SSRGD: Simple Stochastic Recursive Gradient Descent for Escaping Saddle Points

Zhize Li
IIIS, Tsinghua University
zz-li14@mails.tsinghua.edu.cn

Abstract

We analyze stochastic gradient algorithms for optimizing nonconvex problems. In particular, our goal is to find local minima (second-order stationary points) instead of just finding first-order stationary points which may be some bad unstable saddle points. We show that a simple perturbed version of stochastic recursive gradient descent algorithm (called SSRGD) can find an (ϵ, δ)-second-order stationary point with $\widetilde{O}(\sqrt{n}/\epsilon^2 + \sqrt{n}/\delta^4 + n/\delta^3)$ stochastic gradient complexity for nonconvex finite-sum problems. As a by-product, SSRGD finds an ϵ-first-order stationary point with $O(n + \sqrt{n}/\epsilon^2)$ stochastic gradients. These results are almost optimal since Fang et al. [2018] provided a lower bound $\Omega(\sqrt{n}/\epsilon^2)$ for finding even just an ϵ-first-order stationary point. We emphasize that SSRGD algorithm for finding second-order stationary points is as simple as for finding first-order stationary points just by adding a uniform perturbation sometimes, while all other algorithms for finding second-order stationary points with similar gradient complexity need to combine with a negative-curvature search subroutine (e.g., Neon2 [Allen-Zhu and Li, 2018]). Moreover, the simple SSRGD algorithm gets a simpler analysis. Besides, we also extend our results from nonconvex finite-sum problems to nonconvex online (expectation) problems, and prove the corresponding convergence results.

1 Introduction

Nonconvex optimization is ubiquitous in machine learning applications especially for deep neural networks. For convex optimization, every local minimum is a global minimum and it can be achieved by any first-order stationary point, i.e., $\nabla f(x) = 0$. However, for nonconvex problems, the point with zero gradient can be a local minimum, a local maximum or a saddle point. To avoid converging to bad saddle points (including local maxima), we want to find a second-order stationary point, i.e., $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$ (this is a necessary condition for x to be a local minimum). All second-order stationary points indeed are local minima if function f satisfies strict saddle property [Ge et al., 2015]. Note that finding the global minimum in nonconvex problems is NP-hard in general. Also note that it was shown that all local minima are also global minima for some nonconvex problems, e.g., matrix sensing [Bhojanapalli et al., 2016], matrix completion [Ge et al., 2016], and some neural networks [Ge et al., 2017]. Thus, our goal in this paper is to find an approximate second-order stationary point (local minimum) with proved convergence.

There has been extensive research for finding ϵ-first-order stationary point (i.e., $\|\nabla f(x)\| \leq \epsilon$), e.g., GD, SGD and SVRG. See Table 1 for an overview. Although Xu et al. [2018] and Allen-Zhu and Li [2018] independently proposed reduction algorithms Neon/Neon2 that can be combined with previous ϵ-first-order stationary points finding algorithms to find an (ϵ, δ)-second-order stationary point (i.e., $\|\nabla f(x)\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \geq -\delta$). However, algorithms obtained by this reduction are very complicated in practice, and they need 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 smallest eigenvector of $\nabla^2 f(x)$. This also involves a more complicated analysis. Note that in practice, standard first-order stationary point finding algorithms can often work (escape bad saddle points) in nonconvex setting without a negative curvature search subroutine. The reason may be that the saddle points are usually not very stable. So there is a natural question "Is there any simple modification to allow first-order stationary point finding algorithms to get a theoretical second-order guarantee?". For gradient descent (GD), Jin et al. [2017] showed that a simple perturbation step is enough to escape saddle points for finding a second-order stationary point, and this is necessary [Du et al., 2017]. Very recently, Ge et al. [2019] showed that a simple perturbation step is also enough to

find a second-order stationary point for SVRG algorithm [Li and Li, 2018]. Moreover, Ge et al. [2019] also developed a stabilized trick to further improve the dependency of Hessian Lipschitz parameter.

Table 1: Stochastic gradient complexity of optimization algorithms for nonconvex finite-sum problem (1)

Algorithm	Stochastic gradient complexity	Guarantee	Negative-curvature search subroutine
GD [Nesterov, 2004]	$O(\frac{n}{\epsilon^2})$	1st-order	No
SVRG [Reddi et al., 2016], [Allen-Zhu and Hazan, 2016]; SCSG [Lei et al., 2017]; SVRG+ [Li and Li, 2018]	$O(n+\frac{n^{2/3}}{\epsilon^2})$	1st-order	No
SNVRG [Zhou et al., 2018b]; SPIDER [Fang et al., 2018]; SpiderBoost [Wang et al., 2018]; SARAH [Pham et al., 2019]	$O(n+\frac{n^{1/2}}{\epsilon^2})$	1st-order	No
SSRGD (this paper)	$O(n+\frac{n^{1/2}}{\epsilon^2})$	1st-order	No
PGD [Jin et al., 2017]	$\widetilde{O}(\frac{n}{\epsilon^2}+\frac{n}{\delta^4})$	2nd-order	No
Neon2+FastCubic/CDHS [Agarwal et al., 2016, Carmon et al., 2016]	$\widetilde{O}(\frac{n}{\epsilon^{1.5}}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\epsilon^{1.75}}+\frac{n^{3/4}}{\delta^{3.5}})$	2nd-order	Needed
Neon2+SVRG [Allen-Zhu and Li, 2018]	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\delta^{3.5}})$	2nd-order	Needed
Stabilized SVRG [Ge et al., 2019]	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{2/3}}{\delta^4})$	2nd-order	No
SNVRG$^+$+Neon2 [Zhou et al., 2018a]	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n}{\delta^3}+\frac{n^{3/4}}{\delta^{3.5}})$	2nd-order	Needed
SPIDER-SFO$^+$(+Neon2) [Fang et al., 2018]	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon\delta^2}+\frac{1}{\epsilon\delta^3}+\frac{1}{\delta^5})$	2nd-order	Needed
SSRGD (this paper)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n^{1/2}}{\delta^4}+\frac{n}{\delta^3})$	2nd-order	No

Table 2: Stochastic gradient complexity of optimization algorithms for nonconvex online (expectation) problem (2)

Algorithm	Stochastic gradient complexity	Guarantee	Negative-curvature search subroutine
SGD [Ghadimi et al., 2016]	$O(\frac{1}{\epsilon^4})$	1st-order	No
SCSG [Lei et al., 2017]; SVRG+ [Li and Li, 2018]	$O(\frac{1}{\epsilon^{3.5}})$	1st-order	No
SNVRG [Zhou et al., 2018b]; SPIDER [Fang et al., 2018]; SpiderBoost [Wang et al., 2018]; SARAH [Pham et al., 2019]	$O(\frac{1}{\epsilon^3})$	1st-order	No
SSRGD (this paper)	$O(\frac{1}{\epsilon^3})$	1st-order	No
Perturbed SGD [Ge et al., 2015]	poly$(d,\frac{1}{\epsilon},\frac{1}{\delta})$	2nd-order	No
CNC-SGD [Daneshmand et al., 2018]	$\widetilde{O}(\frac{1}{\epsilon^4}+\frac{1}{\delta^{10}})$	2nd-order	No
Neon2+SCSG [Allen-Zhu and Li, 2018]	$\widetilde{O}(\frac{1}{\epsilon^{10/3}}+\frac{1}{\epsilon^2\delta^3}+\frac{1}{\delta^5})$	2nd-order	Needed
Neon2+Natasha2 [Allen-Zhu, 2018]	$\widetilde{O}(\frac{1}{\epsilon^{3.25}}+\frac{1}{\epsilon^3\delta}+\frac{1}{\delta^5})$	2nd-order	Needed
SNVRG$^+$+Neon2 [Zhou et al., 2018a]	$\widetilde{O}(\frac{1}{\epsilon^3}+\frac{1}{\epsilon^2\delta^3}+\frac{1}{\delta^5})$	2nd-order	Needed
SPIDER-SFO$^+$(+Neon2) [Fang et al., 2018]	$\widetilde{O}(\frac{1}{\epsilon^3}+\frac{1}{\epsilon^2\delta^2}+\frac{1}{\delta^5})$	2nd-order	Needed
SSRGD (this paper)	$\widetilde{O}(\frac{1}{\epsilon^3}+\frac{1}{\epsilon^2\delta^3}+\frac{1}{\epsilon\delta^4})$	2nd-order	No

Note: 1. Guarantee (see Definition 1): ϵ-first-order stationary point $\|\nabla f(x)\|\le\epsilon$; (ϵ,δ)-second-order stationary point $\|\nabla f(x)\|\le\epsilon$ and $\lambda_{\min}(\nabla^2 f(x))\ge-\delta$.

2. In the classical setting where $\delta=O(\sqrt{\epsilon})$ [Nesterov and Polyak, 2006, Jin et al., 2017], our simple SSRGD is always (no matter what n and ϵ are) not worse than all other algorithms (in both Table 1 and 2) except FastCubic/CDHS (which need to compute Hessian-vector product) and SPIDER-SFO$^+$. Moreover, our simple SSRGD is not worse than FastCubic/CDHS if $n>1/\epsilon$ and is better than SPIDER-SFO$^+$ if δ is very small (e.g., $\delta\le1/\sqrt{n}$) in Table 1.

Algorithm 1 Simple Stochastic Recursive Gradient Descent (SSRGD)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius r, threshold gradient g_{thres}

for $s = 0, 1, 2, \ldots$ **do**
  if not currently in a super epoch and $\|\nabla f(x_{sm})\| \le g_{\mathrm{thres}}$ **then**
    $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(r)$, start a super epoch
    // we use super epoch since we do not want to add the perturbation too often near a saddle point
  end if
  $v_{sm} \leftarrow \nabla f(x_{sm})$
  for $k = 1, 2, \ldots, m$ **do**
    $t \leftarrow sm + k$
    $x_t \leftarrow x_{t-1} - \eta v_{t-1}$
    $v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}$ // I_b are i.i.d. uniform samples with $|I_b| = b$
    if meet stop condition **then** stop super epoch
  end for
end for

1.1 Our Contributions

In this paper, we propose a simple SSRGD algorithm (described in Algorithm 1) showed that a simple perturbation step is enough to find a second-order stationary point for stochastic recursive gradient descent algorithm. Our results and previous results are summarized in Table 1 and 2. We would like to highlight the following points:

- We improve the result in [Ge et al., 2019] to the almost optimal one (i.e., from $n^{2/3}/\epsilon^2$ to $n^{1/2}/\epsilon^2$) since Fang et al. [2018] provided a lower bound $\Omega(\sqrt{n}/\epsilon^2)$ for finding even just an ϵ-first-order stationary point. Note that for other two $n^{1/2}$ algorithms (i.e., SNVRG$^+$ and SPIDER-SFO$^+$), they both need the negative curvature search subroutine thus are more complicated in practice and in analysis compared with their first-order guarantee algorithms (SNVRG and SPIDER), while our SSRGD is as simple as its first-order guarantee algorithm.
- For more general nonconvex online (expectation) problems (2), we obtain the first algorithm which is as simple as finding first-order stationary points for finding a second-order stationary point with similar state-of-the-art convergence result. See the last column of Table 2.
- Our simple SSRGD algorithm gets simpler analysis. Also, the result for finding a first-order stationary point is a by-product from our analysis. We also give a clear interpretation to show why our analysis for SSRGD algorithm can improve the original SVRG from $n^{2/3}$ to $n^{1/2}$ in Section 5.1. We believe it is very useful for better understanding these two algorithms.

2 Preliminaries

Notation: Let $[n]$ denote the set $\{1, 2, \cdots, n\}$ and $\|\cdot\|$ denote the Eculidean norm for a vector and 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 $O(\cdot)$ to hide the constant and $\widetilde{O}(\cdot)$ to hide the polylogarithmic factor.

In this paper, we consider two types of nonconvex problems. The finite-sum problem has the form

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where $f(x)$ and all individual $f_i(x)$ are possibly nonconvex. This form usually models the empirical risk minimization in machine learning problems.

The online (expectation) problem has the form

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\zeta\sim D}[F(x,\zeta)], \tag{2}$$

where $f(x)$ and $F(x,\zeta)$ are possibly nonconvex. This form usually models the population risk minimization in machine learning problems.

Now, we make standard smoothness assumptions for these two problems.

Assumption 1 (Gradient Lipschitz) *1. For finite-sum problem (1), each $f_i(x)$ is differentiable and has L-Lipschitz continuous gradient, i.e.,*

$$\|\nabla f_i(x_1) - \nabla f_i(x_2)\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d. \tag{3}$$

2. For online problem (2), $F(x,\zeta)$ is differentiable and has L-Lipschitz continuous gradient, i.e.,

$$\|\nabla F(x_1,\zeta) - \nabla F(x_2,\zeta)\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d. \tag{4}$$

Assumption 2 (Hessian Lipschitz) *1. For finite-sum problem (1), each $f_i(x)$ is twice-differentiable and has ρ-Lipschitz continuous Hessian, i.e.,*

$$\|\nabla^2 f_i(x_1) - \nabla^2 f_i(x_2)\| \leq \rho\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d. \tag{5}$$

2. For online problem (2), $F(x,\zeta)$ is twice-differentiable and has ρ-Lipschitz continuous Hessian, i.e.,

$$\|\nabla^2 F(x_1,\zeta) - \nabla^2 F(x_2,\zeta)\| \leq \rho\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d. \tag{6}$$

These two assumptions are standard for finding first-order stationary points (Assumption 1) and second-order stationary points (Assumption 1 and 2) for all algorithms in both Table 1 and 2.

Now we define the approximate first-order stationary points and approximate first-order stationary points.

Definition 1 *x is an ϵ-first-order stationary point for a differentiable function f if*

$$\|\nabla f(x)\| \leq \epsilon. \tag{7}$$

x is an (ϵ,δ)-second-order stationary point for a twice-differentiable function f if

$$\|\nabla f(x)\| \leq \epsilon \ \ and \ \ \lambda_{\min}(\nabla^2 f(x)) \geq -\delta. \tag{8}$$

The definition of (ϵ,δ)-second-order stationary point is the same as [Allen-Zhu and Li, 2018, Daneshmand et al., 2018, Zhou et al., 2018a, Fang et al., 2018] and it generalizes the classical version where $\delta = \sqrt{\rho\epsilon}$ used in [Nesterov and Polyak, 2006, Jin et al., 2017, Ge et al., 2019].

3 Simple Stochastic Recursive Gradient Descent

In this section, we propose the simple stochastic recursive gradient descent algorithm called SSRGD. The high-level description (which omits the stop condition details in Line 10) of this algorithm is in Algorithm 1 and the full algorithm (containing the stop condition) is described in Algorithm 2. Note that we call each outer loop (i.e., Line 2–11 of algorithm 1) an ***epoch***, i.e., iterations t from sm to $(s+1)m$ for an epoch s. We call the iterations between the beginning of perturbation and end of perturbation a ***super epoch***.

The SSRGD algorithm is based on the stochastic recursive gradient descent which is introduced in [Nguyen et al., 2017] for convex optimization. In particular, Nguyen et al. [2017] want to save the storage of past gradients in SAGA [Defazio et al., 2014] by using the recursive gradient. However, this stochastic recursive gradient descent is widely used in recent work for nonconvex optimization such as SPIDER [Fang et al., 2018], SpiderBoost [Wang et al., 2018] and some variants of SARAH (e.g., ProxSARAH [Pham et al., 2019]).

Recall that in the well-known SVRG, Johnson and Zhang [2013] reused a fixed snapshot full gradient $\nabla f(\widetilde{x})$ (which is computed at the beginning of each epoch) in the update step:

$$v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(\widetilde{x})\big) + \nabla f(\widetilde{x}), \tag{9}$$

while the stochastic recursive gradient descent uses the recursive update step:

$$v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}. \tag{10}$$

Algorithm 2 Simple Stochastic Recursive Gradient Descent (SSRGD)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius r, threshold gradient g_{thres}, threshold function value f_{thres}, super epoch length t_{thres}

$super\_epoch \leftarrow 0$
for $s = 0, 1, 2, \ldots$ **do**
  if $super\_epoch = 0$ and $\|\nabla f(x_{sm})\| \le g_{\text{thres}}$ **then**
    $super\_epoch \leftarrow 1$
    $\widetilde{x} \leftarrow x_{sm}, t_{\text{init}} \leftarrow sm$
    $x_{sm} \leftarrow \widetilde{x} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(r)$
  end if
  $v_{sm} \leftarrow \nabla f(x_{sm})$
  for $k = 1, 2, \ldots, m$ **do**
    $t \leftarrow sm + k$
    $x_t \leftarrow x_{t-1} - \eta v_{t-1}$
    $v_t \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1}$ // I_b are i.i.d. uniform samples with $|I_b| = b$
    if $super\_epoch = 1$ and $(f(\widetilde{x}) - f(x_t) \ge f_{\text{thres}}$ or $t - t_{\text{init}} \ge t_{\text{thres}})$ **then**
      $super\_epoch \leftarrow 0$; break
    else if $super\_epoch = 0$ **then**
      break with probability $\frac{1}{m-k+1}$
      // we use random stop since we want to randomly choose a point as the starting point of the next epoch
    end if
  end for
  $x_{(s+1)m} \leftarrow x_t$
end for

4 Convergence Results

Similar to the perturbed GD [Jin et al., 2017] and perturbed SVRG [Ge et al., 2019], we add simple perturbations to the stochastic recursive gradient descent algorithm to escape saddle points efficiently. Besides, we also consider the more general online case. In the following theorems, we provide the convergence results of SSRGD for finding an ϵ-first-order stationary point and an (ϵ, δ)-second-order stationary point for the nonconvex finite-sum problem (1) and online problem (2). The proofs are provided in Appendix B. We give an overview of the proofs in next Section 5.

4.1 Nonconvex Finite-sum Problem

Theorem 1 *Under Assumption 1 (i.e. (3)), let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \frac{\sqrt{5}-1}{2L}$, epoch length $m = \sqrt{n}$ and minibatch size $b = \sqrt{n}$, SSRGD will find an ϵ-first-order stationary point in expectation using*

$$O\Big(n + \frac{L\Delta f\sqrt{n}}{\epsilon^2}\Big)$$

stochastic gradients for nonconvex finite-sum problem (1).

Theorem 2 *Under Assumption 1 and 2 (i.e. (3) and (5)), let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \widetilde{O}(\frac{1}{L})$, epoch length $m = \sqrt{n}$, minibatch size $b = \sqrt{n}$, perturbation radius $r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big)$, threshold gradient $g_{\text{thres}} = \epsilon$, threshold function value $f_{\text{thres}} = \widetilde{O}(\frac{\delta^3}{\rho^2})$ and super epoch length $t_{\text{thres}} = \widetilde{O}(\frac{1}{\eta\delta})$, SSRGD will at least once get to an (ϵ, δ)-second-order stationary point with high probability using*

$$\widetilde{O}\Big(\frac{L\Delta f\sqrt{n}}{\epsilon^2} + \frac{L\rho^2\Delta f\sqrt{n}}{\delta^4} + \frac{\rho^2\Delta f n}{\delta^3}\Big)$$

stochastic gradients for nonconvex finite-sum problem (1).

4.2 Nonconvex Online (Expectation) Problem

In this online case, the following bounded variance assumption is needed. To simplify the presentation, let $\nabla f_i(x) := \nabla F(x, \zeta_i)$ denote a stochastic gradient for online problem (2).

Assumption 3 (Bounded Variance) *For $\forall x \in \mathbb{R}^d$, $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] := \mathbb{E}_{\zeta_i}[\|\nabla F(x, \zeta_i) - \nabla f(x)\|^2] \leq \sigma^2$, where $\sigma > 0$ is a constant.*

Note that this assumption is standard and necessary for this online case since the full gradients are not available (see e.g., [Ghadimi et al., 2016, Lei et al., 2017, Li and Li, 2018, Zhou et al., 2018b, Fang et al., 2018, Wang et al., 2018, Pham et al., 2019]). Moreover, we need to modify the full gradient computation step at the beginning of each epoch to a large batch stochastic gradient computation step (similar to [Lei et al., 2017, Li and Li, 2018]), i.e., change $v_{sm} \leftarrow \nabla f(x_{sm})$ (Line 8 of Algorithm 2) to

$$v_{sm} \leftarrow \frac{1}{B} \sum_{j \in I_B} \nabla f_j(x_{sm}), \tag{11}$$

where I_B are i.i.d. samples with $|I_B| = B$. We call B the batch size and b the minibatch size. Also, we need to change $\|\nabla f(x_{sm})\| \leq g_{\text{thres}}$ (Line 3 of Algorithm 2) to $\|v_{sm}\| \leq g_{\text{thres}}$.

Theorem 3 *Under Assumption 1 (i.e. (4)) and Assumption 3, let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \frac{\sqrt{5}-1}{2L}$, batch size $B = \frac{4\sigma^2}{\epsilon^2}$, minibatch size $b = \sqrt{B} = \frac{\sigma}{\epsilon}$ and epoch length $m = b$, SSRGD will find an ϵ-first-order stationary point in expectation using*

$$O\Big(\frac{\sigma^2}{\epsilon^2} + \frac{L\Delta f\sigma}{\epsilon^3}\Big)$$

stochastic gradients for nonconvex online problem (2).

For achieving a high probability for finding second-order stationary points (i.e., Theorem 4), we need a stronger version of Assumption 3 as in the following Assumption 4.

Assumption 4 (Bounded Variance) *For $\forall i, x$, $\|\nabla f_i(x) - \nabla f(x)\|^2 := \|\nabla F(x, \zeta_i) - \nabla f(x)\|^2 \leq \sigma^2$, where $\sigma > 0$ is a constant.*

We want to point out that Assumption 4 can be relaxed such that $\|\nabla f_i(x) - \nabla f(x)\|$ has sub-Gaussian tail, i.e., $\mathbb{E}[\exp(\lambda\|\nabla f_i(x) - \nabla f(x)\|)] \leq \exp(\lambda^2\sigma^2/2)$. Then it is sufficient for us to get a high probability bound by using Hoeffding bound on these sub-Gaussian variables. Note that Assumption 4 (or the relaxed sub-Gaussian version) is also standard in second-order stationary point finding algorithms (see e.g., [Allen-Zhu and Li, 2018, Zhou et al., 2018a, Fang et al., 2018]).

Theorem 4 *Under Assumption 1, 2 (i.e. (4) and (6)) and Assumption 4, let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \widetilde{O}(\frac{1}{L})$, batch size $B = \widetilde{O}(\frac{\sigma^2}{g_{\text{thres}}^2}) = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$, minibatch size $b = \sqrt{B} = \widetilde{O}(\frac{\sigma}{\epsilon})$, epoch length $m = b$, perturbation radius $r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big)$, threshold gradient $g_{\text{thres}} = \epsilon \leq \delta^2/\rho$, threshold function value $f_{\text{thres}} = \widetilde{O}(\frac{\delta^3}{\rho^2})$ and super epoch length $t_{\text{thres}} = \widetilde{O}(\frac{1}{\eta\delta})$, SSRGD will at least once get to an (ϵ, δ)-second-order stationary point with high probability using*

$$\widetilde{O}\Big(\frac{L\Delta f\sigma}{\epsilon^3} + \frac{\rho^2\Delta f\sigma^2}{\epsilon^2\delta^3} + \frac{L\rho^2\Delta f\sigma}{\epsilon\delta^4}\Big)$$

stochastic gradients for nonconvex online problem (2).

5 Overview of the Proofs

5.1 Finding First-order Stationary Points

In this section, we first show that why this stochastic recursive gradient descent algorithm can improve previous SVRG type algorithm (see e.g., [Li and Li, 2018, Ge et al., 2019]) from $n^{2/3}/\epsilon^2$ to $n^{1/2}/\epsilon^2$. Then we give a simple high-level proof for achieving the $n^{1/2}/\epsilon^2$ convergence result (i.e., Theorem 1).

Why it can be improved from $n^{2/3}/\epsilon^2$ to $n^{1/2}/\epsilon^2$: First, we need a key relation between $f(x_t)$ and $f(x_{t-1})$, where $x_t := x_{t-1} - \eta v_{t-1}$,

$$f(x_t) \leq f(x_{t-1}) - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 + \frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2, \tag{12}$$

where (12) holds since f has L-Lipschitz continuous gradient (Assumption 1). The details for obtaining (12) can be found in Appendix B.1 (see (25)).

Note that (12) is very meaningful and also very important for the proofs. The first term $-\frac{\eta}{2}\|\nabla f(x_{t-1})\|^2$ indicates that the function value will decrease a lot if the gradient $\nabla f(x_{t-1})$ is large. The second term $-\big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2$ indicates that the function value will also decrease a lot if the moving distance $x_t - x_{t-1}$ is large (note that here we require the step size $\eta \leq \frac{1}{L}$). The additional third term $+\frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2$ exists since we use v_{t-1} as a estimator of the actual gradient $\nabla f(x_{t-1})$ (i.e., $x_t := x_{t-1} - \eta v_{t-1}$). So it may increase the function value if v_{t-1} is a bad direction in this step.

To get an ϵ-first-order stationary point, we want to cancel the last two terms in (12). Firstly, we want to bound the last variance term. Recall the variance bound (see Equation (29) in [Li and Li, 2018]) for SVRG algorithm, i.e., estimator (9):

$$\mathbb{E}\big[\|\nabla f(x_{t-1}) - v_{t-1}\|^2\big] \leq \frac{L^2}{b}\mathbb{E}[\|x_{t-1} - \widetilde{x}\|^2]. \tag{13}$$

In order to connect the last two terms in (12), we use Young's inequality for the second term $\|x_t - x_{t-1}\|^2$, i.e., $-\|x_t - x_{t-1}\|^2 \leq \frac{1}{\alpha}\|x_{t-1} - \widetilde{x}\|^2 - \frac{1}{1+\alpha}\|x_t - \widetilde{x}\|^2$ (for any $\alpha > 0$). By plugging this Young's inequality and (13) into (12), we can cancel the last two terms in (12) by summing up (12) for each epoch s (i.e., iterations $sm + 1 \leq t \leq sm + m$), i.e., for each epoch s, we have (see Equation (35) in [Li and Li, 2018])

$$\mathbb{E}[f(x_{(s+1)m})] \leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{sm+m}\mathbb{E}[\|\nabla f(x_{j-1})\|^2]. \tag{14}$$

However, due to the Young's inequalities, we need to let $b \geq m^2$ to cancel the last two terms in (12) for obtaining (14), where b denotes minibatch size and m denotes the epoch length. According to (14), it is not hard to see that $\hat{x}$ is a ϵ-first-order stationary point in expectation (i.e., $\mathbb{E}[\|\nabla f(\hat{x})\|] \leq \epsilon$) if $\hat{x}$ is chosen uniformly randomly from $\{x_{t-1}\}_{t\in[T]}$ and the number of iterations $T = Sm = \frac{2(f(x_0)-f^*)}{\eta\epsilon^2}$. Note that for each iteration we need to compute $b + \frac{n}{m}$ stochastic gradients, where we amortize the full gradient computation of the beginning point of each epoch (n stochastic gradients) into each iteration in its epoch (i.e., n/m) for simple presentation. Thus, the convergence results is $T(b + \frac{n}{m}) \geq \frac{n^{2/3}}{\epsilon^2}$ since $b \geq m^2$, where equality holds if $b = m^2 = n^{2/3}$. Note that here we ignore the $f(x_0) - f^*$ and $\eta = O(1/L)$.

However, for stochastic recursive gradient descent estimator (10), we can bound the last variance term in (12) as (see Equation (31) in Appendix B.1):

$$\mathbb{E}\big[\|\nabla f(x_{t-1}) - v_{t-1}\|^2\big] \leq \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2]. \tag{15}$$

Now, the advantage of (15) compared with (13) is that it is already connected to the second term in (12), i.e., distances $\{x_t - x_{t-1}\}$. Thus we do not need an additional Young's inequality to transform the second term as before. This

makes the function value decrease bound tighter. Similarly, we plug (15) into (12) and sum it up for each epoch to cancel the last two terms in (12), i.e., for each epoch s, we have (see Equation (33) in Appendix B.1)

$$\mathbb{E}[f(x_{(s+1)m})] \leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{sm+m}\mathbb{E}[\|\nabla f(x_{j-1})\|^2]. \tag{16}$$

Compared with (14) (which requires $b \geq m^2$), here (16) only requires $b \geq m$ due to the tighter function value decrease bound since it does not involve the additional Young's inequalities.

High-level proof for achieving $n^{1/2}/\epsilon^2$ result: Now, according to (16), we can use the same above SVRG arguments to show the $n^{1/2}/\epsilon^2$ convergence result, i.e., $\hat{x}$ is a ϵ-first-order stationary point in expectation (i.e., $\mathbb{E}[\|\nabla f(\hat{x})\|] \leq \epsilon$) if $\hat{x}$ is chosen uniformly randomly from $\{x_{t-1}\}_{t\in[T]}$ and the number of iterations $T = Sm = \frac{2(f(x_0)-f^*)}{\eta\epsilon^2}$. Also, for each iteration, we compute $b + \frac{n}{m}$ stochastic gradients. The only difference is that now the convergence results is $T(b+\frac{n}{m}) \geq \frac{n^{1/2}}{\epsilon^2}$ since $b \geq m$ (rather than $b \geq m^2$), where equality holds if $b = m = n^{1/2}$. Here we ignore the $f(x_0) - f^*$ and $\eta = O(1/L)$. Thus our Theorem 1 is obtained.

5.2 Finding Second-order Stationary Points

In this section, we give the high-level proof ideas for finding a second-order stationary point with high probability. Note that our proof is different from that in [Ge et al., 2019] due to the different estimators (9) and (10). Ge et al. [2019] based on the estimator (9) and its first-order analysis in [Li and Li, 2018]. Here, our SSRGD uses the estimator (10). The difference of the first-order analysis between estimator (9) ([Li and Li, 2018]) and estimator (10) (this paper) is already discussed in the last Section 5.1. For the second-order analysis, since the estimator (10) in our SSRGD is more correlated than (9), thus we will use martingales to handle it. Besides, different relations will incur more differences in the detailed proof of second-order guarantee analysis than that of first-order guarantee analysis.

We divide the proof into two situations, i.e., large gradients and around saddle points. According to (16), a natural way to prove the convergence result is that the function value will decrease at a *desired rate* with high probability. Note that the amount for function value decrease is at most $\Delta f := f(x_0) - f^*$.

Large gradients: $\|\nabla f(x)\| \geq g_{\text{thres}}$
In this situation, due to the large gradients, it is sufficient to adjust the first-order analysis to show that the function value will decrease a lot in an epoch. Concretely, we want to show the function value decrease bound (16) holds with high probability. It is not hard to see that the desired rate of function value decrease is $O(\eta g_{\text{thres}}^2) = \widetilde{O}(\frac{\epsilon^2}{L})$ per each iteration (recall the parameters $g_{\text{thres}} = \epsilon$ and $\eta = \widetilde{O}(1/L)$ in our Theorem 2). Also note that we compute $b + \frac{n}{m} = 2\sqrt{n}$ stochastic gradients at each iteration (recall $m = b = \sqrt{n}$ in our Theorem 2). Here we amortize the full gradient computation of the beginning point of each epoch (n stochastic gradients) into each iteration in its epoch (i.e., n/m) for simple presentation (we will analyze this more rigorous in the detailed proofs in Appendices). Thus the number of stochastic gradient computation is at most $\widetilde{O}(\sqrt{n}\frac{\Delta f}{\epsilon^2/L}) = \widetilde{O}(\frac{L\Delta f\sqrt{n}}{\epsilon^2})$ for this large gradients situation.

For the proof, to show the function value decrease bound (16) holds with high probability, we need to show that the bound for variance term ($\|v_k - \nabla f(x_k)\|^2$) holds with high probability. Note that the estimator v_k defined in (10) is correlated with previous v_{k-1}. Fortunately, let $y_k := v_k - \nabla f(x_k)$, then it is not hard to see that $\{y_k\}$ is a martingale vector sequence with respect to a filtration $\{\mathscr{F}_k\}$ such that $\mathbb{E}[y_k|\mathscr{F}_{k-1}] = y_{k-1}$. Moreover, let $\{z_k\}$ denote the associated martingale difference sequence with respect to the filtration $\{\mathscr{F}_k\}$, i.e., $z_k := y_k - \mathbb{E}[y_k|\mathscr{F}_{k-1}] = y_k - y_{k-1}$ and $\mathbb{E}[z_k|\mathscr{F}_{k-1}] = 0$. Thus to bound the variance term $\|v_k - \nabla f(x_k)\|^2$ with high probability, it is sufficient to bound the martingale sequence $\{y_k\}$. This can be bounded with high probability by using the martingale Azuma-Hoeffding inequality. Note that in order to apply Azuma-Hoeffding inequality, we first need to use the Bernstein inequality to bound the associated difference sequence $\{z_k\}$. In sum, we will get the high probability function value decrease bound by applying these two inequalities (see (42) in Appendix B.1).

Note that (42) only guarantees function value decrease 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

3 of Algorithm 2). Similar to [Ge et al., 2019], we achieve this by stopping the epoch at a uniformly random point (see Line 16 of Algorithm 2). We use the following lemma to connect these two situations (large gradients and around saddle points):

Lemma 1 (Connection of Two Situations) *For any epoch s, let x_t be a point uniformly sampled from this epoch $\{x_j\}_{j=sm}^{(s+1)m}$. Moreover, let the step size $\eta \leq \frac{\sqrt{4C'^2+1}-1}{2C'^2L}$ (where $C' = O(\log \frac{dn}{\zeta}) = \widetilde{O}(1)$) and the minibatch size $b \geq m$, there are two cases:*

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

Moreover, $f(x_t) \leq f(x_{sm})$ holds with high probability no matter which case happens.

Note that if Case 2 happens, the function value already decreases a lot in this epoch s (as we already discussed at the beginning of this situation). Otherwise, Case 1 happens, we know the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 19 of Algorithm 2), then we know $\|\nabla f(x_{(s+1)m})\| = \|\nabla f(x_t)\| \leq g_{\rm thres}$. Then we will start a super epoch (see Line 3 of Algorithm 2). This corresponds to the following second situation around saddle points. Note that if $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) > -\delta$, this point $x_{(s+1)m}$ is already an (ϵ, δ)-second-order stationary point (recall $g_{\rm thres} = \epsilon$ in our Theorem 2).

Around saddle points: $\|\nabla f(\widetilde{x})\| \leq g_{\rm thres}$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$ at the initial point $\widetilde{x}$ of a certain super epoch
In this situation, we want to show that the function value will decrease a lot in a *super epoch* (instead of an epoch as in the first situation) with high probability by adding 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 6 in Algorithm 2). Following the classical widely used *two-point analysis* developed in [Jin et al., 2017], 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 $\nabla^2 f(\widetilde{x})$. Then we get two coupled sequences $\{x_t\}$ and $\{x_t'\}$ by running SSRGD update steps (Line 8–12 of Algorithm 2) with the same choice of minibatches (i.e., I_b's in Line 12 of Algorithm 2) for a super epoch. We will show that at least one of these two coupled sequences will decrease the function value a lot (escape the saddle point) with high probability, i.e.,

$$\exists t \leq t_{\rm thres}, \ \text{ such that } \ \max\{f(x_0) - f(x_t), f(x_0') - f(x_t')\} \geq 2f_{\rm thres}. \tag{17}$$

Similar to the classical argument in [Jin et al., 2017], according to (17), we know that in the random perturbation ball, the stuck points can only be a short interval in the e_1 direction, i.e., at least one of two points in the e_1 direction will 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 ζ' (see (48) in Appendix B.1). By a union bound (x_0 is not in a stuck region and (17) holds), with high probability, we have

$$\exists t \leq t_{\rm thres}\ , f(x_0) - f(x_t) \geq 2f_{\rm thres}. \tag{18}$$

Note that the initial point of this super epoch is $\widetilde{x}$ before the perturbation (see Line 6 of Algorithm 2), thus we also 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.,

$$\begin{aligned} f(x_0) &\leq f(\widetilde{x}) + \langle \nabla f(\widetilde{x}), x_0 - \widetilde{x}\rangle + \frac{L}{2}\|x_0 - \widetilde{x}\|^2 \\ &\leq f(\widetilde{x}) + g_{\rm thres} \cdot r + \frac{L}{2} r^2 \\ &= f(\widetilde{x}) + f_{\rm thres}, \end{aligned} \tag{19}$$

where the last inequality holds since the initial point $\widetilde{x}$ satisfying $\|\nabla f(\widetilde{x})\| \leq g_{\rm thres}$ and the perturbation radius is r , and the last equality holds by letting the perturbation radius r small enough. By combining (18) and (19), we obtain with high probability

$$f(\widetilde{x}) - f(x_t) = f(\widetilde{x}) - f(x_0) + f(x_0) - f(x_t) \geq -f_{\rm thres} + 2f_{\rm thres} = f_{\rm thres}. \tag{20}$$

Now, we can obtain the desired rate of function value decrease in this situation is $\frac{f_{\rm thres}}{t_{\rm thres}} = \widetilde{O}(\frac{\delta^3/\rho^2}{1/(\eta\delta)}) = \widetilde{O}(\frac{\delta^4}{L\rho^2})$ per each iteration (recall the parameters $f_{\rm thres} = \widetilde{O}(\delta^3/\rho^2)$, $t_{\rm thres} = \widetilde{O}(1/(\eta\delta))$ and $\eta = \widetilde{O}(1/L)$ in our Theorem 2). Same as before, we compute $b + \frac{n}{m} = 2\sqrt{n}$ stochastic gradients at each iteration (recall $m = b = \sqrt{n}$ in our Theorem 2). Thus the number of stochastic gradient computation is at most $\widetilde{O}(\sqrt{n}\frac{\Delta f}{\delta^4/(L\rho^2)}) = \widetilde{O}(\frac{L\rho^2\Delta f\sqrt{n}}{\delta^4})$ for this around saddle points situation.

Now, the remaining thing is to prove (17). It can be proved by contradiction. Assume the contrary, $f(x_0) - f(x_t) < 2f_{\rm thres}$ and $f(x_0') - f(x_t') < 2f_{\rm thres}$. 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.

Lemma 2 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 8–12 of Algorithm 2) from x_0. Moreover, let the step size $\eta \leq \frac{1}{2C'L}$ and minibatch size $b \geq m$, with probability $1-\zeta$, we have*

$$\forall t, \quad \|x_t - x_0\| \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L}}, \tag{21}$$

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

Then we will show that the stuck region is relatively small in the random perturbation ball, i.e., at least one of x_t and x_t' will go far away from their starting point x_0 and x_0' with high probability.

Lemma 3 (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 8–12 of Algorithm 2) with the same choice of minibatches (i.e., I_b's in Line 12) 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 $t_{\rm thres} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \leq \min\big(\frac{1}{8\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}, \frac{1}{4C_2 L\log t_{\rm thres}}\big) = \widetilde{O}(\frac{1}{L})$, minibatch size $b \geq m$ and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, then with probability $1-\zeta$, we have*

$$\exists T \leq t_{\rm thres}, \quad \max\{\|x_T - x_0\|, \|x_T' - x_0'\|\} \geq \frac{\delta}{C_1\rho}, \tag{22}$$

where $C_1 \geq \frac{20C_2}{\eta L}$ and $C_2 = O(\log\frac{dt_{\rm thres}}{\zeta}) = \widetilde{O}(1)$.

Based on these two lemmas, we are ready to show that (17) holds with high probability. Without loss of generality, we assume $\|x_T - x_0\| \geq \frac{\delta}{C_1\rho}$ in (22) (note that (21) holds for both $\{x_t\}$ and $\{x_t'\}$), then by plugging it into (21) to obtain

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

where the last inequality is due to $T \leq t_{\text{thres}}$ and the first equality holds by letting $C_1' = \frac{8C_1^2 \log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta C' L} = \widetilde{O}(1)$ (recall the parameters $f_{\text{thres}} = \widetilde{O}(\delta^3/\rho^2)$ and $\eta = \widetilde{O}(1/L)$ in our Theorem 2). Now, the high-level proof for this situation is finished.

In sum, the number of stochastic gradient computation is at most $\widetilde{O}(\frac{L\Delta_f\sqrt{n}}{\epsilon^2})$ for the large gradients situation and is at most $\widetilde{O}(\frac{L\rho^2\Delta_f\sqrt{n}}{\delta^4})$ for the around saddle points situation. Moreover, for the classical version where $\delta = \sqrt{\rho\epsilon}$ [Nesterov and Polyak, 2006, Jin et al., 2017], then $\widetilde{O}(\frac{L\rho^2\Delta_f\sqrt{n}}{\delta^4}) = \widetilde{O}(\frac{L\Delta_f\sqrt{n}}{\epsilon^2})$, i.e., both situations get the same stochastic gradient complexity. It also matches the convergence result for finding first-order stationary points (see our Theorem 1) if we ignore the logarithmic factor.

Finally, we point out that there is an extra term $\frac{\rho^2\Delta_f n}{\delta^3}$ in Theorem 2 beyond these two terms obtained from the above two situations. The reason is that we amortize the full gradient computation of the beginning point of each epoch (n stochastic gradients) into each iteration in its epoch (i.e., n/m) for simple presentation. We will analyze this more rigorous in the appendices, which incurs the term $\frac{\rho^2\Delta_f n}{\delta^3}$. For the more general online problem (2), the high-level proofs are almost the same as the finite-sum problem (1). The difference is that we need to use more concentration bounds in the detailed proofs since the full gradients are not available in online case.

6 Conclusion

In this paper, we focus on developing simple algorithm that has theoretical second-order guarantee for nonconvex finite-sum problems and more general nonconvex online problems. Concretely, we propose a simple perturbed version of stochastic recursive gradient descent algorithm (called SSRGD), which is as simple as its first-order stationary point finding algorithm and thus can be simply applied in practice for escaping saddle points (finding local minima). Moreover, the theoretical convergence results of SSRGD for finding second-order stationary points (local minima) almost match the theoretical results for finding first-order stationary points and these results are near-optimal as they almost match the lower bound.

Acknowledgments

The author would like to thank Rong Ge since the author learned a lot under his genuine guidance during the visit at Duke.

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

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

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.

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.

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. *arXiv preprint arXiv:1803.05999*, 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.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsłonline 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*, 2019.

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.

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 *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

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.

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.

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

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

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 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.

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, 2016.

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: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

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.

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. *arXiv preprint arXiv:1806.07811*, 2018b.

A Tools

In this appendix, we recall some classical concentration bounds for matrices and vectors.

Proposition 1 (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*

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

Define

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

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

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

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 martingale concentration bounds, i.e., Azuma-Hoffding inequality. Now, we will only write the vector version not repeat the more general matrix version.

Proposition 3 (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*

$$\|z_k\| = \|y_k - y_{k-1}\| \leq c_k \ \text{ almost surely.} \tag{23}$$

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 (23) with probability one sometime fails. Fortunately, the Azuma-Hoffding inequality also holds with a slackness if $\|z_k\| \leq c_k$ with high probability.

Proposition 4 (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*

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

Then, for all $t \geq 0$,

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

B Missing Proofs

In this appendix, we provide the detailed proofs for Theorem 1–4.

B.1 Proofs for Finite-sum Problem

In this section, we provide the detailed proofs for finite-sum problem (1) (i.e., Theorem 1–2).

First, we obtain the relation between $f(x_t)$ and $f(x_{t-1})$ as follows similar to [Li and Li, 2018, Ge et al., 2019], where we let $x_t := x_{t-1} - \eta v_{t-1}$ and $\bar{x}_t := x_{t-1} - \eta \nabla f(x_{t-1})$,

$$
\begin{aligned}
f(x_t) \leq & f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1} \rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 \qquad (24)\\
= & f(x_{t-1}) + \langle \nabla f(x_{t-1}) - v_{t-1}, x_t - x_{t-1} \rangle + \langle v_{t-1}, x_t - x_{t-1} \rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \langle \nabla f(x_{t-1}) - v_{t-1}, -\eta v_{t-1} \rangle - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \eta\langle \nabla f(x_{t-1}) - v_{t-1}, \nabla f(x_{t-1}) \rangle - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \frac{1}{\eta}\langle x_t - \bar{x}_t, x_{t-1} - \bar{x}_t \rangle - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
& - \frac{1}{2\eta}\big(\|x_t - \bar{x}_t\|^2 + \|x_{t-1} - \bar{x}_t\|^2 - \|x_t - x_{t-1}\|^2\big) \\
= & f(x_{t-1}) + \frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2, \qquad (25)
\end{aligned}
$$

where (24) holds since f has L-Lipschitz continuous gradient (Assumption 1). Now, we bound the variance term as follows, where we take expectations with the history:

$$
\begin{aligned}
& \mathbb{E}[\|v_{t-1} - \nabla f(x_{t-1})\|^2] \\
& = \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) + v_{t-2} - \nabla f(x_{t-1})\Big\|^2\Big] \\
& = \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big) + v_{t-2} - \nabla f(x_{t-2})\Big\|^2\Big] \\
& = \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big)\Big\|^2\Big] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \qquad (26)\\
& = \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\Big\|\big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\big) - \big(\nabla f(x_{t-1}) - \nabla f(x_{t-2})\big)\Big\|^2\Big] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \qquad (27)\\
& \leq \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_b}\Big\|\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-2})\Big\|^2\Big] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2] \qquad (28)\\
& \leq \frac{L^2}{b}\mathbb{E}[\|x_{t-1} - x_{t-2}\|^2] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2], \qquad (29)
\end{aligned}
$$

where (26) and (27) use the law of total expectation and $\mathbb{E}[\|x_1 + x_2 + \cdots + x_k\|^2] = \sum_{i=1}^k \mathbb{E}[\|x_i\|^2]$ if $x_1, x_2, \ldots, x_k$ are independent and of mean zero, (28) uses the fact $\mathbb{E}[\|x - \mathbb{E}x\|^2] \leq \mathbb{E}[\|x\|^2]$, and (29) holds due to the gradient Lipschitz Assumption 1.

Note that for $\mathbb{E}[\|v_{t-2}-\nabla f(x_{t-2})\|^2]$ in (29), we can reuse the same computation above. Thus we can sum up (29) from the beginning of this epoch sm to the point $t-1$,

$$\mathbb{E}[\|v_{t-1}-\nabla f(x_{t-1})\|^2] \le \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j-x_{j-1}\|^2] + \mathbb{E}[\|v_{sm}-\nabla f(x_{sm})\|^2] \tag{30}$$

$$\le \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j-x_{j-1}\|^2], \tag{31}$$

where (31) holds since we compute the full gradient at the beginning point of this epoch, i.e., $v_{sm}=\nabla f(x_{sm})$ (see Line 5 of Algorithm 1). Now, we take expectations for (25) and then sum it up from the beginning of this epoch s, i.e., iterations from sm to t, by plugging the variance (31) into them to get:

$$\begin{aligned}
\mathbb{E}[f(x_t)] &\le \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - (\frac{1}{2\eta}-\frac{L}{2})\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \\
&\quad + \frac{\eta L^2}{2b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\mathbb{E}[\|x_j-x_{j-1}\|^2] \\
&\le \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - (\frac{1}{2\eta}-\frac{L}{2})\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \\
&\quad + \frac{\eta L^2(t-1-sm)}{2b}\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \\
&\le \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - (\frac{1}{2\eta}-\frac{L}{2})\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \\
&\quad + \frac{\eta L^2}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j-x_{j-1}\|^2] \qquad (32) \\
&\le \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2], \qquad (33)
\end{aligned}$$

where (32) holds if the minibatch size $b \ge m$ (note that here $t \le (s+1)m$), and (33) holds if the step size $\eta \le \frac{\sqrt{5}-1}{2L}$.

Proof of Theorem 1. Let $b=m=\sqrt{n}$ and step size $\eta=\frac{\sqrt{5}-1}{2L}$, then (33) holds. Now, the proof is directly obtained by summing up (33) for all epochs $0 \le s \le S$ as follows:

$$\mathbb{E}[f(x_T)] \le \mathbb{E}[f(x_0)] - \frac{\eta}{2}\sum_{j=1}^{T}\mathbb{E}[\|\nabla f(x_{j-1})\|^2]$$

$$\mathbb{E}[\|\nabla f(\hat{x})\|] \le \sqrt{\mathbb{E}[\|\nabla f(\hat{x})\|^2]} \le \sqrt{\frac{2(f(x_0)-f^*)}{\eta T}} = \epsilon, \tag{34}$$

where (34) holds by choosing $\hat{x}$ uniformly from $\{x_{t-1}\}_{t\in[T]}$ and letting $Sm \le T = \frac{2(f(x_0)-f^*)}{\eta\epsilon^2} = O(\frac{L(f(x_0)-f^*)}{\epsilon^2})$. Note that the total number of computation of stochastic gradients equals to

$$Sn + Smb \le \Big\lceil\frac{T}{m}\Big\rceil n + Tb \le \Big(\frac{T}{\sqrt{n}}+1\Big)n + T\sqrt{n} = n + 2T\sqrt{n} = O\Big(n + \frac{L(f(x_0)-f^*)\sqrt{n}}{\epsilon^2}\Big).$$

□

B.1.1 Proof of Theorem 2

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 will decrease a lot since the gradients are large (see (33)). For the second situation (**around saddle points**), we will 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 (31). Then we use it to get a high probability bound of (33) for function value decrease. 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 9 of Algorithm 1), 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 difference sequence $\{z_k\}$. We use the Bernstein inequality to bound the differences as follows.

$$\begin{aligned} z_k = y_k - y_{k-1} &= v_k - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_k) - \nabla f_i(x_{k-1})\big) + v_{k-1} - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\Big). \end{aligned} \tag{35}$$

We define $u_i := \nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))$, and then we have

$$\|u_i\| = \|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\| \le 2\|x_k - x_{k-1}\|, \tag{36}$$

where the last inequality holds due to the gradient Lipschitz Assumption 1. Then, consider the variance term σ^2

$$\begin{aligned} \sigma^2 &= \sum_{i\in I_b}\mathbb{E}[\|u_i\|^2] \\ &= \sum_{i\in I_b}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\|^2] \\ &\le \sum_{i\in I_b}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] \\ &\le bL^2\|x_k - x_{k-1}\|^2, \end{aligned} \tag{37}$$

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 1. According to (36) and (37), we can bound the difference z_k by Bernstein inequality (Proposition 2) as

$$\begin{aligned} \mathbb{P}\Big\{\|z_k\| \ge \frac{t}{b}\Big\} &\le (d+1)\exp\Big(\frac{-t^2/2}{\sigma^2 + Rt/3}\Big) \\ &= (d+1)\exp\Big(\frac{-t^2/2}{bL^2\|x_k - x_{k-1}\|^2 + 2\|x_k - x_{k-1}\|t/3}\Big) \\ &= \zeta_k, \end{aligned}$$

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

$$\|z_k\| \le \frac{CL\|x_k - x_{k-1}\|}{\sqrt{b}} \quad \text{with probability } 1 - \zeta_k. \tag{38}$$

Now, we are ready to get a high probability bound for our original variance term (31) by using the martingale Azuma-Hoeffding inequality. Consider in a specifical 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

$$\|z_t\| \leq c_t = \frac{CL\|x_t - x_{t-1}\|}{\sqrt{b}} \text{ for all } sm+1 \leq t \leq sm+k \text{ with probability } 1-\zeta. \tag{39}$$

Then according to Azuma-Hoeffding inequality (Proposition 4) and noting that $\zeta_k = \zeta/m$, we have

$$\begin{aligned}\mathbb{P}\Big\{\big\|y_{sm+k} - y_{sm}\big\| \geq \beta\Big\} &\leq (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{t=sm+1}^{sm+k} c_t^2}\Big) + \zeta \\ &= 2\zeta,\end{aligned}$$

where the last equality holds by letting $\beta = \sqrt{8\sum_{t=sm+1}^{sm+k} c_t^2 \log\frac{d}{\zeta}} = \frac{C'L\sqrt{\sum_{t=sm+1}^{sm+k}\|x_t - x_{t-1}\|^2}}{\sqrt{b}}$, where $C' = O(C\sqrt{\log\frac{d}{\zeta}}) = \widetilde{O}(1)$. Recall that $y_k := v_k - \nabla f(x_k)$ and at the beginning point of this epoch $y_{sm} = 0$ due to $v_{sm} = \nabla f(x_{sm})$ (see Line 5 of Algorithm 1), thus we have

$$\|v_{t-1} - \nabla f(x_{t-1})\| = \|y_{t-1}\| \leq \frac{C'L\sqrt{\sum_{j=sm+1}^{t-1}\|x_j - x_{j-1}\|^2}}{\sqrt{b}} \tag{40}$$

with probability $1-2\zeta$, where t belongs to $[sm+1, (s+1)m]$.

Now, we use this high probability version (40) instead of the expectation one (31) to obtain the high probability bound for function value decrease (see (33)). We sum up (25) from the beginning of this epoch s, i.e., iterations from sm to t, by plugging (40) into them to get:

$$\begin{aligned}
f(x_t) &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta}{2}\sum_{k=sm+1}^{t-1}\frac{C'^2L^2\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2}{b} \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2}{2b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2(t-1-sm)}{2b}\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2L^2}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \qquad (41) \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2, \qquad (42)
\end{aligned}$$

where (41) holds if the minibatch size $b \geq m$ (note that here $t \leq (s+1)m$), and (42) holds if the step size $\eta \leq \frac{\sqrt{4C'^2+1}-1}{2C'^2L}$.

Note that (42) only guarantees function value decrease 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 3 of Algorithm 2). Similar to [Ge et al., 2019], we achieve this by stopping the epoch at a uniformly random point (see Line 16 of Algorithm 2).

Now we recall Lemma 1 to connect these two situations (large gradients and around saddle points):

Lemma 1 (Connection of Two Situations) *For any epoch s, let x_t be a point uniformly sampled from this epoch $\{x_j\}_{j=sm}^{(s+1)m}$. Moreover, let the step size $\eta \leq \frac{\sqrt{4C'^2+1}-1}{2C'^2L}$ (where $C' = O(\log \frac{dn}{\zeta}) = \widetilde{O}(1)$) and the minibatch size $b \geq m$, there are two cases:*

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

Moreover, $f(x_t) \leq f(x_{sm})$ holds with high probability no matter which case happens.

Proof of Lemma 1. There are two cases in this epoch:

1. If at least half of points of in this epoch $\{x_j\}_{j=sm}^{(s+1)m}$ have gradient norm no larger than $g_{\rm thres}$, then it is easy to see that a uniformly sampled point x_t has gradient norm $\|\nabla f(x_t)\| \leq g_{\rm thres}$ with probability at least $1/2$.
2. Otherwise, at least half of points have gradient norm larger than $g_{\rm thres}$. Then, as long as the sampled point x_t falls into the last quarter of $\{x_j\}_{j=sm}^{(s+1)m}$, we know $\sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2 \geq \frac{m g_{\rm thres}^2}{4}$. This holds with probability at least $1/4$ since x_t is uniformly sampled. Then combining with (42), i.e., $f(x_{sm}) - f(x_t) \geq \frac{\eta}{2}\sum_{j=sm+1}^{t} \|\nabla f(x_{j-1})\|^2$, we obtain the function value decrease $f(x_{sm}) - f(x_t) \geq \frac{\eta m g_{\rm thres}^2}{8}$. Note that (42) holds with high probability if we choose the minibatch size $b \geq m$ and the step size $\eta \leq \frac{\sqrt{4C'^2+1}-1}{2C'^2L}$. By a union bound, the function value decrease $f(x_{sm}) - f(x_t) \geq \frac{\eta m g_{\rm thres}^2}{8}$ with probability at least $1/5$.

Again according to (42), $f(x_t) \leq f(x_{sm})$ always holds with high probability. □

Note that if Case 2 happens, the function value already decreases a lot in this epoch s (corresponding to the first situation large gradients). Otherwise, Case 1 happens, we know the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 19 of Algorithm 2), then we know $\|\nabla f(x_{(s+1)m})\| = \|\nabla f(x_t)\| \leq g_{\rm thres}$. Then we will start a super epoch (corresponding to the second situation around saddle points). Note that if $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) > -\delta$, this point $x_{(s+1)m}$ is already an (ϵ, δ)-second-order stationary point (recall that $g_{\rm thres} = \epsilon$ in our Theorem 2).

Around Saddle Points $\|\nabla f(\widetilde{x})\| \leq g_{\rm thres}$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$: In this situation, we will show that the function value decreases a lot in a *super epoch* (instead of an epoch as in the first situation) with high probability by adding 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 6 in Algorithm 2). Following the classical widely used *two-point analysis* developed in [Jin et al., 2017], 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 8–12 of Algorithm 2) with the same choice of minibatches (i.e., I_b's in Line 12 of Algorithm 2) for a super epoch. We will show that at least one of these two coupled sequences will decrease the function value a lot (escape the saddle point), i.e.,

$$\exists t \leq t_{\rm thres}, \ \text{ such that } \ \max\{f(x_0) - f(x_t), f(x_0') - f(x_t')\} \geq 2f_{\rm thres}. \tag{43}$$

We will prove (43) by contradiction. Assume the contrary, $f(x_0) - f(x_t) < 2f_{\rm thres}$ and $f(x_0') - f(x_t') < 2f_{\rm thres}$. 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 will show that the stuck region is relatively small in the random perturbation ball, i.e., at least one of x_t and x_t' will go far away from their starting point x_0 and x_0' with high probability. Thus there is a contradiction. We recall these two lemmas here and their proofs are deferred to the end of this section.

Lemma 2 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 8–12 of Algorithm 2) from x_0. Moreover, let the step size $\eta \leq \frac{1}{2C'L}$ and minibatch size $b \geq m$, with probability $1-\zeta$, we have*

$$\forall t, \ \ \|x_t - x_0\| \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L}}, \tag{44}$$

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

Lemma 3 (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 8–12 of Algorithm 2) with the same choice of minibatches (i.e., I_b's in Line 12) 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 $t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \leq \min\big(\frac{1}{8\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}, \frac{1}{4C_2 L \log t_{\text{thres}}}\big) = \widetilde{O}(\frac{1}{L})$, minibatch size $b \geq m$ and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, then with probability $1-\zeta$, we have*

$$\exists T \leq t_{\text{thres}}, \quad \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \geq \frac{\delta}{C_1\rho}, \tag{45}$$

where $C_1 \geq \frac{20C_2}{\eta L}$ and $C_2 = O(\log \frac{dt_{\text{thres}}}{\zeta}) = \widetilde{O}(1)$.

Based on these two lemmas, we are ready to show that (43) holds with high probability. Without loss of generality, we assume $\|x_T - x_0\| \geq \frac{\delta}{C_1\rho}$ in (45) (note that (44) holds for both $\{x_t\}$ and $\{x'_t\}$), then plugging it into (44) to obtain

$$\begin{aligned}
\sqrt{\frac{4T(f(x_0) - f(x_T))}{C'L}} &\geq \frac{\delta}{C_1\rho} \\
f(x_0) - f(x_T) &\geq \frac{C'L\delta^2}{4C_1^2\rho^2 T} \\
&\geq \frac{\eta C'L\delta^3}{8C_1^2\rho^2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})} \\
&= \frac{\delta^3}{C_1'\rho^2} \\
&= 2f_{\text{thres}},
\end{aligned} \tag{46}$$

where the last inequality is due to $T \leq t_{\text{thres}}$ and (46) holds by letting $C_1' = \frac{8C_1^2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta C'L}$. Thus, we already prove that at least one of sequences $\{x_t\}$ and $\{x'_t\}$ escapes the saddle point with high probability, i.e.,

$$\exists T \leq t_{\text{thres}} \ , \max\{f(x_0) - f(x_T), f(x'_0) - f(x'_T)\} \geq 2f_{\text{thres}}, \tag{47}$$

if their starting points x_0 and x'_0 satisfying $w_0 := x_0 - x'_0 = r_0 e_1$, where $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. Similar to the classical argument in [Jin et al., 2017], we know that in the random perturbation ball, the stuck points can only be a short interval in the e_1 direction, i.e., at least one of two points in the e_1 direction will 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

$$\frac{r_0 V_{d-1}(r)}{V_d(r)} = \frac{r_0\Gamma(\frac{d}{2}+1)}{\sqrt{\pi}r\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{r_0}{\sqrt{\pi}r}\big(\frac{d}{2}+1\big)^{1/2} \leq \frac{r_0\sqrt{d}}{r} = \zeta', \tag{48}$$

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 (48) and (46) (holds with high probability if x_0 is not in a stuck region), we know

$$f(x_0) - f(x_T) \geq 2f_{\text{thres}} = \frac{\delta^3}{C_1'\rho^2} \tag{49}$$

with high probability. Note that the initial point of this super epoch is $\widetilde{x}$ before the perturbation (see Line 6 of Algorithm 2), 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.,

$$\begin{aligned} f(x_0) &\leq f(\widetilde{x}) + \langle \nabla f(\widetilde{x}), x_0 - \widetilde{x}\rangle + \frac{L}{2}\|x_0 - \widetilde{x}\|^2 \\ &\leq f(\widetilde{x}) + \|\nabla f(\widetilde{x})\|\|x_0 - \widetilde{x}\| + \frac{L}{2}\|x_0 - \widetilde{x}\|^2 \\ &\leq f(\widetilde{x}) + g_{\text{thres}} \cdot r + \frac{L}{2}r^2 \\ &\leq f(\widetilde{x}) + \frac{\delta^3}{2C_1'\rho^2} \\ &= f(\widetilde{x}) + f_{\text{thres}}, \end{aligned} \tag{50}$$

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

Now we combine with (49) and (50) to obtain with high probability

$$f(\widetilde{x}) - f(x_T) = f(\widetilde{x}) - f(x_0) + f(x_0) - f(x_T) \geq -f_{\text{thres}} + 2f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}. \tag{51}$$

Thus we have finished the proof for the second situation (around saddle points), i.e., we show that the function value decrease a lot ($f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}$) in a *super epoch* (recall that $T \leq t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$) by adding a random perturbation $\xi \sim \mathbb{B}_0(r)$ at the initial point $\widetilde{x}$.

Combing these two situations (large gradients and around saddle points) to prove Theorem 2: First, we recall Theorem 2 here since we want to recall the parameter setting.

Theorem 2 *Under Assumption 1 and 2 (i.e. (3) and (5)), let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \widetilde{O}(\frac{1}{L})$, epoch length $m = \sqrt{n}$, minibatch size $b = \sqrt{n}$, perturbation radius $r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big)$, threshold gradient $g_{\text{thres}} = \epsilon$, threshold function value $f_{\text{thres}} = \widetilde{O}(\frac{\delta^3}{\rho^2})$ and super epoch length $t_{\text{thres}} = \widetilde{O}(\frac{1}{\eta\delta})$, SSRGD will at least once get to an (ϵ, δ)-second-order stationary point with high probability using*

$$\widetilde{O}\Big(\frac{L\Delta f\sqrt{n}}{\epsilon^2} + \frac{L\rho^2\Delta f\sqrt{n}}{\delta^4} + \frac{\rho^2\Delta f n}{\delta^3}\Big)$$

stochastic gradients for nonconvex finite-sum problem (1).

Proof of Theorem 2. Now, we prove this 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 g_{thres} (Case 2 of Lemma 1);

2. *Wasted epoch*: If at least half of points in this epoch have gradient norm no larger than g_{thres} and the starting point of the next epoch has gradient norm larger than g_{thres} (it means that this epoch does not guarantee decreasing the function value a lot as the large gradients situation, also it cannot connect to the second super epoch situation since the starting point of the next epoch has gradient norm larger than g_{thres});

3. *Type-2 useful super epoch*: If at least half of points in this epoch have gradient norm no larger than g_{thres} and the starting point of the next epoch (here we denote this point as $\widetilde{x}$) has gradient norm no larger than g_{thres} (i.e., $\|\nabla f(\widetilde{x})\| \leq g_{\text{thres}}$) (Case 1 of Lemma 1), according to Line 3 of Algorithm 2, we will 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 1 and Line 16 of Algorithm 2) and different wasted epoch are independent. Thus, with high probability, there are at most $\widetilde{O}(1)$ 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. Recall that $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. Also recall that the function value always does not increase with high probability (see Lemma 1).

For type-1 useful epoch, according to Case 2 of Lemma 1, we know that the function value decreases at least $\frac{\eta m g_{\text{thres}}^2}{8}$ with probability at least $1/5$. Using a standard concentration, we know that with high probability N_1 type-1 useful epochs will decrease the function value at least $\frac{\eta m g_{\text{thres}}^2 N_1}{80}$, note that the function value can decrease at most Δf. So $\frac{\eta m g_{\text{thres}}^2 N_1}{80} \leq \Delta f$, we get $N_1 \leq \frac{80\Delta f}{\eta m g_{\text{thres}}^2}$.

For type-2 useful super epoch, first we know that the starting point of the super epoch $\widetilde{x}$ has gradient norm $\|\nabla f(\widetilde{x})\| \leq g_{\text{thres}}$. Now if $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \geq -\delta$, then $\widetilde{x}$ is already a (ϵ, δ)-second-order stationary point. Otherwise, $\|\nabla f(\widetilde{x})\| \leq g_{\text{thres}}$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$, this is exactly our second situation (around saddle points). According to (51), 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 high probability. Similar to type-1 useful epoch, we know $N_2 \leq \frac{C_1''\rho^2\Delta f}{\delta^3}$ by a union bound (so we change C_1' to C_1'', anyway we also have $C_1'' = \widetilde{O}(1)$).

Now, we are ready to compute the convergence results to finish the proof for Theorem 2.

$$
\begin{aligned}
&N_1(\widetilde{O}(1)n + n + mb) + N_2(\widetilde{O}(1)n + \lceil\frac{t_{\text{thres}}}{m}\rceil n + t_{\text{thres}} b) \\
&\leq \widetilde{O}\Big(\frac{\Delta f n}{\eta m g_{\text{thres}}^2} + \frac{\rho^2\Delta f}{\delta^3}(n + \frac{\sqrt{n}}{\eta\delta})\Big) \\
&\leq \widetilde{O}\Big(\frac{L\Delta f\sqrt{n}}{\epsilon^2} + \frac{L\rho^2\Delta f\sqrt{n}}{\delta^4} + \frac{\rho^2\Delta f n}{\delta^3}\Big)
\end{aligned} \tag{52}
$$

□

Now, the only remaining thing is to prove Lemma 2 and 3. We provide these two proofs as follows.

Lemma 2 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 8–12 of Algorithm 2) from x_0. Moreover, let the step size $\eta \leq \frac{1}{2C'L}$ and minibatch size $b \geq m$, with probability $1-\zeta$, we have*

$$
\forall t,\ \ \|x_t - x_0\| \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{C'L}},
$$

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

Proof of Lemma 2. First, we assume the variance bound (40) holds for all $0 \leq j \leq t-1$ (this is true with high probability using a union bound by letting $C' = O(\log\frac{dt}{\zeta})$). Then, according to (41), we know for any $\tau \leq t$ in some epoch s

$$
\begin{aligned}
f(x_\tau) &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{\tau}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2 L^2}{2}\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \big(\frac{1}{2\eta} - \frac{L}{2} - \frac{\eta C'^2 L^2}{2}\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 \\
&\leq f(x_{sm}) - \frac{C'L}{4}\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2,
\end{aligned} \tag{53}
$$

where the last inequality holds since the step size $\eta \leq \frac{1}{2C'L}$ and assuming $C' \geq 1$. Now, we sum up (53) for all epochs before iteration t,

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

Then, the proof is finished as

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

□

Lemma 3 (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 8–12 of Algorithm 2) with the same choice of minibatches (i.e., I_b's in Line 12) 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 $t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \leq \min\big(\frac{1}{8\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}, \frac{1}{4C_2 L\log t_{\text{thres}}}\big) = \widetilde{O}(\frac{1}{L})$, minibatch size $b \geq m$ and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, then with probability $1-\zeta$, we have*

$$\exists T \leq t_{\text{thres}}, \quad \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \geq \frac{\delta}{C_1\rho},$$

where $C_1 \geq \frac{20C_2}{\eta L}$ and $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta}) = \widetilde{O}(1)$.

Proof of Lemma 3. We prove this lemma by contradiction. Assume the contrary,

$$\forall t \leq t_{\text{thres}} \ , \|x_t - x_0\| \leq \frac{\delta}{C_1\rho} \ \text{ and } \ \|x'_t - x'_0\| \leq \frac{\delta}{C_1\rho} \tag{54}$$

We will show that the distance between these two coupled sequences $w_t := x_t - x'_t$ will grow 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 (54) and the perturbation radius r. It is not hard to see that the exponential increase will break this upper bound, thus we get a contradiction.

In the following, we prove the exponential increase of w_t by induction. First, we need the expression of w_t (recall that $x_t = x_{t-1} - \eta v_{t-1}$ (see Line 11 of Algorithm 2)):

$$\begin{aligned}
w_t &= w_{t-1} - \eta(v_{t-1} - v'_{t-1}) \\
&= w_{t-1} - \eta\big(\nabla f(x_{t-1}) - \nabla f(x'_{t-1}) + v_{t-1} - \nabla f(x_{t-1}) - v'_{t-1} + \nabla f(x'_{t-1})\big) \\
&= w_{t-1} - \eta\Big(\int_0^1 \nabla^2 f(x'_{t-1} + \theta(x_{t-1} - x'_{t-1}))d\theta(x_{t-1} - x'_{t-1}) + v_{t-1} - \nabla f(x_{t-1}) - v'_{t-1} + \nabla f(x'_{t-1})\Big) \\
&= (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + y_{t-1}) \\
&= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau + y_\tau)
\end{aligned} \tag{55}$$

where $\Delta_\tau := \int_0^1(\nabla^2 f(x'_\tau + \theta(x_\tau - x'_\tau)) - \mathcal{H})d\theta$ and $y_\tau := v_\tau - \nabla f(x_\tau) - v'_\tau + \nabla f(x'_\tau)$. Note that the first term of (55) is in the e_1 direction and is exponential with respect to t, i.e., $(1+\eta\gamma)^t r_0 e_1$, where $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$. To prove the exponential increase of w_t, it is sufficient to show that the first term of (55) will dominate the second term. We inductively prove the following two bounds

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

First, check the base case $t=0$, $\|w_0\| = \|r_0 e_1\| = r_0$ and $\|y_0\| = \|v_0 - \nabla f(x_0) - v_0' + \nabla f(x_0')\| = \|\nabla f(x_0) - \nabla f(x_0) - \nabla f(x_0') + \nabla f(x_0')\| = 0$. Assume they hold for all $\tau \le t-1$, we now prove they hold for t one by one. For Bound 1, it is enough to show the second term of (55) is dominated by half of the first term.

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau)\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|\Delta_\tau\|\|w_\tau\| \\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\|\Delta_\tau\| && (56)\\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\rho D_\tau^x && (57)\\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 t\rho\big(\frac{\delta}{C_1\rho}+r\big) && (58)\\
&\le \frac{3}{C_1}\eta\delta t(1+\eta\gamma)^{t-1} r_0 && (59)\\
&\le \frac{6\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{C_1}(1+\eta\gamma)^{t-1} r_0 && (60)\\
&\le \frac{1}{4}(1+\eta\gamma)^t r_0, && (61)
\end{aligned}$$

where (56) uses the induction for w_τ with $\tau \le t-1$, (57) uses the definition $D_\tau^x := \max\{\|x_\tau - \widetilde{x}\|, \|x_\tau' - \widetilde{x}\|\}$, (58) follows from $\|x_t - \widetilde{x}\| \le \|x_t - x_0\| + \|x_0 - \widetilde{x}\| = \frac{\delta}{C_1\rho} + r$ due to (54) and the perturbation radius r, (59) holds by letting the perturbation radius $r \le \frac{\delta}{C_1\rho}$, (60) holds since $t \le t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, and (61) holds by letting $C_1 \ge 24\log(\frac{8\delta\sqrt{d}}{\rho\zeta' r})$.

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau} y_\tau\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|y_\tau\| \\
&\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\eta\gamma L(1+\eta\gamma)^{\tau} r_0 && (62)\\
&= \eta\eta\gamma L t(1+\eta\gamma)^{t-1} r_0 \\
&\le \eta\eta\gamma L\frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}(1+\eta\gamma)^{t-1} r_0 && (63)\\
&\le 2\eta\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L(1+\eta\gamma)^{t-1} r_0 && (64)\\
&\le \frac{1}{4}(1+\eta\gamma)^t r_0, && (65)
\end{aligned}$$

where (62) uses the induction for y_τ with $\tau \le t-1$, (63) holds since $t \le t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, (64) holds $\gamma \ge \delta$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$), and (65) holds by letting $\eta \le \frac{1}{8\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}$.

Combining (61) and (65), we proved the second term of (55) is dominated by half of the first term. Note that the

first term of (55) is $\|(I-\eta\mathcal{H})^t w_0\| = (1+\eta\gamma)^t r_0$. Thus, we have

$$\frac{1}{2}(1+\eta\gamma)^t r_0 \le \|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0 \tag{66}$$

Now, the remaining thing is to prove the second bound $\|y_t\| \le \eta\gamma L(1+\eta\gamma)^t r_0$. First, we write the concrete expression of y_t:

$$\begin{aligned}
y_t &= v_t - \nabla f(x_t) - v_t' + \nabla f(x_t') \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1} - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t') - \nabla f_i(x_{t-1}')\big) - v_{t-1}' + \nabla f(x_t') \qquad (67)\\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + \nabla f(x_{t-1}) - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t') - \nabla f_i(x_{t-1}')\big) - \nabla f(x_{t-1}') + \nabla f(x_t') \\
&\quad + v_{t-1} - \nabla f(x_{t-1}) - v_{t-1}' + \nabla f(x_{t-1}') \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_t') - \nabla f_i(x_{t-1}) + \nabla f_i(x_{t-1}')\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x_t') - \nabla f(x_{t-1}) + \nabla f(x_{t-1}')\big) + y_{t-1},
\end{aligned}$$

where (67) due to the definition of the estimator v_t (see Line 12 of Algorithm 2). 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 will apply the Azuma-Hoeffding inequalities to get an upper bound for $\|y_t\|$ and then we prove $\|y_t\| \le \eta\gamma L(1+\eta\gamma)^t r_0$ based on that upper bound. In order to apply the Azuma-Hoeffding inequalities for martingale sequence $\|y_t\|$, we first need to bound the difference sequence $\{z_t\}$. We use the Bernstein inequality to bound the differences as follows.

$$\begin{aligned}
z_t = y_t - y_{t-1} &= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_t') - \nabla f_i(x_{t-1}) + \nabla f_i(x_{t-1}')\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x_t') - \nabla f(x_{t-1}) + \nabla f(x_{t-1}')\big) \\
&= \frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) \\
&\qquad - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)\Big). \qquad (68)
\end{aligned}$$

We define $u_i := \big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)$, and then we have

$$\begin{aligned}
\|u_i\| &= \|\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)\| \\
&\le \Big\| \int_0^1 \nabla^2 f_i(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') - \int_0^1 \nabla^2 f_i(x_{t-1}' + \theta(x_{t-1} - x_{t-1}'))d\theta(x_{t-1} - x_{t-1}') \\
&\quad - \int_0^1 \nabla^2 f(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') + \int_0^1 \nabla^2 f(x_{t-1}' + \theta(x_{t-1} - x_{t-1}'))d\theta(x_{t-1} - x_{t-1}')\Big\| \\
&= \|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1}) - (\mathcal{H} w_t + \Delta_t w_t) + (\mathcal{H} w_{t-1} + \Delta_{t-1} w_{t-1})\| \qquad (69)\\
&\le \|(\mathcal{H}_i - \mathcal{H})(w_t - w_{t-1})\| + \|(\Delta_t^i - \Delta_t)w_t - (\Delta_{t-1}^i - \Delta_{t-1})w_{t-1}\| \\
&\le 2L\|w_t - w_{t-1}\| + 2\rho D_t^x \|w_t\| + 2\rho D_{t-1}^x \|w_{t-1}\|, \qquad (70)
\end{aligned}$$

where (69) 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 Lipschitz Assumption 1 and Hessian Lipschitz Assumption 2 (recall $D_t^x := \max\{\|x_t - \widetilde{x}\|, \|x_t' - \widetilde{x}\|\}$). Then, consider the variance term σ^2

$$\begin{aligned}
\sigma^2 &= \sum_{i \in I_b} \mathbb{E}[\|u_i\|^2] \\
&\leq \sum_{i \in I_b} \mathbb{E}[\|\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big)\|^2] \\
&= \sum_{i \in I_b} \mathbb{E}[\|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1})\|^2] \\
&\leq b(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2, \qquad (71)
\end{aligned}$$

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 Lipschitz Assumption 1 and Hessian Lipschitz Assumption 2. According to (70) and (71), we can bound the difference z_k by Bernstein inequality (Proposition 2) 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 $\sigma^2 = b(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2$)

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

where the last equality holds by letting $\alpha = C_4\sqrt{b}(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)$, where $C_4 = O(\log \frac{d}{\zeta_k}) = \widetilde{O}(1)$.

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

$$\|z_k\| \leq c_k = \frac{C_4(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)}{\sqrt{b}} \quad \text{with probability } 1 - \zeta_k. \qquad (72)$$

Now, we are ready to get 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 $(s+1)m$. According to Azuma-Hoeffding inequality (Proposition 4) and letting $\zeta_k = \zeta/m$, we have

$$\begin{aligned}
\mathbb{P}\Big\{\|y_t - y_{sm}\| \geq \beta\Big\} &\leq (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{k=sm+1}^t c_k^2}\Big) + \zeta \\
&= 2\zeta,
\end{aligned}$$

where the last equality is due to $\beta = \sqrt{8\sum_{k=sm+1}^t c_k^2 \log\frac{d}{\zeta}} = \frac{C_3\sqrt{\sum_{k=sm+1}^t (L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2}}{\sqrt{b}}$, where $C_3 = O(C_4\sqrt{\log\frac{d}{\zeta}}) = \widetilde{O}(1)$. 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 $v_{sm}' = \nabla f(x_{sm}')$ (see Line 5 of Algorithm 1), thus we have

$$\|y_t\| = \|y_t - y_{sm}\| \leq \frac{C_3\sqrt{\sum_{k=sm+1}^t (L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2}}{\sqrt{b}} \qquad (73)$$

with probability $1 - 2\zeta$, where t belongs to $[sm+1, (s+1)m]$. Note that we can further relax the parameter C_3 in (73) to $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta})$ (see (74)) for making sure the above arguments hold with probability $1 - \zeta$ for all $t \leq t_{\text{thres}}$ by using a union bound for ζ_t's:

$$\|y_t\| = \|y_t - y_{sm}\| \leq \frac{C_2\sqrt{\sum_{k=sm+1}^t (L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2}}{\sqrt{b}}. \qquad (74)$$

Now, we will show how to bound the right-hand-side of (74) to finish the proof, i.e., prove the remaining second bound $\|y_t\| \le \eta\gamma L(1+\eta\gamma)^t r_0$.

First, we show that the last two terms in the right-hand-side of (74) can be bounded as

$$\begin{aligned}\rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\| &\le \rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^t r_0 + \rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1}r_0\\ &\le 3\rho\big(\frac{\delta}{C_1\rho}+r\big)(1+\eta\gamma)^t r_0\\ &\le \frac{6\delta}{C_1}(1+\eta\gamma)^t r_0, \end{aligned} \tag{75}$$

where the first inequality follows from the induction of $\|w_{t-1}\| \le \frac{3}{2}(1+\eta\gamma)^{t-1}r_0$ and the already proved $\|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0$ in (66), and the last inequality holds by letting the perturbation radius $r \le \frac{\delta}{C_1\rho}$.

Now, we show that the first term of right-hand-side of (74) can be bounded as

$$\begin{aligned} L\|w_t - w_{t-1}\| &= L\Big\| -\eta\mathcal{H}(I-\eta\mathcal{H})^{t-1}w_0 - \eta\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau) + \eta(\Delta_{t-1}w_{t-1}+y_{t-1})\Big\|\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\Big\|\eta\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau)\Big\| + L\|\eta(\Delta_{t-1}w_{t-1}+y_{t-1})\|\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\Big\|\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\Big\|\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\|\\ &\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\| \qquad (76)\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\sum_{\tau=0}^{t-2}\frac{1}{t-1-\tau}\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\|\\ &\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\| \qquad (77)\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\|\\ &\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\|\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\|\\ &\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1}r_0 + L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 \qquad (78)\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\Big(\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-2}r_0 + \eta\gamma L(1+\eta\gamma)^{t-2}r_0\Big)\\ &\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1}r_0 + L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 \qquad (79)\\ &\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\Big(\frac{3\delta}{C_1}(1+\eta\gamma)^{t-2}r_0 + \eta\gamma L(1+\eta\gamma)^{t-2}r_0\Big)\\ &\quad + \frac{3L\eta\delta}{C_1}(1+\eta\gamma)^{t-1}r_0 + L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 \qquad (80)\\ &\le \Big(\frac{4}{C_1}\log t + 2L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0, \qquad (81)\end{aligned}$$

where the first equality follows from (55), (76) holds from the following (82),

$$\|\Delta_t\| \leq \rho D_t^x \leq \rho\big(\frac{\delta}{C_1\rho} + r\big), \tag{82}$$

where (82) holds due to Hessian Lipschitz Assumption 2, (54) and the perturbation radius r (recall that $\Delta_t := \int_0^1(\nabla^2 f(x_t' + \theta(x_t - x_t')) - \mathcal{H})d\theta$, $\mathcal{H} := \nabla^2 f(\widetilde{x})$ and $D_t^x := \max\{\|x_t - \widetilde{x}\|, \|x_t' - \widetilde{x}\|\}$), (77) holds due to $\|\eta\mathcal{H}(I - \eta\mathcal{H})^t\| \leq \frac{1}{t+1}$, (78) holds by plugging the induction $\|w_{t-1}\| \leq \frac{3}{2}(1+\eta\gamma)^{t-1}r_0$ and $\|y_{t-1}\| \leq \eta\gamma L(1+\eta\gamma)^{t-1}r_0$, (79) follows from (82), the induction $\|w_k\| \leq \frac{3}{2}(1+\eta\gamma)^k r_0$ and $\|y_k\| \leq \eta\gamma L(1+\eta\gamma)^k r_0$ (hold for all $k \leq t-1$), (80) holds by letting the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, and the last inequality holds due to $\gamma \geq \delta$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$).

By plugging (75) and (81) into (74), we have

$$\begin{aligned}
\|y_t\| &\leq C_2\left(\frac{6\delta}{C_1}(1+\eta\gamma)^t r_0 + \Big(\frac{4}{C_1}\log t + 2L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0\right) \\
&\leq C_2\Big(\frac{6}{C_1\eta L} + \frac{4}{C_1}\log t + 2L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0 \\
&\leq \eta\gamma L(1+\eta\gamma)^t r_0,
\end{aligned} \tag{83}$$

where the second inequality holds due to $\gamma \geq \delta$, and the last inequality holds by letting $C_1 \geq \frac{20C_2}{\eta L}$ and $\eta \leq \frac{1}{4C_2 L\log t}$. Recall that $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta})$ is enough to let the arguments in this proof hold with probability $1-\zeta$ for all $t \leq t_{\text{thres}}$.

From (66) and (83), we know that the two induction bounds hold for t. We recall the first induction bound here:

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

Thus, we know that $\|w_t\| \geq \frac{1}{2}(1+\eta\gamma)^t r_0 = \frac{1}{2}(1+\eta\gamma)^t\frac{\zeta' r}{\sqrt{d}}$. However, $\|w_t\| := \|x_t - x_t'\| \leq \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x_t' - x_0'\| + \|x_0' - \widetilde{x}\| \leq 2r + 2\frac{\delta}{C_1\rho} \leq \frac{4\delta}{C_1\rho}$ according to (54) and the perturbation radius r. The last inequality is due to the perturbation radius $r \leq \frac{\delta}{C_1\rho}$ (we already used this condition in the previous arguments). This will give a contradiction for (54) if $\frac{1}{2}(1+\eta\gamma)^t\frac{\zeta' r}{\sqrt{d}} \geq \frac{4\delta}{C_1\rho}$ and it will happen if $t \geq \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$.

So the proof of this lemma is finished by contradiction if we let $t_{\text{thres}} := \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, i.e., we have

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

□

B.2 Proofs for Online Problem

In this section, we provide the detailed proofs for online problem (2) (i.e., Theorem 3–4). We will reuse some parts of our previous proofs for finite-sum problem (1) in previous Section B.1.

First, we recall the previous key relation (25) between $f(x_t)$ and $f(x_{t-1})$ as follows (recall $x_t := x_{t-1} - \eta v_{t-1}$):

$$f(x_t) \leq f(x_{t-1}) + \frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2. \tag{84}$$

Next, we recall the previous bound (29) for the variance term:

$$\mathbb{E}[\|v_{t-1} - \nabla f(x_{t-1})\|^2] \leq \frac{L^2}{b}\mathbb{E}[\|x_{t-1} - x_{t-2}\|^2] + \mathbb{E}[\|v_{t-2} - \nabla f(x_{t-2})\|^2]. \tag{85}$$

Now, the following bound for the variance term will be different from the previous finite-sum case. Similar to (30), we sum up (85) from the beginning of this epoch sm to the point $t-1$,

$$\begin{aligned}
\mathbb{E}[\|v_{t-1} - \nabla f(x_{t-1})\|^2] &\leq \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \mathbb{E}[\|v_{sm} - \nabla f(x_{sm})\|^2] &(86)\\
&= \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_{sm}) - \nabla f(x_{sm})\Big\|^2\Big] &(87)\\
&\leq \frac{L^2}{b}\sum_{j=sm+1}^{t-1}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{\sigma^2}{B}, &(88)
\end{aligned}$$

where (86) is the same as (30), (87) uses the modification (11) (i.e., $v_{sm} = \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_{sm})$ instead of the full gradient computation $v_{sm} = \nabla f(x_{sm})$ in the finite-sum case), and the last inequality (88) follows from the bounded variance Assumption 3.

Now, we take expectations for (84) and then sum it up from the beginning of this epoch s, i.e., iterations from sm to t, by plugging the variance (88) into them to get:

$$\begin{aligned}
\mathbb{E}[f(x_t)] &\leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2]\\
&\quad + \frac{\eta L^2}{2b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{\sigma^2}{B}\\
&\leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2]\\
&\quad + \frac{\eta L^2(t-1-sm)}{2b}\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{(t-sm)\eta\sigma^2}{2B}\\
&\leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2]\\
&\quad + \frac{\eta L^2}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|x_j - x_{j-1}\|^2] + \frac{(t-sm)\eta\sigma^2}{2B} &(89)\\
&\leq \mathbb{E}[f(x_{sm})] - \frac{\eta}{2}\sum_{j=sm+1}^{t}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] + \frac{(t-sm)\eta\sigma^2}{2B}, &(90)
\end{aligned}$$

where (89) holds if the minibatch size $b \geq m$ (note that here $t \leq (s+1)m$), (90) holds if the step size $\eta \leq \frac{\sqrt{5}-1}{2L}$.

Proof of Theorem 3. Let $b = m = \frac{2\sigma}{\epsilon}$ and step size $\eta = \frac{\sqrt{5}-1}{2L}$, then (90) holds. Now, the proof is directly obtained by summing up (90) for all epochs $0 \leq s \leq S$ as follows:

$$\mathbb{E}[f(x_T)] \leq \mathbb{E}[f(x_0)] - \frac{\eta}{2}\sum_{j=1}^{T}\mathbb{E}[\|\nabla f(x_{j-1})\|^2] + \frac{T\eta\sigma^2}{2B}$$

$$\mathbb{E}[\|\nabla f(\hat{x})\|] \leq \sqrt{\mathbb{E}[\|\nabla f(\hat{x})\|^2]} \leq \sqrt{\frac{2(f(x_0)-f^*)}{\eta T} + \frac{\sigma^2}{B}} = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \tag{91}$$

where (91) holds by choosing $\hat{x}$ uniformly from $\{x_{t-1}\}_{t\in[T]}$ and letting $Sm \leq T = \frac{8(f(x_0)-f^*)}{\eta\epsilon^2} = O(\frac{L(f(x_0)-f^*)}{\epsilon^2})$ and $B = \frac{4\sigma^2}{\epsilon^2}$. Note that the total number of computation of stochastic gradients equals to

$$SB + Smb \leq \Big\lceil \frac{T}{m} \Big\rceil B + Tb \leq \Big(\frac{T}{2\sigma/\epsilon} + 1\Big)\frac{4\sigma^2}{\epsilon^2} + T\frac{2\sigma}{\epsilon} = \frac{4\sigma^2}{\epsilon^2} + 2T\frac{2\sigma}{\epsilon} = O\Big(\frac{\sigma^2}{\epsilon^2} + \frac{L(f(x_0)-f^*)\sigma}{\epsilon^3}\Big).$$

□

B.2.1 Proof of Theorem 4

Similar to the proof of Theorem 2, for proving the second-order guarantee, we will divide the proof into two situations. The first situation (**large gradients**) is also almost the same as the above arguments for first-order guarantee, where the function value will decrease a lot since the gradients are large (see (90)). For the second situation (**around saddle points**), we will 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 (88). Then we use it to get a high probability bound of (90) for function value decrease. 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 (see (11)) instead of $v_{sm} = \nabla f(x_{sm})$ in the previous finite-sum case. Thus we first need a high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$. According to Assumption 4, we have

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

By applying Bernstein inequality (Proposition 2), we get the high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$ as follows:

$$\mathbb{P}\Big\{\big\|v_{sm} - \nabla f(x_{sm})\big\| \geq \frac{t}{B}\Big\} \leq (d+1)\exp\Big(\frac{-t^2/2}{B\sigma^2 + \sigma t/3}\Big) = \zeta,$$

where the last equality holds by letting $t = C\sqrt{B}\sigma$, where $C = O(\log\frac{d}{\zeta}) = \widetilde{O}(1)$. Now, we have a high probability bound for $\|v_{sm} - \nabla f(x_{sm})\|$, i.e.,

$$\big\|v_{sm} - \nabla f(x_{sm})\big\| \leq \frac{C\sigma}{\sqrt{B}} \quad \text{with probability } 1-\zeta. \tag{92}$$

Now we will try to 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 9 of Algorithm 1), 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 difference sequence $\{z_k\}$. We use the Bernstein inequality to bound the differences as follows.

$$\begin{aligned} z_k = y_k - y_{k-1} &= v_k - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_k) - \nabla f_i(x_{k-1})\big) + v_{k-1} - \nabla f(x_k) - (v_{k-1} - \nabla f(x_{k-1})) \\ &= \frac{1}{b}\sum_{i\in I_b}\Big(\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\Big). \end{aligned} \tag{93}$$

We define $u_i := \nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))$, and then we have

$$\|u_i\| = \|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\| \le 2\|x_k - x_{k-1}\|, \tag{94}$$

where the last inequality holds due to the gradient Lipschitz Assumption 1. Then, consider the variance term

$$\begin{aligned} &\sum_{i\in I_b}\mathbb{E}[\|u_i\|^2] \\ &= \sum_{i\in I_b}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))\|^2] \\ &\le \sum_{i\in I_b}\mathbb{E}[\|\nabla f_i(x_k) - \nabla f_i(x_{k-1})\|^2] \\ &\le bL^2\|x_k - x_{k-1}\|^2, \end{aligned} \tag{95}$$

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 1. According to (94) and (95), we can bound the difference z_k by Bernstein inequality (Proposition 2) as

$$\begin{aligned} \mathbb{P}\Big\{\|z_k\| \ge \frac{t}{b}\Big\} &\le (d+1)\exp\Big(\frac{-t^2/2}{\sigma^2 + Rt/3}\Big) \\ &= (d+1)\exp\Big(\frac{-t^2/2}{bL^2\|x_k - x_{k-1}\|^2 + 2\|x_k - x_{k-1}\|t/3}\Big) \\ &= \zeta_k, \end{aligned}$$

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

$$\|z_k\| \le c_k = \frac{CL\|x_k - x_{k-1}\|}{\sqrt{b}} \quad \text{with probability } 1 - \zeta_k. \tag{96}$$

Now, we are ready to get a high probability bound for our original variance term (88) by using the martingale Azuma-Hoeffding inequality. Consider in a specifical epoch s, i.e, iterations t from $sm+1$ to current $sm+k$, where k is less than m. According to Azuma-Hoeffding inequality (Proposition 4) and letting $\zeta_k = \zeta/m$, we have

$$\begin{aligned} \mathbb{P}\Big\{\|y_{sm+k} - y_{sm}\| \ge \beta\Big\} &\le (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{t=sm+1}^{sm+k} c_t^2}\Big) + \zeta \\ &= 2\zeta, \end{aligned}$$

where the last equality holds by letting $\beta = \sqrt{8\sum_{t=sm+1}^{sm+k} c_t^2 \log\frac{d}{\zeta}} = \frac{C'L\sqrt{\sum_{t=sm+1}^{sm+k}\|x_t - x_{t-1}\|^2}}{\sqrt{b}}$, where $C' = O(C\sqrt{\log\frac{d}{\zeta}}) = \widetilde{O}(1)$. Recall that $y_k := v_k - \nabla f(x_k)$ and at the beginning point of this epoch $\|y_{sm}\| = \|v_{sm} -$

$\nabla f(x_{sm})\| \leq C\sigma/\sqrt{B}$ with probability $1-\zeta$, where $C = O(\log \frac{d}{\zeta}) = \widetilde{O}(1)$ (see (92)). Combining with (92) and using a union bound, we have

$$\|v_{t-1} - \nabla f(x_{t-1})\| = \|y_{t-1}\| \leq \beta + \|y_{sm}\| \leq \frac{C'L\sqrt{\sum_{j=sm+1}^{t-1}\|x_j - x_{j-1}\|^2}}{\sqrt{b}} + \frac{C\sigma}{\sqrt{B}} \tag{97}$$

with probability $1-3\zeta$, where t belongs to $[sm+1, (s+1)m]$.

Now, we use this high probability version (97) instead of the expectation one (88) to obtain the high probability bound for function value decrease (see (90)). We sum up (84) from the beginning of this epoch s, i.e., iterations from sm to t, by plugging (97) into them to get:

$$\begin{aligned}
f(x_t) &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta}{2}\sum_{k=sm+1}^{t-1}\frac{2C'^2L^2\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2}{b} + \frac{\eta}{2}\sum_{j=sm+1}^{t}\frac{2C^2\sigma^2}{B} \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2}{b}\sum_{k=sm+1}^{t-1}\sum_{j=sm+1}^{k}\|x_j - x_{j-1}\|^2 + \frac{(t-sm)\eta C^2\sigma^2}{B} \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{\eta C'^2L^2(t-1-sm)}{b}\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 + \frac{(t-sm)\eta C^2\sigma^2}{B} \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2L^2\big)\sum_{j=sm+1}^{t}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{(t-sm)\eta C^2\sigma^2}{B} \qquad (98) \\
&\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 + \frac{(t-sm)\eta C^2\sigma^2}{B}, \qquad (99)
\end{aligned}$$

where (98) holds if the minibatch size $b \geq m$ (note that here $t \leq (s+1)m$), and (99) holds if the step size $\eta \leq \frac{\sqrt{8C'^2+1}-1}{4C'^2L}$.

Similar to the previous finite-sum case, (99) only guarantees function value decrease 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 3 of Algorithm 2). As we discussed in previous Section B.1.1, we achieve this by stopping the epoch at a uniformly random point (see Line 16 of Algorithm 2).

We want to point out that the second situation will have a bit difference due to (11), i.e., the full gradient of the starting point is not available (see Line 3 of Algorithm 2). Thus some modifications are needed for previous Lemma 1, we use the following lemma to connect these two situations (large gradients and around saddle points):

Lemma 4 (Connection of Two Situations) *For any epoch s, let x_t be a point uniformly sampled from this epoch $\{x_j\}_{j=sm}^{(s+1)m}$. Moreover, let the step size $\eta \leq \frac{\sqrt{8C'^2+1}-1}{4C'^2L}$ (where $C' = O(\log \frac{dm}{\zeta}) = \widetilde{O}(1)$), the minibatch size $b \geq m$ and batch size $B \geq \frac{256C^2\sigma^2}{g_{\text{thres}}^2}$ (where $C = O(\log \frac{d}{\zeta}) = \widetilde{O}(1)$), there are two cases:*

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

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

Moreover, $f(x_t) \leq f(x_{sm}) + \frac{(t-sm)\eta C^2\sigma^2}{B}$ *holds with high probability no matter which case happens.*

Proof of Lemma 4. There are two cases in this epoch:

1. If at least half of points of in this epoch $\{x_j\}_{j=sm}^{(s+1)m}$ have gradient norm no larger than $\frac{g_{\rm thres}}{2}$, then it is easy to see that a uniformly sampled point x_t has gradient norm $\|\nabla f(x_t)\| \leq \frac{g_{\rm thres}}{2}$ 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 19 of Algorithm 2), thus we have $\|\nabla f(x_{(s+1)m})\| \leq \frac{g_{\rm thres}}{2}$ with probability $1/2$. According to (92), we have $\|v_{(s+1)m} - \nabla f(x_{(s+1)m})\| \leq \frac{C\sigma}{\sqrt{B}}$ with probability $1-\zeta$, where $C = O(\log\frac{d}{\zeta}) = \widetilde{O}(1)$. By a union bound, with probability at least $1/3$, we have

$$\|v_{(s+1)m}\| \leq \frac{C\sigma}{\sqrt{B}} + \frac{g_{\rm thres}}{2} \leq \frac{g_{\rm thres}}{16} + \frac{g_{\rm thres}}{2} \leq g_{\rm thres}.$$

2. Otherwise, at least half of points have gradient norm larger than $\frac{g_{\rm thres}}{2}$. Then, as long as the sampled point x_t falls into the last quarter of $\{x_j\}_{j=sm}^{(s+1)m}$, we know $\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 \geq \frac{m g_{\rm thres}^2}{16}$. This holds with probability at least $1/4$ since x_t is uniformly sampled. Then by combining with (99), we obtain the function value decrease

$$f(x_{sm}) - f(x_t) \geq \frac{\eta}{2}\sum_{j=sm+1}^{t}\|\nabla f(x_{j-1})\|^2 - \frac{(t-sm)\eta C^2\sigma^2}{B} \geq \frac{\eta m g_{\rm thres}^2}{32} - \frac{\eta m g_{\rm thres}^2}{256} = \frac{7\eta m g_{\rm thres}^2}{256},$$

where the last inequality is due to $B \geq \frac{256C^2\sigma^2}{g_{\rm thres}^2}$. Note that (99) holds with high probability if we choose the minibatch size $b \geq m$ and the step size $\eta \leq \frac{\sqrt{8C'^2+1}-1}{4C'^2L}$. By a union bound, the function value decrease $f(x_{sm}) - f(x_t) \geq \frac{7\eta m g_{\rm thres}^2}{256}$ with probability at least $1/5$.

Again according to (99), $f(x_t) \leq f(x_{sm}) + \frac{(t-sm)\eta C^2\sigma^2}{B}$ always holds with high probability. □

Note that if Case 2 happens, the function value already decreases a lot in this epoch s (corresponding to the first situation large gradients). Otherwise, Case 1 happens, we know the starting point of the next epoch $x_{(s+1)m} = x_t$ (i.e., Line 19 of Algorithm 2), then we know $\|\nabla f(x_{(s+1)m})\| \leq \frac{g_{\rm thres}}{2}$ and $\|v_{(s+1)m}\| \leq g_{\rm thres}$. Then we will start a super epoch (corresponding to the second situation around saddle points). Note that if $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) > -\delta$, this point $x_{(s+1)m}$ is already an (ϵ,δ)-second-order stationary point (recall that $g_{\rm thres} \leq \epsilon$ in our Theorem 4).

Around Saddle Points $\|v_{(s+1)m}\| \leq g_{\rm thres}$ and $\lambda_{\min}(\nabla^2 f(x_{(s+1)m})) \leq -\delta$: In this situation, we will show that the function value decreases a lot in a *super epoch* (instead of an epoch as in the first situation) with high probability by adding a random perturbation at the initial point $\widetilde{x} = x_{(s+1)m}$. 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 6 in Algorithm 2). Following the classical widely used *two-point analysis* developed in [Jin et al., 2017], 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 8–12 of Algorithm 2) with the same choice of batches and minibatches (i.e., I_B's (see (11) and Line 8) and I_b's (see Line 12))for a super epoch. We will show that at least one of these two coupled sequences will decrease the function value a lot (escape the saddle point), i.e.,

$$\exists t \leq t_{\rm thres}, \ \text{ such that } \ \max\{f(x_0) - f(x_t), f(x_0') - f(x_t')\} \geq 2f_{\rm thres}. \tag{100}$$

We will prove (100) by contradiction. Assume the contrary, $f(x_0) - f(x_t) < 2f_{\rm thres}$ and $f(x_0') - f(x_t') < 2f_{\rm thres}$. 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 will show that the stuck region is relatively small in the random perturbation ball, i.e., at least one of x_t and x'_t will go far away from their starting point x_0 and x'_0 with high probability. Thus there is a contradiction. Similar to Lemma 2 and Lemma 3, we need the following two lemmas. Their proofs are deferred to the end of this section.

Lemma 5 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 8–12 of Algorithm 2) from x_0. Moreover, let the step size $\eta \le \frac{1}{4C'L}$ and minibatch size $b \ge m$, with probability $1-\zeta$, we have*

$$\forall t, \ \ \|x_t - x_0\| \le \sqrt{\frac{4t(f(x_0)-f(x_t))}{5C'L} + \frac{4t^2\eta C^2\sigma^2}{5C'LB}}, \tag{101}$$

where $C' = O(\log\frac{dt}{\zeta}) = \widetilde{O}(1)$ and $C = O(\log\frac{dt}{\zeta m}) = \widetilde{O}(1)$.

Lemma 6 (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 8–12 of Algorithm 2) with the same choice of batches and minibatches (i.e., I_B's (see (11) and Line 8) and I_b's (see Line 12)) 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 $t_{\rm thres} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \le \min\big(\frac{1}{16\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}, \frac{1}{8C_2 L\log t_{\rm thres}}\big) = \widetilde{O}(\frac{1}{L})$, minibatch size $b \ge m$, batch size $B = \widetilde{O}(\frac{\sigma^2}{g_{\rm thres}^2})$ and the perturbation radius $r \le \frac{\delta}{C_1\rho}$, then with probability $1-\zeta$, we have*

$$\exists T \le t_{\rm thres}, \ \ \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \ge \frac{\delta}{C_1\rho}, \tag{102}$$

where $C_1 \ge \frac{20C_2}{\eta L}$, $C_2 = O(\log\frac{dt_{\rm thres}}{\zeta}) = \widetilde{O}(1)$ and $C'_2 = O(\log\frac{dt_{\rm thres}}{\zeta m}) = \widetilde{O}(1)$.

Based on these two lemmas, we are ready to show that (100) holds with high probability. Without loss of generality, we assume $\|x_T - x_0\| \ge \frac{\delta}{C_1\rho}$ in (102) (note that (101) holds for both $\{x_t\}$ and $\{x'_t\}$), then plugging it into (101) to obtain

$$\begin{aligned}
\sqrt{\frac{4T(f(x_0)-f(x_T))}{5C'L} + \frac{4T^2\eta C^2\sigma^2}{5C'LB}} &\ge \frac{\delta}{C_1\rho} \\
f(x_0) - f(x_T) &\ge \frac{5C'L\delta^2}{4C_1^2\rho^2 T} - \frac{T\eta C^2\sigma^2}{B} \\
&\ge \frac{5\eta C'L\delta^3}{8C_1^2\rho^2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})} - \frac{2C^2\sigma^2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{B\delta} \qquad (103)\\
&\ge \frac{\delta^3}{C_1'\rho^2} \qquad (104)\\
&= 2f_{\rm thres},
\end{aligned}$$

where (103) is due to $T \le t_{\rm thres}$ and (104) holds by letting $C_1' = \frac{8C_1^2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{4\eta C'L}$. Recall that $B = \widetilde{O}(\frac{\sigma^2}{g_{\rm thres}^2})$ and $g_{\rm thres} \le \delta^2/\rho$. Thus, we already prove that at least one of sequences $\{x_t\}$ and $\{x'_t\}$ escapes the saddle point with high probability, i.e.,

$$\exists T \le t_{\rm thres} \ \ , \max\{f(x_0) - f(x_T), f(x'_0) - f(x'_T)\} \ge 2f_{\rm thres}, \tag{105}$$

if their starting points x_0 and x'_0 satisfying $w_0 := x_0 - x'_0 = r_0 e_1$, where $r_0 = \frac{\zeta' r}{\sqrt{d}}$ and e_1 denotes the smallest eigenvector direction of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. Similar to the classical argument in [Jin et al., 2017], we know that in

the random perturbation ball, the stuck points can only be a short interval in the e_1 direction, i.e., at least one of two points in the e_1 direction will 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

$$\frac{r_0 V_{d-1}(r)}{V_d(r)} = \frac{r_0 \Gamma(\frac{d}{2}+1)}{\sqrt{\pi} r \Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{r_0}{\sqrt{\pi} r}\big(\frac{d}{2}+1\big)^{1/2} \leq \frac{r_0\sqrt{d}}{r} = \zeta', \tag{106}$$

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 (106) and (104) (holds with high probability if x_0 is not in a stuck region), we know

$$f(x_0) - f(x_T) \geq 2f_{\text{thres}} = \frac{\delta^3}{C_1' \rho^2} \tag{107}$$

with high probability. Note that the initial point of this super epoch is $\widetilde{x}$ before the perturbation (see Line 6 of Algorithm 2), 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.,

$$\begin{aligned} f(x_0) &\leq f(\widetilde{x}) + \langle \nabla f(\widetilde{x}), x_0 - \widetilde{x}\rangle + \frac{L}{2}\|x_0 - \widetilde{x}\|^2 \\ &\leq f(\widetilde{x}) + \|\nabla f(\widetilde{x})\|\|x_0 - \widetilde{x}\| + \frac{L}{2}\|x_0 - \widetilde{x}\|^2 \\ &\leq f(\widetilde{x}) + g_{\text{thres}} \cdot r + \frac{L}{2} r^2 \\ &\leq f(\widetilde{x}) + \frac{\delta^3}{2C_1'\rho^2} \\ &= f(\widetilde{x}) + f_{\text{thres}}, \end{aligned} \tag{108}$$

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

Now we combine with (107) and (108) to obtain with high probability

$$f(\widetilde{x}) - f(x_T) = f(\widetilde{x}) - f(x_0) + f(x_0) - f(x_T) \geq -f_{\text{thres}} + 2f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}. \tag{109}$$

Thus we have finished the proof for the second situation (around saddle points), i.e., we show that the function value decrease a lot ($f_{\text{thres}} = \frac{\delta^3}{2C_1'\rho^2}$) in a *super epoch* (recall that $T \leq t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$) by adding a random perturbation $\xi \sim \mathbb{B}_0(r)$ at the initial point $\widetilde{x}$.

Combing these two situations (large gradients and around saddle points) to prove Theorem 4: First, we recall Theorem 4 here since we want to recall the parameter setting.

Theorem 4 *Under Assumption 1, 2 (i.e. (4) and (6)) and Assumption 4, let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. By letting step size $\eta = \widetilde{O}(\frac{1}{L})$, batch size $B = \widetilde{O}(\frac{\sigma^2}{g_{\text{thres}}^2}) = \widetilde{O}(\frac{\sigma^2}{\epsilon^2})$, minibatch size $b = \sqrt{B} = \widetilde{O}(\frac{\sigma}{\epsilon})$, epoch length $m = b$, perturbation radius $r = \widetilde{O}\big(\min(\frac{\delta^3}{\rho^2\epsilon}, \frac{\delta^{3/2}}{\rho\sqrt{L}})\big)$, threshold gradient $g_{\text{thres}} = \epsilon \leq \delta^2/\rho$, threshold function value $f_{\text{thres}} = \widetilde{O}(\frac{\delta^3}{\rho^2})$ and super epoch length $t_{\text{thres}} = \widetilde{O}(\frac{1}{\eta\delta})$, SSRGD will at least once get to an (ϵ, δ)-second-order stationary point with high probability using*

$$\widetilde{O}\Big(\frac{L\Delta f\sigma}{\epsilon^3} + \frac{\rho^2\Delta f\sigma^2}{\epsilon^2\delta^3} + \frac{L\rho^2\Delta f\sigma}{\epsilon\delta^4}\Big)$$

stochastic gradients for nonconvex online problem (2).

Proof of Theorem 4. Now, we prove this 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 $g_{\rm thres}$ (Case 2 of Lemma 4);

2. *Wasted epoch*: If at least half of points in this epoch have gradient norm no larger than $g_{\rm thres}$ and the starting point of the next epoch has estimated gradient norm larger than $g_{\rm thres}$ (it means that this epoch does not guarantee decreasing the function value a lot as the large gradients situation, also it cannot connect to the second super epoch situation since the starting point of the next epoch has estimated gradient norm larger than $g_{\rm thres}$);

3. *Type-2 useful super epoch*: If at least half of points in this epoch have gradient norm no larger than $g_{\rm thres}$ and the starting point of the next epoch (here we denote this point as $x_{(s+1)m}$) has estimated gradient norm no larger than $g_{\rm thres}$ (i.e., $\|v_{(s+1)m}\| \leq g_{\rm thres}$) (Case 1 of Lemma 4), according to Line 3 of Algorithm 2, we will 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 $2/3$ due to the random stop (see Case 1 of Lemma 4 and Line 16 of Algorithm 2) and different wasted epoch are independent. Thus, with high probability, there are at most $\widetilde{O}(1)$ 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. Recall that $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f.

For type-1 useful epoch, according to Case 2 of Lemma 4, we know that the function value decreases at least $\frac{7\eta m g_{\rm thres}^2}{256}$ with probability at least $1/5$. Using a standard concentration, we know that with high probability N_1 type-1 useful epochs will decrease the function value at least $\frac{7\eta m g_{\rm thres}^2 N_1}{1536}$, note that the function value can decrease at most Δf. So $\frac{7\eta m g_{\rm thres}^2 N_1}{1536} \leq \Delta f$, we get $N_1 \leq \frac{1536\Delta f}{7\eta m g_{\rm thres}^2}$.

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 $\|\nabla f(\widetilde{x})\| \leq g_{\rm thres}/2$ and estimated gradient norm $\|v_{(s+1)m}\| \leq g_{\rm thres}$. Now if $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \geq -\delta$, then $\widetilde{x}$ is already a (ϵ, δ)-second-order stationary point. Otherwise, $\|v_{(s+1)m}\| \leq g_{\rm thres}$ and $\lambda_{\min}(\nabla^2 f(\widetilde{x})) \leq -\delta$, this is exactly our second situation (around saddle points). According to (109), we know that the the function value decrease $(f(\widetilde{x}) - f(x_T))$ is at least $f_{\rm thres} = \frac{\delta^3}{2C_1'\rho^2}$ with high probability. Similar to type-1 useful epoch, we know $N_2 \leq \frac{C_1''\rho^2\Delta f}{\delta^3}$ by a union bound (so we change C_1' to C_1'', anyway we also have $C_1'' = \widetilde{O}(1)$).

Now, we are ready to compute the convergence results to finish the proof for Theorem 4.

$$N_1(\widetilde{O}(1)B + B + mb) + N_2(\widetilde{O}(1)B + \lceil \frac{t_{\rm thres}}{m} \rceil B + t_{\rm thres} b) \tag{110}$$

$$\leq \widetilde{O}\Big(\frac{\Delta f \sigma}{\eta g_{\rm thres}^2 \epsilon} + \frac{\rho^2 \Delta f}{\delta^3}(\frac{\sigma^2}{\epsilon^2} + \frac{\sigma}{\eta\delta\epsilon})\Big)$$

$$\leq \widetilde{O}\Big(\frac{L\Delta f\sigma}{\epsilon^3} + \frac{\rho^2\Delta f\sigma^2}{\epsilon^2\delta^3} + \frac{L\rho^2\Delta f\sigma}{\epsilon\delta^4}\Big) \tag{111}$$

□

Now, the only remaining thing is to prove Lemma 5 and 6. We provide these two proofs as follows.

Lemma 5 (Localization) *Let $\{x_t\}$ denote the sequence by running SSRGD update steps (Line 8–12 of Algorithm 2) from x_0. Moreover, let the step size $\eta \leq \frac{1}{4C'L}$ and minibatch size $b \geq m$, with probability $1-\zeta$, we have*

$$\forall t, \ \ \|x_t - x_0\| \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{5C'L} + \frac{4t^2\eta C^2\sigma^2}{5C'LB}},$$

where $C' = O(\log\frac{dt}{\zeta}) = \widetilde{O}(1)$ and $C = O(\log\frac{dt}{\zeta m}) = \widetilde{O}(1)$.

Proof of Lemma 5. First, we assume the variance bound (97) holds for all $0 \leq j \leq t-1$ (this is true with high probability using a union bound by letting $C' = O(\log\frac{dt}{\zeta})$ and $C = O(\log\frac{dt}{\zeta m})$). Then, according to (98), we know

for any $\tau \leq t$ in some epoch s

$$
\begin{aligned}
f(x_\tau) &\leq f(x_{sm}) - \frac{\eta}{2}\sum_{j=sm+1}^{\tau}\|\nabla f(x_{j-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2 L^2\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 \\
&\quad + \frac{(\tau - sm)\eta C^2\sigma^2}{B} \\
&\leq f(x_{sm}) - \big(\frac{1}{2\eta} - \frac{L}{2} - \eta C'^2 L^2\big)\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 + \frac{(\tau - sm)\eta C^2\sigma^2}{B} \\
&\leq f(x_{sm}) - \frac{5C'L}{4}\sum_{j=sm+1}^{\tau}\|x_j - x_{j-1}\|^2 + \frac{(\tau - sm)\eta C^2\sigma^2}{B},
\end{aligned} \tag{112}
$$

where the last inequality holds since the step size $\eta \leq \frac{1}{4C'L}$ and assuming $C' \geq 1$. Now, we sum up (112) for all epochs before iteration t,

$$
f(x_t) \leq f(x_0) - \frac{5C'L}{4}\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2 + \frac{t\eta C^2\sigma^2}{B}.
$$

Then, the proof is finished as

$$
\|x_t - x_0\| \leq \sum_{j=1}^{t}\|x_j - x_{j-1}\| \leq \sqrt{t\sum_{j=1}^{t}\|x_j - x_{j-1}\|^2} \leq \sqrt{\frac{4t(f(x_0) - f(x_t))}{5C'L} + \frac{4t^2\eta C^2\sigma^2}{5C'LB}}.
$$

□

Lemma 6 (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 8–12 of Algorithm 2) with the same choice of batches and minibatches (i.e., I_B's (see (11) and Line 8) and I_b's (see Line 12)) 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 $t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta} = \widetilde{O}(\frac{1}{\eta\delta})$, the step size $\eta \leq \min\big(\frac{1}{16\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}, \frac{1}{8C_2 L\log t_{\text{thres}}}\big) = \widetilde{O}(\frac{1}{L})$, minibatch size $b \geq m$, batch size $B = \widetilde{O}(\frac{\sigma^2}{g_{\text{thres}}^2})$ and the perturbation radius $r \leq \frac{\delta}{C_1\rho}$, then with probability $1 - \zeta$, we have*

$$
\exists T \leq t_{\text{thres}}, \quad \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \geq \frac{\delta}{C_1\rho},
$$

where $C_1 \geq \frac{20C_2}{\eta L}$, $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta}) = \widetilde{O}(1)$ and $C'_2 = O(\log\frac{dt_{\text{thres}}}{\zeta m}) = \widetilde{O}(1)$.

Proof of Lemma 6. We prove this lemma by contradiction. Assume the contrary,

$$
\forall t \leq t_{\text{thres}} \ , \|x_t - x_0\| \leq \frac{\delta}{C_1\rho} \ \text{ and } \ \|x'_t - x'_0\| \leq \frac{\delta}{C_1\rho} \tag{113}
$$

We will show that the distance between these two coupled sequences $w_t := x_t - x'_t$ will grow 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 (113) and the perturbation radius r. It is not hard to see that the exponential increase will break this upper bound, thus we get a contradiction.

In the following, we prove the exponential increase of w_t by induction. First, we need the expression of w_t (recall that $x_t = x_{t-1} - \eta v_{t-1}$ (see Line 11 of Algorithm 2)):

$$\begin{aligned}
w_t &= w_{t-1} - \eta(v_{t-1} - v'_{t-1}) \\
&= w_{t-1} - \eta\big(\nabla f(x_{t-1}) - \nabla f(x'_{t-1}) + v_{t-1} - \nabla f(x_{t-1}) - v'_{t-1} + \nabla f(x'_{t-1})\big) \\
&= w_{t-1} - \eta\Big(\int_0^1 \nabla^2 f(x'_{t-1} + \theta(x_{t-1} - x'_{t-1}))d\theta(x_{t-1} - x'_{t-1}) + v_{t-1} - \nabla f(x_{t-1}) - v'_{t-1} + \nabla f(x'_{t-1})\Big) \\
&= (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + y_{t-1}) \\
&= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau + y_\tau) \qquad (114)
\end{aligned}$$

where $\Delta_\tau := \int_0^1(\nabla^2 f(x'_\tau + \theta(x_\tau - x'_\tau)) - \mathcal{H})d\theta$ and $y_\tau := v_\tau - \nabla f(x_\tau) - v'_\tau + \nabla f(x'_\tau)$. Note that the first term of (114) is in the e_1 direction and is exponential with respect to t, i.e., $(1+\eta\gamma)^t r_0 e_1$, where $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$. To prove the exponential increase of w_t, it is sufficient to show that the first term of (114) will dominate the second term. We inductively prove the following two bounds

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

First, check the base case $t = 0$, $\|w_0\| = \|r_0 e_1\| = r_0$ holds for Bound 1. However, for Bound 2, we use Bernstein inequality (Proposition 2) 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 (11), 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., I_B's and I_b's). Now, we have

$$\begin{aligned}
y_0 &= v_0 - \nabla f(x_0) - v'_0 + \nabla f(x'_0) \\
&= \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x_0) - \nabla f(x_0) - \frac{1}{B}\sum_{j\in I_B}\nabla f_j(x'_0) + \nabla f(x'_0) \\
&= \frac{1}{B}\sum_{j\in I_B}\Big(\nabla f_j(x_0) - \nabla f_j(x'_0) - (\nabla f(x_0) - \nabla f(x'_0))\Big). \qquad (115)
\end{aligned}$$

We first bound each individual term of (115):

$$\|\nabla f_j(x_0) - \nabla f_j(x'_0) - (\nabla f(x_0) - \nabla f(x'_0))\| \le 2L\|x_0 - x'_0\| = 2L\|w_0\| = 2Lr_0, \qquad (116)$$

where the inequality holds due to the gradient Lipschitz Assumption 1. Then, consider the variance term of (115):

$$\begin{aligned}
&\sum_{j\in I_B}\mathbb{E}[\|\nabla f_j(x_0) - \nabla f_j(x'_0) - (\nabla f(x_0) - \nabla f(x'_0))\|^2] \\
&\le \sum_{j\in I_B}\mathbb{E}[\|\nabla f_j(x_0) - \nabla f_j(x'_0)\|^2] \\
&\le BL^2\|x_0 - x'_0\|^2 \\
&= BL^2\|w_0\|^2 = BL^2 r_0^2, \qquad (117)
\end{aligned}$$

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 1. According to (116) and (117), we can bound y_0 by Bernstein inequality (Proposition 2) 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 $\alpha = C_5 L\sqrt{B} r_0$, where $C_5 = O(\log \frac{d}{\zeta})$. Note that we can further relax the parameter C_5 to $C_2' = O(\log \frac{dt_{\text{thres}}}{\zeta m}) = \widetilde{O}(1)$ for making sure the above arguments hold with probability $1-\zeta$ for all epoch starting points y_{sm} with $sm \le t_{\text{thres}}$. Thus, we have with probability $1-\zeta$,

$$\|y_0\| \le \frac{C_2' L r_0}{\sqrt{B}} \le \eta\gamma L r_0, \tag{118}$$

where the last inequality holds due to $B = \widetilde{O}(\frac{\sigma^2}{g_{\text{thres}}^2})$ (recall that $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$ and $g_{\text{thres}} \le \delta^2/\rho$).

Now, we know that Bound 1 and Bound 2 hold for the base case $t=0$ with high probability. Assume they hold for all $\tau \le t-1$, we now prove they hold for t one by one. For Bound 1, it is enough to show the second term of (114) is dominated by half of the first term.

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau}(\Delta_\tau w_\tau)\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|\Delta_\tau\|\|w_\tau\| \\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\|\Delta_\tau\| && (119)\\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 \sum_{\tau=0}^{t-1}\rho D_\tau^x && (120)\\
&\le \frac{3}{2}\eta(1+\eta\gamma)^{t-1} r_0 t\rho\big(\frac{\delta}{C_1\rho}+r\big) && (121)\\
&\le \frac{3}{C_1}\eta\delta t(1+\eta\gamma)^{t-1} r_0 && (122)\\
&\le \frac{6\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{C_1}(1+\eta\gamma)^{t-1} r_0 && (123)\\
&\le \frac{1}{4}(1+\eta\gamma)^t r_0, && (124)
\end{aligned}$$

where (119) uses the induction for w_τ with $\tau \le t-1$, (120) uses the definition $D_\tau^x := \max\{\|x_\tau - \widetilde{x}\|, \|x_\tau' - \widetilde{x}\|\}$, (121) follows from $\|x_t - \widetilde{x}\| \le \|x_t - x_0\| + \|x_0 - \widetilde{x}\| = \frac{\delta}{C_1\rho} + r$ due to (113) and the perturbation radius r, (122) holds by letting the perturbation radius $r \le \frac{\delta}{C_1\rho}$, (123) holds since $t \le t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, and (124) holds by letting $C_1 \ge 24\log(\frac{8\delta\sqrt{d}}{\rho\zeta' r})$.

$$\begin{aligned}
\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-1-\tau} y_\tau\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau}\|y_\tau\| \\
&\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-1-\tau} 2\eta\gamma L(1+\eta\gamma)^{\tau} r_0 && (125)\\
&= 2\eta\eta\gamma L t(1+\eta\gamma)^{t-1} r_0 \\
&\le 2\eta\eta\gamma L\frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}(1+\eta\gamma)^{t-1} r_0 && (126)\\
&\le 4\eta\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r}) L(1+\eta\gamma)^{t-1} r_0 && (127)\\
&\le \frac{1}{4}(1+\eta\gamma)^t r_0, && (128)
\end{aligned}$$

where (125) uses the induction for y_τ with $\tau \le t-1$, (126) holds since $t \le t_{\text{thres}} = \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, (127) holds $\gamma \ge \delta$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$), and (128) holds by letting $\eta \le \frac{1}{16\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})L}$.

Combining (124) and (128), we proved the second term of (114) is dominated by half of the first term. Note that the first term of (114) is $\|(I-\eta\mathcal{H})^t w_0\| = (1+\eta\gamma)^t r_0$. Thus, we have

$$\frac{1}{2}(1+\eta\gamma)^t r_0 \le \|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0 \tag{129}$$

Now, the remaining thing is to prove the second bound $\|y_t\| \le \eta\gamma L(1+\eta\gamma)^t r_0$. First, we write the concrete expression of y_t:

$$\begin{aligned}
y_t &= v_t - \nabla f(x_t) - v_t' + \nabla f(x_t') \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + v_{t-1} - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t') - \nabla f_i(x_{t-1}')\big) - v_{t-1}' + \nabla f(x_t') \qquad (130)\\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_{t-1})\big) + \nabla f(x_{t-1}) - \nabla f(x_t) \\
&\quad - \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t') - \nabla f_i(x_{t-1}')\big) - \nabla f(x_{t-1}') + \nabla f(x_t') \\
&\quad + v_{t-1} - \nabla f(x_{t-1}) - v_{t-1}' + \nabla f(x_{t-1}') \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_t') - \nabla f_i(x_{t-1}) + \nabla f_i(x_{t-1}')\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x_t') - \nabla f(x_{t-1}) + \nabla f(x_{t-1}')\big) + y_{t-1},
\end{aligned}$$

where (130) due to the definition of the estimator v_t (see Line 12 of Algorithm 2). 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 will apply the Azuma-Hoeffding inequalities to get an upper bound for $\|y_t\|$ and then we prove $\|y_t\| \le 2\eta\gamma L(1+\eta\gamma)^t r_0$ based on that upper bound. In order to apply the Azuma-Hoeffding inequalities for martingale sequence $\|y_t\|$, we first need to bound the difference sequence $\{z_t\}$. We use the Bernstein inequality to bound the differences as follows.

$$\begin{aligned}
z_t &= y_t - y_{t-1} \\
&= \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_t) - \nabla f_i(x_t') - \nabla f_i(x_{t-1}) + \nabla f_i(x_{t-1}')\big) \\
&\quad - \big(\nabla f(x_t) - \nabla f(x_t') - \nabla f(x_{t-1}) + \nabla f(x_{t-1}')\big) \\
&= \frac{1}{b}\sum_{i\in I_b}\Big(\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) \\
&\qquad\qquad - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)\Big). \qquad (131)
\end{aligned}$$

We define $u_i := \big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)$,

and then we have

$$\begin{aligned}
\|u_i\| &= \|\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big) - \big(\nabla f(x_t) - \nabla f(x_t')\big) + \big(\nabla f(x_{t-1}) - \nabla f(x_{t-1}')\big)\| \\
&\le \Big\| \int_0^1 \nabla^2 f_i(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') - \int_0^1 \nabla^2 f_i(x_{t-1}' + \theta(x_{t-1} - x_{t-1}'))d\theta(x_{t-1} - x_{t-1}') \\
&\quad - \int_0^1 \nabla^2 f(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') + \int_0^1 \nabla^2 f(x_{t-1}' + \theta(x_{t-1} - x_{t-1}'))d\theta(x_{t-1} - x_{t-1}')\Big\| \\
&= \|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1}) - (\mathcal{H} w_t + \Delta_t w_t) + (\mathcal{H} w_{t-1} + \Delta_{t-1} w_{t-1})\| \qquad (132) \\
&\le \|(\mathcal{H}_i - \mathcal{H})(w_t - w_{t-1})\| + \|(\Delta_t^i - \Delta_t)w_t - (\Delta_{t-1}^i - \Delta_{t-1})w_{t-1}\| \\
&\le 2L\|w_t - w_{t-1}\| + 2\rho D_t^x \|w_t\| + 2\rho D_{t-1}^x \|w_{t-1}\|, \qquad (133)
\end{aligned}$$

where (132) 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 Lipschitz Assumption 1 and Hessian Lipschitz Assumption 2 (recall $D_t^x := \max\{\|x_t - \widetilde{x}\|, \|x_t' - \widetilde{x}\|\}$). Then, consider the variance term

$$\begin{aligned}
&\sum_{i \in I_b} \mathbb{E}[\|u_i\|^2] \\
&\le \sum_{i \in I_b} \mathbb{E}[\|\big(\nabla f_i(x_t) - \nabla f_i(x_t')\big) - \big(\nabla f_i(x_{t-1}) - \nabla f_i(x_{t-1}')\big)\|^2] \\
&= \sum_{i \in I_b} \mathbb{E}[\|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{t-1} + \Delta_{t-1}^i w_{t-1})\|^2] \\
&\le b(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2, \qquad (134)
\end{aligned}$$

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 1 and Hessian Lipschitz Assumption 2. According to (133) and (134), we can bound the difference z_k by Bernstein inequality (Proposition 2) 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 $\sigma^2 = b(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)^2$)

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

where the last equality holds by letting $\alpha = C_4\sqrt{b}(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)$, where $C_4 = O(\log \frac{d}{\zeta_k}) = \widetilde{O}(1)$.

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

$$\|z_k\| \le c_k = \frac{C_4(L\|w_t - w_{t-1}\| + \rho D_t^x \|w_t\| + \rho D_{t-1}^x \|w_{t-1}\|)}{\sqrt{b}} \quad \text{with probability } 1 - \zeta_k. \qquad (135)$$

Now, we are ready to get an upper bound for y_t by using the martingale Azuma-Hoeffding inequality. Note that we only need to focus on the current epoch that contains the iteration t since the martingale sequence $\{y_t\}$ starts with a new point y_{sm} for each epoch s due to the estimator v_{sm}. Also note that the starting point y_{sm} can be bounded with the same upper bound (118) for all epoch s. Let s denote the current epoch, i.e, iterations from $sm+1$ to current t, where t is no larger than $(s+1)m$. According to Azuma-Hoeffding inequality (Proposition 4) and letting $\zeta_k = \zeta/m$, we have

$$\begin{aligned}
\mathbb{P}\Big\{\|y_t - y_{sm}\| \ge \beta\Big\} &\le (d+1)\exp\Big(\frac{-\beta^2}{8\sum_{k=sm+1}^t c_k^2}\Big) + \zeta \\
&= 2\zeta,
\end{aligned}$$

where the last equality is due to $\beta = \sqrt{8\sum_{k=sm+1}^{t} c_k^2 \log\frac{d}{\zeta}} = \frac{C_3\sqrt{\sum_{k=sm+1}^{t}(L\|w_t-w_{t-1}\|+\rho D_t^x\|w_t\|+\rho D_{t-1}^x\|w_{t-1}\|)^2}}{\sqrt{b}}$, where $C_3 = O(C_4\sqrt{\log\frac{d}{\zeta}}) = \widetilde{O}(1)$. Recall that $y_k := v_k - \nabla f(x_k) - v_k' + \nabla f(x_k')$ and at the beginning point of this epoch $y_{sm} = \|v_{sm} - \nabla f(x_{sm}) - v_{sm}' + \nabla f(x_{sm}')\| \le \eta\gamma L r_0$ with probability $1-\zeta$ (see (118)). Combining with (118) and using a union bound, we have

$$\|y_t\| \le \beta + \|y_{sm}\| \le \frac{C_3\sqrt{\sum_{k=sm+1}^{t}(L\|w_t-w_{t-1}\|+\rho D_t^x\|w_t\|+\rho D_{t-1}^x\|w_{t-1}\|)^2}}{\sqrt{b}} + \eta\gamma L r_0 \tag{136}$$

with probability $1-3\zeta$, where t belongs to $[sm+1,(s+1)m]$. Note that we can further relax the parameter C_3 in (136) to $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta})$ (see (137)) for making sure the above arguments hold with probability $1-\zeta$ for all $t \le t_{\text{thres}}$ by using a union bound for ζ_t's:

$$\|y_t\| \le \frac{C_2\sqrt{\sum_{k=sm+1}^{t}(L\|w_t-w_{t-1}\|+\rho D_t^x\|w_t\|+\rho D_{t-1}^x\|w_{t-1}\|)^2}}{\sqrt{b}} + \eta\gamma L r_0, \tag{137}$$

where t belongs to $[sm+1,(s+1)m]$.

Now, we will show how to bound the right-hand-side of (137) to finish the proof, i.e., prove the remaining second bound $\|y_t\| \le 2\eta\gamma L(1+\eta\gamma)^t r_0$.

First, we show that the last two terms in the first term of right-hand-side of (137) can be bounded as

$$\begin{aligned}
\rho D_t^x\|w_t\| + \rho D_{t-1}^x\|w_{t-1}\| &\le \rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^t r_0 + \rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1} r_0 \\
&\le 3\rho\big(\frac{\delta}{C_1\rho}+r\big)(1+\eta\gamma)^t r_0 \\
&\le \frac{6\delta}{C_1}(1+\eta\gamma)^t r_0,
\end{aligned} \tag{138}$$

where the first inequality follows from the induction of $\|w_{t-1}\| \le \frac{3}{2}(1+\eta\gamma)^{t-1} r_0$ and the already proved $\|w_t\| \le \frac{3}{2}(1+\eta\gamma)^t r_0$ in (129), and the last inequality holds by letting the perturbation radius $r \le \frac{\delta}{C_1\rho}$.

Now, we show that the first term in (137) can be bounded as

$$\begin{aligned}
L\|w_t - w_{t-1}\| &= L\Big\| -\eta\mathcal{H}(I-\eta\mathcal{H})^{t-1} w_0 - \eta\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau) + \eta(\Delta_{t-1}w_{t-1}+y_{t-1})\Big\| \\
&\le L\eta\gamma(1+\eta\gamma)^{t-1} r_0 + L\Big\|\eta\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}(\Delta_\tau w_\tau + y_\tau)\Big\| + L\|\eta(\Delta_{t-1}w_{t-1}+y_{t-1})\| \\
&\le L\eta\gamma(1+\eta\gamma)^{t-1} r_0 + L\eta\Big\|\sum_{\tau=0}^{t-2}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\Big\| \max_{0\le k\le t-2}\|\Delta_k w_k + y_k\| \\
&\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\|
\end{aligned} \tag{139}$$

$$
\begin{aligned}
&\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\sum_{\tau=0}^{t-2}\frac{1}{t-1-\tau}\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\| \\
&\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\| && (140)\\
&\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\| \\
&\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\|w_{t-1}\| + L\eta\|y_{t-1}\| \\
&\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\max_{0\le k\le t-2}\|\Delta_k w_k + y_k\| \\
&\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1}r_0 + 2L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 && (141)\\
&\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\Big(\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-2}r_0 + 2\eta\gamma L(1+\eta\gamma)^{t-2}r_0\Big) \\
&\quad + L\eta\rho\big(\frac{\delta}{C_1\rho}+r\big)\frac{3}{2}(1+\eta\gamma)^{t-1}r_0 + 2L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 && (142)\\
&\le L\eta\gamma(1+\eta\gamma)^{t-1}r_0 + L\eta\log t\Big(\frac{3\delta}{C_1}(1+\eta\gamma)^{t-2}r_0 + 2\eta\gamma L(1+\eta\gamma)^{t-2}r_0\Big) \\
&\quad + \frac{3L\eta\delta}{C_1}(1+\eta\gamma)^{t-1}r_0 + 2L\eta\eta\gamma L(1+\eta\gamma)^{t-1}r_0 && (143)\\
&\le \Big(\frac{4}{C_1}\log t + 4L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0, && (144)
\end{aligned}
$$

where the first equality follows from (114), (139) holds from the following (145),

$$
\|\Delta_t\| \le \rho D_t^x \le \rho\big(\frac{\delta}{C_1\rho}+r\big), \tag{145}
$$

where (145) holds due to Hessian Lipschitz Assumption 2, (113) and the perturbation radius r (recall that $\Delta_t := \int_0^1(\nabla^2 f(x_t'+\theta(x_t-x_t'))-\mathcal{H})d\theta$, $\mathcal{H}:=\nabla^2 f(\widetilde{x})$ and $D_t^x := \max\{\|x_t-\widetilde{x}\|, \|x_t'-\widetilde{x}\|\}$), (140) holds due to $\|\eta\mathcal{H}(I-\eta\mathcal{H})^t\| \le \frac{1}{t+1}$, (141) holds by plugging the induction $\|w_{t-1}\| \le \frac{3}{2}(1+\eta\gamma)^{t-1}r_0$ and $\|y_{t-1}\| \le 2\eta\gamma L(1+\eta\gamma)^{t-1}r_0$, (142) follows from (145), the induction $\|w_k\| \le \frac{3}{2}(1+\eta\gamma)^k r_0$ and $\|y_k\| \le 2\eta\gamma L(1+\eta\gamma)^k r_0$ (hold for all $k \le t-1$), (143) holds by letting the perturbation radius $r \le \frac{\delta}{C_1\rho}$, and the last inequality holds due to $\gamma \ge \delta$ (recall $-\gamma := \lambda_{\min}(\mathcal{H}) = \lambda_{\min}(\nabla^2 f(\widetilde{x})) \le -\delta$).

By plugging (138) and (144) into (137), we have

$$
\begin{aligned}
\|y_t\| &\le C_2\left(\frac{6\delta}{C_1}(1+\eta\gamma)^t r_0 + \Big(\frac{4}{C_1}\log t + 4L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0\right) + \eta\gamma L r_0 \\
&\le C_2\Big(\frac{6}{C_1\eta L} + \frac{4}{C_1}\log t + 4L\eta\log t\Big)\eta\gamma L(1+\eta\gamma)^t r_0 + \eta\gamma L r_0 \\
&\le 2\eta\gamma L(1+\eta\gamma)^t r_0, && (146)
\end{aligned}
$$

where the second inequality holds due to $\gamma \ge \delta$, and the last inequality holds by letting $C_1 \ge \frac{20C_2}{\eta L}$ and $\eta \le \frac{1}{8C_2 L\log t}$. Recall that $C_2 = O(\log\frac{dt_{\text{thres}}}{\zeta})$ is enough to let the arguments in this proof hold with probability $1-\zeta$ for all $t \le t_{\text{thres}}$.

From (129) and (146), we know that the two induction bounds hold for t. We recall the first induction bound here:

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

Thus, we know that $\|w_t\| \ge \frac{1}{2}(1+\eta\gamma)^t r_0 = \frac{1}{2}(1+\eta\gamma)^t\frac{\zeta' r}{\sqrt{d}}$. However, $\|w_t\| := \|x_t - x_t'\| \le \|x_t - x_0\| + \|x_0 - \widetilde{x}\| + \|x_t' - x_0'\| + \|x_0' - \widetilde{x}\| \le 2r + 2\frac{\delta}{C_1\rho} \le \frac{4\delta}{C_1\rho}$ according to (113) and the perturbation radius r. The last inequality

is due to the perturbation radius $r \leq \frac{\delta}{C_1\rho}$ (we already used this condition in the previous arguments). This will give a contradiction for (113) if $\frac{1}{2}(1+\eta\gamma)^t \frac{\zeta' r}{\sqrt{d}} \geq \frac{4\delta}{C_1\rho}$ and it will happen if $t \geq \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$.

So the proof of this lemma is finished by contradiction if we let $t_{\text{thres}} := \frac{2\log(\frac{8\delta\sqrt{d}}{C_1\rho\zeta' r})}{\eta\delta}$, i.e., we have

$$\exists T \leq t_{\text{thres}}, \quad \max\{\|x_T - x_0\|, \|x'_T - x'_0\|\} \geq \frac{\delta}{C_1\rho}.$$

□