \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{6}$$

Assumption 2 (Average L-smoothness) *A function $f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)$ is average L-smooth if*

$$\exists L > 0, \text{ such that } \mathbb{E}_i[\|\nabla f_i(x) - \nabla f_i(y)\|^2] \leq L^2\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{7}$$

It is not hard to see that Assumption 2 implies Assumption 1.

Assumption 3 (Bounded variance) *The stochastic gradient has bounded variance if*

$$\exists \sigma > 0, \text{ such that } \mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] \leq \sigma^2, \quad \forall x \in \mathbb{R}^d. \tag{8}$$

PL setting: We also prove faster linear convergence rates for nonconvex functions under the Polyak-Łojasiewicz (PL) condition (Polyak, 1963), i.e., $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$. Similar to Definition 1, due to the nonsmooth term $h(x)$ in problem (1), we use the gradient mapping $\mathcal{G}_\eta(x)$ (see (5)) to define a more general form of PL condition as follows:

Assumption 4 (PL condition) *A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ satisfies PL condition* [13] *if*

$$\begin{aligned} &\exists \mu > 0, \text{ such that } \|\mathcal{G}_\eta(x)\|^2 \geq 2\mu(\Phi(x) - \Phi^*), \ \forall x \in \mathbb{R}^d \\ &\qquad (\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \text{ if } h(x) \equiv 0 \text{ in (1)}). \end{aligned} \tag{9}$$

When Assumption 4 holds, we say that it is the PL setting. In the PL setting, we can show linear convergence to the global minimum. Here, we directly use the optimality gap $\Phi(x) - \Phi^*$ as the convergence criterion (see e.g., Reddi et al., 2016b; Lei et al., 2017; Li and Li, 2018; Zhou et al., 2018b), i.e., we use the following Definition 3 in place of Definition 1 for the PL setting.

Definition 3 *$\widehat{x}$ is called an ϵ-approximate solution for problem* (1) *under PL condition (Assumption 4) if $\mathbb{E}[\Phi(\widehat{x}) - \Phi^*] \leq \epsilon$.*

Local minima: Finally, we define the approximate local minimum. Note that in this setting, we do not consider the nonsmooth term, i.e., $h(x) \equiv 0$ in (1). Otherwise the second-order guarantee in Definition 4 is not well-defined for the nonsmooth term.

Definition 4 *$\widehat{x}$ is called an (ϵ, δ)-local minimum for a twice-differentiable function f if*

$$\|\nabla f(\widehat{x})\| \leq \epsilon \ \text{ and } \ \lambda_{\min}(\nabla^2 f(\widehat{x})) \geq -\delta. \tag{10}$$

13. It is worth noting that the PL condition does not imply convexity of the function. For example, $f(x) = x^2 + 3\sin^2 x$ is a nonconvex function but f satisfies PL condition with $\mu = 1/32$.

For finding an approximate local minimum instead of finding an approximate first-order stationary point, we usually need the extra smoothness assumption for the Hessians of f_is.

Assumption 5 (Gradient and Hessian Lipschitz) *A function $f_i : \mathbb{R}^d \to \mathbb{R}$ has an L-Lipschitz continuous gradient if*

$$\exists L > 0, \text{ such that } \|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d, \tag{11}$$

and has a ρ-Lipschitz continuous Hessian if

$$\exists \rho > 0, \text{ such that } \|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq \rho\|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{12}$$

Definition 4 and Assumption 5 are also standard in the literature for finding local minima (see e.g., Ge et al., 2015; Jin et al., 2017; Xu et al., 2018; Allen-Zhu and Li, 2018; Zhou et al., 2018a; Fang et al., 2018; Ge et al., 2019; Li, 2019).

For achieving a high probability result of finding the (ϵ, δ)-local minimum in the online case (i.e., Case 2 in Theorem 9), we need a slightly stronger version of bounded variance Assumption 6 in place of Assumption 3.

Assumption 6 (Bounded Variance) *$\exists \sigma > 0$, such that $\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma^2$, $\forall i, x$.*

We want to point out that Assumption 6 can be relaxed such that $\|\nabla f_i(x) - \nabla f(x)\|$ has sub-Gaussian tail. Then it is sufficient for us to get a high probability bound by using Hoeffding bound on these sub-Gaussian variables. Again, Assumption 6 (or the relaxed sub-Gaussian version) is also standard in the online case for finding approximate local minima (see e.g., Allen-Zhu and Li, 2018; Zhou et al., 2018a; Fang et al., 2018; Jin et al., 2019; Fang et al., 2019; Li, 2019).

If we further assume that f has L_3-Lipschitz continuous third-order derivative, it is possible to achieve even better convergence rate.

Assumption 7 (Third-order Derivative Lipschitz) *A function $f : \mathbb{R}^d \to \mathbb{R}$ has an L_3-Lipschitz continuous third-order derivative if*

$$\exists L_3 > 0, \text{ such that } \|\nabla^3 f(x) - \nabla^3 f(y)\|_F \leq L_3\|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{13}$$

We note such smoothness assumption has already been used in other previous works such as (Anandkumar and Ge, 2016; Carmon et al., 2017; Yu et al., 2017) for escaping higher order saddle points or for achieving better results.

4. ProxSVRG+

In this section, we propose a proximal stochastic gradient algorithm called ProxSVRG+ (Li and Li, 2018). The details of ProxSVRG+ are described in Algorithm 1. We call B the batch size and b the minibatch size.

We note that our algorithm is similar to nonconvex ProxSVRG (Reddi et al., 2016b) and convex Prox-SVRG (Xiao and Zhang, 2014). Prox-SVRG (Xiao and Zhang, 2014) only focused on convex problems, while nonconvex ProxSVRG (Reddi et al., 2016b) analyzed nonconvex problems. The major difference of our ProxSVRG+ from Prox-SVRG and nonconvex ProxSVRG is that we avoid the computation of the full gradient at the beginning of each epoch, i.e., B may not equal to n (see Line 4 of Algorithm 1) while Prox-SVRG and nonconvex ProxSVRG used $B = n$. Our

Algorithm 1: ProxSVRG+

1 **Input:** initial point x_0, batch size B, minibatch size b, epoch length m, step size η
2 $\widetilde{x}^0 = x_0$
3 **for** $s = 0, 1, 2, \ldots$ **do**
4 $\quad g^s = \frac{1}{B}\sum_{j\in I_B} \nabla f_j(\widetilde{x}^s)$ [14]
5 $\quad$ **for** $k = 1, 2, \ldots, m$ **do**
6 $\quad\quad t = sm + k$
7 $\quad\quad v_{t-1} = \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\big) + g^s$
8 $\quad\quad x_t = \mathrm{prox}_{\eta h}(x_{t-1} - \eta v_{t-1})$
9 $\quad$ **end**
10 $\quad \widetilde{x}^{s+1} = x_{(s+1)m}$
11 **end**

contribution mainly lies in the analysis, which is tighter. Note that even if we choose $B = n$, our analysis is stronger than ProxSVRG (Reddi et al., 2016b) (see Table 1). Also, our ProxSVRG+ shows that the "stochastically controlled" trick of SCSG (Lei et al., 2017) (i.e., the epoch length is a geometrically distributed random variable) is not really necessary for achieving the desired convergence bound.[15] As a result, our ProxSVRG+ generalizes the result of SCSG to the more general nonsmooth nonconvex setting and yields simpler analysis.

4.1 Convergence results of ProxSVRG+

Now, we present the main convergence results for ProxSVRG+.

Theorem 5 *Let the step size $\eta \le \frac{1}{(1+2m/\sqrt{b})L}$, where b denotes the minibatch size (Line 7 of Algorithm 1) and m denotes the epoch length (Line 5 of Algorithm 1). Then Algorithm 1 can find an ϵ-approximate solution for problem* (1)*, i.e., $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|] \le \epsilon$ (see Definition 1). We distinguish the following two cases:*

1. *(Finite-sum) Suppose Assumption 2 holds. Let batch size $B = n$ and $m = \sqrt{b}$. Then the number of SFO calls is at most*

$$B + 12L(\Phi(x_0) - \Phi^*)\Big(\frac{B}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\Big) = n + O\Big(\frac{n}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\Big).^{16}$$

14. If $B = n$, ProxSVRG+ is almost the same as ProxSVRG (Reddi et al., 2016b) (i.e., $g^s = \frac{1}{n}\sum_{j=1}^n \nabla f_j(\widetilde{x}^{s-1}) = \nabla f(\widetilde{x}^{s-1})$) except some detailed parameter settings (e.g., step size, epoch length).
15. A similar observation was also made in Natasha1.5 (Allen-Zhu, 2018). However, Natasha1.5 divides each epoch into multiple sub-epochs and randomly chooses the iteration point at the end of each sub-epoch. In our ProxSVRG+ (Algorithm 1), the epoch length is deterministic and it directly uses the point in the last iteration at the end of each epoch.
16. In case the number of SFO calls is less than B (i.e., if the total number of epochs $S < 1$), we may add an explicit term B to the number of SFO calls since the algorithm uses B SFO calls at the beginning of the first epoch $s = 0$ at Line 4 of Algorithm 1. In this situation, ProxSVRG+ (Algorithm 1) terminates within the first epoch $s = 0$, and the first term B is dominating.

2. *(Finite-sum or online) Suppose Assumptions 2 and 3 hold. Let batch size* $B = \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$ *and* $m = \sqrt{b}$*. Then the number of SFO calls is at most*

$$B + 12L(\Phi(x_0) - \Phi^*)\Big(\frac{B}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\Big) = \min\Big\{n, \frac{2\sigma^2}{\epsilon^2}\Big\} + O\Big(\min\Big\{n, \frac{\sigma^2}{\epsilon^2}\Big\}\frac{1}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\Big).$$

In both cases, the number of PO calls equals to the total number of iterations $T = Sm$*, which is at most*

$$\frac{12L(\Phi(x_0) - \Phi^*)}{\epsilon^2} = O\left(\frac{1}{\epsilon^2}\right).$$

Remark: For simplicity of presentation and better comparison with previous bounds, the bounds in Theorem 5 are stated under condition $m = \sqrt{b}$. In fact, our convergence analysis allows for more general values of m and b, and the bounds would depend on both m and b. Please see the proof of Theorem 5 for the details.

The proof for Theorem 5 is notably different from that of ProxSVRG (Reddi et al., 2016b). Reddi et al. (2016b) used a Lyapunov function $R_t^s = \Phi(x_t) + c_t\|x_t - \widetilde{x}^s\|^2$ and showed that it decreases by the accumulated gradient mapping $\sum_{t=sm}^{sm+m-1}\|\mathcal{G}_\eta(x_t)\|^2$ in epoch s (i.e., $R_{(s+1)m}^s \leq R_{sm}^s - \sum_{t=sm}^{(s+1)m-1}\|\mathcal{G}_\eta(x_t)\|^2$). In our proof, we *directly* show that $\Phi(x_t)$ decreases by the accumulated gradient mapping (i.e., $\Phi(x_{(s+1)m}) \leq \Phi(x_{sm}) - \sum_{t=sm}^{(s+1)m-1}\|\mathcal{G}_\eta(x_t)\|^2$) using a different analysis. This is made possible by tightening the inequalities using Young's inequality and the relation between the variance of stochastic gradient estimator and the inner product of the gradient difference and point difference. Also, our convergence result holds for any minibatch size $b \in [1, n]$ unlike ProxSVRG which requires $b \leq n^{2/3}$. Moreover, our ProxSVRG+ uses much less proximal oracle calls than ProxSVRG (see Table 1).

For the online/stochastic Case 2, we avoid the computation of the full gradient at the beginning of each epoch, i.e., B may be less than n. Then, we use the similar idea in SCSG (Lei et al., 2017) to bound the variance term, but we do not need the "stochastically controlled" trick of SCSG (as we discussed before) to achieve the desired convergence bound which yields a much simpler analysis for our ProxSVRG+.

We defer the proof of Theorem 5 to Appendix A. We want to mention that the proof in this paper simplifies our previous proof provided in (Li and Li, 2018) and allows for a larger step size and leads to a better constant in the convergence result (i.e., 12 vs. 36).

5. SSRGD

Now, we present our new SSRGD algorithm to solve the *nonsmooth* nonconvex problems (1). The orginal version of SSRGD in (Li, 2019) was designed to solve the smooth nonconvex problems (i.e., $h(x) \equiv 0$ in (1)) and to find the approximate local minima by escaping saddle points. The new SSRGD algorithm in this paper can be seen as a proximal version of the original SSRGD algorithm.

In this section, we first focus on finding an ϵ-approximate solution. Hence, we ignore the super epoch part (Line 3–5 and Line 11 of Algorithm 2) which is used for escaping saddle points, and add the proximal operator (Line 9 of Algorithm 2) for dealing with this nonsmooth setting. Line 3–5 and Line 11 will be useful in the next Section 7 when we aim to find an (ϵ, δ)-local minimum (see Definition 4). Here, we show that SSRGD (Algorithm 2) achieves the optimal convergence

Algorithm 2: SSRGD

1 **Input:** initial point x_0, batch size B, minibatch size b, epoch length m, step size η
2 **for** $s = 0, 1, 2, \ldots$ **do**
3 **if** $\|\nabla f(x_{sm})\| \le \epsilon$ and not currently in a super epoch **then**
4 $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(r)$, start a super epoch
   // we use super epoch to avoid adding the perturbation steps too often near a saddle point
5 **end**
6 $v_{sm} \leftarrow \frac{1}{B}\sum_{j\in I_B} \nabla f_j(x_{sm})$
7 **for** $k = 1, 2, \ldots, m$ **do**
8 $t \leftarrow sm + k$
9 $x_t \leftarrow \mathrm{prox}_{\eta h}(x_{t-1} - \eta v_{t-1})$
10 $v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}$
11 **if** meet stop condition **then** stop super epoch
12 **end**
13 **end**

results for finding the ϵ-approximate (first-order) solution for the nonsmooth nonconvex problems (1). The main update step (Line 10) adopts the recursive formula which was originally proposed in (Nguyen et al., 2017), and also used in several previous papers on nonconvex problems such as SPIDER (Fang et al., 2018), SpiderBoost (Wang et al., 2019), ProxSARAH (Pham et al., 2019). Our main contribution in this section is a simple and clean analysis that is inspired by our analysis of ProxSVRG+.

5.1 Convergence results of SSRGD

Now, we present the main theorem for SSRGD which can lead to the optimal convergence results.

Theorem 6 *Let the step size $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L}$, where b denotes the minibatch size (Line 10 of Algorithm 2) and m denotes the epoch length (Line 7 of Algorithm 2). Then Algorithm 2 can find an ϵ-approximate solution for problem* (1)*, i.e., $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|] \le \epsilon$ (see Definition 1). We distinguish the following two cases:*

1. *(Finite-sum) Suppose Assumption 2 holds. We let batch size $B = n$ and $m = b$. Then the number of SFO calls is at most*
$$B + 8L(\Phi(x_0) - \Phi^*)\Big(\frac{B}{\epsilon^2 b} + \frac{b}{\epsilon^2}\Big) = n + O\Big(\frac{n}{\epsilon^2 b} + \frac{b}{\epsilon^2}\Big).$$
2. *(Finite-sum or online) Suppose Assumptions 2 and 3 hold. We let batch size $B = \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$ and $m = b$. Then the number of SFO calls is at most*
$$B + 8L(\Phi(x_0) - \Phi^*)\Big(\frac{B}{\epsilon^2 b} + \frac{b}{\epsilon^2}\Big) = \min\Big\{n, \frac{2\sigma^2}{\epsilon^2}\Big\} + O\Big(\min\Big\{n, \frac{\sigma^2}{\epsilon^2}\Big\}\frac{1}{\epsilon^2 b} + \frac{b}{\epsilon^2}\Big).$$

In both cases, the number of PO calls equals to the total number of iterations $T = Sm$, which is at most
$$\frac{8L(\Phi(x_0) - \Phi^*)}{\epsilon^2} = O\left(\frac{1}{\epsilon^2}\right).$$

Remark: Similar to Theorem 5, our analysis allows for more general value of m and b. Compared with the convergence results of ProxSVRG+ (Theorem 5), SSRGD improves the factor $\sqrt{b}$ to b, i.e., $O(\frac{B}{\epsilon^2\sqrt{b}}+\frac{b}{\epsilon^2})$ in Theorem 5 to $O(\frac{B}{\epsilon^2 b}+\frac{b}{\epsilon^2})$ in Theorem 6. In particular, in the finite-sum Case 1, the best result for ProxSVRG+ is $\frac{n^{2/3}}{\epsilon^2}$ where minibatch $b=n^{2/3}$, while the best result for SSRGD is $\frac{\sqrt{n}}{\epsilon^2}$ where minibatch $b=\sqrt{n}$. Moreover, SSRGD can achieve the optimal upper bounds, matching lower bounds $\Omega(n+\frac{\sqrt{n}}{\epsilon^2})$ for the finite-sum case and $\Omega(B+\frac{\sqrt{B}}{\epsilon^2})$ for the online case, shown in (Fang et al., 2018; Li et al., 2021). We defer the proof of Theorem 6 to Appendix B.

6. Faster Linear Convergence under PL Condition

In this section, we show that better convergence can be achieved if the objective function $\Phi(x)$ satisfies the PL condition (Assumption 4).

$$\exists \mu>0, \text{ such that } \|\mathcal{G}_\eta(x)\|^2 \geq 2\mu(\Phi(x)-\Phi^*),\ \forall x\in\mathbb{R}^d.$$

Karimi et al. (2016) showed that PL condition is weaker than many conditions (e.g., strong convexity (SC), restricted strong convexity (RSC) and weak strong convexity (WSC) (Necoara et al., 2015)). Also, if Φ is convex, PL condition is equivalent to the error bounds (EB) and quadratic growth (QG) condition (Luo and Tseng, 1993; Anitescu, 2000).

Note that under the PL condition, one can obtain a faster linear convergence $O(\cdot\log\frac{1}{\epsilon})$ (see Theorem 7 and 8) rather than the sublinear convergence $O(\cdot\frac{1}{\epsilon^2})$ (see Theorem 5 and 6). See Tables 3–4 for an overview of convergence results in this PL setting. In many cases, although the objective function is globally nonconvex, some local regions (e.g., large gradient regions) may satisfy the PL condition. We prove that ProxSVRG+ (Algorithm 1) and SSRGD (Algorithm 2) with same parameter settings for the finite-sum case can automatically switch to the faster linear convergence rate in these regions where PL condition is satisfied. Also note that under the PL condition, we can use the optimality gap $\Phi(x)-\Phi^*$ as the convergence criterion (see Definition 3) instead of $\|\mathcal{G}_\eta(x)\|$ (see Definition 1). Besides, we can directly use the final iteration x_{Sm} as the output point in this PL setting instead of the randomly chosen one $\widehat{x}$. Similar to (Reddi et al., 2016b; Li and Li, 2018), we mainly consider the case where the condition number $\kappa\geq\sqrt{n}$ in the following subsections. Note that if $\kappa<\sqrt{n}$, the SFO complexity of both ProxSVRG+ and SSRGD can be bounded by $O(n\log\frac{1}{\epsilon})$, i.e., independent with κ. The detailed proofs of Theorem 7–8 are deferred to Appendix C.

6.1 ProxSVRG+ under PL Condition

Similar to Theorem 5, we provide the convergence result of ProxSVRG+ (Algorithm 1) under PL condition in the following theorem.

Theorem 7 *Let the step size $\eta\leq\frac{1}{(1+2m/\sqrt{b})L}$, where b denotes the minibatch size (Line 7 of Algorithm 1) and m denotes the epoch length (Line 5 of Algorithm 1). Then the final iteration point x_{Sm} in Algorithm 1 satisfies $\mathbb{E}[\Phi(x_{Sm})-\Phi^*]\leq\epsilon$ under PL condition. We distinguish the following two cases:*

1. *(Finite-sum) Suppose Assumptions 2 and 4 hold. We let batch size $B = n$ and $m = \sqrt{b}$. Then the number of SFO calls can be bounded by*

$$\Big(\frac{B}{\sqrt{b}} + b\Big)\frac{3L}{\mu}\log\frac{2(\Phi(x_0) - \Phi^*)}{\epsilon} = O\Big(\big(\frac{n}{\sqrt{b}} + b\big)\kappa\log\frac{1}{\epsilon}\Big).$$

2. *(Finite-sum or online) Suppose Assumptions 2, 3 and 4 hold. We let batch size $B = \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$ $m = \sqrt{b}$. Then the number of SFO calls can be bounded by*

$$\Big(\frac{B}{\sqrt{b}} + b\Big)\frac{3L}{\mu}\log\frac{2(\Phi(x_0) - \Phi^*)}{\epsilon} = O\Big(\big(\frac{\min\{n, \frac{\sigma^2}{\mu\epsilon}\}}{\sqrt{b}} + b\big)\kappa\log\frac{1}{\epsilon}\Big).$$

In both cases, the number of PO calls equals to the total number of iterations $T = Sm$ which is bounded by

$$\frac{3L}{\mu}\log\frac{2(\Phi(x_0) - \Phi^*)}{\epsilon} = O\Big(\kappa\log\frac{1}{\epsilon}\Big),$$

where $\kappa := \frac{L}{\mu}$.

Remark: From the above theorem, we can see that under the PL condition, ProxSVRG+ (Algorithm 1) can achieve a faster linear convergence $O(\cdot\log\frac{1}{\epsilon})$ rather than the sublinear convergence $O(\cdot\frac{1}{\epsilon^2})$ (see Theorem 5). We would like to mention that Theorem 7 uses exactly the same parameter setting as in Theorem 5 for the finite-sum case. Hence, ProxSVRG+ can automatically switch to this faster linear convergence rate instead of the previous sublinear convergence as long as the objective function $\Phi(x)$ satisfies the PL condition in these regions.

6.2 SSRGD under PL Condition

Similar to Theorem 6, we provide the convergence result of SSRGD (Algorithm 2) under PL condition in the following theorem.

Theorem 8 *Let the step size $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L}$, where b denotes the minibatch size (Line 10 of Algorithm 2) and m denotes the epoch length (Line 7 of Algorithm 2). Then the final iteration point x_{Sm} in Algorithm 2 satisfies $\mathbb{E}[\Phi(x_{Sm}) - \Phi^*] \le \epsilon$ under PL condition. We distinguish the following two cases:*

1. *(Finite-sum) Suppose Assumptions 2 and 4 hold. We let batch size $B = n$ and $m = b$. Then the number of SFO calls can be bounded by*

$$\Big(\frac{B}{b} + b\Big)\frac{2L}{\mu}\log\frac{2(\Phi(x_0) - \Phi^*)}{\epsilon} = O\Big(\big(\frac{n}{b} + b\big)\kappa\log\frac{1}{\epsilon}\Big).$$

2. *(Finite-sum or online) Suppose Assumptions 2, 3 and 4 hold. We let batch size $B = \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$ and $m = b$. Then the number of SFO calls can be bounded by*

$$\Big(\frac{B}{b} + b\Big)\frac{2L}{\mu}\log\frac{2(\Phi(x_0) - \Phi^*)}{\epsilon} = O\Big(\big(\frac{\min\{n, \frac{\sigma^2}{\mu\epsilon}\}}{b} + b\big)\kappa\log\frac{1}{\epsilon}\Big).$$

In both cases, the number of PO calls equals to the total number of iterations $T = Sm$ which is bounded by

$$\frac{2L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon} = O\Big(\kappa\log\frac{1}{\epsilon}\Big),$$

where $\kappa := \frac{L}{\mu}$.

Remark: Similarly, under the PL condition, SSRGD (Algorithm 2) also achieves a faster linear convergence $O(\cdot\log\frac{1}{\epsilon})$ rather than the previous sublinear convergence $O(\cdot\frac{1}{\epsilon^2})$ (see previous Theorem 6). Theorem 8 also uses the same parameter setting as in Theorem 6 for the finite-sum case and hence SSRGD can also switch to this faster linear convergence rate when PL condition is satisfied as ProxSVRG+. Compared with the convergence results of ProxSVRG+ (Theorem 7), SSRGD improves the factor $\sqrt{b}$ to b, i.e., $O((\frac{B}{\sqrt{b}}+b)\kappa\log\frac{1}{\epsilon})$ in Theorem 7 to $O((\frac{B}{b}+b)\kappa\log\frac{1}{\epsilon})$ in Theorem 8. In particular, the best result for ProxSVRG+ is $O(n^{2/3}\kappa\log\frac{1}{\epsilon})$ where minibatch $b = n^{2/3}$, while the best result for SSRGD is $O(\sqrt{n}\kappa\log\frac{1}{\epsilon})$ where minibatch $b = \sqrt{n}$.

7. Finding Approximate Local Minima

In this section, we show that our SSRGD algorithm (Li, 2019) can further find the approximate local minima. SSRGD (Algorithm 2) in Section 5 is just to finding an ϵ-approximate (first-order) solution (see Definition 1) not the (ϵ,δ)-local minimum (see Definition 4), we ignored the super epoch part. In this section, we present the details of the algorithm which can be found in Algorithm 3. In particular, our algorithm is either in a normal epoch ($super\_epoch = 0$) or in a super epoch ($super\_epoch = 1$). We call each inner loop (m iterations) a normal epoch (Line 10–19 of Algorithm 3), i.e., iterations from $t = sm+1$ to $t = sm+m$ consist of the epoch s. A super epoch may contains multiple normal epochs. We enter a super epoch if we are currently in a normal epoch and v_{sm} has a small norm (i.e., near a saddle point) (Line 4 of Algorithm 3). When we enter a super epoch, we add a random perturbation to the current point $\widetilde{x}$ (Line 7 of Algorithm 3). We exit a super epoch if the function value decrease significantly ($f(\widetilde{x}) - f(x_t) \geq f_{\rm thres}$) or the number of iterations exceeds a threshold ($t - t_{\rm init} \geq t_{\rm thres}$). We exit a normal epoch (not in a super epoch) by stopping at a uniformly randomly chosen iteration out of m iterations (Line 17 of Algorithm 3).

7.1 Convergence results of SSRGD for finding approximate local minima

Now, we present the main theorem for SSRGD (Algorithm 3) for finding approximate local minima which corresponds to the convergence results listed in Table 5 and 6. We would like to point out that in this local minima setting, we consider the *smooth* nonconvex case $\Phi(x) = f(x)$ in problem (1), i.e., the nonsmooth term $h(x) \equiv 0$. Otherwise the second-order guarantee in the definition of (ϵ,δ)-local minimum (Definition 4) is not well-defined for the nonsmooth term. Also note that our SSRGD for finding an (ϵ,δ)-local minimum is as simple as its counterpart for finding an ϵ-approximate first-order solution ($\|\nabla f(x)\| \leq \epsilon$) just by adding a random perturbation sometimes, without requiring a negative curvature search subroutine (such as Neon/Neon2) which is typically required by other algorithms. Thus our SSRGD can be simply applied in practice for finding approximate local minimum, and also it leads to simpler convergence analysis.

Algorithm 3: SSRGD (full version for finding approximate local minima)

Input: initial point x_0, batch size B, minibatch size b, epoch length m, step size η, perturbation radius r, threshold function value $f_{\rm thres}$, super epoch length $t_{\rm thres}$
$super\_epoch \leftarrow 0$
for $s = 0, 1, 2, \ldots$ **do**
  if $super\_epoch = 0$ and $\|v_{sm}\| \le \epsilon$ **then**
    $super\_epoch \leftarrow 1$
    $\widetilde{x} \leftarrow x_{sm}, t_{\rm init} \leftarrow sm$
    $x_{sm} \leftarrow \widetilde{x} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(r)$
    // we use super epoch to avoid adding the perturbation steps too often near a saddle point
  end
  $v_{sm} \leftarrow \frac{1}{B}\sum_{j\in I_B} \nabla f_j(x_{sm})$
  for $k = 1, 2, \ldots, m$ **do**
    $t \leftarrow sm + k$
    $x_t \leftarrow x_{t-1} - \eta v_{t-1}$
    $v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}$
    if $super\_epoch = 1$ and $(f(\widetilde{x}) - f(x_t) \ge f_{\rm thres}$ or $t - t_{\rm init} \ge t_{\rm thres})$ **then**
      $super\_epoch \leftarrow 0$; break
    else if $super\_epoch = 0$ **then**
      break with probability $\frac{1}{m-k+1}$
      // we stop this epoch by randomly choosing a point as the starting point of the next epoch
    end
  end
  $x_{(s+1)m} \leftarrow x_t$
end

Theorem 9 *Suppose f satisfies Assumption 5, i.e., f has an L-Lipshitz gradient and a ρ-Lipschitz Hessian. Let step size $\eta = \widetilde{O}(\frac{1}{L})$, epoch length $m = b = \sqrt{B}$, where B, b denote the batch and minibatch size. Moreover, let perturbation radius $r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big)$, threshold function value $f_{\rm thres} = \widetilde{O}(\frac{\delta^3}{\rho^2})$ and super epoch length $t_{\rm thres} = \widetilde{O}(\frac{1}{\eta\delta})$. Denote $\Delta_0 := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. Then Algorithm 3 reaches to an (ϵ, δ)-local minimum at least once with high probability $1 - \zeta$. We distinguish the following two cases:*

1. *(Finite-sum) Let batch size $B = n$. Then the number of stochastic gradient computations is at most*
$$\widetilde{O}\Big(\frac{L\Delta_0\sqrt{n}}{\epsilon^2} + \frac{L\rho^2\Delta_0\sqrt{n}}{\delta^4} + \frac{\rho^2\Delta_0 n}{\delta^3}\Big).$$
2. *(Online) We further assume Assumptions 6 holds. Let batch size $B = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$. Then the number of stochastic gradient computations is at most*
$$\widetilde{O}\Big(\frac{L\Delta_0\sigma}{\epsilon^3} + \frac{L\rho^2\Delta_0\sigma}{\epsilon\delta^4} + \frac{\rho^2\Delta_0\sigma^2}{\epsilon^2\delta^3}\Big).$$

Remark: Note that we can also write Case 2 of Theorem 9 as

$$\widetilde{O}\Big(\frac{L\Delta_0\sqrt{\min\{n,\frac{\sigma^2}{\epsilon^2}\}}}{\epsilon^2}+\frac{L\rho^2\Delta_0\sqrt{\min\{n,\frac{\sigma^2}{\epsilon^2}\}}}{\delta^4}+\frac{\rho^2\Delta_0\min\{n,\frac{\sigma^2}{\epsilon^2}\}}{\delta^3}\Big)$$

by letting $B=\min\{n,\widetilde{O}(\frac{\sigma^2}{\epsilon^2})\}$ in a similar way to Case 2 of Theorem 5 and Theorem 6. Due to the second-order guarantee, the proofs of the finite-sum case and the online case have more difference than previous first-order guarantee methods, so we split the proof of Theorem 9 into two parts, one for case $B=n$ and one for $B\neq n$ (see Appendix D for more details). Also note that if we ignore δ (second-order guarantee $\lambda_{\min}(\nabla^2 f(\widehat{x}))\geq-\delta$), e.g., $\delta=\infty$, then the convergence result provided in Theorem 9 (i.e., $\frac{\sqrt{n}}{\epsilon^2}$ or $\frac{1}{\epsilon^3}$) matches its corresponding result with first-order guarantee in Theorem 6 (which is optimal for finding the ϵ-approximate first-order solution $\|\nabla f(\widehat{x})\|\leq\epsilon$).

Finally, we show that better convergence rate can be achieved if we further assume that f has L_3-Lipschitz continuous third-order derivative (i.e., Assumption 7). This can be achieved by replacing the super epoch part of Algorithm 3 by a negative-curvature search step (e.g., Neon2 (Allen-Zhu and Li, 2018)). Our convergence result matches the best known result by Zhou et al. (2018a), which also uses a negative-curvature search procedure. In particular, we obtain the following theorem.

Theorem 10 (Online case under third-order Lipschitz) *Suppose that Assumptions 5, 6 and 7 hold. Let step size $\eta=\widetilde{O}(\frac{1}{L})$ and batch size $B=\widetilde{O}(\frac{\sigma^2}{\epsilon^2})$, epoch length $m=b=\sqrt{B}$, where B,b denote the batch and minibatch size. Denote $\Delta_0:=f(x_0)-f^*$, where x_0 is the initial point and f^* is the optimal value of f. If we replace the super epoch part of Algorithm 3 by a negative-curvature search step (e.g., Neon2 (Allen-Zhu and Li, 2018)), then it reaches to an (ϵ,δ)-local minimum at least once with high probability $1-\zeta$. The number of stochastic gradient computations is at most*

$$\widetilde{O}\Big(\frac{L\Delta_0\sigma}{\epsilon^3}+\frac{L_3\Delta_0\sigma^2}{\epsilon^2\delta^2}+\frac{L_3L^2\Delta_0}{\delta^4}\Big).$$

Acknowledgements

The authors would like to thank Rong Ge, Chi Jin, Cong Fang for useful discussions and clarifications of their results, and anonymous reviewers for many helpful and constructive suggestions. The research is supported in part by the National Natural Science Foundation of China Grant 62161146004, Turing AI Institute of Nanjing and Xi'an Institute for Interdisciplinary Information Core Technology.

Appendix A. Missing Proofs for Section 4 ProxSVRG+

In this section, we provide the analysis for ProxSVRG+. Our new proof simplifies our original proof in Li and Li (2018). Before proving Theorem 5, we need a useful lemma for the proximal operator. Here we use the following lemma in Lan et al. (2019), instead of the previous Lemma 1 in Li and Li (2018).

Lemma 11 (Lan et al., 2019) *Let $x^+:=\mathrm{prox}_{\eta h}(x-\eta v)$. We have*

$$h(x^+)\leq h(z)+\langle v,z-x^+\rangle+\frac{1}{2\eta}\|z-x\|^2-\frac{1}{2\eta}\|x^+-x\|^2-\frac{1}{2\eta}\|z-x^+\|^2,\ \ \forall z\in\mathbb{R}^d. \quad (14)$$

Proof of Theorem 5. Let $x_t := \mathrm{prox}_{\eta h}(x_{t-1} - \eta v_{t-1})$ and $\bar{x}_t := \mathrm{prox}_{\eta h}\big(x_{t-1} - \eta \nabla f(x_{t-1})\big)$. By letting $x^+ = x_t, x = x_{t-1}, v = v_{t-1}$ and $z = \bar{x}_t$ in (14), we have

$$h(x_t) \le h(\bar{x}_t) + \langle v_{t-1}, \bar{x}_t - x_t\rangle + \frac{1}{2\eta}\|\bar{x}_t - x_{t-1}\|^2 - \frac{1}{2\eta}\|x_t - x_{t-1}\|^2 - \frac{1}{2\eta}\|\bar{x}_t - x_t\|^2. \quad (15)$$

Besides, by letting $x^+ = \bar{x}_t, x = x_{t-1}, v = \nabla f(x_{t-1})$ and $z = x = x_{t-1}$ in (14), we have

$$h(\bar{x}_t) \le h(x_{t-1}) + \langle \nabla f(x_{t-1}), x_{t-1} - \bar{x}_t\rangle - \frac{1}{2\eta}\|\bar{x}_t - x_{t-1}\|^2 - \frac{1}{2\eta}\|x_{t-1} - \bar{x}_t\|^2. \quad (16)$$

Moreover, in view of L-smoothness of f, we have

$$f(x_t) \le f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2. \quad (17)$$

We add (15)–(17) to obtain (recall that $\Phi(x) := f(x) + h(x)$)

$$\begin{aligned}
\Phi(x_t) &\le \Phi(x_{t-1}) - \frac{1}{2\eta}\|x_{t-1} - \bar{x}_t\|^2 - (\frac{1}{2\eta} - \frac{L}{2})\|x_t - x_{t-1}\|^2 \\
&\quad + \langle v_{t-1} - \nabla f(x_{t-1}), \bar{x}_t - x_t\rangle - \frac{1}{2\eta}\|\bar{x}_t - x_t\|^2 \\
&\le \Phi(x_{t-1}) - \frac{1}{2\eta}\|x_{t-1} - \bar{x}_t\|^2 - (\frac{1}{2\eta} - \frac{L}{2})\|x_t - x_{t-1}\|^2 + \frac{\eta}{2}\|v_{t-1} - \nabla f(x_{t-1})\|^2 \quad (18) \\
&= \Phi(x_{t-1}) - \frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2 - (\frac{1}{2\eta} - \frac{L}{2})\|x_t - x_{t-1}\|^2 + \frac{\eta}{2}\|v_{t-1} - \nabla f(x_{t-1})\|^2, \quad (19)
\end{aligned}$$

where (18) uses Young's inequality, and (19) uses the definition of gradient mapping $\mathcal{G}_\eta(x_{t-1})$ (see (5)) and recall $\bar{x}_t := \mathrm{prox}_{\eta h}\big(x_{t-1} - \eta \nabla f(x_{t-1})\big)$.

Now, we bound the variance term in (19) as follows, where the expectations are over I_b and I_B:

$$\begin{aligned}
&\mathbb{E}\Big[\|v_{t-1} - \nabla f(x_{t-1})\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\Big) - \big(\nabla f(x_{t-1}) - g^s\big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\Big) - \Big(\nabla f(x_{t-1}) - \frac{1}{B}\sum_{j\in I_B}\nabla f_j(\widetilde{x}^s)\Big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\big) - \big(\nabla f(x_{t-1}) - \nabla f(\widetilde{x}^s)\big)\Big) + \frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(\widetilde{x}^s) - \nabla f(\widetilde{x}^s)\Big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\big) - \big(\nabla f(x_{t-1}) - \nabla f(\widetilde{x}^s)\big)\Big)\Big\|^2\Big] \\
&\qquad + \mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(\widetilde{x}^s) - \nabla f(\widetilde{x}^s)\Big)\Big\|^2\Big] \quad (20) \\
&= \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\Big\|\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(\widetilde{x}^s)\big) - \big(\nabla f(x_{t-1}) - \nabla f(\widetilde{x}^s)\big)\Big)\Big\|^2\Big]
\end{aligned}$$

$$+\mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(\widetilde{x}^s)-\nabla f(\widetilde{x}^s)\Big)\Big\|^2\Big] \tag{21}$$

$$\le \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\big\|\nabla f_i(x_{t-1})-\nabla f_i(\widetilde{x}^s)\big\|^2\Big]+\mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(\widetilde{x}^s)-\nabla f(\widetilde{x}^s)\Big)\Big\|^2\Big] \tag{22}$$

$$\le \frac{L^2}{b}\mathbb{E}[\|x_{t-1}-\widetilde{x}^s\|^2]+\frac{I\{B<n\}\sigma^2}{B}, \tag{23}$$

where (20) holds due to the independence of I_b and I_B, (21) holds since $\mathbb{E}[\|x_1+x_2+\cdots+x_k\|^2]=\sum_{i=1}^k\mathbb{E}[\|x_i\|^2]$ if $x_1,x_2,\ldots,x_k$ are independent and of mean zero, (22) uses the fact that $\mathbb{E}[\|x-\mathbb{E}x\|^2]\le\mathbb{E}[\|x\|^2]$, for any random variable x, and the last inequality (23) holds due to the average L-smoothness Assumption 2 and bounded variance Assumption 3. Note that the second term $\frac{I\{B<n\}\sigma^2}{B}$ in (23) can be deleted (i.e., Assumption 3 is not needed) if we choose $B=n$.

Now, we plug (23) into (19) to obtain

$$\begin{aligned}
&\mathbb{E}[\Phi(x_t)]\\
&\le \mathbb{E}\Big[\Phi(x_{t-1})-\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2-(\frac{1}{2\eta}-\frac{L}{2})\|x_t-x_{t-1}\|^2+\frac{\eta L^2}{2b}\|x_{t-1}-\widetilde{x}^s\|^2+\frac{I\{B<n\}\eta\sigma^2}{2B}\Big]\\
&\le \mathbb{E}\Big[\Phi(x_{t-1})-\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2-(\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t+1}\|x_t-\widetilde{x}^s\|^2+(\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t}\|x_{t-1}-\widetilde{x}^s\|^2\\
&\qquad+\frac{\eta L^2}{2b}\|x_{t-1}-\widetilde{x}^s\|^2+\frac{I\{B<n\}\eta\sigma^2}{2B}\Big],
\end{aligned} \tag{24}$$

where (24) uses Young's inequality $\|x_t-\widetilde{x}^s\|^2\le(1+\alpha_t)\|x_t-x_{t-1}\|^2+\big(1+\frac{1}{\alpha_t}\big)\|x_{t-1}-\widetilde{x}^s\|^2$, i.e., $-\|x_t-x_{t-1}\|^2\le-\frac{1}{\alpha_t+1}\|x_t-\widetilde{x}^s\|^2+\frac{1}{\alpha_t}\|x_{t-1}-\widetilde{x}^s\|^2$. Also let step size $\eta\le 1/L$ (so that $\frac{1}{2\eta}-\frac{L}{2}\ge 0$).

Adding (24) for all iteration in epoch s, i.e., $t=sm+1$ to $t=sm+m$, we get

$$\begin{aligned}
&\mathbb{E}[\Phi(x_{(s+1)m})]\\
&\le \mathbb{E}\Big[\Phi(x_{sm})-\sum_{t=sm+1}^{sm+m}\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2-\sum_{t=sm+1}^{sm+m}(\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t+1}\|x_t-\widetilde{x}^s\|^2\\
&\qquad+\sum_{t=sm+1}^{sm+m}\Big((\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t}+\frac{\eta L^2}{2b}\Big)\|x_{t-1}-\widetilde{x}^s\|^2+\sum_{t=sm+1}^{sm+m}\frac{I\{B<n\}\eta\sigma^2}{2B}\Big]\\
&\le \mathbb{E}\Big[\Phi(x_{sm})-\sum_{t=sm+1}^{sm+m}\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2-\sum_{t=sm+1}^{sm+m-1}(\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t+1}\|x_t-\widetilde{x}^s\|^2\\
&\qquad+\sum_{t=sm+2}^{sm+m}\Big((\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t}+\frac{\eta L^2}{2b}\Big)\|x_{t-1}-\widetilde{x}^s\|^2+\sum_{t=sm+1}^{sm+m}\frac{I\{B<n\}\eta\sigma^2}{2B}\Big]\\
&= \mathbb{E}\Big[\Phi(x_{sm})-\sum_{t=sm+1}^{sm+m}\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2+\sum_{t=sm+1}^{sm+m}\frac{I\{B<n\}\eta\sigma^2}{2B}\\
&\qquad-\sum_{t=sm+1}^{sm+m-1}\Big((\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_t+1}-(\frac{1}{2\eta}-\frac{L}{2})\frac{1}{\alpha_{t+1}}-\frac{\eta L^2}{2b}\Big)\|x_t-\widetilde{x}^s\|^2\Big]
\end{aligned} \tag{25}$$

$$\leq \mathbb{E}\Big[\Phi(x_{sm}) - \sum_{t=sm+1}^{sm+m} \frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2 + \sum_{t=sm+1}^{sm+m} \frac{I\{B<n\}\eta\sigma^2}{2B}\Big] \tag{26}$$

where (25) holds since $\|\cdot\|^2$ is always non-negative and $\widetilde{x}^s = x_{sm}$, and (26) holds by choosing α_t and η such that $(\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t+1} - (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_{t+1}} - \frac{\eta L^2}{2b} \geq 0$ for $sm+1 \leq t \leq sm+m-1$. Concretely, if we choose $\alpha_t = 2(t\%m) - 1$ and $\eta \leq \frac{1}{L}$, then for any $sm+1 \leq t \leq sm+m-1$, we have that

$$(\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t+1} - (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_{t+1}} - \frac{\eta L^2}{2b} \geq \frac{1-\eta L}{2\eta}(\frac{1}{2(m-1)} - \frac{1}{2m-1}) - \frac{\eta L^2}{2b} \geq 0.$$

Note that the last inequality is quadratic in η. We can verify that choosing $\eta \leq \frac{1}{(1+2m/\sqrt{b})L}$ suffices to make the inequality hold.

Now, we sum up (26) for all epochs $0 \leq s \leq S-1$ as follows:

$$\mathbb{E}[\Phi(x_{Sm}) - \Phi^*] \leq \mathbb{E}\Big[\Phi(x_0) - \Phi^* - \sum_{s=0}^{S-1}\sum_{t=sm+1}^{sm+m} \frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2 + \sum_{s=0}^{S-1}\sum_{t=sm+1}^{sm+m} \frac{I\{B<n\}\eta\sigma^2}{2B}\Big]$$

$$\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|^2] \leq \frac{2\big(\Phi(x_0) - \Phi^*\big)}{Sm\eta} + \frac{I\{B<n\}\sigma^2}{B} \tag{27}$$

$$\leq \frac{\epsilon^2}{2} + \frac{\epsilon^2}{2} = \epsilon^2. \tag{28}$$

Note that $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|] \leq \sqrt{\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|^2]} \leq \epsilon$. The first inequality in (28) holds by randomly choose $\widehat{x}$ from $\{x_{t-1}\}_{t\in[Sm]}$, and the second in (28) holds by choosing $Sm \geq \frac{4(\Phi(x_0)-\Phi^*)}{\epsilon^2\eta}$ and $B \geq \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$.

Now, we can see that the total number of iterations is

$$T = Sm = \frac{4(\Phi(x_0) - \Phi^*\big)}{\epsilon^2\eta}.$$

Choosing $\eta = \frac{1}{(1+2m/\sqrt{b})L}$, we can see that the number of PO calls equals to

$$T = Sm = \frac{4(\Phi(x_0) - \Phi^*\big)}{\epsilon^2\eta} = \frac{4(\Phi(x_0) - \Phi^*\big)(1+2m/\sqrt{b})L}{\epsilon^2}.$$

The number of SFO calls equals to

$$SB + Smb = \frac{4L(\Phi(x_0) - \Phi^*)(1+2m/\sqrt{b})}{\epsilon^2}\big(\frac{B}{m} + b\big).$$

If we choose $m = \sqrt{b}$ (then $\eta \leq \frac{1}{(1+2m/\sqrt{b})L} = \frac{1}{3L}$), the total number of PO calls equals to $T = Sm = \frac{12L(\Phi(x_0)-\Phi^*)}{\epsilon^2}$. The number of SFO calls is $12L(\Phi(x_0) - \Phi^*)\big(\frac{n}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\big)$ if $B = n$ (In this case, the second term in (27) is 0 and thus Assumption 3 is not needed), and $12L(\Phi(x_0) - \Phi^*)\big(\frac{B}{\epsilon^2\sqrt{b}} + \frac{b}{\epsilon^2}\big)$ if $B \geq \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$.

In case the number of SFO calls is less than B (i.e., if the total number of epochs $S < 1$), we may add an explicit term B to the SFO result since the algorithm at least uses B SFO calls in the first epoch $s = 0$ at Line 4 of Algorithm 1. In this situation, ProxSVRG+ (Algorithm 1) terminates within the first epoch $s = 0$. □

Appendix B. Missing Proofs for Section 5 SSRGD

Now, we provide the detailed proofs for Theorem 6.

Proof of Theorem 6. First, according to the update step $x_t := \mathrm{prox}_{\eta h}(x_{t-1} - \eta v_{t-1})$, we recall the key inequality (19):

$$\Phi(x_t) \le \Phi(x_{t-1}) - \frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2 - (\frac{1}{2\eta} - \frac{L}{2})\|x_t - x_{t-1}\|^2 + \frac{\eta}{2}\|v_{t-1} - \nabla f(x_{t-1})\|^2. \quad (29)$$

Now, we bound the variance term in (29) as follows:

$$\begin{aligned}
&\mathbb{E}[\|v_{t-1} - \nabla f(x_{t-1})\|^2] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) + v_{t-2} - \nabla f(x_{t-1})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big) + v_{t-2} - \nabla f(x_{t-2})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big)\Big\|^2\Big] \\
&\qquad + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \quad (30) \\
&= \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\Big\|\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big\|^2\Big] \\
&\qquad + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \quad (31) \\
&\le \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\Big\|\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\Big\|^2\Big] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \quad (32) \\
&\le \frac{L^2}{b}\mathbb{E}[\|x_{t-1} - x_{t-2}\|^2] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2], \quad (33)
\end{aligned}$$

where (30) and (31) use the law of total expectation and $\mathbb{E}[\|y_1 + y_2 + \cdots + y_k\|^2] = \sum_{i=1}^k \mathbb{E}[\|y_i\|^2]$ if $y_1, y_2, \ldots, y_k$ are independent and of mean zero, (32) uses the fact $\mathbb{E}[\|x - \mathbb{E}x\|^2] \le \mathbb{E}[\|x\|^2]$, and (33) holds due to the average L-smoothness Assumption 2.

Note that for $\mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2]$ in (33), we can reuse the same computation above. Thus we can sum up (33) from the beginning of this epoch sm to the point $t-1$,

$$\begin{aligned}
\mathbb{E}[\|v_{t-1} - \nabla f(x_{t-1})\|^2] &\le \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \mathbb{E}[\|v_{sm} - \nabla f(x_{sm})\|^2] \quad (34) \\
&\le \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{I\{B<n\}\sigma^2}{B}, \quad (35)
\end{aligned}$$

Now, we take expectations for (29) and then sum it up from the beginning of this epoch s, i.e., iterations from $sm+1$ to t, by plugging the variance (35) into them to get:

$$\mathbb{E}[\Phi(x_t)] \le \mathbb{E}[\Phi(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\mathcal{G}_\eta(x_{j-1})\|^2] - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2]$$

$$
\begin{aligned}
&+\frac{\eta L^2}{2b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\mathbb{E}[\|x_j-x_{j-1}\|^2]+\frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{I\{B<n\}\sigma^2}{B}\\
&\le \mathbb{E}[\Phi(x_{sm})]-\frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\mathcal{G}_\eta(x_{j-1})\|^2]-\big(\frac{1}{2\eta}-\frac{L}{2}\big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2]\\
&\quad+\frac{\eta L^2(t-1-sm)}{2b}\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2]+\frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{I\{B<n\}\sigma^2}{B}\\
&\le \mathbb{E}[\Phi(x_{sm})]-\frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\mathcal{G}_\eta(x_{j-1})\|^2]+\frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{I\{B<n\}\sigma^2}{B}\\
&\quad-\Big(\big(\frac{1}{2\eta}-\frac{L}{2}\big)-\frac{\eta L^2(m-1)}{2b}\Big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \qquad (36)\\
&\le \mathbb{E}[\Phi(x_{sm})]-\frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\mathcal{G}_\eta(x_{j-1})\|^2]+\frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{I\{B<n\}\sigma^2}{B}, \qquad (37)
\end{aligned}
$$

where (36) holds due to here $t \le sm+m$ in epoch s, (37) holds if the step size $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L}$.

Now, we sum up (37) for all epochs $0 \le s \le S-1$ to finish the proof as follows:

$$
\begin{aligned}
\mathbb{E}[\Phi(x_{Sm})-\Phi^*] &\le \mathbb{E}\Big[\Phi(x_0)-\Phi^*-\frac{\eta}{2}\sum_{s=0}^{S-1}\sum_{t=sm+1}^{sm+m}\|\mathcal{G}_\eta(x_{t-1})\|^2+\frac{\eta}{2}\sum_{s=0}^{S-1}\sum_{t=sm+1}^{sm+m}\frac{I\{B<n\}\sigma^2}{B}\Big]\\
\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|^2] &\le \frac{2\big(\Phi(x_0)-\Phi^*\big)}{Sm\eta}+\frac{I\{B<n\}\sigma^2}{B} \qquad (38)\\
&\le \frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}=\epsilon^2. \qquad (39)
\end{aligned}
$$

Note that $\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|] \le \sqrt{\mathbb{E}[\|\mathcal{G}_\eta(\widehat{x})\|^2]} \le \epsilon$. The inequality (38) holds by randomly choose $\widehat{x}$ from $\{x_{t-1}\}_{t\in[Sm]}$, and (39) holds by choosing $Sm \ge \frac{4(\Phi(x_0)-\Phi^*)}{\epsilon^2\eta}$ and $B \ge \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$.

By choosing $\eta = \frac{1}{(1+\sqrt{(m-1)/b})L}$, the total number of iterations is

$$
T = Sm = \frac{4(\Phi(x_0)-\Phi^*)}{\epsilon^2\eta} = \frac{4(\Phi(x_0)-\Phi^*)(1+\sqrt{(m-1)/b})L}{\epsilon^2},
$$

which is also the number of PO calls. The number of SFO calls is

$$
SB+Smb = 4L(\Phi(x_0)-\Phi^*)(1+\sqrt{(m-1)/b})\left(\frac{B}{\epsilon^2 m}+\frac{b}{\epsilon^2}\right).
$$

If we choose $m = b$ (then $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L} = \frac{1}{2L}$), the total number of PO calls is $T = \frac{8L(\Phi(x_0)-\Phi^*)}{\epsilon^2}$. The number of SFO calls equals to $Sn + Smb = 8L(\Phi(x_0)-\Phi^*)\big(\frac{n}{\epsilon^2 b}+\frac{b}{\epsilon^2}\big)$ if $B = n$ (i.e., the second term in (38) is 0 and thus Assumption 3 is not needed), or equals to $SB + Smb = 8L(\Phi(x_0)-\Phi^*)\big(\frac{B}{\epsilon^2 b}+\frac{b}{\epsilon^2}\big)$ if $B \ge \min\{n, \frac{2\sigma^2}{\epsilon^2}\}$.

In case the number of SFO calls is less than B (i.e., if the total number of epochs $S < 1$), we may add an explicit term B to the SFO result since the algorithm at least uses B SFO calls in the first epoch $s = 0$ at Line 6 of Algorithm 2. In this situation, SSRGD (Algorithm 2) terminates within the first epoch $s = 0$. □

Appendix C. Missing Proofs for Section 6 PL Condition

Now we provide the proofs for ProxSVRG+ (Theorem 7) and SSRGD (Theorem 8) under PL condition.

C.1 Proof for ProxSVRG+ under PL condition

Proof of Theorem 7. First, we recall a key inequality (24) from the proof of Theorem 5, i.e.,

$$\mathbb{E}[\Phi(x_t)] \le \mathbb{E}\Big[\Phi(x_{t-1}) - \frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2 - (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t+1}\|x_t - \widetilde{x}^s\|^2 + \Big((\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t} + \frac{\eta L^2}{2b}\Big)\|x_{t-1} - \widetilde{x}^s\|^2 + \frac{I\{B<n\}\eta\sigma^2}{2B}\Big].$$

Then we plug the PL inequality (9), i.e., $\|\mathcal{G}_\eta(x)\|^2 \ge 2\mu(\Phi(x) - \Phi^*)$ into it to obtain

$$\mathbb{E}[\Phi(x_t) - \Phi^*] \le \mathbb{E}\Big[(1-\mu\eta)(\Phi(x_{t-1}) - \Phi^*) - (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t+1}\|x_t - \widetilde{x}^s\|^2 + \Big((\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t} + \frac{\eta L^2}{2b}\Big)\|x_{t-1} - \widetilde{x}^s\|^2 + \frac{I\{B<n\}\eta\sigma^2}{2B}\Big].$$

Now, we reorder it as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\Phi(x_t) - \Phi^* + (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t+1}\|x_t - \widetilde{x}^s\|^2\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)(\Phi(x_{t-1}) - \Phi^*) + \Big((\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t} + \frac{\eta L^2}{2b}\Big)\|x_{t-1} - \widetilde{x}^s\|^2 + \frac{I\{B<n\}\eta\sigma^2}{2B}\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)\Big((\Phi(x_{t-1}) - \Phi^*) + \frac{(\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t} + \frac{\eta L^2}{2b}}{1-\mu\eta}\|x_{t-1} - \widetilde{x}^s\|^2\Big) + \frac{I\{B<n\}\eta\sigma^2}{2B}\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)\Big((\Phi(x_{t-1}) - \Phi^*) + (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_{t-1}+1}\|x_{t-1} - \widetilde{x}^s\|^2\Big) + \frac{I\{B<n\}\eta\sigma^2}{2B}\Big], \quad (40)
\end{aligned}$$

where (40) holds by choosing α_ts and η to satisfy $(\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_t} + \frac{\eta L^2}{2b} \le (\frac{1}{2\eta} - \frac{L}{2})\frac{1-\mu\eta}{\alpha_{t-1}+1}$. Similar to the proof of Theorem 5, we can choose $\alpha_t = 2(t\%m) - 1$ and $\eta \le \frac{1}{(1+2m/\sqrt{b})L}$.

Telescoping (40) for all iterations $sm+1 \le t \le sm+m$ in epoch s, we have

$$\begin{aligned}
&\mathbb{E}[\Phi(x_{(s+1)m}) - \Phi^*] \\
&\le \mathbb{E}\Big[\Phi(x_{(s+1)m}) - \Phi^* + (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_{(s+1)m}+1}\|x_{(s+1)m} - \widetilde{x}^s\|^2\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)^m\Big((\Phi(x_{sm}) - \Phi^*) + (\frac{1}{2\eta} - \frac{L}{2})\frac{1}{\alpha_{sm}+1}\|x_{sm} - \widetilde{x}^s\|^2\Big)
\end{aligned}$$

$$+\frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{j=0}^{m-1}(1-\mu\eta)^j\Big]$$

$$=\mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm})-\Phi^*)+\frac{I\{B<n\}\eta\sigma^2}{2B}\frac{(1-(1-\mu\eta)^m)}{\mu\eta}\Big], \tag{41}$$

where the last equation (41) holds due to $\widetilde{x}^s = x_{sm}$ (see Line 10 of Algorithm 1).

Similarly, we telescope (41) for all epochs $0 \le s \le S-1$ to finish the proof:

$$\begin{aligned}
&\mathbb{E}[\Phi(x_{Sm})-\Phi^*]\\
&\le \mathbb{E}\Big[(1-\mu\eta)^{Sm}(\Phi(x_0)-\Phi^*)+\frac{I\{B<n\}\eta\sigma^2}{2B}\frac{(1-(1-\mu\eta)^m)}{\mu\eta}\frac{(1-(1-\mu\eta)^{Sm})}{1-(1-\mu\eta)^m}\Big]\\
&\le (1-\mu\eta)^{Sm}(\Phi(x_0)-\Phi^*)+\frac{I\{B<n\}\sigma^2}{2\mu B} \qquad (42)\\
&\le \frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon, \qquad (43)
\end{aligned}$$

where (43) holds by choosing $Sm \ge \frac{1}{\mu\eta}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ and $B \ge \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$.

In the following, for simple presentation, we choose $m=\sqrt{b}$ (then $\eta \le \frac{1}{(1+2m/\sqrt{b})L}=\frac{1}{3L}$). Note that there is no constraint for m and b in our convergence proof. The total number of iterations is

$$T = Sm = \frac{1}{\mu\eta}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon} = \frac{3L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}.$$

The number of PO calls equals to $T = Sm = \frac{3L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$. The proof is finished since the number of SFO calls equals to $Sn + Smb = \big(\frac{n}{\sqrt{b}}+b\big)\frac{3L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ if $B=n$ (i.e., the second term in (42) is 0 and thus Assumption 3 is not needed), or equals to $SB + Smb = \big(\frac{B}{\sqrt{b}}+b\big)\frac{3L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ if $B \ge \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$. □

C.2 Proof for SSRGD under PL condition

Proof of Theorem 8. Similar to the proof of Theorem 7, we first recall a key inequality from the proof of Theorem 6 which combines (29) and (35), i.e.,

$$\mathbb{E}[\Phi(x_t)] \le \mathbb{E}\Big[\Phi(x_{t-1})-\frac{\eta}{2}\|\mathcal{G}_\eta(x_{t-1})\|^2-(\frac{1}{2\eta}-\frac{L}{2})\|x_t-x_{t-1}\|^2 + \frac{\eta L^2}{2b}\sum_{j=sm+1}^{t-1}\|x_j-x_{j-1}\|^2+\frac{I\{B<n\}\eta\sigma^2}{2B}\Big].$$

Then we plug the PL inequality (9), i.e., $\|\mathcal{G}_\eta(x)\|^2 \ge 2\mu(\Phi(x)-\Phi^*)$ into it to obtain

$$\mathbb{E}[\Phi(x_t)-\Phi^*] \le \mathbb{E}\Big[(1-\mu\eta)(\Phi(x_{t-1})-\Phi^*)-(\frac{1}{2\eta}-\frac{L}{2})\|x_t-x_{t-1}\|^2 + \frac{\eta L^2}{2b}\sum_{j=sm+1}^{t-1}\|x_j-x_{j-1}\|^2+\frac{I\{B<n\}\eta\sigma^2}{2B}\Big]. \tag{44}$$

We sum it up ((44) $\times \frac{1}{(1-\mu\eta)^k}$ for iteration $t = sm + k$) for all iterations in epoch s, i.e., $t = sm + k$ where k from 1 to m:

$$\mathbb{E}\Big[\frac{\Phi(x_{(s+1)m}) - \Phi^*}{(1-\mu\eta)^m}\Big] \le \mathbb{E}\Big[\Phi(x_{sm}) - \Phi^* - (\frac{1}{2\eta} - \frac{L}{2})\sum_{k=1}^{m}\frac{1}{(1-\mu\eta)^k}\|x_{sm+k} - x_{sm+k-1}\|^2 + \frac{\eta L^2}{2b}\sum_{k=1}^{m}\Big(\frac{1}{(1-\mu\eta)^k}\sum_{j=1}^{k-1}\|x_{sm+j} - x_{sm+j-1}\|^2\Big) + \frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{k=1}^{m}\frac{1}{(1-\mu\eta)^k}\Big]. \tag{45}$$

Then we deduce it as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\Phi(x_{(s+1)m}) - \Phi^*\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm}) - \Phi^*) - (\frac{1}{2\eta} - \frac{L}{2})\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\|x_{sm+k} - x_{sm+k-1}\|^2 \\
&\quad + \frac{\eta L^2}{2b}\sum_{k=1}^{m}\Big(\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\sum_{j=1}^{k-1}\|x_{sm+j} - x_{sm+j-1}\|^2\Big) \\
&\quad + \frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\Big] \\
&= \mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm}) - \Phi^*) - (\frac{1}{2\eta} - \frac{L}{2})\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\|x_{sm+k} - x_{sm+k-1}\|^2 \\
&\quad + \frac{\eta L^2}{2b}\sum_{k=1}^{m}\Big(\sum_{j=k+1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^j}\Big)\|x_{sm+k} - x_{sm+k-1}\|^2 \\
&\quad + \frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\Big] \\
&\le \mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm}) - \Phi^*) - (\frac{1}{2\eta} - \frac{L}{2})\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\|x_{sm+k} - x_{sm+k-1}\|^2 \\
&\quad + \frac{\eta L^2(m-1)}{2b}\sum_{k=1}^{m}\|x_{sm+k} - x_{sm+k-1}\|^2 + \frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\Big] \qquad (46) \\
&\le \mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm}) - \Phi^*) + \frac{I\{B<n\}\eta\sigma^2}{2B}\sum_{k=1}^{m}\frac{(1-\mu\eta)^m}{(1-\mu\eta)^k}\Big] \qquad (47) \\
&\le \mathbb{E}\Big[(1-\mu\eta)^m(\Phi(x_{sm}) - \Phi^*) + \frac{I\{B<n\}\eta\sigma^2}{2B}\frac{(1-(1-\mu\eta)^m)}{\mu\eta}\Big], \qquad (48)
\end{aligned}$$

where (46) uses the fact $\sum_{i=0}^{m-2}(1-\mu\eta)^i \le \sum_{i=0}^{m-2} 1 = m-1$ (here $\mu\eta \le 1$ due to $\mu \le L$ and $\eta \le \frac{1}{L}$), and (47) holds by choosing appropriate η to cancel the point distance terms $\|x_{sm+k} - x_{sm+k-1}\|^2$. Similar to the proof of Theorem 6, we can choose $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L}$.

Now, we telescope (48) for all epochs $0 \le s \le S-1$ to finish the proof:

$$\begin{aligned}
&\mathbb{E}[\Phi(x_{Sm}) - \Phi^*] \\
&\le \mathbb{E}\Big[(1-\mu\eta)^{Sm}(\Phi(x_0)-\Phi^*) + \frac{I\{B<n\}\eta\sigma^2}{2B}\frac{(1-(1-\mu\eta)^m)}{\mu\eta}\frac{(1-(1-\mu\eta)^{Sm})}{1-(1-\mu\eta)^m}\Big] \\
&\le (1-\mu\eta)^{Sm}(\Phi(x_0)-\Phi^*) + \frac{I\{B<n\}\sigma^2}{2\mu B} \qquad (49)\\
&\le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \qquad (50)
\end{aligned}$$

where (50) holds by choosing $Sm \ge \frac{1}{\mu\eta}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ and $B \ge \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$.

In the following, for simple presentation, we choose $m = b$ (then $\eta \le \frac{1}{(1+\sqrt{(m-1)/b})L} = \frac{1}{2L}$). Note that there is no constraint for m and b in our convergence proof. The total number of iterations is

$$T = Sm = \frac{1}{\mu\eta}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon} = \frac{2L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}.$$

The number of PO calls equals to $T = Sm = \frac{2L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$. The proof is finished since the number of SFO calls equals to $Sn + Smb = \big(\frac{n}{b}+b\big)\frac{2L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ if $B = n$ (i.e., the second term in (49) is 0 and thus Assumption 3 is not needed), or equals to $SB + Smb = \big(\frac{B}{b}+b\big)\frac{2L}{\mu}\log\frac{2(\Phi(x_0)-\Phi^*)}{\epsilon}$ if $B \ge \min\{n, \frac{\sigma^2}{\mu\epsilon}\}$. □

Appendix D. Missing Proofs for Section 7 Local Minima

Now, we provide the detailed proofs for Theorem 9. Note that due to the second-order guarantee, the proofs of the finite-sum case and the online case have more difference than previous first-order guarantee methods (e.g., proof of Theorem 5 and 6). One of the reason is that for the perturbation condition $\|v_{sm}\| \le \epsilon$ in Line 4 of Algorithm 3, $v_{sm} = \nabla f(x_{sm})$ for finite-sum case ($B = n$) while $v_{sm} = \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_{sm})$ for the online case. So we need an extra high probability bound $\|\nabla f(x_{sm})\| \le \epsilon$ in the online case. In the following, we divide the proof of Theorem 9 into two parts, i.e., finite-sum (Section D.2) and online (Section D.3). Before the proof, we recall some standard concentration bounds in Section D.1.

D.1 Tools

Here, we recall some classical concentration bounds for matrices and vectors.

Proposition 12 (Bernstein Inequality (Tropp, 2012)) *Consider a finite sequence $\{Z_k\}$ of independent, random matrices with dimension $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E}[Z_k] = 0 \ \text{ and } \ \|Z_k\| \le R \ \text{ almost surely.}$$

Define

$$\sigma^2 := \max\Big\{\big\|\sum_k \mathbb{E}[Z_k Z_k^T]\big\|, \big\|\sum_k \mathbb{E}[Z_k^T Z_k]\big\|\Big\}.$$

Then, for all $t \geq 0$*,*

$$\mathbb{P}\Big\{\|\sum_k Z_k\| \geq t\Big\} \leq (d_1+d_2)\exp\Big(\frac{-t^2/2}{\sigma^2+Rt/3}\Big).$$

In our proof, we only need its special case vector version as follows, where $z_k = v_k - \mathbb{E}[v_k]$.

Proposition 13 (Bernstein Inequality (Tropp, 2012)) *Consider a finite sequence* $\{v_k\}$ *of independent, random vectors with dimension* d*. Assume that each random matrix satisfies*

$$\|v_k - \mathbb{E}[v_k]\| \leq R \ \ almost\ surely.$$

Define

$$\sigma^2 := \sum_k \mathbb{E}\|v_k - \mathbb{E}[v_k]\|^2.$$

Then, for all $t \geq 0$*,*

$$\mathbb{P}\Big\{\|\sum_k (v_k - \mathbb{E}[v_k])\| \geq t\Big\} \leq (d+1)\exp\Big(\frac{-t^2/2}{\sigma^2+Rt/3}\Big).$$

Moreover, we also need the following martingale concentration bounds, i.e., Azuma-Hoeffding inequality. Now, we only state the vector version (the more general matrix version is not needed).

Proposition 14 (Azuma-Hoeffding Inequality (Hoeffding, 1963; Tropp, 2011)) *Consider a martingale vector sequence* $\{y_k\}$ *with dimension* d*, and let* $\{z_k\}$ *denote the associated martingale difference sequence with respect to a filtration* $\{\mathscr{F}_k\}$*, i.e.,* $z_k := y_k - \mathbb{E}[y_k|\mathscr{F}_{k-1}] = y_k - y_{k-1}$ *and* $\mathbb{E}[z_k|\mathscr{F}_{k-1}] = 0$*. Suppose that* $\{z_k\}$ *satisfies*

$$\|z_k\| = \|y_k - y_{k-1}\| \leq c_k \ \ almost\ surely. \tag{51}$$

Then, for all $t \geq 0$*,*

$$\mathbb{P}\Big\{\|y_k - y_0\| \geq t\Big\} \leq (d+1)\exp\Big(\frac{-t^2}{8\sum_{i=1}^k c_i^2}\Big).$$

However, the assumption that $\|z_k\| \leq c_k$ in (51) with probability 1 is too strict and it may fail sometimes. Fortunately, the Azuma-Hoeffding inequality also holds with a slackness if $\|z_k\| \leq c_k$ with high probability.

Proposition 15 (Azuma-Hoeffding Inequality with High Probability (Chung and Lu, 2006; Tao and Vu, 2015)) *Consider a martingale vector sequence* $\{y_k\}$ *with dimension* d*, and let* $\{z_k\}$ *denote the associated martingale difference sequence with respect to a filtration* $\{\mathscr{F}_k\}$*, i.e.,* $z_k := y_k - \mathbb{E}[y_k|\mathscr{F}_{k-1}] = y_k - y_{k-1}$ *and* $\mathbb{E}[z_k|\mathscr{F}_{k-1}] = 0$*. Suppose that* $\{z_k\}$ *satisfies*

$$\|z_k\| = \|y_k - y_{k-1}\| \leq c_k \ \ with\ high\ probability\ 1-\zeta_k.$$

Then, for all $t \geq 0$*,*

$$\mathbb{P}\Big\{\|y_k - y_0\| \geq t\Big\} \leq (d+1)\exp\Big(\frac{-t^2}{8\sum_{i=1}^k c_i^2}\Big) + \sum_{i=1}^k \zeta_k.$$

D.2 Proof of Theorem 9 (finite-sum)

For proving the second-order guarantee, we divide the proof into two situations. The first situation (**large gradients**) is almost the same as the above arguments for first-order guarantee, where the function value decreases significantly since the gradients are large (see (37)). For the second situation (**around saddle points**), we show that the function value can also decrease a lot by adding a random perturbation. The reason is that saddle points are usually unstable and the stuck region is relatively small in a random perturbation ball.

Large Gradients: First, we need a high probability bound for the variance term instead of the expectation one (35) (note that here $B = n$ in the finite-sum case). Then we use it to get a high probability bound of (37) for the decrease of the function value. Recall that $v_k = \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_k) - \nabla f_i(x_{k-1})\big) + v_{k-1}$ (see Line 13 of Algorithm 3). We let $y_k := v_k - \nabla f(x_k)$ and $z_k := y_k - y_{k-1}$. It is not hard to verify that $\{y_k\}$ is a martingale sequence and $\{z_k\}$ is the associated martingale difference sequence. In order to apply the Azuma-Hoeffding inequalities to get a high probability bound, we first need to bound the martingale difference sequence $\{z_k\}$. We use the Bernstein inequality to bound the differences as follows.

$$\begin{aligned} z_k = y_k - y_{k-1} &= v_k - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_k) - \nabla f_i(x_{k-1})\big) + v_{k-1} - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\Big) \\ &= \frac{1}{b}\sum_{i\in I_b} u_i, \end{aligned} \tag{52}$$

where we define $u_i := \nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))$ in (52). Then we have

$$\|u_i\| = \|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\| \le 2L\|x_k - x_{k-1}\|, \tag{53}$$

where the last inequality holds due to the gradient Lipschitz Assumption 5. Then, consider the variance term

$$\begin{aligned} \mathbb{E}\Big[\sum_{i\in I_b}\|u_i\|^2\Big] &= b\mathbb{E}_i[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\|^2] \\ &\le b\mathbb{E}_i[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] \\ &\le bL^2\|x_k - x_{k-1}\|^2, \end{aligned} \tag{54}$$

where the first inequality uses the fact $\mathbb{E}[\|x - \mathbb{E}x\|^2] \le \mathbb{E}[\|x\|^2]$, and the last inequality uses the gradient Lipschitz Assumption 5. According to (53) and (54), we can bound the difference z_k by Bernstein inequality (Proposition 13) as

$$\begin{aligned} \mathbb{P}\Big\{\big\|z_k\big\| \ge \frac{t}{b}\Big\} &\le (d+1)\exp\Big(\frac{-t^2/2}{\mathbb{E}[\sum_{i\in I_b}\|u_i\|^2] + Rt/3}\Big) \\ &= (d+1)\exp\Big(\frac{-t^2/2}{bL^2\|x_k - x_{k-1}\|^2 + 2L\|x_k - x_{k-1}\|t/3}\Big) \end{aligned}$$

$$= \zeta_k,$$

where the last equality holds by letting $t = CL\sqrt{b}\|x_k - x_{k-1}\|$, where $C = O(\log \frac{d}{\zeta_k})$. Now, we have a high probability bound for the difference sequence $\{z_k\}$, i.e.,

$$\|z_k\| \leq \frac{CL}{\sqrt{b}}\|x_k - x_{k-1}\| \quad \text{with probability } 1 - \zeta_k. \tag{55}$$

Now, we are ready to get a high probability bound for our original variance term (35) by using the martingale Azuma-Hoeffding inequality. Consider in a specific epoch s, i.e, iterations t from $sm + 1$ to current $sm + k$, where k is less than m (note that we only need to consider the current epoch since each epoch we start with $y = 0$). We use a union bound for the difference sequence $\{z_t\}$ by letting $\zeta_k = \zeta'/m$ such that

$$\|z_t\| \leq c_t = \frac{CL}{\sqrt{b}}\|x_t - x_{t-1}\| \text{ for all } t \in [sm+1, sm+k] \text{ with probability } 1 - \zeta'. \tag{56}$$

Define $\beta := \sqrt{8\sum_{t=sm+1}^{sm+k} c_t^2 \log\frac{d}{\zeta'}} = \frac{C'L}{\sqrt{b}}\sqrt{\sum_{t=sm+1}^{sm+k}\|x_t - x_{t-1}\|^2}$, where $C' = O(C\sqrt{\log\frac{d}{\zeta'}}) = O(\log\frac{d}{\zeta_k}\sqrt{\log\frac{d}{\zeta'}}) = O(\log\frac{dm}{\zeta'}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$. According to Azuma-Hoeffding inequality (Proposition 15) and noting that $\zeta_k = \zeta'/m$, we have that

$$\mathbb{P}\Big\{\big\|y_{sm+k} - y_{sm}\big\| \geq \beta\Big\} \leq (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{t=sm+1}^{sm+k} c_t^2}\Big) + \zeta' = 2\zeta'.$$

Recall that $y_k := v_k - \nabla f(x_k)$ and at the beginning point of this epoch $y_{sm} = 0$ due to $v_{sm} = \nabla f(x_{sm})$ since $B = n$ in this finite-sum case (see Line 9 of Algorithm 3). Thus, for any $t \in [sm+1, sm+m]$, we have that

$$\|v_{t-1} - \nabla f(x_{t-1})\| = \|y_{t-1}\| \leq \beta := \frac{C'L}{\sqrt{b}}\sqrt{\sum_{j=sm+1}^{t-1}\|x_j - x_{j-1}\|^2}, \tag{57}$$

holds with probability $1 - 2\zeta'$, where $C' = O(\log\frac{dm}{\zeta'}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$.

Now, we use this high probability version (57) instead of the expectation one (35) to obtain the high probability bound for the decrease of the function value (see (37)). We sum up (29) from the beginning of this epoch s, i.e., iterations from $sm + 1$ to t, by plugging (57) into them to get:

$$\begin{aligned} f(x_t) &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\ &\quad + \frac{\eta}{2}\sum_{k=sm+1}^{t-1}\frac{C'^2L^2\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2}{b} \\ &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \end{aligned} \tag{58}$$

$$
\begin{aligned}
&+ \frac{\eta C'^2 L^2}{2b} \sum_{k=sm+1}^{t-1} \sum_{j=sm+1}^{k} \|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \frac{\eta}{2} \sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big) \sum_{j=sm+1}^{t} \|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2 L^2 (t-1-sm)}{2b} \sum_{j=sm+1}^{t} \|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \frac{\eta}{2} \sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2 L^2}{2}\big) \sum_{j=sm+1}^{t} \|x_j - x_{j-1}\|^2 \quad (59) \\
&\leq f(x_{sm}) - \frac{\eta}{2} \sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2, \quad (60)
\end{aligned}
$$

where (59) holds if the minibatch size $b \geq m$ (note that here $t \leq (s+1)m$), and (60) holds if the step size $\eta \leq \frac{1}{(1+C')L}$, where $C' = O(\log \frac{dm}{\zeta'} \sqrt{\log \frac{d}{\zeta'}})$. Note that (58) uses (57) which holds with probability $1 - 2\zeta'$. Thus by a union bound, we know that (60) holds with probability at least $1 - 2m\zeta'$.

Note that (60) only guarantees that the function value decreases significantly only when the summation of gradients in this epoch is large. However, in order to connect the guarantees between first situation (large gradients) and second situation (around saddle points), we need to show guarantees that are related to the *gradient of the starting point* of each epoch (see Line 4 of Algorithm 3). Similar to Ge et al. (2019), we achieve this by stopping the epoch at a uniformly random point (see Line 17 of Algorithm 3).

Now we prove Lemma 16 to distinguish these two situations (large gradients and around saddle points):

Lemma 16 (Two Situations) *For any epoch s, let x_t be a point uniformly sampled from this epoch $\{x_j\}_{j=sm+1}^{(s+1)m}$. We choose the step size $\eta \leq \frac{1}{(1+C')L}$ (where $C' = O(\log \frac{dm}{\zeta'} \sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$) and the minibatch size $b \geq m$. Then for any $\epsilon > 0$, either of the following two cases happens:*

1. *(Small gradient, possibly around a saddle point) If at least half of points in this epoch have gradient norm no larger than ϵ, then $\|\nabla f(x_t)\| \leq \epsilon$ holds with probability at least $1/2$;*
2. *(Large gradient) Otherwise, we know $f(x_{sm}) - f(x_t) \geq \frac{\eta m \epsilon^2}{8}$ holds with probability at least $1/5$.*

Moreover, $f(x_t) \leq f(x_{sm})$ holds with high probability $1 - 2m\zeta'$ no matter which case happens.

Proof of Lemma 16. There are two cases in this epoch:

1. If at least half of points in this epoch $\{x_j\}_{j=sm+1}^{(s+1)m}$ have gradient norm no larger than ϵ, then it is easy to see that a uniformly sampled point x_t has gradient norm $\|\nabla f(x_t)\| \leq \epsilon$ with probability at least $1/2$.

2. Otherwise, at least half of points have gradient norm larger than ϵ. Then, as long as the sampled point x_t falls into the last quarter of $\{x_j\}_{j=sm+1}^{(s+1)m}$, we know $\sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2 \geq \frac{m\epsilon^2}{4}$. This holds with probability at least $1/4$ since x_t is uniformly sampled. Then combining with (60), i.e., $f(x_{sm}) - f(x_t) \geq \frac{\eta}{2}\sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2$, we can see that the function value decreases $f(x_{sm}) - f(x_t) \geq \frac{\eta m\epsilon^2}{8}$. Note that (60) holds with high probability $1-2m\zeta'$ if we choose the minibatch size $b \geq m$ and the step size $\eta \leq \frac{1}{(1+C')L}$. By a union bound, the function value decrease $f(x_{sm}) - f(x_t) \geq \frac{\eta m\epsilon^2}{8}$ with probability at least $1/5$ (e.g., choose $\zeta' \leq 1/40m$).

Again according to (60), $f(x_t) \leq f(x_{sm})$ holds with high probability $1 - 2m\zeta'$. □

Note that if Case 2 happens, the function value would decrease significantly in this epoch s (corresponding to the first situation large gradients). Otherwise if Case 1 happens, we know the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 20 of Algorithm 3), and we know $\|\nabla f(x_{(s+1)m})\| = \|\nabla f(x_t)\| \leq \epsilon$. In this case, we start a super epoch (corresponding to the second situation around saddle points). Note that if $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) > -\delta$, the point $x_{(s+1)m}$ is already an (ϵ, δ)-local minimum.

Around Saddle Points $\|\nabla f(\widetilde{x})\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$: In this situation, we show that the function value decreases significantly in a *super epoch* with high probability. Recall that we add a random perturbation at the initial point $\widetilde{x}$. To simplify the presentation, we use $x_0 := \widetilde{x} + \xi$ to denote the starting point of the super epoch after the perturbation, where ξ uniformly $\sim \mathbb{B}_0(r)$ and the perturbation radius is r (see Line 7 of Algorithm 3). We follow the *two-point analysis* developed in Jin et al. (2017). The high level idea is as follows: one can divide the perturbation ball $\mathbb{B}_0(r)$ into two disjoint regions: (1) an escaping region which consists of all the points whose function value decreases by at least f_{thres} after t_{thres} steps; (2) the rest which we call the stuck region. The key insight in Jin et al. (2017) is that the stuck region occupies a very small proportion of the volume of perturbation ball. In particular, they show that the stuck region looks like "thin pancake" (see Figure 1 and 2 in Jin et al. (2017)). Let e_1 be the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$ For any two points along e_1 direction that are not very close, one can show at least one of them must not be in the stuck regoin. This implies that the intersection of the line along e_1 direction and the stuck region can be at most an interval of a small length, which indicates that the pancake is thin in the e_1 direction, which can be turned into an upper bound on the volume of the stuck region by standard calculus. Since we use a more involved update rule, our analysis is somewhat more technical.

In particular, we consider two coupled points x_0 and x_0' with $w_0 := x_0 - x_0' = r_0 e_1$, where r_0 is a scalar and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. Then we get two coupled sequences $\{x_t\}$ and $\{x_t'\}$ by running SSRGD update steps (Line 9–13 of Algorithm 3) with the same choice of minibatches (i.e., I_b's in Line 13 of Algorithm 3) for a super epoch. We show that the function value decreases significantly for at least one of these two coupled sequences (escape the saddle point), i.e.,

$$\exists t \leq t_{\text{thres}}, \ \text{such that} \ \ \max\{f(x_0) - f(x_t), f(x_0') - f(x_t')\} \geq 2f_{\text{thres}}. \tag{61}$$

Now, we prove (61) by contradiction. Assume the contrary, $f(x_0) - f(x_t) < 2f_{\text{thres}}$ and $f(x_0') - f(x_t') < 2f_{\text{thres}}$. First, we show that if function value does not decrease a lot, then all

iteration points are not far from the starting point with high probability. Then we show at least one of x_t and x'_t should go far away from their starting point x_0 and x'_0 with high probability, rendering a contradiction. We need the following two technical lemmas and their proofs are deferred to the end of this section.

Lemma 17 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 9–13 of Algorithm 3) from x_0. Moreover, let the step size $\eta \le \frac{1}{(1+2C')L}$ and minibatch size $b \ge m$. With probability $1-2t\zeta'$, we have*

$$\forall t \ge 0, \ \ \|x_t - x_0\| \le \sqrt{\frac{4t(f(x_0)-f(x_t))}{C'L}}, \tag{62}$$

where $C' = O(\log \frac{dm}{\zeta'}\sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$.

Lemma 18 (Small Stuck Region) *If the initial point $\widetilde{x}$ satisfies $-\gamma := \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$, then let $\{x_t\}$ and $\{x'_t\}$ be two coupled sequences by running SSRGD update steps (Line 9–13 of Algorithm 3) with the same choice of minibatches (i.e., I_b's in Line 13 of Algorithm 3) from x_0 and x'_0 with $w_0 := x_0 - x'_0 = r_0 e_1$, where $x_0 \in \mathbb{B}_{\widetilde{x}}(r)$, $x'_0 \in \mathbb{B}_{\widetilde{x}}(r)$, $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\nabla^2 f(\widetilde{x})$. Moreover, let the super epoch length $t_{\rm thres} = \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \le \frac{1}{15(1+\log t_{\rm thres})C'L} = \widetilde{O}(\frac{1}{L})$, minibatch size $b \ge m$ and the perturbation radius $r \le \frac{\delta}{C_1\rho}$. With probability $1-2T\zeta'$, we have*

$$\exists T \le t_{\rm thres}, \ \ \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \ge \frac{\delta}{C_1\rho}, \tag{63}$$

where $C' = O(\log \frac{dm}{\zeta'}\sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$ and $C_1 \ge 1 + 48C'\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) = \widetilde{O}(1)$.

Based on these two lemmas, we are ready to show that (61) holds with high probability. Without loss of generality, we assume $\|x_T - x_0\| \ge \frac{\delta}{C_1\rho}$ in (63) (note that (62) holds for both $\{x_t\}$ and $\{x'_t\}$). Then plugging it into (62), we obtain

$$\begin{aligned}
\sqrt{\frac{4T(f(x_0)-f(x_T))}{C'L}} &\ge \frac{\delta}{C_1\rho} \\
f(x_0) - f(x_T) &\ge \frac{C'L\delta^2}{4C_1^2\rho^2 T} \\
&\ge \frac{C'L\eta\delta^3}{4C_1^2\rho^2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})} \\
&= \frac{\delta^3}{C_1'\rho^2} \\
&\overset{\rm def}{=} 2f_{\rm thres},
\end{aligned} \tag{64}$$

where the last inequality is due to $T \le t_{\rm thres} := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta}$, (64) holds by letting $C_1' = \frac{4C_1^2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{C'L\eta} = \widetilde{O}(1)$, and the last equality is due to the definition of $f_{\rm thres} := \frac{\delta^3}{2C_1'\rho^2} = \widetilde{O}(\frac{\delta^3}{\rho^2})$. Thus, we have already proved that at least one of sequences $\{x_t\}$ and $\{x'_t\}$ escapes the saddle point with probability

$1-4T\zeta'$ (by union bound of (62) and (63)), i.e.,

$$\exists T \le t_{\text{thres}}, \quad \max\{f(x_0)-f(x_T), f(x_0')-f(x_T')\} \ge 2f_{\text{thres}}, \tag{65}$$

if their starting points x_0 and x_0' satisfying $w_0 := x_0 - x_0' = r_0 e_1$. Now, using the same argument as in Jin et al. (2017), we know that in the random perturbation ball, the stuck points can only be a short interval in each line along the e_1 direction, i.e., at least one of two points in the e_1 direction would escape the saddle point if their distance is larger than $r_0 = \frac{\zeta' r}{\sqrt{d}}$. Thus, we know that the probability of the starting point $x_0 = \widetilde{x} + \xi$ (where ξ uniformly $\sim \mathbb{B}_0(r)$) located in the stuck region is less than

$$\frac{r_0 V_{d-1}(r)}{V_d(r)} = \frac{r_0 \Gamma(\frac{d}{2}+1)}{\sqrt{\pi} r \Gamma(\frac{d}{2}+\frac{1}{2})} \le \frac{r_0}{\sqrt{\pi} r}\big(\frac{d}{2}+1\big)^{1/2} \le \frac{r_0\sqrt{d}}{r} = \zeta', \tag{66}$$

where $V_d(r)$ denotes the volume of a Euclidean ball with radius r in d dimension, and the first inequality holds due to Gautschi's inequality. By a union bound for (64) and (66) (x_0 is not in a stuck region), we know that

$$f(x_0) - f(x_T) \ge 2f_{\text{thres}} = \frac{\delta^3}{C_1'\rho^2} \tag{67}$$

holds with probability $1-(4T+1)\zeta'$. Note that the initial point of this super epoch is $\widetilde{x}$ before the perturbation (see Line 7 of Algorithm 3), thus we need to show that the perturbation step $x_0 = \widetilde{x} + \xi$ (where ξ uniformly $\sim \mathbb{B}_0(r)$) does not increase the function value a lot, i.e.,

$$\begin{aligned} f(x_0) &\le f(\widetilde{x}) + \langle \nabla f(\widetilde{x}), x_0 - \widetilde{x}\rangle + \frac{L}{2}\|x_0-\widetilde{x}\|^2 \\ &\le f(\widetilde{x}) + \|\nabla f(\widetilde{x})\|\|x_0-\widetilde{x}\| + \frac{L}{2}\|x_0-\widetilde{x}\|^2 \\ &\le f(\widetilde{x}) + \epsilon\cdot r + \frac{L}{2}r^2 \\ &\le f(\widetilde{x}) + \frac{\delta^3}{2C_1'\rho^2} \\ &= f(\widetilde{x}) + f_{\text{thres}}, \end{aligned} \tag{68}$$

where the last inequality holds by letting the perturbation radius $r \le \min\{\frac{\delta^3}{4C_1'\rho^2\epsilon}, \sqrt{\frac{\delta^3}{2C_1'\rho^2 L}}\}$.

Now we combine with (67) and (68) to obtain that

$$f(\widetilde{x}) - f(x_T) = f(\widetilde{x}) - f(x_0) + f(x_0) - f(x_T) \ge -f_{\text{thres}} + 2f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2} \tag{69}$$

holds with probability at least $1-(4T+1)\zeta' \ge 1 - 5t_{\text{thres}}\zeta'$, where $C_1' = \widetilde{O}(1)$.

Thus we have finished the proof for the second situation (around saddle points), i.e., we show that the function value decreases a lot ($f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}$) in a *super epoch* (recall that $T \le t_{\text{thres}} = \frac{\log(\frac{8\delta\sqrt{d}}{\rho r \zeta'})}{\eta\delta}$).

Combing these two situations (large gradients and around saddle points) to prove Theorem 9: Now, we prove the theorem by distinguishing the epochs into three types as follows:

1. *Type-1 useful epoch*: If at least half of points in this epoch have gradient norm larger than ϵ (Case 2 of Lemma 16);

2. *Wasted epoch*: If at least half of points in this epoch have gradient norm no larger than ϵ and the starting point of the next epoch has gradient norm larger than ϵ (it means that in this epoch one can not guarantee a significant decrease of the function value as in the large gradients situation, and it does not lead to a super epoch (the second situation) since the starting point of the next epoch has gradient norm larger than ϵ);

3. *Type-2 useful super epoch*: If at least half of points in this epoch have gradient norm no larger than ϵ and the starting point of the next epoch (here we denote this point as $\widetilde{x}$) has gradient norm no larger than ϵ (i.e., $\|\nabla f(\widetilde{x})\| \leq \epsilon$) (Case 1 of Lemma 16), according to Line 4 of Algorithm 3, we start a super epoch. So here we denote this epoch along with its following super epoch as a type-2 useful super epoch.

First, it is easy to see that the probability of a wasted epoch happened is less than $1/2$ due to the random stop (see Case 1 of Lemma 16). Note for different wasted epochs, returned points are independently sampled. Thus, with high probability $1-\zeta'$, there are at most $\log \frac{1}{\zeta'} = \widetilde{O}(1)$ wasted epochs happened before a type-1 useful epoch or type-2 useful super epoch. Now, we use N_1 and N_2 to denote the number of type-1 useful epochs and type-2 useful super epochs that the algorithm is needed. Also recall that the function value always does not increase with high probability $1-2m\zeta'$ for any epoch (see Lemma 16).

For type-1 useful epoch, according to Case 2 of Lemma 16, we know that the function value decreases at least $\frac{\eta m \epsilon^2}{8}$ with probability at least $1/5$. Using a union bound, we know that with probability $1-4N_1/5$, N_1 type-1 useful epochs will decrease the function value at least $\frac{\eta m \epsilon^2 N_1}{40}$. Note that the function value can decrease at most $\Delta_0 := f(x_0) - f^*$ and also recall that the function value always does not increase with high probability $1-2m\zeta'$ for any epoch (see Lemma 16). So let $\frac{\eta m \epsilon^2 N_1}{40} \leq \Delta_0$, we get $N_1 \leq \frac{40\Delta_0}{\eta m \epsilon^2}$ with probability at least $1 - \widetilde{O}(N_1 m \zeta')$ by a union bound. We can let $\zeta' \leq \widetilde{O}(1/N_1 m)$.

For type-2 useful super epoch, first we know that the starting point of the super epoch $\widetilde{x}$ has gradient norm $\|\nabla f(\widetilde{x})\| \leq \epsilon$. Now if $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \geq -\delta$, then $\widetilde{x}$ is already a (ϵ, δ)-local minimum. Otherwise, $\|\nabla f(\widetilde{x})\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$, this is exactly our second situation (around saddle points). According to (69), we know that the the function value decrease ($f(\widetilde{x}) - f(x_T)$) is at least $f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}$ with probability at least $1 - 5t_{\text{thres}}\zeta' \geq 1/2$ (let $\zeta' \leq 1/10t_{\text{thres}}$), where $C_1' = \widetilde{O}(1)$. Similar to type-1 useful epoch, we know $N_2 \leq \frac{4C_1'\rho^2\Delta_0}{\delta^3}$ with probability at least $1 - \widetilde{O}(N_2 t_{\text{thres}}\zeta')$ by a union bound. We can let $\zeta' \leq \widetilde{O}(1/N_2 t_{\text{thres}})$.

Now, we are ready to bound the number of SFO calls in Theorem 9 (finite-sum) as follows:

$$
\begin{aligned}
&N_1(\widetilde{O}(1)n + n + mb) + N_2\Big(\widetilde{O}(1)n + \left\lceil \frac{t_{\text{thres}}}{m} \right\rceil n + t_{\text{thres}} b\Big) \\
&\leq \widetilde{O}\Big(\frac{\Delta_0 n}{\eta m \epsilon^2} + \frac{\rho^2 \Delta_0}{\delta^3}(n + \frac{\sqrt{n}}{\eta\delta})\Big) \\
&\leq \widetilde{O}\Big(\frac{L\Delta_0\sqrt{n}}{\epsilon^2} + \frac{L\rho^2\Delta_0\sqrt{n}}{\delta^4} + \frac{\rho^2\Delta_0 n}{\delta^3}\Big). \qquad (70)
\end{aligned}
$$

By a union bound of these types and set $\zeta = \widetilde{O}(N_1 m + N_2 t_{\rm thres})\zeta'$ (note that ζ' only appears in the log term $\log(\frac{1}{\zeta'})$, so it can be chosen as small as we want), we know that the SFO calls of SSRGD can be bounded by (70) with probability $1-\zeta$. This finishes the proof of Theorem 9. Now, it remains to prove Lemma 17 and 18.

Proof of Lemma 17. First, we know the variance bound (57) holds with probability $1-2\zeta'$. Then by a union bound, it holds with probability $1-2t\zeta'$ for all $0 \le j \le t-1$. Then, according to (59), we know for any $\tau \le t$ in some epoch s

$$\begin{aligned} f(x_\tau) &\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{\tau}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2 L^2}{2}\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 \\ &\le f(x_{sm}) - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2 L^2}{2}\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 \\ &\le f(x_{sm}) - \frac{C'L}{4}\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2, \end{aligned} \tag{71}$$

where (71) holds by setting the step size $\eta \le \frac{1}{(1+2C')L}$. Recall that $C' = O(\log\frac{dm}{\zeta'}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$. Now, we sum up (71) for all epochs before iteration t,

$$f(x_t) \le f(x_0) - \frac{C'L}{4}\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2.$$

Then, the proof is finished as

$$\|x_t - x_0\| \le \sum_{j=1}^{t}\|x_j - x_{j-1}\| \le \sqrt{t\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2} \le \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L}}.$$

□

Proof of Lemma 18. We prove this lemma by contradiction. Assume the contrary,

$$\forall t \le T,\ \ \|x_t - x_0\| \le \frac{\delta}{C_1\rho}\ \text{ and }\ \|x'_t - x'_0\| \le \frac{\delta}{C_1\rho}, \tag{72}$$

where $T := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\gamma} \le t_{\rm thres} := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta}$ (note that $\gamma \ge \delta$ due to $-\gamma := \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$). We show that the distance between these two coupled sequences $w_t := x_t - x'_t$ grows exponentially if they are not very close in the e_1 direction at the beginning, i.e., $w_0 := x_0 - x'_0 = r_0 e_1$. Recall that $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. However, $\|w_t\| = \|x_t - x'_t\| \le \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x'_t - x'_0\| + \|x'_0 - \widetilde{x}\| \le 2r + 2\frac{\delta}{C_1\rho}$ according to (72) and the perturbation radius r. It is not hard to see that if $\|w_t\|$ increases exponentially, this inequality cannot be true for reasonably large t, rendering a contradiction.

In the following, we prove that $\|w_t\|$ increases exponentially by induction on t. First, we need the expression of w_t. Define $\Delta_\tau := \int_0^1(\nabla^2 f(x'_\tau + \theta(x_\tau - x'_\tau)) - \mathcal{H})d\theta$ and $y_\tau := v_\tau - \nabla f(x_\tau) -$

$v'_\tau + \nabla f(x'_\tau)$. Recall that $x_t = x_{t-1} - \eta v_{t-1}$ (see Line 12 of Algorithm 3). Hence one can easily see that

$$\begin{aligned}
w_t &= w_{t-1} - \eta(v_{t-1} - v'_{t-1}) \\
&= w_{t-1} - \eta\big(\nabla f(x_{t-1}) - \nabla f(x'_{t-1}) + v_{t-1} - \nabla f(x_{t-1}) - v'_{t-1} + \nabla f(x'_{t-1})\big) \\
&= w_{t-1} - \eta\Big(\int_0^1 \nabla^2 f(x'_{t-1} + \theta(x_{t-1} - x'_{t-1}))d\theta(x_{t-1} - x'_{t-1}) + y_{t-1}\Big) \\
&= w_{t-1} - \eta\big((\mathcal{H} + \Delta_{t-1})w_{t-1} + y_{t-1}\big) \\
&= (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + y_{t-1}) \qquad (73) \\
&= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau + y_\tau). \qquad (74)
\end{aligned}$$

First, one can see that the first term of (74) is in the e_1 direction and it increases exponentially with respect to t, i.e., $(1+\eta\gamma)^t r_0 e_1$, where $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$. Hence, to prove that $\|w_t\|$ increases exponentially, it suffices to show that the norm of the first term of (74) dominate that of the second term. For this purpose, we need the following bounds for $\|w_t\|$ and $\|y_t\|$, stated in the following lemma.

Lemma 19 *Suppose $w_0 := x_0 - x'_0 = r_0 e_1$ where $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 is the eigenvector corresponding to the smallest eigenvalue of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. If (72) holds, then with probability $1 - 2T\zeta'$, the following bounds hold for all $t \le T$:*

1. $\frac{1}{2}(1+\eta\gamma)^t r_0 \le \|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0$;
2. $\|y_t\| \le \frac{\gamma}{4C_2}(1+\eta\gamma)^t r_0$.

where $C_2 := \log(\frac{8\delta\sqrt{d}}{\rho r \zeta'})$.

Proof of Lemma 19. We prove this lemma inductively. First, check the base case $t = 0$, $\|w_0\| = \|r_0 e_1\| = r_0$ and $\|y_0\| = \|v_0 - \nabla f(x_0) - v'_0 + \nabla f(x'_0)\| = \|\nabla f(x_0) - \nabla f(x_0) - \nabla f(x'_0) + \nabla f(x'_0)\| = 0$. Now, assuming they hold for all $\tau \le t-1$, we now prove they hold for t. For the bounds of $\|w_t\|$, it suffices to show that the second term of (74) is dominated by half of the first term. Now, we first consider the first part of the second term:

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau)\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|\Delta_\tau\|\|w_\tau\| \\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\|\Delta_\tau\| \qquad (75) \\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\rho D_\tau^x \qquad (76) \\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 t\rho\big(\frac{\delta}{C_1\rho} + r\big) \qquad (77)
\end{aligned}$$

$$\leq \frac{3}{C_1}\eta\delta t(1+\eta\gamma)^{t-1}r_0 \tag{78}$$

$$\leq \frac{3\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{C_1}(1+\eta\gamma)^{t-1}r_0 \tag{79}$$

$$\leq \frac{1}{4}(1+\eta\gamma)^t r_0, \tag{80}$$

where (75) uses the induction hypothesis for w_τ with $\tau \leq t-1$, (76) uses Assumption 5 and the definition $D_\tau^x := \max\{\|x_\tau - \widetilde{x}\|, \|x'_\tau - \widetilde{x}\|\}$, (77) follows from $\|x_t - \widetilde{x}\| \leq \|x_t - x_0\| + \|x_0 - \widetilde{x}\| = \frac{\delta}{C_1\rho} + r$ due to (72) and the perturbation radius r, (78) holds by letting the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, (79) holds since $t \leq T \leq t_{\mathrm{thres}} := \frac{1}{\eta\delta}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$, and (80) holds due to the definition of $C_1 \geq 12\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$.

Now, the second part can be bounded as follows:

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}y_\tau\| &\leq \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|y_\tau\| \\
&\leq \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\frac{\gamma}{4C_2}(1+\eta\gamma)^{\tau}r_0 && (81)\\
&= \frac{\eta\gamma}{4C_2}t(1+\eta\gamma)^{t-1}r_0 \\
&\leq \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{4C_2}(1+\eta\gamma)^{t-1}r_0 && (82)\\
&= \frac{1}{4}(1+\eta\gamma)^t r_0, && (83)
\end{aligned}$$

where (81) uses the induction for y_τ with $\tau \leq t-1$, (82) holds since $t \leq T := \frac{2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\gamma}$, and (83) holds due to the definition of $C_2 = \log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$.

Combining (80) and (83), we can see that the norm of the second term of (74) is at most one half of that of the first term. Note that the norm of the first term of (74) is $\|(I-\eta\mathcal{H})^t w_0\| = (1+\eta\gamma)^t r_0$. Thus, we have

$$\frac{1}{2}(1+\eta\gamma)^t r_0 \leq \|w_t\| \leq \frac{3}{2}(1+\eta\gamma)^t r_0. \tag{84}$$

Now, the remaining thing is to prove the second bound $\|y_t\| \leq \frac{\gamma}{4C_2}(1+\eta\gamma)^t r_0$, which is somewhat technical. First, we write the concrete expression of y_t:

$$\begin{aligned}
y_t &= v_t - \nabla f(x_t) - v'_t + \nabla f(x'_t) \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1} - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x'_t) - \nabla f_i(x'_{t-1})\big) - v'_{t-1} + \nabla f(x'_t) && (85)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + \nabla f(x_{t-1}) - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x'_t) - \nabla f_i(x'_{t-1})\big) - \nabla f(x'_{t-1}) + \nabla f(x'_t) + y_{t-1} \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x'_t) - \nabla f_i(x_{t-1}) + \nabla f_i(x'_{t-1})\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x'_t) - \nabla f(x_{t-1}) + \nabla f(x'_{t-1})\big) + y_{t-1},
\end{aligned}
$$

where (85) is due to the definition of the estimator v_t (see Line 13 of Algorithm 3). We further define the difference $z_t := y_t - y_{t-1}$. It is not hard to verify that $\{y_t\}$ is a martingale sequence and $\{z_t\}$ is the associated martingale difference sequence. We can apply the Azuma-Hoeffding inequality to get an upper bound for $\|y_t\|$ and then we prove $\|y_t\| \le \frac{\gamma}{4C_2}(1+\eta\gamma)^t r_0$ based on that upper bound. In order to apply the Azuma-Hoeffding inequality for martingale sequence $\|y_t\|$, we first need to bound the difference sequence $\{z_t\}$.

$$
\begin{aligned}
z_t = y_t - y_{t-1} &= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x'_t) - \nabla f_i(x_{t-1}) + \nabla f_i(x'_{t-1})\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x'_t) - \nabla f(x_{t-1}) + \nabla f(x'_{t-1})\big) \\
&= \frac{1}{b}\sum_{i\in I_b} u_i, \qquad (86)
\end{aligned}
$$

where we define $u_i := \big(\nabla f_i(x_t) - \nabla f_i(x'_t)\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x'_{t-1})\big) - \big(\nabla f(x_t) - \nabla f(x'_t)\big) + \big(\nabla f(x_{t-1}) - \nabla f(x'_{t-1})\big)$ in the last equality (86). Then we have

$$
\begin{aligned}
\|u_i\| &\le \Big\| \int_0^1 \nabla^2 f_i(x'_t + \theta(x_t - x'_t))d\theta(x_t - x'_t) - \int_0^1 \nabla^2 f_i(x'_{t-1} + \theta(x_{t-1} - x'_{t-1}))d\theta(x_{t-1} - x'_{t-1}) \\
&\quad - \int_0^1 \nabla^2 f(x'_t + \theta(x_t - x'_t))d\theta(x_t - x'_t) + \int_0^1 \nabla^2 f(x'_{t-1} + \theta(x_{t-1} - x'_{t-1}))d\theta(x_{t-1} - x'_{t-1})\Big\| \\
&= \|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1}) - (\mathcal{H} w_t + \Delta_t w_t) + (\mathcal{H} w_{t-1} + \Delta_{t-1} w_{t-1})\| \\
&\le \|(\mathcal{H}_i - \mathcal{H})(w_t - w_{t-1})\| + \|(\Delta_t^i - \Delta_t)w_t - (\Delta_{t-1}^i - \Delta_{t-1})w_{t-1}\| \\
&\le 2L\|w_t - w_{t-1}\| + 2\rho D_t^x \|w_t\| + 2\rho D_{t-1}^x \|w_{t-1}\|, \qquad (87)
\end{aligned}
$$

where the equality holds since we define $\Delta_t := \int_0^1 (\nabla^2 f(x'_t + \theta(x_t - x'_t)) - \mathcal{H})d\theta$ and $\Delta_t^i := \int_0^1 (\nabla^2 f_i(x'_t + \theta(x_t - x'_t)) - \mathcal{H}_i)d\theta$, and the last inequality holds due to the gradient and Hessian Lipschitz Assumption 5 (recall $D_t^x := \max\{\|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|\}$). Then, consider the variance term:

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{i\in I_b}\|u_i\|^2\Big] &\le b\mathbb{E}_i[\|\big(\nabla f_i(x_t) - \nabla f_i(x'_t)\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x'_{t-1})\big)\|^2] \\
&= b\mathbb{E}_i[\|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1})\|^2] \\
&\le b(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2, \qquad (88)
\end{aligned}
$$

where the first inequality uses the fact $\mathbb{E}[\|x - \mathbb{E}x\|^2] \leq \mathbb{E}[\|x\|^2]$, and the last inequality uses the gradient and Hessian Lipschitz Assumption 5. According to (87) and (88), we can bound the difference z_k by Bernstein inequality (Proposition 13) as (where $R = 2L\|w_t - w_{t-1}\| + 2\rho D_t^x\|w_t\| + 2\rho D_{t-1}^x\|w_{t-1}\|$ and $\sigma^2 = b(L\|w_t - w_{t-1}\| + \rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\|)^2$)

$$\mathbb{P}\Big\{\big\|z_t\big\| \geq \frac{\alpha}{b}\Big\} \leq (d+1)\exp\Big(\frac{-\alpha^2/2}{\sigma^2 + R\alpha/3}\Big) = \zeta_k,$$

where the last equality holds by letting $\alpha = C\sqrt{b}(L\|w_t - w_{t-1}\| + \rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\|)$, where $C = O(\log\frac{d}{\zeta_k})$.

Now, we have a high probability bound for the difference sequence $\{z_k\}$, i.e.,

$$\|z_k\| \leq c_k = \frac{C}{\sqrt{b}}(L\|w_t - w_{t-1}\| + \rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\|) \text{ with probability } 1 - \zeta_k.$$

Next, we provide an upper bound for $\|y_t\|$ by using the martingale Azuma-Hoeffding inequality. Note that we only need to consider the current epoch that contains the iteration t since each epoch we start with $y = 0$. Let s denote the current epoch, i.e, iterations from $sm+1$ to current t, where t is no larger than $(s+1)m$. Define

$$\beta := \sqrt{8\sum_{k=sm+1}^{t} c_k^2 \log\frac{d}{\zeta'}} = \frac{C'}{\sqrt{b}}\sqrt{\sum_{k=sm+1}^{t}(L\|w_t - w_{t-1}\| + \rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\|)^2},$$

where $C' = O(C\sqrt{\log\frac{d}{\zeta'}}) = O(\log\frac{d}{\zeta_k}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$. According to Azuma-Hoeffding inequality (Proposition 15) and letting $\zeta_k = \zeta'/m$, we have

$$\mathbb{P}\Big\{\big\|y_t - y_{sm}\big\| \geq \beta\Big\} \leq (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{k=sm+1}^{t} c_k^2}\Big) + \zeta' = 2\zeta'.$$

Recall that $y_k := v_k - \nabla f(x_k) - v_k' + \nabla f(x_k')$ and at the beginning point of this epoch $y_{sm} = 0$ due to $v_{sm} = \nabla f(x_{sm})$ and $v_{sm}' = \nabla f(x_{sm}')$ (note that batch size $B = n$ in this finite-sum case). Thus. for any $t \in [sm+1, (s+1)m]$, we have

$$\|y_t\| = \|y_t - y_{sm}\| \leq \beta := \frac{C'}{\sqrt{b}}\sqrt{\sum_{k=sm+1}^{t}(L\|w_t - w_{t-1}\| + \rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\|)^2} \quad (89)$$

holds with high probability $1 - 2\zeta'$. Furthermore, by a union bound, we know that (89) holds with probability at least $1 - 2T\zeta'$ for all $t \leq T$.

Now, we show how to bound the right-hand-side of (89). First, we show that the last two terms in the right-hand-side of (89) can be bounded as

$$\begin{aligned}\rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\| &\leq \rho\big(\frac{\delta}{C_1\rho} + r\big)\frac{3}{2}(1+\eta\gamma)^t r_0 + \rho\big(\frac{\delta}{C_1\rho} + r\big)\frac{3}{2}(1+\eta\gamma)^{t-1} r_0 \\ &\leq 3\rho\big(\frac{\delta}{C_1\rho} + r\big)(1+\eta\gamma)^t r_0\end{aligned}$$

$$\leq \frac{6\delta}{C_1}(1+\eta\gamma)^t r_0, \tag{90}$$

where the first inequality follows from the induction hypothesis of $\|w_{t-1}\| \leq \frac{3}{2}(1+\eta\gamma)^{t-1}r_0$ and the bound $\|w_t\| \leq \frac{3}{2}(1+\eta\gamma)^t r_0$ in (84) which we have already proved, and the last inequality holds by letting the perturbation radius $r \leq \frac{\delta}{C_1\rho}$.

Now, we bound the first term of right-hand-side of (89). According to (73), we have

$$\begin{aligned}
L\|w_t - w_{t-1}\| &= L\|\eta\mathcal{H}w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + y_{t-1})\| \\
&\leq L\eta\|\mathcal{H}w_{t-1}\| + L\eta\|\Delta_{t-1}w_{t-1} + y_{t-1}\| \\
&\leq L\eta\|\text{Proj}_{S_-}\mathcal{H}w_{t-1}\| + L\eta\|\text{Proj}_{S_+}\mathcal{H}w_{t-1}\| + L\eta\|\Delta_{t-1}w_{t-1} + y_{t-1}\| \qquad (91) \\
&\leq L\eta\gamma\|w_{t-1}\| + L\eta\|\text{Proj}_{S_+}\mathcal{H}w_{t-1}\| + L\eta\|\Delta_{t-1}\|\|w_{t-1}\| + L\eta\|y_{t-1}\| \\
&\leq (1+\frac{2}{C_1\rho})L\eta\gamma\|w_{t-1}\| + L\eta\|\text{Proj}_{S_+}\mathcal{H}w_{t-1}\| + L\eta\|y_{t-1}\| \qquad (92) \\
&\leq (\frac{3}{2} + \frac{3}{C_1} + \frac{1}{4C_2})L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\|\text{Proj}_{S_+}\mathcal{H}w_{t-1}\|, \qquad (93)
\end{aligned}$$

where (91) holds by splitting the space into two subspace: 1) subspace S_- spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $[-\gamma, 0]$; 2) subspace S_+ spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $(0, L]$, (92) holds from the following (94), and the last inequality (93) follows from the induction hypothesis of $\|w_{t-1}\| \leq \frac{3}{2}(1+\eta\gamma)^{t-1}r_0$ and $\|y_{t-1}\| \leq \frac{\gamma}{4C_2}(1+\eta\gamma)^{t-1}r_0$.

$$\forall t \leq T, \quad \|\Delta_t\| \leq \rho D_t^x \leq \rho\big(\frac{\delta}{C_1\rho} + r\big) \leq \frac{2\delta}{C_1} \leq \frac{2\gamma}{C_1}, \tag{94}$$

which holds by letting the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, and noting that $\gamma \geq \delta$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$).

Now, we bound the second term of (93) as follows:

$$\begin{aligned}
&L\eta\|\text{Proj}_{S_+}\mathcal{H}w_{t-1}\| \\
&= L\eta\big\| - \text{Proj}_{S_+}\mathcal{H}(I-\eta\mathcal{H})^{t-1}w_0 - \sum_{\tau=0}^{t-2}\text{Proj}_{S_+}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau)\big\| \qquad (95) \\
&= L\eta\big\| - \sum_{\tau=0}^{t-2}\text{Proj}_{S_+}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau)\big\| \qquad (96) \\
&\leq L\eta\sum_{\tau=0}^{t-2}\big\|\text{Proj}_{S_+}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\big\|\|\Delta_\tau w_\tau + y_\tau\| \\
&\leq L\eta\sum_{\tau=0}^{t-2}\frac{1}{t-1-\tau}\|\Delta_\tau w_\tau + y_\tau\| \qquad (97) \\
&\leq L\eta\log t\max_{0\leq k\leq t-2}\|\Delta_k w_k + y_k\| \\
&\leq L\eta\log t\Big(\frac{2\gamma}{C_1}\frac{3}{2}(1+\eta\gamma)^{t-2}r_0 + \frac{\gamma}{4C_2}(1+\eta\gamma)^{t-2}r_0\Big) \qquad (98) \\
&= \Big(\frac{3}{C_1}\log t + \frac{1}{4C_2}\log t\Big)L\eta\gamma(1+\eta\gamma)^{t-2}r_0, \qquad (99)
\end{aligned}$$

where the first equality (95) follows from (74), (96) holds since $w_0 = r_0 e_1$ is in the e_1 direction, (98) uses (94) and the induction hypothesis of $\|w_k\| \le \frac{3}{2}(1+\eta\gamma)^k r_0$ and $\|y_k\| \le \frac{\gamma}{4C_2}(1+\eta\gamma)^k r_0$, for all $k \le t-1$. The inequality (97) follows from the fact $\max_{x\in[0,1]} x(1-x)^t \le \frac{1}{t+1}$. Note that S_+ denotes the subspace spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $(0, L]$, thus $\left\|\mathrm{Proj}_{S_+}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\right\| \le \max_{\lambda\in(0,L]}\eta\lambda(1-\eta\lambda)^{t-2-\tau} \le \frac{1}{t-1-\tau}$. Also note that $\eta\lambda \le 1$ due to $\eta \le \frac{1}{L}$ and $\lambda \in (0, L]$.

By plugging (99) into (93), we have

$$L\|w_t - w_{t-1}\| \le \Big(\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2}\Big) L\eta\gamma(1+\eta\gamma)^{t-1} r_0. \tag{100}$$

Now we can bound $\|y_t\|$ by plugging (90) and (100) into (89) and noting that $t - sm \le m \le b$:

$$\begin{aligned}
\|y_t\| &\le C'\left(\frac{6\delta}{C_1}(1+\eta\gamma)^t r_0 + \Big(\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2}\Big) L\eta\gamma(1+\eta\gamma)^{t-1} r_0\right) \\
&\le \Big(\frac{6C'}{C_1} + (\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2})C'L\eta\Big)\gamma(1+\eta\gamma)^t r_0 \\
&\le \Big(\frac{1}{8C_2} + \frac{1}{8C_2}\Big)\gamma(1+\eta\gamma)^t r_0 \\
&= \frac{1}{4C_2}\gamma(1+\eta\gamma)^t r_0,
\end{aligned} \tag{101}$$

where the second inequality holds due to $\delta \le \gamma$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$), and the last inequality holds by letting $C_1 \ge 48C'C_2$ (recall that $C_2 := \log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$ defined in Lemma 19), and $\eta \le \frac{1}{15(1+\log t)C'L}$. Recall that (89) holds with probability at least $1 - 2T\zeta'$ for all $t \le T$. This finishes the proof of Lemma 19. □

From the Lemma 19, one can see that $\|w_t\| \ge \frac{1}{2}(1+\eta\gamma)^t r_0 = \frac{1}{2}(1+\eta\gamma)^t \frac{\zeta' r}{\sqrt{d}}$. On the other hand, $\|w_t\| := \|x_t - x_t'\| \le \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x_t' - x_0'\| + \|x_0' - \widetilde{x}\| \le 2r + 2\frac{\delta}{C_1\rho} \le \frac{4\delta}{C_1\rho}$ according to (72) and the perturbation radius $r \le \frac{\delta}{C_1\rho}$. Hence, for any $t \ge T = \frac{1}{\eta\gamma}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$, we get a contradiction to (72), i.e., $\|w_t\| \ge \frac{1}{2}(1+\eta\gamma)^T \frac{\zeta' r}{\sqrt{d}} \ge \frac{4\delta}{\rho} \ge \frac{4\delta}{C_1\rho}$, where $C_1 \ge 1 + 48C'\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) \ge 1$ defined in Lemma 18. Also note that $T = \frac{1}{\eta\gamma}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) \le t_{\mathrm{thres}} := \frac{1}{\eta\delta}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$ due to $\delta \le \gamma$. This contradiction finishes the proof of Lemma 18.

□

D.3 Proof of Theorem 9 (online)

The proof for the online case follows almost the same framework as in the finite-sum case in Section D.2. Although the only difference in the algorithm is that here we compute a large batch of stochastic gradient ($v_{sm} \ne \nabla f(x_{sm})$ in Line 4 and 9 of Algorithm 3), instead of a full gradient, it leads to many changes in the analysis. Hence, we present the full proof for the online case as well. Again, we distinguish two situations, the *large gradients* case, in which the function value decreases significantly, and the *around saddle points* case, in which we add a random perturbation.

Large Gradients: First, we provide a high probability bound for the variance term, and then use it to get a high probability bound for the decrease of the function. Note that in this online case, $v_{sm} = \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_{sm})$ at the beginning of each epoch instead of $v_{sm} = \nabla f(x_{sm})$

(where $B = n$) in the previous finite-sum case. Thus we first need a high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$. According to Assumption 6, we have

$$\|\nabla f_j(x) - \nabla f(x)\| \le \sigma,$$
$$\sum_{j\in I_B} \|\nabla f_j(x) - \nabla f(x)\|^2 \le B\sigma^2.$$

By applying Bernstein inequality (Proposition 13), we get the high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$ as follows:

$$\mathbb{P}\Big\{\big\|v_{sm} - \nabla f(x_{sm})\big\| \ge \frac{t}{B}\Big\} \le (d+1)\exp\Big(\frac{-t^2/2}{B\sigma^2 + \sigma t/3}\Big) = \zeta',$$

where the last equality holds by letting $t = C_3\sqrt{B}\sigma$, where $C_3 = O(\log\frac{d}{\zeta'}) = \widetilde{O}(1)$. Now, we have a high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$, i.e.,

$$\big\|v_{sm} - \nabla f(x_{sm})\big\| \le \frac{C_3\sigma}{\sqrt{B}} \quad \text{with probability } 1-\zeta'. \tag{102}$$

Now we obtain a high probability bound for the variance term of other points beyond the starting points. Recall that $v_k = \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_k) - \nabla f_i(x_{k-1})\big) + v_{k-1}$ (see Line 13 of Algorithm 3), and the martingale sequence $y_k := v_k - \nabla f(x_k)$, $z_k := y_k - y_{k-1}$, which is the associated martingale difference sequence, and $u_i := \nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))$. By (52), we know that

$$z_k = y_k - y_{k-1} = \frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\Big) = \frac{1}{b}\sum_{i\in I_b} u_i. \tag{103}$$

Using the same argument as in Section D.2 (See (53),(54),(55)), one can see that $\|u_i\| \le 2L\|x_k - x_{k-1}\|$ and $\mathbb{E}[\sum_{i\in I_b}\|u_i\|^2] \le bL^2\|x_k - x_{k-1}\|^2$, and then one can apply Bernstein inequality (Proposition 13) to see that

$$\|z_k\| \le c_k = \frac{CL}{\sqrt{b}}\|x_k - x_{k-1}\| \quad \text{with probability } 1-\zeta_k, \tag{104}$$

where $C = O(\log\frac{d}{\zeta_k}) = \widetilde{O}(1)$.

Now, we are ready to get a high probability bound for the variance term using the martingale Azuma-Hoeffding inequality. Consider in a specific epoch s, i.e, iterations t from $sm+1$ to current $sm+k$, where k is less than m. Let $\beta := \sqrt{8\sum_{t=sm+1}^{sm+k} c_t^2 \log\frac{d}{\zeta'}} = \frac{C'L}{\sqrt{b}}\sqrt{\sum_{t=sm+1}^{sm+k}\|x_t - x_{t-1}\|^2}$, where $C' = O(C\sqrt{\log\frac{d}{\zeta'}}) = O(\log\frac{d}{\zeta_k}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$. According to Azuma-Hoeffding inequality (Proposition 15) and letting $\zeta_k = \zeta'/m$, we have

$$\mathbb{P}\Big\{\big\|y_{sm+k} - y_{sm}\big\| \ge \beta\Big\} \le (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{t=sm+1}^{sm+k} c_t^2}\Big) + \zeta' = 2\zeta'.$$

Recall that $y_k := v_k - \nabla f(x_k)$ and at the beginning point of this epoch $\|y_{sm}\| = \|v_{sm} - \nabla f(x_{sm})\| \le C_3\sigma/\sqrt{B}$ with probability $1-\zeta'$, where $C = O(\log \frac{d}{\zeta'}) = \widetilde{O}(1)$ (see (102)). Combining with (102) and using a union bound, for any $t \in [sm+1, (s+1)m]$, we have that

$$\|v_{t-1} - \nabla f(x_{t-1})\| = \|y_{t-1}\| \le \beta + \|y_{sm}\| \le \frac{C'L\sqrt{\sum_{j=sm+1}^{t-1}\|x_j - x_{j-1}\|^2}}{\sqrt{b}} + \frac{C_3\sigma}{\sqrt{B}} \quad (105)$$

holds with probability $1-3\zeta'$.

Now, we use it to obtain a high probability bound for the decrease of the function value. We sum up (29) from the beginning of this epoch s, i.e., iterations from $sm+1$ to t, by plugging (105) into them to get:

$$\begin{aligned}
f(x_t) &\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta}{2}\sum_{k=sm+1}^{t-1}\frac{2C'^2L^2\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2}{b} + \frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{2C_3^2\sigma^2}{B} \qquad (106)\\
&\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2}{b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2 + \frac{(t-sm)\eta C_3^2\sigma^2}{B} \\
&\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2(t-1-sm)}{b}\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 + \frac{(t-sm)\eta C_3^2\sigma^2}{B} \\
&\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2L^2\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{(t-sm)\eta C^2\sigma^2}{B} \qquad (107)\\
&\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 + \frac{(t-sm)\eta C_3^2\sigma^2}{B}, \qquad (108)
\end{aligned}$$

where (107) holds if the minibatch size $b \ge m$ (note that here $t \le (s+1)m$), and (108) holds if the step size $\eta \le \frac{1}{(1+2C')L}$, where $C' = O(\log\frac{dm}{\zeta'}\sqrt{\log\frac{d}{\zeta'}})$. Note that (106) uses (105) which holds with probability $1-3\zeta'$. Thus by a union bound, we know that (108) holds with probability at least $1-3m\zeta'$.

Next, we show an analogue of Lemma 16 which connects the guarantees between first situation (large gradients) and second situation (around saddle points) by relating to the *gradient of the starting point* of each epoch (see Line 4 of Algorithm 3). This proof requires several modifications since we use stochastic gradients for v_{sm}.

Lemma 20 (Two Situations) *For any epoch s, let x_t be a point uniformly sampled from this epoch $\{x_j\}_{j=sm+1}^{(s+1)m}$ and choose the step size $\eta \le \frac{1}{(1+2C')L}$ (where $C' = O(\log \frac{dm}{\zeta'}\sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$) and the minibatch size $b \ge m$. Then for any $\epsilon > 0$, by letting batch size $B \ge \frac{256C_3^2\sigma^2}{\epsilon^2}$ (where $C_3 = O(\log \frac{d}{\zeta'}) = \widetilde{O}(1)$), we have two cases:*

1. *If at least half of points in this epoch have gradient norm no larger than $\frac{\epsilon}{2}$, then $\|\nabla f(x_{(s+1)m})\| \le \frac{\epsilon}{2}$ and $\|v_{(s+1)m}\| \le \epsilon$ hold with probability at least $1/3$;*

2. *Otherwise, we know $f(x_{sm}) - f(x_t) \ge \frac{7\eta m\epsilon^2}{256}$ holds with probability at least $1/5$.*

Moreover, $f(x_t) \le f(x_{sm}) + \frac{(t-sm)\eta C_3^2\sigma^2}{B}$ holds with high probability $1 - 3m\zeta'$ no matter which case happens.

Proof of Lemma 20. There are two cases in this epoch:

1. If at least half of points in this epoch $\{x_j\}_{j=sm+1}^{(s+1)m}$ have gradient norm no larger than $\frac{\epsilon}{2}$, then it is easy to see that a uniformly sampled point x_t has gradient norm $\|\nabla f(x_t)\| \le \frac{\epsilon}{2}$ with probability at least $1/2$. Moreover, note that the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 20 of Algorithm 3), thus we have $\|\nabla f(x_{(s+1)m})\| \le \frac{\epsilon}{2}$ with probability $1/2$. According to (102), we have $\|v_{(s+1)m} - \nabla f(x_{(s+1)m})\| \le \frac{C_3\sigma}{\sqrt{B}}$ with probability $1 - \zeta'$, where $C = O(\log \frac{d}{\zeta'}) = \widetilde{O}(1)$. By a union bound, with probability at least $1/3$ (e.g., choose $\zeta' \le 1/6$), we have
$$\|v_{(s+1)m}\| \le \frac{C_3\sigma}{\sqrt{B}} + \frac{\epsilon}{2} \le \frac{\epsilon}{16} + \frac{\epsilon}{2} \le \epsilon.$$

2. Otherwise, at least half of points have gradient norm larger than $\frac{\epsilon}{2}$. Then, as long as the sampled point x_t falls into the last quarter of $\{x_j\}_{j=sm+1}^{(s+1)m}$, we know $\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 \ge \frac{m\epsilon^2}{16}$. This holds with probability at least $1/4$ since x_t is uniformly sampled. Then by combining with (108), we obtain that the function value decreases
$$f(x_{sm}) - f(x_t) \ge \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \frac{(t-sm)\eta C_3^2\sigma^2}{B} \ge \frac{\eta m\epsilon^2}{32} - \frac{\eta m\epsilon^2}{256} = \frac{7\eta m\epsilon^2}{256},$$
where the last inequality is due to $B \ge \frac{256C_3^2\sigma^2}{\epsilon^2}$. Note that (108) holds with high probability $1 - 3m\zeta'$ if we choose the minibatch size $b \ge m$ and the step size $\eta \le \frac{1}{(1+2C')L}$. By a union bound, the function value decrease $f(x_{sm}) - f(x_t) \ge \frac{\eta m\epsilon^2}{64}$ with probability at least $1/5$ (e.g., choose $\zeta' \le 1/60m$).

Again according to (108), $f(x_t) \le f(x_{sm}) + \frac{(t-sm)\eta C_3^2\sigma^2}{B}$holds with high probability $1 - 3m\zeta'$. □

Note that if Case 2 happens, the function value would decrease significantly in this epoch s (corresponding to the first situation large gradients). Otherwise if Case 1 happens, we know the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 20 of Algorithm 3), then we know $\|\nabla f(x_{(s+1)m})\| \le \frac{\epsilon}{2}$ and $\|v_{(s+1)m}\| \le \epsilon$. In this case, we start a super epoch (corresponding

to the second situation around saddle points). Note that if $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) > -\delta$, the point $x_{(s+1)m}$ is already an (ϵ, δ)-local minimum.

Around Saddle Points $\|v_{(s+1)m}\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) \leq -\delta$: In this situation, we show that the function value decreases significantly in a *super epoch* with high probability by adding a random perturbation at the initial point $\widetilde{x} = x_{(s+1)m}$. We denote $x_0 := \widetilde{x} + \xi$ to denote the starting point of the super epoch after the perturbation, where ξ uniformly $\sim \mathbb{B}_0(r)$ and the perturbation radius is r (see Line 7 of Algorithm 3). Again, we follow the *two-point analysis* developed in Jin et al. (2017). In particular, consider two coupled points x_0 and x_0' with $w_0 := x_0 - x_0' = r_0 e_1$, where r_0 is a scalar and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. We show that at least for one of these two coupled sequences $\{x_t\}$ and $\{x_t'\}$, the function value decrease a lot (escape the saddle point), i.e.,

$$\exists t \leq t_{\text{thres}}, \ \text{ such that } \ \max\{f(x_0) - f(x_t), f(x_0') - f(x_t')\} \geq 2f_{\text{thres}}. \tag{109}$$

The proof outline of (109) is the same as that in Section D.2. We assume by contradiction that $f(x_0) - f(x_t) < 2f_{\text{thres}}$ and $f(x_0') - f(x_t') < 2f_{\text{thres}}$. Similar to Lemma 17 and Lemma 18, we need the following two technical lemmas in the online setting. Their proofs are deferred to the end of this section.

Lemma 21 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 9–13 of Algorithm 3) from x_0. Moreover, let the step size $\eta \leq \frac{1}{(1+2C')L}$ and minibatch size $b \geq m$. With probability $1 - 3t\zeta'$, we have*

$$\forall t \geq 0, \ \ \|x_t - x_0\| \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L} + \frac{4t^2\eta C_3^2\sigma^2}{C'LB}}, \tag{110}$$

where $C' = O(\log \frac{dm}{\zeta'}\sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$ and $C_3 = O(\log \frac{d}{\zeta'}) = \widetilde{O}(1)$.

Lemma 22 (Small Stuck Region) *If the initial point $\widetilde{x}$ satisfies $-\gamma := \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$, then let $\{x_t\}$ and $\{x_t'\}$ be two coupled sequences by running SSRGD update steps (Line 9–13 of Algorithm 3) with the same choice of batches and minibatches (i.e., I_B's in Line 9 of Algorithm 3 and I_b's in Line 13 of Algorithm 3) from x_0 and x_0' with $w_0 := x_0 - x_0' = r_0 e_1$, where $x_0 \in \mathbb{B}_{\widetilde{x}}(r)$, $x_0' \in \mathbb{B}_{\widetilde{x}}(r)$, $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\nabla^2 f(\widetilde{x})$. Moreover, let the super epoch length $t_{\text{thres}} = \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \leq \frac{1}{30(1+\log t_{\text{thres}})C'L} = \widetilde{O}(\frac{1}{L})$, minibatch size $b \geq m$, batch size $B = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$ and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$. Then with probability $1 - 3T\zeta'$, we have*

$$\exists T \leq t_{\text{thres}}, \ \ \max\{\|x_T - x_0\|, \|x_T' - x_0'\|\} \geq \frac{\delta}{C_1\rho}, \tag{111}$$

where where $C' = O(\log \frac{dm}{\zeta'}\sqrt{\log \frac{d}{\zeta'}}) = \widetilde{O}(1)$ and $C_1 \geq 1 + 96C'\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) = \widetilde{O}(1)$.

Based on these two lemmas, we are ready to show that (109) holds with high probability. Without loss of generality, we assume $\|x_T - x_0\| \geq \frac{\delta}{C_1\rho}$ in (111) (note that (110) holds for both $\{x_t\}$

and $\{x_t'\}$). Then plugging it into (110), we obtain

$$\sqrt{\frac{4T(f(x_0)-f(x_T))}{C'L}+\frac{4T^2\eta C_3^2\sigma^2}{C'LB}} \geq \frac{\delta}{C_1\rho} \tag{112}$$

Hence, we can see that

$$\begin{aligned} f(x_0)-f(x_T) &\geq \frac{C'L\delta^2}{4C_1^2\rho^2T}-\frac{T\eta C_3^2\sigma^2}{B} \\ &\geq \frac{C'L\eta\delta^3}{4C_1^2\rho^2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}-\frac{C_3^2\sigma^2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{B\delta} \qquad (113)\\ &\geq \frac{\delta^3}{C_1'\rho^2} \qquad (114)\\ &\overset{\text{def}}{=} 2f_{\text{thres}}, \end{aligned}$$

where the last equality is due to the definition of $f_{\text{thres}} := \frac{\delta^3}{2C_1'\rho^2} = \widetilde{O}(\frac{\delta^3}{\rho^2})$, (113) is due to $T \leq t_{\text{thres}} := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta}$, and (114) holds by letting $C_1' = \frac{8C_1^2\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{C'L\eta} = \widetilde{O}(1)$. Recall that $B = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$ and $\epsilon \leq \delta^2/\rho$. Thus, we have already proved that at least one of sequences $\{x_t\}$ and $\{x_t'\}$ escapes the saddle point with probability $1-6T\zeta'$ (by union bound of (110) and (111)), i.e.,

$$\exists T \leq t_{\text{thres}}, \quad \max\{f(x_0)-f(x_T), f(x_0')-f(x_T')\} \geq 2f_{\text{thres}}, \tag{115}$$

if their starting points x_0 and x_0' satisfying $w_0 := x_0 - x_0' = r_0e_1$.

Next, using exactly the same volume argument as in Section D.2, we obtain that

$$f(\widetilde{x})-f(x_T) = f(\widetilde{x})-f(x_0)+f(x_0)-f(x_T) \geq -f_{\text{thres}}+2f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2} \tag{116}$$

holds with probability $1-(6T+1)\zeta' \geq 1-7t_{\text{thres}}\zeta'$, where $C_1' = \widetilde{O}(1)$. Here we use the fact $f(x_0) \leq f(\widetilde{x}) + f_{\text{thres}}$ which follows from (68). Hence, we have finished the proof for the second situation (around saddle points).

Combing these two situations (large gradients and around saddle points) to prove Theorem 9: We distinguishing the epochs into three types, *Type-1 useful epoch*, *Wasted epoch* and *Type-2 useful super epoch* in exactly the same way as in Section D.2. Recall in a Type-1 useful epoch, at least half of points in this epoch have gradient norm larger than $\epsilon/2$ (Case 2 of Lemma 20). If at least half of points in this epoch have gradient norm no larger than $\epsilon/2$ and the starting point of the next epoch has estimated gradient norm larger than ϵ, we say it is a wasted epoch. In a Type-2 useful super epoch, at least half of points in this epoch have gradient norm no larger than ϵ and the starting point of the next epoch has estimated gradient norm no larger than ϵ (Case 1 of Lemma 20). The argument is very similar to the one in Section D.2 as well, except some quantitative details.

First, we can see that the probability of a wasted epoch happened is less than $2/3$ due to the random stop (see Case 1 of Lemma 20). Note for different wasted epochs, returned points are independently sampled. Thus, with high probability $1-\zeta'$, at most $O(\log\frac{1}{\zeta'}) = \widetilde{O}(1)$ wasted

epochs would happen before a type-1 useful epoch or type-2 useful super epoch. We use N_1 and N_2 to denote the number of type-1 useful epochs and type-2 useful super epochs.

For type-1 useful epoch, according to Case 2 of Lemma 20, we know that the function value decreases at least $\frac{7\eta m\epsilon^2}{256}$ with probability at least $1/5$. Using a union bound, we know that with probability $1-4N_1/5$, N_1 type-1 useful epochs will decrease the function value at least $\frac{7\eta m\epsilon^2 N_1}{1280}$. Note that the function value can decrease at most $\Delta_0 := f(x_0) - f^*$ and also recall that the function value can only increase at most $\frac{\eta m C_3^2\sigma^2}{B}$ with high probability $1-3m\zeta'$ for any (wasted) epoch, where $C_3 = O(\log\frac{d}{\zeta'}) = \widetilde{O}(1)$ (see Lemma 20). By choosing $B = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$ and small enough ζ', N_1 type-1 useful epochs will decrease the function value at least $\frac{\eta m\epsilon^2 N_1}{200}$ with probability at least $1-\widetilde{O}(N_1 m\zeta')$ by a union bound. We can let $\zeta' \le \widetilde{O}(1/N_1 m)$. So let $\frac{\eta m\epsilon^2 N_1}{200} \le \Delta_0$, we get $N_1 \le \frac{200\Delta_0}{\eta m\epsilon^2}$.

For type-2 useful super epoch, first we know that the starting point of the super epoch $\widetilde{x} := x_{(s+1)m}$ has gradient norm $\|\nabla f(\widetilde{x})\| \le \epsilon/2$ and estimated gradient norm $\|v_{(s+1)m}\| \le \epsilon$. Now if $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \ge -\delta$, then $\widetilde{x}$ is already a (ϵ,δ)-local minimum. Otherwise, $\|v_{(s+1)m}\| \le \epsilon$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$, this is exactly our second situation (around saddle points). According to (116), we know that the the function value decrease $(f(\widetilde{x}) - f(x_T))$ is at least $f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}$ with probability at least $1 - 7t_{\text{thres}}\zeta' \ge 1/2$ (let $\zeta' \le 1/14t_{\text{thres}}$), where $C_1' = \widetilde{O}(1)$. Similar to type-1 useful epoch, we know $N_2 \le \frac{4C_1'\rho^2\Delta_0}{\delta^3}$ with probability at least $1-\widetilde{O}(N_2 t_{\text{thres}}\zeta')$ by a union bound. We can let $\zeta' \le \widetilde{O}(1/N_2 t_{\text{thres}})$.

Now, we are ready to bound the number of SFO calls in Theorem 9 (online) as follows:

$$
\begin{aligned}
&N_1(\widetilde{O}(1)B + B + mb) + N_2(\widetilde{O}(1)B + \lceil\frac{t_{\text{thres}}}{m}\rceil B + t_{\text{thres}}b)\\
&\le \widetilde{O}\Big(\frac{\Delta_0\sigma}{\eta\epsilon^2\epsilon} + \frac{\rho^2\Delta_0}{\delta^3}(\frac{\sigma^2}{\epsilon^2} + \frac{\sigma}{\eta\delta\epsilon})\Big)\\
&\le \widetilde{O}\Big(\frac{L\Delta_0\sigma}{\epsilon^3} + \frac{\rho^2\Delta_0\sigma^2}{\epsilon^2\delta^3} + \frac{L\rho^2\Delta_0\sigma}{\epsilon\delta^4}\Big). \qquad (117)
\end{aligned}
$$

By a union bound of these types and set $\zeta = \widetilde{O}(N_1 m + N_2 t_{\text{thres}})\zeta'$ (note that ζ' only appears in the log term $\log(\frac{1}{\zeta'})$, so it can be chosen as small as we want), we know that the SFO calls of SSRGD can be bounded by (117) with probability $1-\zeta$. This finishes the proof of Theorem 9 (the online case). Now, the only remaining thing is to prove Lemma 21 and 22.

Proof of Lemma 21. First, we know that the variance bound (105) holds with probability $1-3\zeta'$. Then by a union bound, it holds with probability $1-3t\zeta'$ for all $0 \le j \le t-1$. Then, according to (107), we know for any $\tau \le t$ in some epoch s

$$
\begin{aligned}
f(x_\tau) &\le f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{\tau}\|\nabla f(x_{j-1})\|^2 - (\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2L^2)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2\\
&\quad + \frac{(\tau - sm)\eta C_3^2\sigma^2}{B}\\
&\le f(x_{sm}) - (\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2L^2)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 + \frac{(\tau - sm)\eta C^2\sigma^2}{B}
\end{aligned}
$$

$$\leq f(x_{sm}) - \frac{C'L}{4}\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 + \frac{(\tau - sm)\eta C_3^2\sigma^2}{B}, \tag{118}$$

where the last inequality holds since the step size $\eta \leq \frac{1}{(1+2C')L}$. Recall that $C' = O(\log\frac{dm}{\zeta'}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$ and $C_3 = O(\log\frac{d}{\zeta'}) = \widetilde{O}(1)$. Now, we sum up (118) for all epochs before iteration t,

$$f(x_t) \leq f(x_0) - \frac{C'L}{4}\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2 + \frac{t\eta C^2\sigma^2}{B}.$$

Then, the proof is finished as

$$\|x_t - x_0\| \leq \sum_{j=1}^{t}\|x_j - x_{j-1}\| \leq \sqrt{t\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2} \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L} + \frac{4t^2\eta C_3^2\sigma^2}{C'LB}}.$$

□

Proof of Lemma 22. We prove this lemma by contradiction. Assume the contrary,

$$\forall t \leq T,\ \ \|x_t - x_0\| \leq \frac{\delta}{C_1\rho}\ \text{ and }\ \|x_t' - x_0'\| \leq \frac{\delta}{C_1\rho}, \tag{119}$$

where $T := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\gamma} \leq t_{\text{thres}} := \frac{\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})}{\eta\delta}$ (note that $\gamma \geq \delta$ due to $-\gamma := \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$). We will show that the distance between these two coupled sequences $w_t := x_t - x_t'$ grows exponentially since they have a gap in the e_1 direction at the beginning, i.e., $w_0 := x_0 - x_0' = r_0 e_1$, where $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. However, $\|w_t\| = \|x_t - x_t'\| \leq \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x_t' - x_0'\| + \|x_0' - \widetilde{x}\| \leq 2r + 2\frac{\delta}{C_1\rho}$ according to (119) and the perturbation radius r. It is not hard to see that if $\|w_t\|$ increases exponentially, this inequality cannot be true for reasonably large t, rendering a contradiction.

In the following, we prove the exponential increase of $\|w_t\|$ by induction. First, recall the expression of w_t in (73) and (74):

$$w_t = (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + y_{t-1}) \tag{120}$$

$$= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau + y_\tau), \tag{121}$$

where $\Delta_\tau := \int_0^1(\nabla^2 f(x_\tau' + \theta(x_\tau - x_\tau')) - \mathcal{H})d\theta$ and $y_\tau := v_\tau - \nabla f(x_\tau) - v_\tau' + \nabla f(x_\tau')$.

Again, to show the exponential increase of $\|w_t\|$, it is sufficient to show that the first term of (121) dominates the second term. To this end, we show the following bound, which is almost the same as Lemma 19, except that the succeed probability changes to $1 - 3T\zeta'$.

Lemma 23 *Suppose $w_0 := x_0 - x_0' = r_0 e_1$ where $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 is the eigenvector corresponding to the smallest eigenvalue of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. If (119) holds, then with probability $1 - 3T\zeta'$, the following bounds hold for all $t \leq T$:*

1. $\frac{1}{2}(1+\eta\gamma)^t r_0 \le \|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0$;
2. $\|y_t\| \le \frac{\gamma}{4C_2}(1+\eta\gamma)^t r_0$.

where $C_2 := \log(\frac{8\delta\sqrt{d}}{\rho r \zeta'})$.

Proof of Lemma 23. First, check the base case $t = 0$, $\|w_0\| = \|r_0 e_1\| = r_0$ holds for Bound 1. However, the base case of Bound 2 requires more work. Here, we use Bernstein inequality (Proposition 13) to show that $\|y_0\| = \|v_0 - \nabla f(x_0) - v_0' + \nabla f(x_0')\| \le \eta\gamma L r_0$. According to Line 9 of Algorithm 3, we know that $v_0 = \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_0)$ and $v_0' = \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_0')$ (recall that these two coupled sequence $\{x_t\}$ and $\{x_t'\}$ use the same choice of batches and minibatches (i.e., same I_B's and I_b's). Now, we have

$$\begin{aligned} y_0 &= v_0 - \nabla f(x_0) - v_0' + \nabla f(x_0') \\ &= \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_0) - \nabla f(x_0) - \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_0') + \nabla f(x_0') \\ &= \frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(x_0) - \nabla f_j(x_0') - (\nabla f(x_0) - \nabla f(x_0'))\Big). \end{aligned} \tag{122}$$

We first bound the norm of each individual term of (122):

$$\|\nabla f_j(x_0) - \nabla f_j(x_0') - (\nabla f(x_0) - \nabla f(x_0'))\| \le 2L\|x_0 - x_0'\| = 2L\|w_0\| = 2Lr_0, \tag{123}$$

where the inequality holds due to the gradient Lipschitz Assumption 5. Then, consider the corresponding variance:

$$\begin{aligned} &\mathbb{E}\Big[\sum_{j\in I_B}\|\nabla f_j(x_0) - \nabla f_j(x_0') - (\nabla f(x_0) - \nabla f(x_0'))\|^2\Big] \\ &\le B\mathbb{E}_j[\|\nabla f_j(x_0) - \nabla f_j(x_0')\|^2] \le BL^2\|x_0 - x_0'\|^2 = BL^2\|w_0\|^2 = BL^2 r_0^2, \end{aligned} \tag{124}$$

where the first inequality uses the fact $\mathbb{E}[\|x - \mathbb{E}x\|^2] \le \mathbb{E}[\|x\|^2]$, and the last inequality uses the gradient Lipschitz Assumption 5. According to (123) and (124), we can bound $\|y_0\|$ by Bernstein inequality (Proposition 13) as

$$\begin{aligned} \mathbb{P}\Big\{\|y_0\| \ge \frac{\alpha}{B}\Big\} &\le (d+1)\exp\Big(\frac{-\alpha^2/2}{\sigma^2 + R\alpha/3}\Big) \\ &= (d+1)\exp\Big(\frac{-\alpha^2/2}{BL^2 r_0^2 + 2Lr_0\alpha/3}\Big) \\ &= \zeta', \end{aligned}$$

where the last equality holds by letting $\alpha = C_3 L\sqrt{B}r_0$, where $C_3 = O(\log\frac{d}{\zeta'})$. By further choosing $B = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$, we can see that the base case

$$\|y_0\| \le \frac{C_3 L r_0}{\sqrt{B}} \le \frac{\gamma}{8C_2}r_0, \tag{125}$$

holds with probability $1-\zeta'$.

Now, we proceed to the induction step: Assuming Bound 1 and Bound 2 hold for all $\tau \le t-1$, we now prove they hold for t. For Bound 1, same arguments as in Lemma 19 can show that the second term of (121) is dominated by half of the first term. We do not repeat the proof which are exactly the same. Note that the first term of (121) is $\|(I-\eta\mathcal{H})^t w_0\| = (1+\eta\gamma)^t r_0$. Thus, we have the first bound:

$$\frac{1}{2}(1+\eta\gamma)^t r_0 \le \|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0 \tag{126}$$

Now, we proceed to the second bound $\|y_t\| \le \frac{\gamma}{4C_2}(1+\eta\gamma)^t r_0$. Define

$$\beta := \sqrt{8\sum_{k=sm+1}^{t} c_k^2 \log\frac{d}{\zeta'}} = \frac{C'}{\sqrt{b}}\sqrt{\sum_{k=sm+1}^{t}(L\|w_t-w_{t-1}\|+\rho D_t^x\|w_t\|+\rho D_{t-1}^x\|w_{t-1}\|)^2},$$

where $C' = O(C\sqrt{\log\frac{d}{\zeta'}}) = O(\log\frac{d}{\zeta_k}\sqrt{\log\frac{d}{\zeta'}}) = \widetilde{O}(1)$ and $\zeta_k = \zeta'/m$. The same proof as in Lemma 19 show that

$$\mathbb{P}\Big\{\big\|y_t - y_{sm}\big\| \ge \beta\Big\} \le (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{k=sm+1}^{t} c_k^2}\Big) + \zeta' = 2\zeta'. \tag{127}$$

Recall that $y_k := v_k - \nabla f(x_k) - v_k' + \nabla f(x_k')$ and at the beginning point of this epoch $y_{sm} = \|v_{sm} - \nabla f(x_{sm}) - v_{sm}' + \nabla f(x_{sm}')\| \le \frac{\gamma}{8C_2} r_0$ with probability $1-\zeta'$ (see (125)). Combining with (127) and using a union bound, for any $t \in [sm+1, (s+1)m]$, we have that

$$\|y_t\| \le \beta + \|y_{sm}\| \le \frac{C'}{\sqrt{b}}\sqrt{\sum_{k=sm+1}^{t}(L\|w_t-w_{t-1}\|+\rho D_t^x\|w_t\|+\rho D_{t-1}^x\|w_{t-1}\|)^2} + \frac{\gamma}{8C_2}r_0 \tag{128}$$

holds with probability $1-3\zeta'$. Furthermore, by a union bound, we know that (128) holds with probability at least $1-3T\zeta'$ for all $t \le T$.

Now, we bound the right-hand-side of (128) to finish the proof. The proof will be the same as in Lemma 19. The last two terms inside the square root can be bounded as in (90):

$$\rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\| \le \frac{6\delta}{C_1}(1+\eta\gamma)^t r_0, \tag{129}$$

The first term in the square root can also be bounded in the same way as in (91)–(100):

$$L\|w_t - w_{t-1}\| \le \Big(\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2}\Big)L\eta\gamma(1+\eta\gamma)^{t-1} r_0. \tag{130}$$

By plugging (129) and (130) into (128), we have

$$\|y_t\| \le C'\left(\frac{6\delta}{C_1}(1+\eta\gamma)^t r_0 + \Big(\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2}\Big)L\eta\gamma(1+\eta\gamma)^{t-1} r_0\right) + \frac{\gamma}{8C_2}r_0$$

$$\leq \Big(\frac{6C'}{C_1} + (\frac{3}{2} + \frac{3(1+\log t)}{C_1} + \frac{1+\log t}{4C_2})C'L\eta\Big)\gamma(1+\eta\gamma)^t r_0 + \frac{\gamma}{8C_2}r_0$$
$$\leq \Big(\frac{1}{16C_2} + \frac{1}{16C_2}\Big)\gamma(1+\eta\gamma)^t r_0 + \frac{\gamma}{8C_2}r_0$$
$$= \frac{1}{4C_2}\gamma(1+\eta\gamma)^t r_0, \tag{131}$$

where the second inequality holds due to $\delta \leq \gamma$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$), and the last inequality holds by letting $C_1 \geq 1 + 96C'C_2$ (recall that $C_2 := \log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$ defined in Lemma 23), and $\eta \leq \frac{1}{30(1+\log t)C'L}$. Recall that (128) holds with probability at least $1 - 3T\zeta'$ for all $t \leq T$. This finishes the proof of Lemma 23. □

From the Lemma 23, one can see that $\|w_t\| \geq \frac{1}{2}(1+\eta\gamma)^t r_0 = \frac{1}{2}(1+\eta\gamma)^t \frac{\zeta' r}{\sqrt{d}}$. On the other hand, $\|w_t\| := \|x_t - x_t'\| \leq \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x_t' - x_0'\| + \|x_0' - \widetilde{x}\| \leq 2r + 2\frac{\delta}{C_1\rho} \leq \frac{4\delta}{C_1\rho}$ according to (119) and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$. Hence, for any $t \geq T = \frac{1}{\eta\gamma}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$, we get a contradiction to (119), i.e., $\|w_t\| \geq \frac{1}{2}(1+\eta\gamma)^T \frac{\zeta' r}{\sqrt{d}} \geq \frac{4\delta}{\rho} \geq \frac{4\delta}{C_1\rho}$, where $C_1 \geq 1 + 96C'\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) \geq 1$ defined in Lemma 22. Also note that $T = \frac{1}{\eta\gamma}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) \leq t_{\text{thres}} := \frac{1}{\eta\delta}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'})$ due to $\delta \leq \gamma$. This contradiction finishes the proof of Lemma 22. □

D.4 Proof of Theorem 10 (Under third-order Lipschitzness assumption)

Proof of Theorem 10. The proof is similar to the proof for the online case of Theorem 9 provided in Section D.3. Again, we distinguish two situations, the *large gradients* case, in which the function value decreases significantly, and the *around saddle points* case. The proof for the first case (large gradients) is exactly same as in the first case in Section D.3 (i.e., Lemma 20).

The difference is in the second case (around saddle points). In previous Section D.3, we add a random perturbation at the starting point of the super epoch. Concretely, we show that the function value decreases a lot in this super epoch with high probability (see (116)), i.e.,

$$\exists T \leq t_{\text{thres}}, \;\; f(\widetilde{x}) - f(x_T) \geq f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2} \tag{132}$$

holds with probability at least $1 - 7t_{\text{thres}}\zeta'$, where $C_1' = \widetilde{O}(1)$. Recall that the super epoch length $t_{\text{thres}} := \frac{1}{\eta\delta}\log(\frac{8\delta\sqrt{d}}{\rho r\zeta'}) = \widetilde{O}(\frac{1}{\eta\delta})$ (see Lemma 22) and $\widetilde{x}$ is the starting point of this super epoch. However, for Theorem 10 which further assumes the L_3-Lipschitz of third-order derivative (i.e., Assumption 7), one can show that the function value decreases by a larger amount (*improving a factor of* δ), i.e., $\frac{3\delta^2}{8L_3}$ in (133) instead of $\frac{\delta^3}{2C_1'\rho^2}$ in (132). Finally we can see that the result of Theorem 10 indeed improves the previous online case of Theorem 9 by a factor of δ. Now we formalize the proof of Theorem 10 in this second case (around saddle points). Here we directly reuse the function value decrease lemma provided in Yu et al. (2017). Note that here we can remove the expectation in Lemma 4.6 of Yu et al. (2017) by choosing $y = \arg\min_{y\in\{y_-, y_+\}} f(y)$.

Lemma 24 (Lemma 4.6 in (Yu et al., 2017)) *Suppose that Assumptions 5, 6 and 7 hold. If the start point $\widetilde{x}$ satisfies $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$. Then one can apply a negative curvature search step for finding a direction to decrease the function value. In particular, Neon2^{online} (Allen-Zhu and Li,*

2018) can return a point y such that

$$f(\widetilde{x}) - f(y) \geq \frac{3\delta^2}{8L_3} \tag{133}$$

holds with probability $1-\zeta'$ and the total number of stochastic gradient computations is at most $T = O(\frac{L^2}{\delta^2}\log^2\frac{d}{\zeta'}) = \widetilde{O}(\frac{L^2}{\delta^2})$.

Now, we are ready to combine these two situations (large gradients and around saddle points) to prove Theorem 10. The arguments are similar to that in Section D.3. The only difference is that here we replace super epoch step by the negative curvature search step (i.e., replace (132) by (133)) in the around saddle points situation. Concretely, i) for large gradients situation, N_1 type-1 useful epochs will decrease the function value at least $\frac{\eta m\epsilon^2 N_1}{200}$ with probability at least $1-\widetilde{O}(N_1 m\zeta')$ by a union bound. We can let $\zeta' \leq \widetilde{O}(1/N_1 m)$. So let $\frac{\eta m\epsilon^2 N_1}{200} \leq \Delta_0$, we get $N_1 \leq \frac{200\Delta_0}{\eta m\epsilon^2}$. ii) for around saddle points situation, according to (133), we know that the the function value decrease $(f(\widetilde{x})-f(y))$ is $\frac{3\delta^2}{8L_3}$ with probability at least $1-\zeta'$. Similar to the large gradients situation, we know $N_2 \leq \frac{16L_3\Delta_0}{3\delta^2}$ with probability at least $1-\widetilde{O}(N_2\zeta')$ by a union bound. We can let $\zeta' \leq \widetilde{O}(1/N_2)$. Now, we bound the number of SFO calls in Theorem 10 (online case under third-order Lipschitz) as follows:

$$N_1(\widetilde{O}(1)B + B + mb) + N_2(\widetilde{O}(1)B + T) \leq \widetilde{O}\Big(\frac{L\Delta_0\sigma}{\epsilon^3} + \frac{L_3\Delta_0\sigma^2}{\epsilon^2\delta^2} + \frac{L_3L^2\Delta_0}{\delta^4}\Big). \tag{134}$$

By a union bound of these types and set $\zeta = \widetilde{O}(N_1 m + N_2)\zeta'$ (note that ζ' only appears in the log term $\log(\frac{1}{\zeta'})$, so it can be chosen as small as we want), we know that the SFO calls can be bounded by (134) with probability $1-\zeta$. □

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems*, pages 2680–2691, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pages 3720–3730, 2018.

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on learning theory*, pages 81–102, 2016.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning*, pages 654–663. PMLR, 2017.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In *International Conference on Machine Learning*, pages 1155–1164, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

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.

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 687–697, 2018.

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.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

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.

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

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized SVRG: Simple variance reduction for nonconvex optimization. In *Conference on learning theory*, pages 1394–1448. PMLR, 2019.

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

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2): 267–305, 2016.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster nonconvex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017.

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. PMLR, 2018.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811*, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. *arXiv preprint arXiv:1901.08689*, 2019.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

Lihua Lei and Michael Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *Artificial Intelligence and Statistics*, pages 148–156, 2017.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

Boyue Li, Zhize Li, and Yuejie Chi. DESTRESS: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 4(3):1031–1051, 2022a.

Zhize Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. In *Advances in Neural Information Processing Systems*, pages 1523–1533, 2019.

Zhize Li. ANITA: An optimal loopless accelerated variance-reduced gradient method. *arXiv preprint arXiv:2103.11333*, 2021.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5569–5579, 2018.

Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

Zhize Li and Peter Richtárik. CANITA: Faster rates for distributed convex optimization with communication compression. In *Advances in Neural Information Processing Systems*, pages 13770–13781, 2021a.

Zhize Li and Peter Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021b.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR, 2020.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. SoteriaFL: A unified framework for private federated learning with communication compression. *arXiv preprint arXiv:2206.09888*, 2022b.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv preprint arXiv:1504.06298*, 2015.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

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

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv preprint arXiv:1902.05679*, 2019.

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

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323, 2016a.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016b.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems*, pages 4384–4396, 2021.

Peter Richtárik, Igor Sokolov, Elnur Gasanov, Ilyas Fatkhullin, Zhize Li, and Eduard Gorbunov. 3PC: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. In *International Conference on Machine Learning*, pages 18596–18648. PMLR, 2022.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

Terence Tao and Van Vu. Random matrices: Universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.

Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA, 2011.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2406–2416, 2019.

Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3639–3647, 2016.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Yi Xu, Jing Rong, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5535–5545, 2018.

Yaodong Yu, Pan Xu, and Quanquan Gu. Third-order smoothness helps: Even faster stochastic optimization algorithms for finding local minima. *arXiv preprint arXiv:1712.06585*, 2017.

Haoyu Zhao, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. Faster rates for compressed federated learning with client-variance reduction. *arXiv preprint arXiv:2112.13097*, 2021a.

Haoyu Zhao, Zhize Li, and Peter Richtárik. FedPAGE: A fast local stochastic gradient method for communication-efficient federated learning. *arXiv preprint arXiv:2108.04755*, 2021b.

Haoyu Zhao, Boyue Li, Zhize Li, Peter Richtárik, and Yuejie Chi. BEER: Fast $O(1/T)$ rate for decentralized nonconvex optimization with communication compression. *arXiv preprint arXiv:2201.13320*, 2022.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Finding local minima via stochastic nested variance reduction. *arXiv preprint arXiv:1806.08782*, 2018a.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3921–3932, 2018b.