Proceedings of Machine Learning Research vol 99:1–55, 2019 32nd Annual Conference on Learning Theory

Stabilized SVRG: Simple Variance Reduction for Nonconvex Optimization

Rong Ge RONGGE@CS.DUKE.EDU
Duke University

Zhize Li ZZ-LI14@MAILS.TSINGHUA.EDU.CN
Tsinghua University

Weiyao Wang WEIYAO.WANG1997@GMAIL.COM
Xiang Wang XWANG@CS.DUKE.EDU
Duke University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Variance reduction techniques like SVRG (Johnson and Zhang, 2013) provide simple and fast algorithms for optimizing a convex finite-sum objective. For nonconvex objectives, these techniques can also find a first-order stationary point (with small gradient). However, in nonconvex optimization it is often crucial to find a second-order stationary point (with small gradient and almost PSD hessian). In this paper, we show that Stabilized SVRG – a simple variant of SVRG – can find an ϵ-second-order stationary point using only $\widetilde{O}(n^{2/3}/\epsilon^2 + n/\epsilon^{1.5})$ stochastic gradients. To our best knowledge, this is the first second-order guarantee for a simple variant of SVRG. The running time almost matches the known guarantees for finding ϵ-first-order stationary points.

Keywords: nonconvex optimization, saddle point, variance reduction

1. Introduction

Nonconvex optimization is widely used in machine learning. Recently, for problems like matrix sensing (Bhojanapalli et al., 2016), matrix completion (Ge et al., 2016), and certain objectives for neural networks (Ge et al., 2017b), it was shown that all local minima are also globally optimal, therefore simple local search algorithms can be used to solve these problems.

For a convex function $f(x)$, a local and global minimum is achieved whenever the point has zero gradient: $\nabla f(x) = 0$. However, for nonconvex functions, a point with zero gradient can also be a saddle point. To avoid converging to saddle points, recent results (Ge et al., 2015; Jin et al., 2017a,b) prove stronger results that show local search algorithms converge to ϵ-approximate second-order stationary points – points with small gradients and almost positive semi-definite Hessians (see Definition 1).

In theory, Xu et al. (2018) and Allen-Zhu and Li (2017) independently showed that finding a second-order stationary point is not much harder than finding a first-order stationary point – they give reduction algorithms Neon/Neon2 that can converge to second-order stationary points when combined with algorithms that find first-order stationary points. Algorithms obtained by such reductions are complicated, and they require a negative curvature search subroutine: given a point x, find an approximate smallest eigenvector of $\nabla^2 f(x)$. In practice, standard algorithms for convex optimization work in a nonconvex setting without a negative curvature search subroutine.

What algorithms can be directly adapted to the nonconvex setting, and what are the simplest modifications that allow a theoretical analysis? For gradient descent, Jin et al. (2017a) showed that a simple perturbation step is enough to find a second-order stationary point, and this was later shown to be necessary (Du et al., 2017). For accelerated gradient, Jin et al. (2017b) showed a simple modification would allow the algorithm to work in the nonconvex setting, and escape from saddle points faster than gradient descent. In this paper, we show that there is also a simple modification to the Stochastic Variance Reduced Gradient (SVRG) algorithm (Johnson and Zhang, 2013) that is guaranteed to find a second-order stationary point.

SVRG is designed to optimize a finite sum objective $f(x)$ of the following form:

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

where evaluating f would require evaluating every f_i. In the original result, Johnson and Zhang (2013) showed that when $f_i(x)$'s are L-smooth and $f(x)$ is μ strongly convex, SVRG finds a point with error ϵ in time $O(n\log(1/\epsilon))$ when $L/\mu = O(n)$. The same guarantees were also achieved by algorithms like SAG (Roux et al., 2012), SDCA (Shalev-Shwartz and Zhang, 2013) and SAGA (Defazio et al., 2014), but SVRG is much cleaner both in terms of implementation and analysis.

SVRG was analyzed in nonconvex regimes, Reddi et al. (2016) and Allen-Zhu and Hazan (2016) showed that SVRG can find an ϵ-first-order stationary point using $O(\frac{n^{2/3}}{\epsilon^2} + n)$ stochastic gradients. Li and Li (2018) analyzed a batched-gradient version of SVRG and achieved the same guarantee with much simpler analysis. These results can then be combined with the reduction (Allen-Zhu and Li, 2017; Xu et al., 2018) to give complicated algorithms for finding second-order stationary points. Using more complicated optimization techniques, it is possible to design faster algorithms for finding first-order stationary points, including FastCubic (Agarwal et al., 2016), SNVRG (Zhou et al., 2018b), SPIDER-SFO (Fang et al., 2018). These algorithms can also combine with procedures like Neon2 to give second-order guarantees.

In this paper, we give a variant of SVRG called Stabilized SVRG that is able to find ϵ-second-order stationary points, while maintaining the simplicity of the SVRG algorithm. See Table 1 for a comparison between our algorithm and existing results. The main term $\widetilde{O}(n^{2/3}/\epsilon^2)$ in the running time of our algorithm matches the analysis with first-order guarantees. All other algorithms that achieve second-order guarantees require negative curvature search subroutines like Neon2, and many are more complicated than SVRG even without this subroutine.

2. Preliminaries

2.1. Notations

We use $\mathbb{N}$, $\mathbb{R}$ to denote the set of natural numbers and real numbers respectively. We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. Let I_b be a multi-set of size b whose i-th element ($i = 1, 2, ..., b$) is chosen i.i.d. from $[n]$ uniformly (I_b is used to denote the samples used in a mini-batch for the algorithm). For vectors we use $\langle u, v\rangle$ to denote their inner product, and for matrices we use $\langle A, B\rangle := \sum_{i,j} A_{ij}B_{ij}$ to denote the trace of $AB^\top$. We use $\|\cdot\|$ to denote the Euclidean norm for a vector and spectral norm for a matrix, and $\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$ to denote the largest and the smallest eigenvalue of a real symmetric matrix.

Throughout the paper, we use $\widetilde{O}(f(n))$ and $\widetilde{\Omega}(f(n))$ to hide poly log factors on relevant parameters. We did not try to optimize the poly log factors in the proof.

Algorithm	Stochastic Gradients	Guarantee	Simple
SVRG (Reddi et al., 2016) (Allen-Zhu and Hazan, 2016)	$O(\frac{n^{2/3}}{\epsilon^2}+n)$	1st-Order	✓
Minibatch-SVRG (Li and Li, 2018)	$O(\frac{n^{2/3}}{\epsilon^2}+n)$	1st-Order	✓
Neon2+SVRG (Allen-Zhu and Li, 2017)	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\epsilon^{1.5}}+\frac{n^{3/4}}{\epsilon^{1.75}})$	2nd-Order	×
Neon2+FastCubic/CDHS (Agarwal et al., 2016; Carmon et al., 2016)	$\widetilde{O}(\frac{n}{\epsilon^{1.5}}+\frac{n^{3/4}}{\epsilon^{1.75}})$	2nd-Order	×
SNVRG$^+$+Neon2 (Zhou et al., 2018a,b)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{n}{\epsilon^{1.5}}+\frac{n^{3/4}}{\epsilon^{1.75}})$	2nd-Order	×
SPIDER-SFO$^+$ (Fang et al., 2018)	$\widetilde{O}(\frac{n^{1/2}}{\epsilon^2}+\frac{1}{\epsilon^{2.5}})$	2nd-Order	×
Stabilized SVRG (this paper)	$\widetilde{O}(\frac{n^{2/3}}{\epsilon^2}+\frac{n}{\epsilon^{1.5}})$	2nd-Order	✓

Table 1: Optimization algorithms for non-convex finite-sum objective

2.2. Finite-Sum Objective and Stationary Points

Now we define the objective that we try to optimize. A finite-sum objective has the form

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

where f_i maps a d-dimensional vector to a scalar and n is finite. In our model, both $f_i(x)$ and $f(x)$ can be non-convex. We make standard smoothness assumptions as follows:

Assumption 1 *Each individual function $f_i(x)$ has L-Lipschitz Gradient, that is,*

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

This implies that the average function $f(x)$ also has L-Lipschitz gradient. We assume the average function $f(x)$ and individual functions have Lipschitz Hessian. That is,

Assumption 2 *The average function $f(x)$ has ρ-Lipschitz Hessian, which means*

$$\forall x_1,x_2\in\mathbb{R}^d,\ \|\nabla^2 f(x_1)-\nabla^2 f(x_2)\|\le\rho\|x_1-x_2\|;$$

each individual function $f_i(x)$ has ρ'-Lipschitz Hessian, which means

$$\forall x_1,x_2\in\mathbb{R}^d,\ \|\nabla^2 f_i(x_1)-\nabla^2 f_i(x_2)\|\le\rho'\|x_1-x_2\|.$$

These two assumptions are standard in the literature for finding second-order stationary points (Ge et al., 2015; Jin et al., 2017a,b; Allen-Zhu and Li, 2017). The goal of non-convex optimization algorithms is to converge to an approximate-second-order stationary point.

Definition 1 *For a differentiable function f, x is a first-order stationary point if $\|\nabla f(x)\|=0$; x is an ϵ-first-order stationary point if $\|\nabla f(x)\|\le\epsilon$.*

For twice-differentiable function f, x is a second-order stationary point if

$$\|\nabla f(x)\|=0 \text{ and } \lambda_{\min}(\nabla^2 f(x))\ge 0.$$

If f is ρ-Hessian Lipschitz, x is an ϵ-second-order stationary point if

$$\|\nabla f(x)\|\le\epsilon, \text{ and } \lambda_{\min}(\nabla^2 f(x))\ge-\sqrt{\rho\epsilon}.$$

This definition of ϵ-second-order stationary point is standard in previous literature (Ge et al., 2015; Jin et al., 2017a,b). Note that the definition of second-order stationary point uses the Hessian Lipschitzness parameter ρ of the average function $f(x)$ (instead of ρ' of individual function). It is easy to check that $\rho \leq \rho'$. In Appendix F we show there are natural applications where $\rho' = \Theta(d)\rho$, so in general algorithms that do not depend heavily on ρ'/ρ are preferred.

2.3. SVRG Algorithm

In this section we give a brief overview of the SVRG algorithm. In particular we follow the mini-batch version in Li and Li (2018) which is used for our analysis for simplicity.

SVRG algorithm has an outer loop. We call each iteration of the outer loop an **epoch**. At the beginning of each epoch, define the snapshot vector $\widetilde{x}$ to be the current iterate and compute its full gradient $\nabla f(\widetilde{x})$. Each epoch of SVRG consists of m iterations. In each iteration, the SVRG algorithm picks b random samples (with replacement) from $[n]$ and form a multi-set I_b, and then estimate the gradient as:

$$v_t := \frac{1}{b}\sum_{i\in I_b}(\nabla f_i(x_t) - \nabla f_i(\widetilde{x}) + \nabla f(\widetilde{x}))$$

After estimating the gradient, the SVRG algorithm performs an update $x_{t+1} \leftarrow x_t - \eta v_t$, where η is the step size. The choice of gradient estimate gives an unbiased estimate of the true gradient, and often has much smaller variance compared to stochastic gradient descent. The pseudo-code for minibatch-SVRG is given in Algorithm 1.

Algorithm 1 SVRG(x_0, m, b, η, S)

Input: initial point x_0, epoch length m, minibatch size b, step size η, number of epochs S.
Output: point x_{Sm}.

```
for s = 0, 1, ..., S − 1 do
    Compute ∇f(x_sm).
    for t = 1, 2, ..., m do
        Sample b i.i.d. numbers uniformly from [n] and form a multi-set I_b.
        v_{sm+t−1} ← (1/b) Σ_{i∈I_b} (∇f_i(x_{sm+t−1}) − ∇f_i(x_sm) + ∇f(x_sm))).
        x_{sm+t} ← x_{sm+t−1} − η v_{sm+t−1}.
    end for
end for
return x_Sm.
```

3. Our Algorithms: Perturbed SVRG and Stabilized SVRG

In this paper we give two simple modifications to the original SVRG algorithm. First, similar to perturbed gradient descent (Jin et al., 2017a), we add perturbations to SVRG algorithm to make it escape from saddle points efficiently. We will show that this algorithm finds an ϵ-second-order stationary point in $\widetilde{O}((\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})(1 + (\frac{\rho'}{n^{1/3}\rho})^2))$ time, where $\Delta f := f(x_0) - f^*$ is the difference between initial function value and the optimal function value. This algorithm is efficient as long as $\rho' \leq \rho n^{1/3}$, but can be slower if ρ' is much larger (see Appendix F for an example where

$\rho' = \Theta(d)\rho$). To achieve stronger guarantees, we introduce *Stabilized SVRG*, which is another simple modification on top of Perturbed SVRG that improves the dependency on ρ'.

3.1. Perturbed SVRG

Algorithm 2 Perturbed SVRG($x_0, m, b, \eta, \delta, \mathscr{G}$)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius δ, threshold gradient $\mathscr{G}$

for $s = 0, 1, 2, \cdots$ **do**
  Compute $\nabla f(x_{sm})$.
  if not currently in a super epoch and $\|\nabla f(x_{sm})\| \le \mathscr{G}$ **then**
    $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(\delta)$, start a super epoch
  end if
  for $t = 1, 2, \cdots, m$ **do**
    Sample b i.i.d. numbers uniformly from $[n]$ and form a multi-set I_b.
    $v_{sm+t-1} \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_{sm+t-1}) - \nabla f_i(x_{sm}) + \nabla f(x_{sm})\big)\big)$.
    $x_{sm+t} \leftarrow x_{sm+t-1} - \eta v_{sm+t-1}$.
    if Stopping condition is met **then** Stop super epoch
  end for
end for

Similar to gradient descent, if one starts SVRG exactly at a saddle point, it is easy to check that the algorithm will not move. To avoid this problem, we propose Perturbed SVRG. A high level description is in Algorithm 2. Intuitively, since at the beginning of each epoch in SVRG the gradient of the function is computed, we can add a small perturbation to the current point if the gradient turns out to be small (which means we are either near a saddle point or already at a second-order stationary point). Similar to perturbed gradient descent in Jin et al. (2017a), we also make sure that the algorithm does not add a perturbation very often - the next perturbation can only happen either after many iterations ($T_{\max}$) or if the point travels enough distance ($\mathscr{L}$). The full algorithm is a bit more technical and is given in Algorithm 4 in appendix.

Later, we will call the steps between the beginning of perturbation and end of perturbation a **super epoch**. When the algorithm is not in a super epoch, for technical reasons we also use a version of SVRG that stops at a random iteration (not reflected in Algorithm 2 but is in Algorithm 4).

For perturbed SVRG, we have the following guarantee:

Theorem 2 *Assume the function $f(x)$ is ρ-Hessian Lipschitz, and each individual function $f_i(x)$ is L-smooth and ρ'-Hessian-Lipschitz. Let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. There exist mini-batch size $b = \widetilde{O}(n^{2/3})$, epoch length $m = n/b$, step size $\eta = \widetilde{O}(1/L)$, perturbation radius $\delta = \widetilde{O}(\min(\frac{\rho^{1.5}\sqrt{\epsilon}}{\max(\rho^2,(\rho'/m)^2)}, \frac{\rho^{0.75}\epsilon^{0.75}}{\max(\rho,\rho'/m)\sqrt{L}}))$, super epoch length $T_{\max} = \widetilde{O}(\frac{L}{\sqrt{\rho\epsilon}})$, threshold gradient $\mathscr{G} = \widetilde{O}(\epsilon)$, threshold distance $\mathscr{L} = \widetilde{O}(\frac{\sqrt{\epsilon\rho}}{\max(\rho,\rho'/m)})$, such that Perturbed SVRG (Algorithm 4) will at least once get to an ϵ-second-order stationary point with high probability using*

$$\widetilde{O}\Big(\big(\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}}\big)\big(1 + (\frac{\rho'}{n^{1/3}\rho})^2\big)\Big)$$

stochastic gradients.

3.2. Stabilized SVRG

Algorithm 3 Stabilized SVRG($x_0, m, b, \eta, \delta, \mathscr{G}$)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius δ, threshold gradient $\mathscr{G}$

for $s = 0, 1, 2, \cdots$ **do**
  Compute $\nabla f(x_{sm})$.
  if not currently in a super epoch and $\|\nabla f(x_{sm})\| \le \mathscr{G}$ **then**
    $v_{shift} \leftarrow \nabla f(x_{sm})$.
    $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(\delta)$, start a super epoch
  end if
  for $t = 1, 2, \cdots, m$ **do**
    Sample b i.i.d. numbers uniformly from $[n]$ and form a multi-set I_b.
    $v_{sm+t-1} \leftarrow \frac{1}{b}\sum_{i\in I_b}\big(\nabla f_i(x_{sm+t-1}) - \nabla f_i(x_{sm}) + \nabla f(x_{sm}))\big) - v_{shift}$.
    $x_{sm+t} \leftarrow x_{sm+t-1} - \eta v_{sm+t-1}$.
    if Stopping condition is met **then** Stop super epoch and $v_{shift} \leftarrow 0$.
  end for
end for

In order to relax the dependency on ρ', we further introduce stabilization in the algorithm. Basically, if we encounter a saddle point $\widetilde{x}$, we will run SVRG iterations on a shifted function $\hat{f}(x) := f(x) - \langle \nabla f(\widetilde{x}), x - \widetilde{x}\rangle$, whose gradient at $\widetilde{x}$ is *exactly zero*. Another minor (but important) modification is to perturb the point in a ball with much smaller radius compared to Algorithm 2. We will give more intuitions to show why these modifications are necessary in Section 4.3.

The high level ideas of Stabilized SVRG is given in Algorithm 3. In the pseudo-code, the key observation is that gradient on the shifted function is equal to the gradient of original function plus a stabilizing term. Detailed implementation of Stabilized SVRG is deferred to Algorithm 5. For Stabilized SVRG, the time complexity in the following theorem only has a poly-logarithmic dependency on ρ', which is hidden in $\widetilde{O}(\cdot)$ notation.

Theorem 3 *Assume the function $f(x)$ is ρ-Hessian Lipschitz, and each individual function $f_i(x)$ is L-smooth and ρ'-Hessian Lipschitz. Let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. There exists mini-batch size $b = \widetilde{O}(n^{2/3})$, epoch length $m = n/b$, step size $\eta = \widetilde{O}(1/L)$, perturbation radius $\delta = \widetilde{O}(\min(\frac{\sqrt{\epsilon}}{\sqrt{\rho}}, \frac{m\sqrt{\rho\epsilon}}{\rho'}))$, super epoch length $T_{\max} = \widetilde{O}(\frac{L}{\sqrt{\rho\epsilon}})$, threshold gradient $\mathscr{G} = \widetilde{O}(\epsilon)$, threshold distance $\mathscr{L} = \widetilde{O}(\frac{\sqrt{\epsilon}}{\sqrt{\rho}})$, such that Stabilized SVRG (Algorithm 5) will at least once get to an ϵ-second-order stationary point with high probability using*

$$\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$$

stochastic gradients.

In previous work (Allen-Zhu and Li, 2017), it has been shown that Neon2+SVRG has similar time complexity for finding second-order stationary point, $\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\rho^2\Delta f}{\epsilon^{1.5}} + \frac{n^{3/4}\rho^2\sqrt{L}\Delta f}{\epsilon^{1.75}})$. Our result achieves a slightly better convergence rate using a much simpler variant of SVRG.

4. Overview of Proof Techniques

In this section, we illustrate the main ideas in the proof of Theorems 2 and 3. Similar to many existing proofs for escaping saddle points, we will show that Algorithms 2 and 3 can decrease the function value efficiently either when the current point x_t has a large gradient ($\|\nabla f(x_t)\| \geq \epsilon$) or has a large negative curvature ($\lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho\epsilon}$). Since the function value cannot decrease below the global optimal f^*, the algorithms will be able to find a second-order stationary point within the desired number of iterations.

In the proof, we use $\mathscr{G}$ to denote the threshold of the gradient norm. Starting from a saddle point, the super-epoch ends if the number of steps exceeds the threshold $T_{\max}$ or the distance to the saddle point exceeds the threshold distance $\mathscr{L}$. Throughout the analysis, we use $s(t)$ to denote the index of the snapshot point of iterate x_t. More precisely, $s(t) = m\lfloor t/m \rfloor$.

4.1. Exploiting Large Gradients

There have already been several proofs that show SVRG can converge to a first-order stationary point, and our proof here is very similar. First, we show that the gradient estimate is accurate as long as the current point is close to the snapshot point.

Lemma 4 *For any point x_t, let the gradient estimate be $v_t := \frac{1}{b}\sum_{i\in I_b}(\nabla f_i(x_t) - \nabla f_i(x_{s(t)}) + \nabla f(x_{s(t)}))$, where $x_{s(t)}$ is the snapshot point of the current epoch. Then, with probability at least $1-\zeta$, we have*

$$\|v_t - \nabla f(x_t)\| \leq O\Big(\frac{\log(d/\zeta)L}{\sqrt{b}}\Big)\|x_t - x_{s(t)}\|.$$

This lemma is standard and the version for expected square error was proved in Li and Li (2018). Here we only applied simple concentration inequalities to get a high probability bound.

Next, we show that the function value decrease is lower bounded by the summation of gradient norm squares. The proof of the following lemma is adopted from Li and Li (2018) with minor modifications.

Lemma 5 *For any epoch, suppose the initial point is x_0, which is also the snapshot point for this epoch. Assume for any $0 \leq t \leq m-1$, $\|v_t - \nabla f(x_t)\| \leq \frac{C_1 L}{\sqrt{b}}\|x_t - x_0\|$, where $C_1 = \widetilde{O}(1)$ comes from Lemma 4. Then, given $\eta \leq \frac{1}{3C_1 L}, b \geq m^2$, we have*

$$f(x_0) - f(x_t) \geq \sum_{\tau=0}^{t-1}\frac{\eta}{2}\|\nabla f(x_\tau)\|^2$$

for any $1 \leq t \leq m$.

Using this fact, we can now state the guarantee for exploiting large gradients.

Lemma 6 *For any epoch, suppose the initial point is x_0. Let x_t be a point uniformly sampled from $\{x_\tau\}_{\tau=1}^m$. Then, given $\eta = \widetilde{\Theta}(1/L), b \geq m^2$, for any value of $\mathscr{G}$ we have two cases:*

1. *if at least half of points in $\{x_\tau\}_{\tau=1}^m$ have gradient no larger than $\mathscr{G}$, we know $\|\nabla f(x_t)\| \leq \mathscr{G}$ holds with probability at least $1/2$;*
2. *otherwise, we know $f(x_0) - f(x_t) \geq \frac{\eta}{2}\frac{m\mathscr{G}^2}{4}$ holds with probability at least $1/5$.*

Further, no matter which case happens we always have $f(x_t) \le f(x_0)$ with high probability.

As this lemma suggests, our algorithm will stop at a random iterate when it is not in a super epoch (this is reflected in the detailed Algorithms 4 and 5). In the first case, since there are at least half points with small gradients, by uniform sampling, we know the sampled point must have small gradient with at least half probability. In the second case, the function value decreases significantly. Proofs for lemmas in this section are deferred to Appendix B.

4.2. Exploiting Negative Curvature - Perturbed SVRG

Section 4.1 already showed that if the algorithm is not in a super epoch, with constant probability every epoch of SVRG will either decrease the function value significantly, or end at a point with small gradient. In the latter case, if the point with small gradient also has almost positive semi-definite Hessian, then we have found an approximate-second-order stationary point. Otherwise, the algorithm will enter a super epoch, and we will show that with a reasonable probability Algorithm 2 can decrease the function value significantly within the super epoch.

For simplicity, we will reset the indices for the iterates in the super epoch. Let the initial point be $\widetilde{x}$, the point after the perturbation be x_0, and the iterates in this super epoch be $x_1, ..., x_t$.

The proof for Perturbed SVRG is very similar to the proof of perturbed gradient descent in Jin et al. (2017a). In particular, we perform a *two point analysis*. That is, we consider two coupled samples of the perturbed point x_0, x_0'. Let e_1 be the smallest eigendirection of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. The two perturbed points x_0 and x_0' only differ in the e_1 direction. We couple the two trajectories from x_0 and x_0' by choosing the same mini-batches for both of them. The iterates of the two sequences are denoted by $x_0, ..., x_t$ and $x_0', ..., x_t'$ respectively. Our goal is to show that with good probability one of these two points can escape the saddle point.

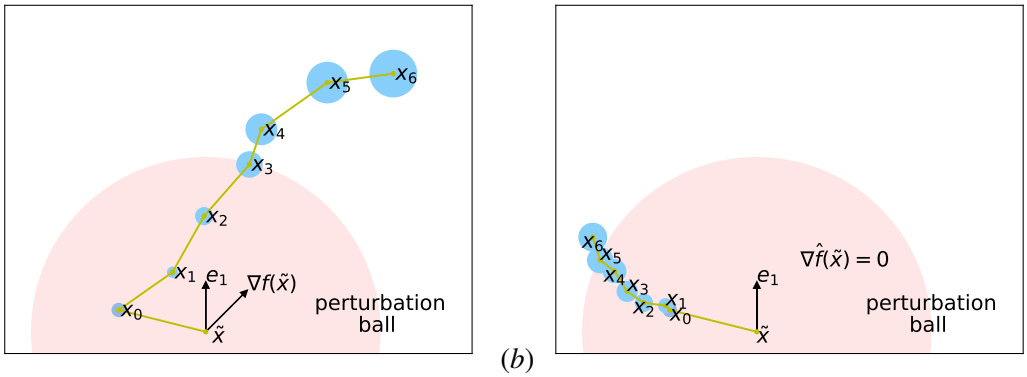
To do that, we will keep track of the difference between the two sequences $w_t = x_t - x_t'$. The key lemma in this section uses Hessian Lipschitz condition to show that the variance of w_t (introduced by the random choice of mini-batch) can actually be much smaller than the variance we observe in Lemma 4. More precisely,

Lemma 7 *Let $\{x_t\}$ and $\{x_t'\}$ be two SVRG sequences running on f that use the same choice of mini-batches. Let $x_{s(t)}$ be the snapshot point for iterate t. Let $w_t := x_t - x_t'$ and $P_t = \max(\|x_{s(t)} - \widetilde{x}\|, \|x'_{s(t)} - \widetilde{x}\|, \|x_t - \widetilde{x}\|, \|x_t' - \widetilde{x}\|)$. Then, with probability at least $1 - \zeta$, we have*

$$\|\xi_t - \xi_t'\| \le O\Big(\frac{\log(d/\zeta)}{\sqrt{b}}\Big) \min\Big(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|), L(\|w_t\| + \|w_{s(t)}\|)\Big).$$

This variance is often much smaller than before as in the extreme case, if $\rho' = 0$ (individual functions are quadratics), the variance is proportional to $\widetilde{O}(L/\sqrt{b})\|w_t - w_{s(t)}\|$. In the proof we will show that w_t cannot change very quickly within a single epoch so $\|w_t - w_{s(t)}\|$ is much smaller than $\|w_t\|$ or $\|w_{s(t)}\|$. Using this new variance bound we can prove:

Lemma 8 (informal) *Let $\{x_t\}$ and $\{x_t'\}$ be two SVRG sequences running on f that use the same choice of mini-batches. Assume $w_0 = x_0 - x_0'$ aligns with e_1 direction and $|\langle e_1, w_0\rangle| \ge \frac{\delta}{4\sqrt{d}}$. Setting the parameters appropriately we know with high probability $\max(\|x_T - \widetilde{x}\|, \|x_T' - \widetilde{x}\|) \ge \mathscr{L}$, for some $T \le \widetilde{O}(1/(\eta\gamma))$.*



(*a*) (*b*)

Figure 1: SVRG trajectories on the original function f and the stabilized function $\hat{f}$. The size of the blue circle at each point indicates the magnitude of the variance.

Intuitively, this lemma is true because at every iterate we expect w_t to be multiplied by a factor of $(1+\eta\gamma)$ if the iterate follows exact gradient, and the variance bound from Lemma 7 is tight enough. The precise statement of the lemma is given in Lemma 16 in Appendix C. The lemma shows that one of the points can escape from a local neighborhood, which by the following lemma is enough to guarantee function value decrease:

Lemma 9 *Let x_0 be the initial point, which is also the snapshot point of the current epoch. Let $\{x_t\}$ be the iterates of SVRG running on f starting from x_0. Fix any $t \geq 1$, suppose for every $0 \leq \tau \leq t-1$, $\|\xi_\tau\| \leq \frac{C_1 L}{\sqrt{b}}\|x_\tau - x_{s(\tau)}\|$, where C_1 comes from Lemma 4. Given $\eta \leq \frac{1}{3C_1 L}$, $b \geq m^2$, we have*

$$\|x_t - x_0\|^2 \leq \frac{4t}{C_1 L}(f(x_0) - f(x_t)).$$

This lemma can be proved using the same technique as Lemma 5. All proofs in this section are deferred to Appendix C.

4.3. Exploiting Negative Curvature - Stabilized SVRG

The main problem in the previous analysis is that when ρ' is large, the variance estimate in Lemma 7 is no longer very strong. To solve this problem, note that the additional term $\rho' P_t(\|w_t\| + \|w_{s(t)}\|)$ is proportional to P_t (the maximum distance of the iterates to the initial point). If we can make sure that the iterates stay very close to the initial point for long enough we will still be able to use Lemma 7 to get a good variance estimate.

However, in Perturbed SVRG, the iterates are not going to stay close to the starting point $\widetilde{x}$, as the initial point $\widetilde{x}$ can have a non-negligible gradient that will make the iterates travel a significant distance (see Figure 1 (a)). To fix this problem, we make a simple change to the function to set the gradient at $\widetilde{x}$ equal to 0. More precisely, define the stabilized function $\hat{f}(x) := f(x) - \langle \nabla f(\widetilde{x}), x - \widetilde{x}\rangle$. After this stabilization, at least the first few iterates will not travel very far (see Figure 1 (b)). Our algorithm will apply SVRG on this stabilized function.

For the stabilized function $\hat{f}(x)$, we have $\nabla \hat{f}(\widetilde{x}) = 0$, so $\widetilde{x}$ is an exact first-order stationary point. In this case, suppose the initial radius of perturbation δ is small, we will show that the behavior of the algorithm has two phases. In Phase 1, the iterates will remain in a ball around $\widetilde{x}$ whose radius is $\widetilde{O}(\delta)$, which allows us to have very tight bounds on the variance and the potential changes

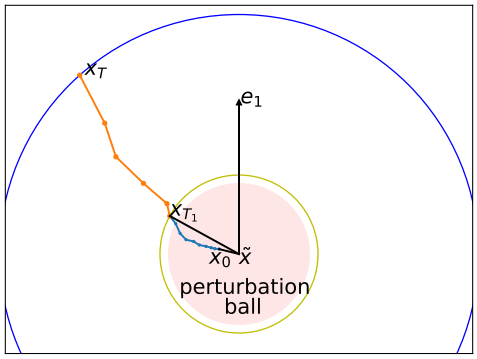


Figure 2: Two phases of a super epoch in Stabilized SVRG

in the Hessian. By the end of Phase 1, we show that the projection in the negative eigendirections of $\mathcal{H} = \nabla^2 f(\widetilde{x})$ is already at least $\widetilde{\Omega}(\delta)$. This means that Phase 1 has basically done a negative curvature search without a separate subroutine! Using the last point of Phase 1 as a good initialization, in Phase 2 we show that the point will eventually escape. See Figure 2 for the two phases.

The rest of the subsection will describe the two phases in more details in order to prove the following main lemma:

Lemma 10 (informal) *Let $\widetilde{x}$ be the initial point with gradient $\|\nabla f(\widetilde{x})\| \le \mathscr{G}$ and $\lambda_{\min}(\mathcal{H}) = -\gamma < 0$. Let $\{x_t\}$ be the iterates of SVRG running on $\hat{f}$ starting from x_0, which is the perturbed point of $\widetilde{x}$. Let T be the length of the current super epoch. Setting the parameters appropriately we know with probability at least $1/8$, $f(x_T) - f(\widetilde{x}) \le -C_5 \frac{\gamma^3}{\rho^2}$; and with high probability, $f(x_T) - f(\widetilde{x}) \le \frac{C_5}{20}\frac{\gamma^3}{\rho^2}$, where $T = \widetilde{O}(\frac{1}{\eta\gamma}), C_5 = \widetilde{\Theta}(1)$.*

Basically, this lemma shows that starting from a saddle point, with constant probability the function value decreases by $\widetilde{\Omega}(\frac{\gamma^3}{\rho^2})$ after a super epoch; with high probability, the function value does not increase by more than $\widetilde{O}(\frac{\gamma^3}{\rho^2})$. The precise statement of this lemma is given in Lemma 22 in Appendix D. Proofs for lemmas in this section are deferred to Appendix D.

4.3.1. ANALYSIS OF PHASE 1

Let S be the subspace spanned by all the eigenvectors of $\mathcal{H}$ with eigenvalues at most $-\frac{\gamma}{\log(d)}$. Our goal is to show that by the end of Phase 1, the projection of $x_t - \widetilde{x}$ on subspace S becomes large while the total movement $\|x_t - \widetilde{x}\|$ is still bounded. To prove this, we use the following conditions to define Phase 1:

Stopping Condition: An iterate x_t is in Phase 1 if (1) $t \le 1/\eta\gamma$ or (2) $\|\text{Proj}_S(x_t - \widetilde{x})\| \le \frac{\delta}{10}$.

If both conditions break, Phase 1 has ended. Intuitively, the second condition guarantees that the projection of $x_t - \widetilde{x}$ on subspace S is large at the end of Phase 1. The first condition makes sure that Phase 1 is long enough such that the projection of $x_t - x_{t-1}$ along positive eigendirections of $\mathcal{H}$ has shrunk significantly, which will be crucial in the analysis of Phase 2.

With the above two conditions, the length of Phase 1 can be defined as

$$T_1 = \sup\left\{t | \forall t' \le t-1, \left(t' \le \frac{1}{\eta\gamma}\right) \vee \left(\|\text{Proj}_S(x_{t'} - \widetilde{x})\| \le \frac{\delta}{10}\right)\right\}. \tag{2}$$

The main lemma for Phase 1 gives the following guarantee:

Lemma 11 (informal) *By choosing $\eta = \widetilde{O}(1/L)$, $b = \widetilde{O}(n^{2/3})$ and $\delta = \widetilde{O}(\min(\frac{\gamma}{\rho}, \frac{m\gamma}{\rho'}))$, with constant probability, the length of the first phase T_1 is $\widetilde{\Theta}(1/\eta\gamma)$ and*

$$\|x_{T_1} - \widetilde{x}\| \leq \widetilde{O}(\delta) \text{ and } \|Proj_S(x_{T_1} - \widetilde{x})\| \geq \frac{1}{10}\delta.$$

We will first show that the iterates in Phase 1 cannot go very far from the initial point:

Lemma 12 (informal) *Let T_1 be the length of Phase 1. Setting parameters appropriately we know with high probability $\|x_t - x_{t-1}\| \leq \widetilde{O}(\frac{1}{t})\delta$ for every $1 \leq t \leq \min(T_1, \frac{\log(d)}{\eta\gamma})$.*

The formal version of the above lemma is in Lemma 19. Taking the sum over all t and note that $\sum_{t=1}^{T} 1/t = \Theta(\log T)$, this implies that the iterates are constrained in a ball whose radius is not much larger than δ. If we choose δ to be small enough, within this ball Lemma 7 will give very sharp bounds on the variance of the gradient estimates. This allows us to repeat the two-point analysis in Section 4.2 and prove that at least one sequence must have a large projection on S subspace within $\frac{\log(d)}{\eta\gamma}$ steps. Recall that in the two point analysis, we consider two coupled samples of the perturbed points x_0, x_0'. The two perturbed points x_0 and x_0' only differ in the e_1 direction. These two sequences $\{x_t\}$ and $\{x_t'\}$ share the same choice of mini-batches at each step. Basically, we prove after $\frac{\log(d)}{\eta\gamma}$ steps, the difference between two sequences along e_1 direction becomes large, which implies that at least one sequence must have large distance to $\widetilde{x}$ on S subspace. The formal version of the following lemma is in Lemma 20.

Lemma 13 (informal) *Let $\{x_t\}$ and $\{x_t'\}$ be two SVRG sequences running on $\hat{f}$ that use the same choice of mini-batches. Assume $w_0 = x_0 - x_0'$ aligns with e_1 direction and $|\langle e_1, w_0\rangle| \geq \frac{\delta}{4\sqrt{d}}$. Let T_1, T_1' be the length of Phase 1 for $\{x_t\}$ and $\{x_t'\}$ respectively. Setting parameters appropriately with high probability we have $\min(T_1, T_1') \leq \frac{\log(d)}{\eta\gamma}$. W.l.o.g., suppose $T_1 \leq \frac{\log(d)}{\eta\gamma}$ and we further have $\|x_{T_1} - \widetilde{x}\| \leq \widetilde{O}(1)\delta$, $\|Proj_S(x_{T_1} - \widetilde{x})\| \geq \frac{1}{10}\delta$.*

Remark 14 *We note that the guarantee of Lemma 13 for Phase 1 is very similar to the guarantee of a negative curvature search subroutine: we find a direction $x_{T_1} - \widetilde{x}$ that has a large projection in subspace S, which contains only the very negative eigenvectors of $\mathcal{H}$.*

4.3.2. ANALYSIS OF PHASE 2

By the guarantee of Phase 1, we know if it is successful $x_{T_1} - \widetilde{x}$ has a large projection in subspace S of very negative eigenvalues. Starting from such a point, in Phase 2 we will show that the projection of $x_t - \widetilde{x}$ in S grows exponentially and exceeds the threshold distance within $\widetilde{O}(\frac{1}{\eta\gamma})$ steps. In order to prove this, we use the following expansion,

$$x_t - \widetilde{x} = (I - \eta\mathcal{H})(x_{t-1} - \widetilde{x}) - \eta\Delta_{t-1}(x_{t-1} - \widetilde{x}) - \eta\xi_{t-1},$$

where $\Delta_{t-1} = \int_0^1 (\nabla^2 \hat{f}(\widetilde{x} + \theta(x_{t-1} - \widetilde{x})) - \mathcal{H})d\theta$. Intuitively, if we only have the first term, it's clear that $\|\text{Proj}_S(x_t - \widetilde{x})\| \geq (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$. The norm in subspace S increases exponentially and will become very far from $\widetilde{x}$ in a small number of iterations. Our proof bounds the Hessian changing term $\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})$ and variance term $\eta\xi_{t-1}$ separately to show that they do not influence the exponential increase. The main lemma that we will prove for Phase 2 is:

Lemma 15 (informal) *Assume Phase 1 is successful in the sense that* $T_1 \leq \frac{\log(d)}{\eta\gamma}$ *and* $\|x_{T_1} - \widetilde{x}\| \leq \widetilde{O}(1)\delta$, $\|Proj_S(x_{T_1} - \widetilde{x})\| \geq \frac{1}{10}\delta$. *Setting parameters appropriately with high probability we know there exists* $T = \widetilde{O}(\frac{1}{\eta\gamma})$ *such that* $\|x_T - \widetilde{x}\| \geq \widetilde{\Omega}(\frac{\gamma}{\rho})$.

The precise version of the above lemma is in Lemma 21 in Appendix D. Similar to Lemma 8, the lemma above shows that the iterates will escape from a local neighborhood if Phase 1 was successful (which happens with at least constant probability). We can then use Lemma 9 to bound the function value decrease.

4.4. Proof of Main Theorems

Finally we are ready to sketch the proof for Theorem 3. For each epoch, if the gradients are large, by Lemma 6 we know with constant probability the function value decreases by at least $\widetilde{\Omega}(n^{1/3}\epsilon^2/L)$. For each super epoch, if the starting point has significant negative curvature, by Lemma 10, we know with constant probability the function value decreases by at least $\widetilde{\Omega}(\epsilon^{1.5}/\sqrt{\rho})$. We also know that the number of stochastic gradient for each epoch is $\widetilde{O}(n)$ and that for each super epoch is $\widetilde{O}(n + n^{2/3}L/\sqrt{\rho\epsilon})$. Thus, we know after

$$\widetilde{O}\left(\frac{L\Delta f}{n^{1/3}\epsilon^2}\cdot n + \frac{\sqrt{\rho}\Delta f}{\epsilon^{1.5}}\cdot(n + \frac{n^{2/3}L}{\sqrt{\rho\epsilon}})\right)$$

stochastic gradients, the function value will decrease below the global optimal f^* with high probability unless we have already met an ϵ-second-order stationary point. Thus, we will at least once get to an ϵ-second-order stationary point within $\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$ stochastic gradients. The formal proof of Theorem 3 is deferred to Appendix E. The proof for Theorem 2 is almost the same except that it uses Lemma 8 instead of Lemma 10 for the guarantee of the super epoch.

5. Conclusion

This paper gives a new algorithm Stabilized SVRG that is able to find an ϵ-second-order stationary point using $\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2} + \frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$ stochastic gradients. To our best knowledge this is the first algorithm that does not rely on a separate negative curvature search subroutine, and it is much simpler than all existing algorithms with similar guarantees. In our proof, we developed the new technique of stabilization (Section 4.3), where we showed if the initial point has exactly 0 gradient and the initial perturbation is small, then the first phase of the algorithm can achieve the guarantee of a negative curvature search subroutine. We believe the stabilization technique can be useful for analyzing other optimization algorithms in nonconvex settings without using an explicit negative curvature search. We hope techniques like this will allow us to develop nonconvex optimization algorithms that are as simple as their convex counterparts.

Acknowledgement

This work was supported by NSF CCF-1704656.

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. *arXiv preprint arXiv:1708.08694*, 2017.

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. *arXiv preprint arXiv:1711.06673*, 2017.

Zhi-Dong Bai and Yong-Qua Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix. *The Annals of Probability*, pages 1729–1741, 1988.

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.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

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

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 pointsonline 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, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017a.

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

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017a.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017b.

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. *arXiv preprint arXiv:1802.04477*, 2018.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in neural information processing systems*, pages 2663–2671, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2904–2913, 2018.

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

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.

Appendix A. Detailed Descriptions of Our Algorithm

In this section, we give the complete descriptions of the Perturbed SVRG and Stabilized SVRG algorithms.

Perturbed SVRG Perturbed SVRG is given in Algorithm 4. The only difference of this algorithm with the high level description in Algorithm 2 is that we have now stated the stopping condition explicitly, and when the algorithm is not running a super epoch, we choose a random iterate as the starting point of the next epoch (this is necessary because of the guarantee in Lemma 5).

In the algorithm, the break probability in Step 16 is used to implement the random stopping. Breaking the loop with this probability is exactly equivalent to finishing the loop and sampling x_{sm+t} for $t = 1, 2, ..., m$ uniformly at random.

Algorithm 4 Perturbed SVRG($x_0, m, b, \eta, \delta, T_{\max}, \mathscr{G}, \mathscr{L}$)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius δ, super-epoch length $T_{\max}$, threshold gradient $\mathscr{G}$, threshold length $\mathscr{L}$

1: $super\_epoch \leftarrow 0$.
2: **for** $s = 0, 1, 2, \cdots$ **do**
3: Compute $\nabla f(x_{sm})$.
4: **if** $super\_epoch = 0 \ \wedge \ \|\nabla f(x_{sm})\| \leq \mathscr{G}$ **then**
5: $super\_epoch \leftarrow 1$.
6: $\widetilde{x} \leftarrow x_{sm}, t_{init} \leftarrow sm$.
7: $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(\delta)$.
8: **end if**
9: **for** $t = 1, 2, \cdots, m$ **do**
10: Sample b i.i.d. numbers uniformly from $[n]$ and form a multi-set I_b.
11: $v_{sm+t-1} \leftarrow \frac{1}{b} \sum_{i \in I_b} \big(\nabla f_i(x_{sm+t-1}) - \nabla f_i(x_{sm}) + \nabla f(x_{sm})\big)\big)$.
12: $x_{sm+t} \leftarrow x_{sm+t-1} - \eta v_{sm+t-1}$.
13: **if** $super\_epoch = 1 \ \wedge \ \big(\|x_{sm+t} - \widetilde{x}\| \geq \mathscr{L} \ \vee \ sm + t - t_{init} \geq T_{\max}\big)$ **then**
14: $super\_epoch \leftarrow 0$; Break.
15: **else if** $super\_epoch = 0$ **then**
16: Break with probability $\frac{1}{m-(t-1)}$.
17: **end if**
18: **end for**
19: $x_{(s+1)m} \leftarrow x_{sm+t}$.
20: **end for**

Stabilized SVRG Stabilized SVRG is given in Algorithm 5. The only differences between Stabilized SVRG and Perturbed SVRG is that Stabilized SVRG adds an additional shift of $-\nabla f(\tilde{x})$ when it is in a super epoch ($stabilizing = 1$ in the algorithm).

Algorithm 5 Stabilized SVRG($x_0, m, b, \eta, \delta, T_{\max}, \mathscr{G}, \mathscr{L}$)

Input: initial point x_0, epoch length m, minibatch size b, step size η, perturbation radius δ, super-epoch length $T_{\max}$, threshold gradient $\mathscr{G}$, threshold length $\mathscr{L}$

$stabilizing \leftarrow 0$.
for $s = 0, 1, 2, \cdots$ **do**
  Compute $\nabla f(x_{sm})$.
  if $stabiling = 0 \ \wedge \ \|\nabla f(x_{sm})\| \leq \mathscr{G}$ **then**
    $stabilizing \leftarrow 1$.
    $\widetilde{x} \leftarrow x_{sm}, t_{init} \leftarrow sm$.
    $x_{sm} \leftarrow x_{sm} + \xi$, where ξ uniformly $\sim \mathbb{B}_0(\delta)$.
  end if
  for $t = 1, 2, \cdots, m$ **do**
    Sample b i.i.d. numbers uniformly from $[n]$ and form a multi-set I_b.
    $v_{sm+t-1} \leftarrow \frac{1}{b} \sum_{i \in I_b} \big(\nabla f_i(x_{sm+t-1}) - \nabla f_i(x_{sm}) + \nabla f(x_{sm}) \big) - stabilizing \times \nabla f(\widetilde{x})$.
    $x_{sm+t} \leftarrow x_{sm+t-1} - \eta v_{sm+t-1}$.
    if $stabilizing = 1 \ \wedge \ \big(\|x_{sm+t} - \widetilde{x}\| \geq \mathscr{L} \ \vee \ sm + t - t_{init} \geq T_{\max} \big)$ **then**
      $stabilizing \leftarrow 0$; Break.
    else if $stabilizing = 0$ **then**
      Break with probability $\frac{1}{m-(t-1)}$.
    end if
  end for
  $x_{(s+1)m} \leftarrow x_{sm+t}$.
end for

Appendix B. Proofs of Exploiting Large Gradients

In this section, we adapt the proof from Li and Li (2018) to show that SVRG can reduce the function value when the gradient is large. First, we give guarantees on the gradient estimate (Lemma 4). Note that previously such bounds were known in the expectation sense, here we convert the bounds to a high probability bound by applying a vector Bernstein's inequality (Lemma 29).

Lemma 4 *For any point x_t, let the gradient estimate be $v_t := \frac{1}{b}\sum_{i\in I_b}(\nabla f_i(x_t) - \nabla f_i(x_{s(t)}) + \nabla f(x_{s(t)}))$, where $x_{s(t)}$ is the snapshot point of the current epoch. Then, with probability at least $1-\zeta$, we have*

$$\|v_t - \nabla f(x_t)\| \le O\Big(\frac{\log(d/\zeta)L}{\sqrt{b}}\Big)\|x_t - x_{s(t)}\|.$$

Proof of Lemma 4. In order to apply Bernstein inequality, we first show for each i, the norm of $(\nabla f_i(x_t) - \nabla f_i(x_{s(t)}) + \nabla f(x_{s(t)}) - \nabla f(x_t))$ is bounded.

$$\begin{aligned}
&\|\nabla f_i(x_t) - \nabla f_i(x_{s(t)}) + \nabla f(x_{s(t)}) - \nabla f(x_t)\| \\
=&\|\nabla f(x_t) - \nabla f(x_{s(t)}) - (\nabla f_i(x_t) - \nabla f_i(x_{s(t)}))\| \\
\le&\|\nabla f(x_t) - \nabla f(x_{s(t)})\| + \|(\nabla f_i(x_t) - \nabla f_i(x_{s(t)}))\| \\
\le&2L\|x_t - x_{s(t)}\|,
\end{aligned}$$

where the last inequality is due to the smoothness of f and f_i.

Then, we bound the summation of variance of each term as follows.

$$\begin{aligned}
\sigma^2 :=& \sum_{i\in I_b}\mathbb{E}[\|\nabla f(x_t) - \nabla f(x_{s(t)}) - (\nabla f_i(x_t) - \nabla f_i(x_{s(t)}))\|^2] \\
\le& \sum_{i\in I_b}\mathbb{E}[\|\nabla f_i(x_t) - \nabla f_i(x_{s(t)})\|^2] \\
\le& \sum_{i\in I_b} L^2\|x_t - x_{s(t)}\|^2 \\
=& bL^2\|x_t - x_{s(t)}\|^2,
\end{aligned}$$

where the first inequality is due to $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \le \mathbb{E}[X^2]$ and the second inequality holds because the gradient of f_i is L-Lipschtiz.

Then, according to the vector version Bernstein inequality (Lemma 29), we have

$$\Pr[\|bv_t - b\nabla f(x_t)\| \ge r] \le (d+1)\exp\Big(\frac{-r^2/2}{bL^2\|x_t - x_{s(t)}\|^2 + \frac{2L\|x_t - x_{s(t)}\|\cdot r}{3}}\Big)$$

Thus, with probability at least $1-\zeta$, we have

$$\|v_t - \nabla f(x_t)\| \le O\Big(\frac{\log(d/\zeta)L}{\sqrt{b}}\Big)\|x_t - x_{s(t)}\|,$$

where $O(\cdot)$ hides constants. □

Using this upperbound on the error of gradient estimates, we can then show that the function value decreases as long as the norms of gradients are large along the path. Note that this part of the proof is also why we require $b \ge m^2$, which results in the $n^{2/3}$ term in the running time.

Lemma 5 *For any epoch, suppose the initial point is x_0, which is also the snapshot point for this epoch. Assume for any $0 \le t \le m-1$, $\|v_t - \nabla f(x_t)\| \le \frac{C_1 L}{\sqrt{b}}\|x_t - x_0\|$, where $C_1 = \widetilde{O}(1)$ comes from Lemma 4. Then, given $\eta \le \frac{1}{3C_1 L}, b \ge m^2$, we have*

$$f(x_0) - f(x_t) \ge \sum_{\tau=0}^{t-1} \frac{\eta}{2}\|\nabla f(x_\tau)\|^2$$

for any $1 \le t \le m$.

Proof of Lemma 5. First, we obtain the relation between $f(x_t)$ and $f(x_{t-1})$ as follows. For any $1 \le t \le m$,

$$\begin{aligned}
f(x_t) \le & f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 && (3)\\
= & f(x_{t-1}) + \langle \nabla f(x_{t-1}) - v_{t-1}, x_t - x_{t-1}\rangle + \langle v_{t-1}, x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \langle \nabla f(x_{t-1}) - v_{t-1}, -\eta v_{t-1}\rangle - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 && (4)\\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \eta\langle \nabla f(x_{t-1}) - v_{t-1}, \nabla f(x_{t-1})\rangle \\
& - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \frac{1}{\eta}\langle x_t - \bar{x}_t, x_{t-1} - \bar{x}_t\rangle - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 && (5)\\
= & f(x_{t-1}) + \eta\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \big(\frac{1}{\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
& - \frac{1}{2\eta}\big(\|x_t - \bar{x}_t\|^2 + \|x_{t-1} - \bar{x}_t\|^2 - \|x_t - x_{t-1}\|^2\big) \\
= & f(x_{t-1}) + \frac{\eta}{2}\|\nabla f(x_{t-1}) - v_{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2, && (6)
\end{aligned}$$

where (3) holds due to smoothness condition, and (4) and (5) follow from these two definitions, i.e., $x_t := x_{t-1} - \eta v_{t-1}$ and $\bar{x}_t := x_{t-1} - \eta\nabla f(x_{t-1})$.

According to the assumption, we have $\|\nabla f(x_{t-1}) - v_{t-1}\|^2 \le \frac{C_1^2 L^2}{b}\|x_{t-1} - x_0\|^2$. Choosing $\eta \le \frac{1}{3C_1 L}$, we have

$$\begin{aligned}
f(x_t) &\le f(x_{t-1}) + \frac{\eta L^2 C_1^2}{2b}\|x_{t-1} - x_0\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \big(\frac{1}{2\eta} - \frac{L}{2}\big)\|x_t - x_{t-1}\|^2 \\
&\le f(x_{t-1}) + \frac{LC_1}{6b}\|x_{t-1} - x_0\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - LC_1\|x_t - x_{t-1}\|^2 \\
&\le f(x_{t-1}) + \big(\frac{L}{6b} + \frac{L}{2t-1}\big)C_1\|x_{t-1} - x_0\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 - \frac{L}{2t}C_1\|x_t - x_0\|^2,
\end{aligned}$$

where the last inequality uses Young's inequality $\|x_t - x_0\|^2 \le \big(1 + \frac{1}{\alpha}\big)\|x_{t-1} - x_0\|^2 + (1+\alpha)\|x_t - x_{t-1}\|^2$ by choosing $\alpha = 2t - 1$.

Now, adding the above inequalities for all iterations $1 \le t \le t'$, where $t' \le m$,

$$\begin{aligned}
f(x_{t'}) \le & f(x_0) - \sum_{t=1}^{t'} \frac{\eta}{2} \|\nabla f(x_{t-1})\|^2 - \sum_{t=1}^{t'} \frac{L}{2t} C_1 \|x_t - x_0\|^2 \\
& + \sum_{t=1}^{t'} \big(\frac{L}{6b} + \frac{L}{2t-1}\big) C_1 \|x_{t-1} - x_0\|^2 \\
= & f(x_0) - \sum_{t=1}^{t'} \frac{\eta}{2} \|\nabla f(x_{t-1})\|^2 - \sum_{t=1}^{t'-1} \big(\frac{L}{2t} - \frac{L}{6b} - \frac{L}{2t+1}\big) C_1 \|x_t - x_0\|^2 \\
& - \frac{L}{2t'} C_1 \|x_{t'} - x_0\|^2 \\
\le & f(x_0) - \sum_{t=1}^{t'} \frac{\eta}{2} \|\nabla f(x_{t-1})\|^2 - \frac{L}{2t'} C_1 \|x_{t'} - x_0\|^2 \qquad (7)
\end{aligned}$$

where (7) holds because $\frac{L}{2t} - \frac{L}{6b} - \frac{L}{2t+1} \ge 0$ for any $1 \le t \le m$ as long as $b \ge m^2$.

Thus, for any $1 \le t' \le m$, we have

$$f(x_0) - f(x_{t'}) \ge \sum_{\tau=0}^{t'-1} \frac{\eta}{2} \|\nabla f(x_\tau)\|^2.$$

□

A limitation of Lemma 5 is that it only guarantees function value decrease when the *sum* of squared gradients is large. However, in order to connect the guarantees between first and second order steps, we want to identify a single iterate that has a small gradient. We achieve this by stopping the SVRG iterations at a uniformly random location.

Lemma 6 *For any epoch, suppose the initial point is x_0. Let x_t be a point uniformly sampled from $\{x_\tau\}_{\tau=1}^m$. Then, given $\eta = \widetilde{\Theta}(1/L), b \ge m^2$, for any value of $\mathscr{G}$, we have two cases:*

1. *if at least half of points in $\{x_\tau\}_{\tau=1}^m$ have gradient no larger than $\mathscr{G}$, we know $\|\nabla f(x_t)\| \le \mathscr{G}$ holds with probability at least $1/2$;*
2. *Otherwise, we know $f(x_0) - f(x_t) \ge \frac{\eta}{2} \frac{m\mathscr{G}^2}{4}$ holds with probability at least $1/5$.*

Further, no matter which case happens we always have $f(x_t) \le f(x_0)$ with high probability.

Proof of Lemma 6. Let $\{x_\tau\}_{\tau=0}^m$ be the iterates of SVRG starting from x_0. Then, there are two cases:

- If at least half of points of $\{x_\tau\}_{\tau=1}^m$ have gradient norm at most $\mathscr{G}$, then it's clear that a uniformly sampled point x_t has gradient norm $\|\nabla f(x_t)\| \le \mathscr{G}$ with probability at least $1/2$.
- Otherwise, we know at least half of points from $\{x_\tau\}_{\tau=1}^m$ has gradient norm larger than $\mathscr{G}$. Then, as long as the sampled point falls into the last quarter of $\{x_\tau\}_{\tau=1}^m$, we know $\sum_{\tau=0}^{t-1} \|\nabla f(x_\tau)\|^2 \ge \frac{m\mathscr{G}^2}{4}$. Thus, for a uniformly sampled point x_t, with probability at least $1/4$, we have $\sum_{\tau=0}^{t-1} \|\nabla f(x_\tau)\|^2 \ge \frac{m\mathscr{G}^2}{4}$.

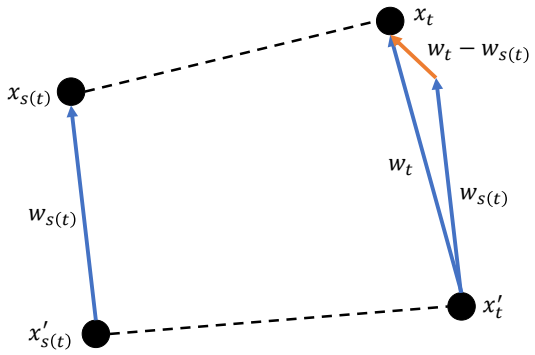


Figure 3: Comparison between $\|w_t - w_{s(t)}\|$ and $\|w_t\| + \|w_{s(t)}\|$

According to Lemma 4 and the union bound, we know there exists $C_1 = \widetilde{O}(1)$ such that with high probability, $\|v_t - \nabla f(x_t)\| \leq \frac{C_1 L}{\sqrt{b}}\|x_t - x_0\|$ holds for every $0 \leq t \leq m-1$. Combining with Lemma 5, we know given $\eta \leq \frac{1}{3C_1 L}, b \geq m^2$, we have $f(x_0) - f(x_t) \geq \sum_{\tau=0}^{t-1} \frac{\eta}{2}\|\nabla f(x_\tau)\|^2$ for any $1 \leq t \leq m$. By another union bound, we know with probability at least $1/5$, $f(x_0) - f(x_t) \geq \frac{\eta}{2}\frac{m\mathscr{G}^2}{4}$.

Again by Lemma 4 and Lemma 5, we know $f(x_t) \leq f(x_0)$ holds with high probability. □

Appendix C. Proofs of Exploiting Negative Curvature - Perturbed SVRG

In this section, we show that starting from a point with negative curvature, Perturbed SVRG can decrease the function value significantly after a super epoch.

As discussed in Section 4.2, we use two point analysis to show that with good probability one of these two points can escape the saddle point. Let $\widetilde{x}$ be the initial point of the super epoch. We consider two coupled samples of the perturbed point x_0, x'_0. The two perturbed points x_0 and x'_0 only differ in the e_1 direction, where e_1 is the smallest eigendirection of Hessian $\mathcal{H} := \nabla^2 f(\widetilde{x})$. Let the SVRG iterates running on f starting from x_0 and x'_0 be $\{x_t\}$ and $\{x'_t\}$ respectively. We will keep track of the difference between the two sequences $w_t = x_t - x'_t$, and show that w_t increases exponentially and becomes large after one super epoch, which means at least one sequence must escape the initial point $\widetilde{x}$.

In the following proof, we first show that the variance of w_t can be well bounded. This is the place where we use the assumption that each individual function is ρ'-Hessian Lipschitz.

Lemma 7 *Let $\{x_t\}$ and $\{x'_t\}$ be two SVRG sequences running on f that use the same choice of mini-batches. Let $x_{s(t)}$ be the snapshot point for iterate t. Let $w_t := x_t - x'_t$ and $P_t = \max(\|x_{s(t)} - \widetilde{x}\|, \|x'_{s(t)} - \widetilde{x}\|, \|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|)$. Then, with probability at least $1-\zeta$, we have*

$$\|\xi_t - \xi'_t\| \leq O\Big(\frac{\log(d/\zeta)}{\sqrt{b}}\Big) \min\Big(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|), L(\|w_t\| + \|w_{s(t)}\|)\Big).$$

In the extreme case, if each individual function f_i is exactly a quadratic function, then we know $\rho' = 0$ and the variance is proportional to $\widetilde{O}(L/\sqrt{b})\|w_t - w_{s(t)}\|$. As illustrated in Figure 3, w_t cannot change very quickly within a single epoch so $\|w_t - w_{s(t)}\|$ is much smaller than $\|w_t\|$ or $\|w_{s(t)}\|$.

Proof of Lemma 7. Similar as the proof in Lemma 4, here we use Bernstein inequality to prove that the difference between the variances of two coupled sequences is also upper bounded.

Recall that,

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

where I_b is a uniformly sampled multi-set of $[n]$ with size b.

Let the Hessian of f at $\widetilde{x}$ be $\mathcal{H}$ and let the Hessian of f_i at $\widetilde{x}$ be $\mathcal{H}_i$ for each i. Let $\xi_{t,i} - \xi_{t,i}'$ be the i-th term in the above sum. In order to apply Bernstein inequality, we first show for each i,

$$\begin{aligned}&\big\|\xi_{t,i} - \xi_{t,i}'\big\|\\ \le&\Big\|(\nabla f_i(x_t) - \nabla f_i(x_t')) - (\nabla f_i(x_{s(t)}) - \nabla f_i(x'_{s(t)}))\Big\|\\ &+\Big\|(\nabla f(x_t) - \nabla f(x_t')) - (\nabla f(x_{s(t)}) - \nabla f(x'_{s(t)}))\Big\|\\ =&\left\|\int_0^1 \nabla^2 f_i(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') - \int_0^1 \nabla^2 f_i(x'_{s(t)} + \theta(x_{s(t)} - x'_{s(t)}))d\theta(x_{s(t)} - x'_{s(t)})\right\|\\ &+\left\|\int_0^1 \nabla^2 f(x_t' + \theta(x_t - x_t'))d\theta(x_t - x_t') - \int_0^1 \nabla^2 f(x'_{s(t)} + \theta(x_{s(t)} - x'_{s(t)}))d\theta(x_{s(t)} - x'_{s(t)})\right\|\\ =&\Big\|\mathcal{H}_i w_t + \Delta_t^i w_t - (\mathcal{H}_i w_{s(t)} + \Delta^i_{s(t)} w_{s(t)})\Big\| + \big\|\mathcal{H} w_t + \Delta_t w_t - (\mathcal{H} w_{s(t)} + \Delta_{s(t)} w_{s(t)})\big\|\\ \le&\|\mathcal{H}_i\|\|w_t - w_{s(t)}\| + \|\Delta_t^i\|\|w_t\| + \|\Delta^i_{s(t)}\|\|w_{s(t)}\|\\ &+\|\mathcal{H}\|\|w_t - w_{s(t)}\| + \|\Delta_t\|\|w_t\| + \|\Delta_{s(t)}\|\|w_{s(t)}\|\\ \le&2L\|w_t - w_{s(t)}\| + 2\rho' P_t(\|w_t\| + \|w_{s(t)}\|)\end{aligned}$$

where $\Delta_t^i = \int_0^1(\nabla^2 f_i(x_t' + \theta(x_t - x_t')) - \mathcal{H}_i)d\theta(x_t - x_t')$ and $\Delta_t = \int_0^1(\nabla^2 f(x_t' + \theta(x_t - x_t')) - \mathcal{H})d\theta(x_t - x_t')$. The last inequality holds since each individual function is L-smooth and ρ' Hessian Lipschitz. Specifically, due to the L-smoothness, we have $\|\mathcal{H}_i\|, \|\mathcal{H}\| \le L$. Because of the Hessian Lipschitz condition and the definition of P_t, we have $\|\Delta_t^i\|, \|\Delta^i_{s(t)}\|, \|\Delta_t\|, \|\Delta_{s(t)}\| \le \rho' P_t$.

Then, we bound the summation of variance of each term as follows.

$$\begin{aligned}&\sigma^2\\ :=&\sum_{i\in I_b}\mathbb{E}\left\|\xi_{t,i} - \xi_{t,i}'\right\|^2\\ \le&\sum_{i\in I_b}\mathbb{E}\left[\left\|(\nabla f_i(x_t) - \nabla f_i(x_t')) - (\nabla f_i(x_{s(t)}) - \nabla f_i(x'_{s(t)}))\right\|^2\right]\\ \le&\sum_{i\in I_b}\big(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|)\big)^2\\ =&b\big(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|)\big)^2,\end{aligned}$$

where the first inequality is due to $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[X^2]$.

Then, according to the vector version Bernstein inequality (Lemma 29), with probability at least $1-\zeta$, we have

$$\|\xi_t - \xi_t'\| \leq O\Big(\frac{\log(d/\zeta)}{\sqrt{b}}\Big)\left(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|)\right),$$

where $O(\cdot)$ hides constants.

In order to prove the other bound for the variance difference, we can use smoothness condition to bound each term as follows.

$$\begin{aligned}
&\|\xi_{t,i} - \xi_{t,i}'\| \\
\leq&\|\nabla f_i(x_t) - \nabla f_i(x_t')\| + \|\nabla f_i(x_{s(t)}) - \nabla f_i(x_{s(t)}')\| \\
&+ \|(\nabla f(x_t) - \nabla f(x_t')\| + \|\nabla f(x_{s(t)}) - \nabla f(x_{s(t)}')\| \\
\leq&2L(\|w_t\| + \|w_{s(t)}\|).
\end{aligned}$$

The summation of variance of each term can be bounded as

$$\sigma^2 \leq L^2(\|w_t\| + \|w_{s(t)}\|)^2.$$

Again, using Bernstein inequality, we know with probability at least $1-\zeta$

$$\|\xi_t - \xi_t'\| \leq O\Big(\frac{\log(d/\zeta)}{\sqrt{b}}\Big)L(\|w_t\| + \|w_{s(t)}\|).$$

By union bound, we know with probability at least $1-2\zeta$,

$$\|\xi_t - \xi_t'\| \leq O\Big(\frac{\log(d/\zeta)}{\sqrt{b}}\Big)\min\Big(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|), L(\|w_t\| + \|w_{s(t)}\|)\Big).$$

□

Suppose the initial point $\widetilde{x}$ of the super epoch has a large negative curvature ($\lambda_{\min}(\mathcal{H}) = -\gamma < 0$). Also assume initially the two sequences has a reasonable distance along e_1 direction, which is the most negative eigendirection of $\mathcal{H}$. Then, using the above bound for the variance of w_t, we are able to prove that the distance between two sequences increases exponentially, and becomes large after $\widetilde{O}(\frac{1}{\eta\gamma})$ steps, which means at least one sequence must escape the initial point $\widetilde{x}$.

Lemma 16 *Let $\{x_t\}$ and $\{x_t'\}$ be two SVRG sequences running on f that use the same choice of mini-batches. Assume $w_0 = x_0 - x_0'$ aligns with e_1 direction and $|\langle e_1, w_0\rangle| \geq \frac{\delta}{4\sqrt{d}}$. Let the threshold distance $\mathscr{L} := \frac{\gamma}{C_3\max(\rho,\rho'/m)}$. Assume for every $0 \leq t \leq \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma} - 1$, $\|\xi_t - \xi_t'\| \leq \frac{C_1'}{\sqrt{b}}\min\left(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|), L(\|w_t\| + \|w_{s(t)}\|)\right)$, where C_1' comes from Lemma 7. Then there exists large enough constant c such that as long as*

$$\eta \leq \frac{1}{c\log(\frac{d\gamma}{\rho\delta})C_1' \cdot L}, \quad C_3 \geq \frac{1}{\eta L}.$$

we have

$$\max(\|x_T - \widetilde{x}\|, \|x_T' - \widetilde{x}\|) \geq \mathscr{L},$$

for some $T \leq \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$.

The proof of this lemma is similar to the analysis in Jin et al. (2017a). However, we make crucial use of Lemma 7. Throughout the proof, the intuition is that at every iteration, w_t is close to a multiple of e_1. Therefore, the next w_{t+1} is close to $(I - \eta H)w_t = (1 + \eta\gamma)w_t$. The difference between w_{t+1} and w_t is therefore only $\eta\gamma w_t$ whose norm is much smaller than either w_t or w_{t+1}. As a result, Lemma 7 gives a much tighter bound on the variance, and allows the proof to go through.

Proof of Lemma 16. For the sake of contradiction, assume for any $t \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$, $\max(\|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|) < \mathscr{L}$. Basically, we will show that the distance between two sequences grows exponentially and will become larger than $2\mathscr{L}$ after $\frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$ steps, which by triangle inequality implies that at least one sequence escapes after $\frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$ steps.

For any $0 \le t \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$, we will inductively prove that

1. $\frac{4}{5}(1+\eta\gamma)^t\|w_0\| \le \|w_t\| \le \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$;
2. $\|\xi_t - \xi'_t\| \le \mu \cdot \eta\gamma C'_1 L(1+\eta\gamma)^t\|w_0\|$, where $\mu = \widetilde{O}(1)$.

The base case trivially holds because $\frac{4}{5}\|w_0\| \le \|w_0\| \le \frac{6}{5}\|w_0\|$ and $\xi_0 = \xi'_0 = 0$. Fix any $t \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$, assume for every $\tau \le t-1$, the two induction hypotheses hold, we prove they still hold for t.

Proving Hypothesis 1. Let's first prove $\frac{4}{5}(1+\eta\gamma)^t\|w_0\| \le \|w_t\| \le \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$. We can expand w_t as follows,

$$\begin{aligned} w_t &= w_{t-1} - \eta(v_{t-1} - v'_{t-1}) \\ &= (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + \xi_{t-1} - \xi'_{t-1}) \\ &= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi'_\tau) \end{aligned}$$

where $\Delta_\tau = \int_0^1(\nabla^2 f(x'_\tau + \theta(x_\tau - x'_\tau)) - \mathcal{H})d\theta$. It's clear that the first term aligns with e direction and has norm $(1+\eta\gamma)^t\|w_0\|$. Thus, we only need to show $\|\eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi'_\tau)\| \le \frac{1}{5}(1+\eta\gamma)^t\|w_0\|$.

We first look at the Hessian changing term. According to the assumptions, we know $\|x_\tau - \widetilde{x}\|, \|x'_\tau - \widetilde{x}\| \le \mathscr{L}$ for any $\tau \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$. Thus,

$$\begin{aligned}
\left\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}\Delta_\tau w_\tau\right\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-\tau-1}\|\Delta_\tau\|\|w_\tau\| \\
&\le \eta\sum_{\tau=0}^{t-1}\rho\max(\|x_\tau-\widetilde{x}\|,\|x'_\tau-\widetilde{x}\|)\frac{6}{5}(1+\eta\gamma)^t\|w_0\| \\
&\le \eta\sum_{\tau=0}^{t-1}\frac{6}{5}\rho\frac{\gamma}{C_3\max(\rho,\rho'/m)}(1+\eta\gamma)^t\|w_0\| \\
&\le \frac{1}{\gamma}\cdot\frac{12}{5}\log(\frac{d\gamma}{\rho\delta})\frac{\gamma}{C_3}(1+\eta\gamma)^t\|w_0\| \\
&\le \frac{1}{10}(1+\eta\gamma)^t\|w_0\|,
\end{aligned}$$

where the second last inequality uses the assumption that $t \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$ and the last inequality holds as long as $C_3 \ge 24\log(\frac{d\gamma}{\rho\delta})$.

For the variance term, we have

$$\begin{aligned}
\left\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\xi_\tau-\xi'_\tau)\right\| &\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-\tau-1}\|\xi_\tau-\xi'_\tau\| \\
&\le \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-\tau-1}\mu\eta\gamma C'_1L(1+\eta\gamma)^\tau\|w_0\| \\
&\le \eta\frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}\mu\eta\gamma C'_1L(1+\eta\gamma)^t\|w_0\| \\
&\le \frac{1}{10}(1+\eta\gamma)^t\|w_0\|,
\end{aligned}$$

where the last inequality holds as long as $\eta \le \frac{1}{20\log(\frac{d\gamma}{\rho\delta})\mu C'_1\cdot L}$.

Overall, we have $\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau+\xi_\tau-\xi'_\tau)\| \le \frac{1}{5}(1+\eta\gamma)^t\|w_0\|$, which implies $\frac{4}{5}(1+\eta\gamma)^t\|w_0\| \le \|w_t\| \le \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$.

Proving Hypothesis 2. Next, we show the second hypothesis also holds, $\|\xi_t-\xi'_t\| \le \mu\cdot\eta\gamma C'_1L(1+\eta\gamma)^t\|w_0\|$. We separately consider two cases when $\frac{1}{\eta\gamma} \le m$ and $\frac{1}{\eta\gamma} > m$.

If $\frac{1}{\eta\gamma} \le m$, we have

$$\begin{aligned}
\|\xi_t-\xi'_t\| &\le \frac{C'_1}{\sqrt{b}}\left(L(\|w_t\|+\|w_{s(t)}\|)\right) \\
&\le \frac{C'_1}{\sqrt{b}}2L\cdot\frac{6}{5}(1+\eta\gamma)^t\|w_0\| \\
&\le \mu\frac{C'_1L}{\sqrt{b}}(1+\eta\gamma)^t\|w_0\| \\
&\le \mu\cdot\eta\gamma C'_1L(1+\eta\gamma)^t\|w_0\|,
\end{aligned}$$

where the third inequality holds as long as $\mu \geq 3$ and the last inequality holds because $\frac{1}{\sqrt{b}} \leq \frac{1}{m} \leq \eta\gamma$.

If $\frac{1}{\eta\gamma} > m$, we need to bound $\|w_t - w_{s(t)}\|$ more carefully. We can write $w_t - w_{s(t)}$ as follows,

$$w_t - w_{s(t)} = \left((I - \eta\mathcal{H})^{t-s(t)} - I\right) w_{s(t)} - \eta \sum_{\tau=s(t)}^{t-1} (I - \eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi'_\tau).$$

For the first term, we have

$$\begin{aligned}\left\|\left((I - \eta\mathcal{H})^{t-s(t)} - I\right) w_{s(t)}\right\| \leq& \|(I - \eta\mathcal{H})^{t-s(t)} - I\|\|w_{s(t)}\| \\ \leq& \left((1+\eta\gamma)^m - 1\right)\frac{6}{5}(1+\eta\gamma)^t\|w_0\| \\ \leq& 3m\eta\gamma \cdot (1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

where the last inequality holds since $(1+\eta\gamma)^m \leq 1 + 2m\eta\gamma$ if $m\eta\gamma < 1$.

For the hessian changing term, we have

$$\begin{aligned}\|\eta \sum_{\tau=s(t)}^{t-1} (I - \eta\mathcal{H})^{t-\tau-1}\Delta_\tau w_\tau\| \leq& \eta \sum_{\tau=s(t)}^{t-1} 2\frac{\gamma}{C_3}(1+\eta\gamma)^t\|w_0\| \\ \leq& \eta m \cdot 2\frac{\gamma}{C_3}(1+\eta\gamma)^t\|w_0\| \\ \leq& m\eta\gamma(1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

assuming $C_3 \geq 2$.

For the variance term, we have

$$\begin{aligned}\|\eta \sum_{\tau=s(t)}^{t-1} (I - \eta\mathcal{H})^{t-\tau-1}(\xi_\tau - \xi'_\tau)\| \leq& \eta \sum_{\tau=s(t)}^{t-1} (1+\eta\gamma)^{t-\tau-1}\|\xi_\tau - \xi'_\tau\| \\ \leq& \eta \sum_{\tau=s(t)}^{t-1} (1+\eta\gamma)^{t-\tau-1}\mu\eta\gamma C'_1 L(1+\eta\gamma)^\tau\|w_0\| \\ \leq& \mu C'_1\eta L \cdot m\eta\gamma(1+\eta\gamma)^t\|w_0\| \\ \leq& m\eta\gamma(1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

where the second inequality uses induction hypothesis and the last inequality assumes $\eta \leq \frac{1}{C'_1\mu\cdot L}$.

Overall, we have $\|w_t - w_{s(t)}\| \leq 5m\eta\gamma(1+\eta\gamma)^t\|w_0\|$. Thus, when $\frac{1}{\eta\gamma} > m$, we can bound $\|\xi_t - \xi_{s(t)}\|$ as follows,

$$\begin{aligned}\|\xi_t - \xi'_t\| \leq& \frac{C'_1}{\sqrt{b}}\left(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|)\right) \\ \leq& \frac{C'_1}{\sqrt{b}}\left(L \cdot 5m\eta\gamma + \rho'\frac{12\gamma}{5C_3\max(\rho, \rho'/m)}\right)(1+\eta\gamma)^t\|w_0\| \\ \leq& \frac{C'_1}{\sqrt{b}}\left(L \cdot 5m\eta\gamma + \frac{12}{5}L \cdot m\eta\gamma\right)(1+\eta\gamma)^t\|w_0\| \\ \leq& \mu \cdot \eta\gamma C'_1 L(1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

where the second last inequality assumes $C_3 \geq \frac{1}{\eta L}$ and the last inequality holds as long as $\mu \geq 8$. Here, we use the fact that $P_t \leq \max(\|x_{s(t)} - \widetilde{x}\|, \|x'_{s(t)} - \widetilde{x}\|, \|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|) \leq \mathscr{L}$.

Overall, we know there exists large enough constant c such that the induction holds as long as

$$\eta \leq \frac{1}{c \log(\frac{d\gamma}{\rho\delta}) C'_1 \cdot L}$$
$$C_3 \geq \frac{1}{\eta L}.$$

Thus, we know $\|w_t\| \geq \frac{4}{5}(1+\eta\gamma)^t \|w_0\|$ for any $t \leq \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$. Specifically, when $t = \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$, we have

$$\begin{aligned}\|w_t\| &\geq \frac{4}{5}(1+\eta\gamma)^t \|w_0\| \\ &\geq \frac{4}{5}(1+\eta\gamma)^{\frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}} \frac{\delta}{4\sqrt{d}} \\ &\geq \frac{\gamma}{5\rho},\end{aligned}$$

which implies $\max(\|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|) \geq \frac{\gamma}{10\rho}$. Assuming $C_3 \geq 10$, this contradicts the assumption that $\max(\|x_t - \widetilde{x}\|, \|x'_t - \widetilde{x}\|) < \frac{\gamma}{C_3 \max(\rho, \rho'/m)} =: \mathscr{L}$, for any $t \leq \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$. Thus, we know there exists $T \leq \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma}$ such that,

$$\max(\|x_T - \widetilde{x}\|, \|x'_T - \widetilde{x}\|) \geq \mathscr{L}.$$

□

In the next lemma, we show that the function value decrease can be lower bounded by the distance to the snapshot point. Combined with the above lemma, this shows that the function value decreases significantly in the super epoch. The proof of this lemma is almost the same as the proof of Lemma 5.

Lemma 9 *Let x_0 be the initial point, which is also the snapshot point of the current epoch. Let $\{x_t\}$ be the iterates of SVRG running on f starting from x_0. Fix any $t \geq 1$, suppose for every $0 \leq \tau \leq t-1, \|\xi_\tau\| \leq \frac{C_1 L}{\sqrt{b}} \|x_\tau - x_{s(\tau)}\|$, where C_1 comes from Lemma 4. Given $\eta \leq \frac{1}{3C_1 L}, b \geq m^2$, we have*

$$\|x_t - x_0\|^2 \leq \frac{4t}{C_1 L}(f(x_0) - f(x_t)).$$

Proof of Lemma 9. From Equation (7) in the proof of Lemma 5, we know for any $t' \leq t$,

$$\|x_{t'} - x_{s(t')}\|^2 \leq \frac{2(t' - s(t'))}{C_1 L}(f(x_{s(t')}) - f(x_{t'})),$$

where $x_{s(t')}$ is the snapshot point of $x_{t'}$.

If $t \leq m$, we know there is only one epoch from x_0 to x_t and

$$\|x_t - x_0\|^2 \leq \frac{2t}{C_1 L}(f(x_0) - f(x_t)).$$

If $t > m$, we need to divide $x_t - x_0$ into multiple epochs and bound them separately. We have

$$\begin{aligned}\|x_t - x_0\|^2 &= \|x_m - x_0 + x_{2m} - x_m + \cdots x_t - x_{s(t)}\|^2 \\ &\leq \lceil \frac{t}{m} \rceil \left(\sum_{\tau=1}^{\lfloor t/m \rfloor} \|x_{\tau m} - x_{(\tau-1)m}\|^2 + \|x_t - x_{s(t)}\|^2 \right) \\ &\leq \frac{2t}{m} \cdot \frac{2m}{C_1 L}(f(x_0) - f(x_t)) \\ &\leq \frac{4t}{C_1 L}(f(x_0) - f(x_t))\end{aligned}$$

Combining two cases, we have

$$\|x_t - x_0\|^2 \leq \frac{4t}{C_1 L}(f(x_0) - f(x_t)).$$

□

Next, we show that starting from a randomly perturbed point, with constant probability the function value decreases a lot within a super epoch.

Lemma 17 *Let $\widetilde{x}$ be the initial point with gradient $\|\nabla f(\widetilde{x})\| \leq \mathscr{G}$ and $\lambda_{\min}(\mathcal{H}) = -\gamma < 0$. Let $\{x_t\}$ be the iterates of SVRG running on f starting from x_0, which is a uniformly perturbed point from $\widetilde{x}$. There exist $\eta = \widetilde{O}(1/L), b = \widetilde{O}(n^{2/3}), \delta = \widetilde{O}(\min(\frac{\rho\gamma}{\max(\rho^2, (\rho'/m)^2)}, \frac{\gamma^{1.5}}{\max(\rho, \rho'/m)\sqrt{L}}))$, $\mathscr{G} = \widetilde{O}(\frac{\gamma^2}{\rho}), \mathscr{L} = \widetilde{O}(\frac{\gamma}{\max(\rho, \rho'/m)}), T_{\max} = \widetilde{O}(\frac{1}{\eta\gamma})$ such that with probability at least $1/8$,*

$$f(x_T) - f(\widetilde{x}) \leq -C_5 \cdot \frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)};$$

and with high probability,

$$f(x_T) - f(\widetilde{x}) \leq \frac{C_5}{20} \cdot \frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)};$$

where $C_5 = \widetilde{\Theta}(1)$ and T is the length of the current super epoch and $T \leq T_{\max}$.

This lemma is basically a combination of Lemma 16 and Lemma 17. Lemma 16 shows that with reasonable probability, one of two random starting points is going to travel a large distance, while Lemma 17 shows such a point would decrease the function value. The only additional thing is to prove is that the function value does not increase by too much when the point does not escape. Intuitively this is true because with high probability the function value can only increase during the initial perturbation.

Proof of Lemma 17. With the help of Lemma 16, we first prove that $\{x_t\}$ escapes the saddle point with a constant probability. Let $\{x_t\}$ and $\{x'_t\}$ be two SVRG sequences starting from x_0 and x'_0 respectively, where x_0 and x'_0 are two perturbed points satisfying $\|x_0 - \widetilde{x}\|, \|x'_0 - \widetilde{x}\| \leq \delta$. According to Lemma 16, we know at least one sequence escapes the saddle point if $x_0 - x'_0$ aligns with e_1 direction and has norm as least $\frac{\delta}{4\sqrt{d}}$.

We first show that, for two coupled random points x_0 and x_0', their distance is at least $\frac{\delta}{4\sqrt{d}}$ with a reasonable probability. Marginally, x_0 and x_0' are both uniformly sampled from the ball centered at $\widetilde{x}$ with radius δ. They are coupled in the sense that they have the same projections onto the orthogonal subspace of e_1. Then, similar as the analysis in Jin et al. (2017a),

$$\Pr\left[\|x_0 - x_0'\| < \frac{\delta}{4\sqrt{d}}\right] \le \frac{1}{2}\frac{\frac{\delta}{\sqrt{d}} \times \mathrm{Vol}(\mathbb{B}_0^{(d-1)}(\delta))}{\mathrm{Vol}(\mathbb{B}_0^{(d)}(\delta))} = \frac{1}{2}\frac{1}{\sqrt{\pi d}}\frac{\Gamma(d/2+1)}{\Gamma(d/2+1/2)} \le \frac{1}{2}.$$

Thus, we know with at least half probability, we have $|\langle x_0 - x_0', e_1\rangle| \ge \frac{\delta}{4\sqrt{d}}$. In order to apply Lemma 16, we still need to make sure $\|\xi_t - \xi_t'\|$ is well bounded for every $0 \le t \le \frac{2\log(\frac{d\gamma}{\rho\delta})}{\eta\gamma} - 1$, which happens with high probability due to Lemma 7. Thus, by the union bound and Lemma 16, we know with probability no less than $1/3$, at least one sequence between $\{x_t\}$ and $\{x_t'\}$ must escape the saddle point. Marginally, we know from a randomly perturbed point x_0, sequence $\{x_t\}$ escapes the saddle point within a super epoch with probability at least $1/6$. Precisely, there exists $\eta = \frac{1}{C_6 \cdot L}, \mathscr{L} = \frac{\gamma}{C_3 \max(\rho, \rho'/m)}, T \le \frac{C_7}{\eta\gamma}$ such that

$$\|x_T - \widetilde{x}\| \ge \mathscr{L}$$

holds with probability at least $1/6$. Here, we have $C_3, C_6, C_7 = \widetilde{O}(1)$.

Combing Lemma 4 and Lemma 9, we also know with high probability

$$\|x_T - x_0\|^2 \le \frac{T}{C_4 L}(f(x_0) - f(x_T))$$

where $C_4 = \widetilde{O}(1)$.

By a union bound, we know with probability at least $1/8$, we have

$$\begin{aligned}
f(x_0) - f(x_T) \ge& \frac{C_4 L}{T}\|x_T - x_0\|^2 \\
\ge& \frac{C_4 L}{T}\left(\|x_T - \widetilde{x}\| - \|x_0 - \widetilde{x}\|\right)^2 \\
\ge& \frac{C_4 L}{T}\left(\frac{\gamma}{C_3 \max(\rho, \rho'/m)} - \delta\right)^2 \\
\ge& \frac{C_4 L \eta\gamma}{C_7}\frac{\gamma^2}{4C_3^2 \max(\rho^2, (\rho'/m)^2)} \\
=& \frac{C_4}{4C_7 C_3^2 C_6}\frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)},
\end{aligned}$$

where the last inequality holds as long as $\delta \le \frac{\gamma}{2C_3 \max(\rho, \rho'/m)}$.

Let the threshold gradient $\mathscr{G} := \frac{\gamma^2}{C_8\rho}$. Since f is L-smooth, we have

$$\begin{aligned}
f(x_0) - f(\widetilde{x}) \le& \|\nabla f(\widetilde{x})\| \cdot \|x_0 - \widetilde{x}\| + \frac{L}{2}\|\widetilde{x} - x_0\|^2 \\
\le& \frac{\gamma^2}{C_8\rho}\delta + \frac{L}{2}\delta^2.
\end{aligned}$$

Thus, with probability at least $1/8$, we know

$$\begin{aligned} f(x_T) - f(\widetilde{x}) =& f(x_T) - f(x_0) + f(x_0) - f(\widetilde{x}) \\ \leq& -\frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)} + \frac{\gamma^2}{C_8\rho}\delta + \frac{L}{2}\delta^2. \end{aligned}$$

If Lemma 16 fails, the function value is not guaranteed to decrease. On the other hand, we know that with high probability the function value does not increase, $f(x_T) - f(x_0) \leq 0$. Thus, with high probability, we know

$$f(x_T) - f(\widetilde{x}) \leq \frac{\gamma^2}{C_8\rho}\delta + \frac{L}{2}\delta^2.$$

Assuming $\delta \leq \min(\frac{C_4C_8}{168C_7C_3^2C_6}\frac{\rho\gamma}{\max(\rho^2,(\rho'/m)^2)}, \sqrt{\frac{C_4}{84C_7C_3^2C_6}}\frac{\gamma^{1.5}}{\max(\rho,\rho'/m)\sqrt{L}})$, we know with probability at least $1/8$,

$$f(x_T) - f(\widetilde{x}) \leq -\frac{20}{21}\cdot\frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)};$$

and with high probability,

$$f(x_T) - f(\widetilde{x}) \leq \frac{1}{21}\cdot\frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\max(\rho^2, (\rho'/m)^2)}.$$

We finish the proof by choosing $C_5 := \frac{20}{21}\frac{C_4}{4C_7C_3^2C_6}$. □

Appendix D. Proofs of Exploiting Negative Curvature - Stabilized SVRG

In this section, we analyze the behavior of Stabilized SVRG when the initial gradient is small. The proofs will depend on Lemma 4, Lemma 7 and Lemma 9, which were proved for f but clearly also holds for shifted function $\hat{f}$.

Let the initial point of the super epoch be $\widetilde{x}$, whose hessian is denoted by $\mathcal{H}$. Assume the initial point has large negative curvature, $\lambda_{\min}(\mathcal{H}) = -\gamma < 0$. Let x_0 be the perturbed point and let $\{x_t\}$ be the SVRG iterates running on $\hat{f}$ starting from $\widetilde{x}$. As we discussed in Section 4.3, there are two phases in the analysis. In the first phase, the distance between the current iterate x_t and the starting point $\widetilde{x}$ remains small (comparable to the random perturbation), while at the end the direction of $x_t - \widetilde{x}$ aligns with the negative eigendirections. In the second phase, the distance to the initial point $\widetilde{x}$ blows up exponentially and the algorithm escapes from saddle points.

To analyze the two phases of the algorithm, we make use of the following expansion for the one-step movement of the algorithm:

Lemma 18 *Let $\widetilde{x}$ be the initial point with Hessian $\mathcal{H}$, and x_0 be its perturbed point. Let $\{x_t\}$ be the iterates of SVRG running on $\hat{f}$ starting from x_0. For any $t \geq 1$, we have the following expansion,*

$$\begin{aligned} x_t - x_{t-1} =& -\eta(I - \eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0) + \eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I - \eta\mathcal{H})^{t-2-\tau}\xi_\tau \\ & - \eta\sum_{\tau=0}^{t-2}(I - \eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1} - x_\tau) - \eta\xi_{t-1}, \end{aligned}$$

where variance term $\xi_\tau = v_\tau - \nabla \hat{f}(x_\tau)$ *and hessian changing term* $\Delta_\tau = \int_0^1 (\nabla^2 \hat{f}(x_\tau + \theta(x_{\tau+1} - x_\tau)) - \mathcal{H}) d\theta$.

Intuitively, the first term $-\eta(I - \eta\mathcal{H})^{t-1}\nabla \hat{f}(x_0)$ corresponds to what happens to the algorithm if the function is quadratic (with Hessian equal to $\mathcal{H}$ at $\widetilde{x}$). The second and the fourth term measures the difference introduced by the error in the gradient updates. The third term measures the difference introduced by the fact that the Hessian is not a constant. Our analysis will bound the last three terms to show that the behavior of the algorithm is very similar to what happens if we only have the first term.

Proof of Lemma 18. According to the algorithm, we know

$$\begin{aligned} x_t - x_{t-1} &= -\eta v_{t-1} \\ &= -\eta(\nabla \hat{f}(x_{t-1}) + \xi_{t-1}), \end{aligned}$$

where $\xi_{t-1} = v_{t-1} - \nabla \hat{f}(x_{t-1})$. We can further expand $\nabla \hat{f}(x_t)$ as follows.

$$\begin{aligned} \nabla \hat{f}(x_t) &= \nabla \hat{f}(x_{t-1}) + \int_0^1 \Big(\nabla^2 \hat{f}\big(x_{t-1} + \theta(x_t - x_{t-1})\big)\Big) d\theta (x_t - x_{t-1}) \\ &= \nabla \hat{f}(x_{t-1}) + \mathcal{H}(x_t - x_{t-1}) + \Delta_{t-1}(x_t - x_{t-1}) \\ &= \nabla \hat{f}(x_{t-1}) - \eta\mathcal{H}(\nabla \hat{f}(x_{t-1}) + \xi_{t-1}) + \Delta_{t-1}(x_t - x_{t-1}) \\ &= (I - \eta\mathcal{H})\nabla \hat{f}(x_{t-1}) - \eta\mathcal{H}\xi_{t-1} + \Delta_{t-1}(x_t - x_{t-1}) \\ &= (I - \eta\mathcal{H})^t \nabla \hat{f}(x_0) - \eta\mathcal{H}\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}\xi_\tau + \sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}\Delta_\tau(x_{\tau+1} - x_\tau), \end{aligned}$$

where $\Delta_\tau = \int_0^1 (\nabla^2 \hat{f}(x_\tau + \theta(x_{\tau+1} - x_\tau)) - \mathcal{H}) d\theta$. Thus, we know

$$\begin{aligned} x_t - x_{t-1} =& -\eta(\nabla \hat{f}(x_{t-1}) + \xi_{t-1}) \\ =& -\eta(I - \eta\mathcal{H})^{t-1}\nabla \hat{f}(x_0) + \eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I - \eta\mathcal{H})^{t-2-\tau}\xi_\tau \\ & - \eta\sum_{\tau=0}^{t-2}(I - \eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1} - x_\tau) - \eta\xi_{t-1} \end{aligned}$$

□

D.1. Proofs of Phase 1

In Phase 1, the goal of the algorithm is to stay close to the original point $\widetilde{x}$, while making $x_t - \widetilde{x}$ aligned with the negative eigendirections of $\mathcal{H}$ (Hessian at $\widetilde{x}$).

Recall the definition of the length of Phase 1 as follows,

$$T_1 = \sup\left\{t | \forall t' \le t-1, \left(t' \le \frac{1}{\eta\gamma}\right) \vee \left(\|\text{Proj}_S(x_{t'} - \widetilde{x})\| \le \frac{\delta}{10}\right)\right\}.$$

We will first show that $x_t - x_{t-1}$ is bounded by $\widetilde{O}(1/t)\delta$ for every $1 \le t \le \min(T_1, \frac{\log(d)}{\eta\gamma})$. This lemma is very technical, and the main idea is to use the expansion in Lemma 18 and bound the terms by considering their projections in different subspaces. Intuitively, the behavior can be separated into several cases based on the eigenvalues of $\mathcal{H}$ in the corresponding subspace:

1. eigenvalue smaller than $-\gamma/\log d$. These directions will grow exponentially, and we will stop the first phase when the projection in this subspace is large.
2. eigenvalue between $-\gamma/\log d$ and 0. These directions will also grow, but they do not grow by more than a constant factor.
3. small positive eigenvalue (smaller than $+\gamma$). These directions don't move much throughout the iterates.
4. large positive eigenvalue (much larger than γ). These directions move very fast at the beginning, but converges very quickly and will not move much later on.

In the proof we will consider the behavior of these separate subspaces (where cases 3 and 4 will be combined). The detailed proof is deferred to Section D.2.

Lemma 19 *Let T_1 be the length of Phase 1. Assume for any $0 \le t \le \min(T_1, \frac{\log(d)}{\eta\gamma}) - 1$, $\|\xi_t\| \le \frac{C_1 L}{\sqrt{b}}\|x_t - x_{s(t)}\|$, where C_1 comes from Lemma 4. Then, there exists large enough constant c such that as long as*

$$\eta \le \frac{1}{cC_1 \log(nd)\log(n\frac{\log(d)}{\eta\gamma}) \cdot L}, \quad \mu \ge c\log(d)\log^2(\frac{\log(d)}{\eta\gamma}), \quad \delta \le \frac{\gamma}{\rho\mu^2},$$

we have for every $1 \le t \le \min(T_1, \frac{\log(d)}{\eta\gamma})$,

$$\|x_t - x_{t-1}\| \le \frac{\mu}{t}\delta.$$

Now we want to prove that Phase 1 is successful with a reasonable probability. That is, at the end of Phase 1, with reasonable probability the distance $x_{T_1} - \widetilde{x}$ is order $\widetilde{O}(\delta)$, while $\mathrm{Proj}_S(x_{T_1} - \widetilde{x})$ is at least $\delta/10$, where δ is the perturbation radius. By the above lemma, actually we only need to show that the length of Phase 1 is bounded by $\frac{\log(d)}{\eta\gamma}$. In the following proof, we show that between a pair of coupled sequences, at least one of them must end the Phase 1 within $\frac{\log(d)}{\eta\gamma}$ steps. Similar as in Lemma 16, we use two point analysis to show the difference between two sequences along e_1 direction increases exponentially and will become very large after $\frac{\log(d)}{\eta\gamma}$ steps, which implies that at least one sequence must have a large projection on S subspace.

Lemma 20 *Let $\{x_t\}$ and $\{x'_t\}$ be two SVRG sequences running on $\hat{f}$ that use the same choice of mini-batches. Assume $w_0 = x_0 - x'_0$ aligns with e_1 direction and $|\langle e_1, w_0\rangle| \ge \frac{\delta}{4\sqrt{d}}$. Let T_1, T'_1 be the length of Phase 1 for $\{x_t\}$ and $\{x'_t\}$ respectively. Assume for every $1 \le t \le \min(T_1, \frac{\log(d)}{\eta\gamma})$, $\|x_t - x_{t-1}\| \le \frac{C_2}{t}\delta$ and for every $1 \le t \le \min(T'_1, \frac{\log(d)}{\eta\gamma})$, $\|x'_t - x'_{t-1}\| \le \frac{C_2}{t}\delta$, where C_2 comes from Lemma 19. Assume for every $0 \le t \le \frac{\log(d)}{\eta\gamma} - 1$, $\|\xi_t - \xi'_t\| \le$*

$\frac{C_1'}{\sqrt{b}} \min\left(L\|w_t - w_{s(t)}\| + \rho' P_t(\|w_t\| + \|w_{s(t)}\|), L(\|w_t\| + \|w_{s(t)}\|)\right)$, *where* C_1' *comes from Lemma 7. Then there exists large enough constant* c *such that as long as*

$$\delta \leq \min\left(\frac{\gamma}{c\log(d)\log(\frac{\log(d)}{\eta\gamma})C_2\rho}, \frac{m\eta L\gamma}{\rho'}\right), \quad \eta \leq \frac{1}{c\log(d)\log(\frac{\log(d)}{\eta\gamma})C_1'C_2 \cdot L},$$

we have $\min(T_1, T_1') \leq \frac{\log(d)}{\eta\gamma}$. *W.l.o.g., suppose* $T_1 \leq \frac{\log(d)}{\eta\gamma}$ *and we further have*

$$\forall 0 \leq t \leq T_1,\ \|x_t - \widetilde{x}\| \leq 3\log(\frac{\log(d)}{\eta\gamma})C_2\delta,$$
$$\|\mathit{Proj}_S(x_{T_1} - \widetilde{x})\| \geq \frac{1}{10}\delta.$$

Proof of Lemma 20. For the sake of contradiction, assume the length of Phase 1 for both sequences are larger than $\frac{\log(d)}{\eta\gamma}$. Basically, we will show that the distance between two sequences along e_1 direction grows exponentially and will become very large after $\frac{\log(d)}{\eta\gamma}$ steps, which implies that at least one sequence has a large projection along e_1 direction after $\frac{\log(d)}{\eta\gamma}$ steps.

For any $0 \leq t \leq \frac{\log(d)}{\eta\gamma}$, we will inductively prove that

1. $\|\text{Proj}_{e_1} w_t\| \geq \frac{4}{5}(1+\eta\gamma)^t\|w_0\|$ and $\|w_t\| \leq \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$;
2. $\|\xi_t - \xi_t'\| \leq \mu \cdot \eta\gamma C_1' L(1+\eta\gamma)^t\|w_0\|$, where $\mu = \widetilde{O}(1)$.

The base case trivially holds. Fix any $t \leq \frac{\log(d)}{\eta\gamma}$, assume for every $\tau \leq t-1$, the two induction hypotheses hold, we prove they still hold for t.

Proving Hypothesis 1. Let's first prove $\|\text{Proj}_{e_1} w_t\| \geq \frac{4}{5}(1+\eta\gamma)^t\|w_0\|$ and $\|w_t\| \leq \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$. We can expand w_t as follows,

$$\begin{aligned} w_t &= w_{t-1} - \eta(v_{t-1} - v_{t-1}') \\ &= (I - \eta\mathcal{H})w_{t-1} - \eta(\Delta_{t-1}w_{t-1} + \xi_{t-1} - \xi_{t-1}') \\ &= (I - \eta\mathcal{H})^t w_0 - \eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi_\tau') \end{aligned}$$

where $\Delta_\tau = \int_0^1(\nabla^2\hat{f}(x_\tau' + \theta(x_\tau - x_\tau')) - \mathcal{H})d\theta$. It's clear that the first term aligns with e direction and has norm $(1+\eta\gamma)^t\|w_0\|$. Thus, we only need to show $\|\eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi_\tau')\| \leq \frac{1}{5}(1+\eta\gamma)^t\|w_0\|$.

We first look at the Hessian changing term. According to the assumptions, we know $\|x_\tau - \widetilde{x}\|, \|x'_\tau - \widetilde{x}\| \leq 3\log(\frac{\log(d)}{\eta\gamma})C_2\delta$ for any $\tau \leq \frac{\log(d)}{\eta\gamma}$. Thus,

$$\begin{aligned}
\left\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}\Delta_\tau w_\tau\right\| &\leq \eta\sum_{\tau=0}^{t-1}(1+\eta\gamma)^{t-\tau-1}\|\Delta_\tau\|\|w_\tau\| \\
&\leq \eta\sum_{\tau=0}^{t-1}\rho\max(\|x_\tau-\widetilde{x}\|,\|x'_\tau-\widetilde{x}\|)\frac{6}{5}(1+\eta\gamma)^t\|w_0\| \\
&\leq \eta\sum_{\tau=0}^{t-1}\frac{18}{5}\log(\frac{\log(d)}{\eta\gamma})C_2\rho\delta(1+\eta\gamma)^t\|w_0\| \\
&\leq \frac{1}{\gamma}\cdot 4\log(d)\log(\frac{\log(d)}{\eta\gamma})C_2\rho\delta(1+\eta\gamma)^t\|w_0\| \\
&\leq \frac{1}{10}(1+\eta\gamma)^t\|w_0\|,
\end{aligned}$$

where the last inequality holds as long as $\delta \leq \frac{\gamma}{40\log(d)\log(\frac{\log(d)}{\eta\gamma})C_2\rho}$.

By the analysis in Lemma 16, we can bound the variance term as follows,

$$\left\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\xi_\tau-\xi'_\tau)\right\| \leq \frac{1}{10}(1+\eta\gamma)^t\|w_0\|,$$

assuming $\eta \leq \frac{1}{10\log(d)\mu C'_1\cdot L}$.

Overall, we have $\|\eta\sum_{\tau=0}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau+\xi_\tau-\xi'_\tau)\| \leq \frac{1}{5}(1+\eta\gamma)^t\|w_0\|$, which implies $\|\text{Proj}_e w_t\| \geq \frac{4}{5}(1+\eta\gamma)^t\|w_0\|$ and $\|w_t\| \leq \frac{6}{5}(1+\eta\gamma)^t\|w_0\|$.

Proving Hypothesis 2. Next, we show the second hypothesis also holds, $\|\xi_t-\xi'_t\| \leq \mu\cdot\eta\gamma C'_1L(1+\eta\gamma)^t\|w_0\|$. We separately consider two cases when $\frac{1}{\eta\gamma} \leq m$ and $\frac{1}{\eta\gamma} > m$. If $\frac{1}{\eta\gamma} \leq m$, the analysis is same as in Lemma 16. We have $\|\xi_t - \xi'_t\| \leq \mu\cdot\eta\gamma C'_1L(1+\eta\gamma)^t\|w_0\|$, as long as $\mu \geq 3$.

If $\frac{1}{\eta\gamma} > m$, we need to bound $\|w_t - w_{s(t)}\|$ more carefully. We can write $w_t - w_{s(t)}$ as follows,

$$w_t - w_{s(t)} = \left((I-\eta\mathcal{H})^{t-s(t)} - I\right)w_{s(t)} - \eta\sum_{\tau=s(t)}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\Delta_\tau w_\tau + \xi_\tau - \xi'_\tau).$$

The analysis for the first term and the variance term is again same as in Lemma 16. We have

$$\left\|\left((I-\eta\mathcal{H})^{t-s(t)} - I\right)w_{s(t)}\right\| + \left\|\eta\sum_{\tau=s(t)}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}(\xi_\tau-\xi'_\tau)\right\| \leq 4m\eta\gamma\cdot(1+\eta\gamma)^t\|w_0\|,$$

assuming $\eta \leq \frac{1}{C'_1\mu\cdot L}$.

For the Hessian changing term, we have

$$\begin{aligned}\left\|\eta\sum_{\tau=s(t)}^{t-1}(I-\eta\mathcal{H})^{t-\tau-1}\Delta_\tau w_\tau\right\| &\le \eta\sum_{\tau=s(t)}^{t-1}3\log(\frac{\log(d)}{\eta\gamma})C_2\rho\delta\frac{6}{5}(1+\eta\gamma)^t\|w_0\|\\ &\le \eta m\cdot 4\log(\frac{\log(d)}{\eta\gamma})C_2\rho\delta(1+\eta\gamma)^t\|w_0\|\\ &\le m\eta\gamma(1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

assuming $\delta\le\frac{\gamma}{4\log(\frac{\log(d)}{\eta\gamma})C_2\rho}$.

Overall, we have $\|w_t-w_{s(t)}\|\le 5m\eta\gamma(1+\eta\gamma)^t\|w_0\|$. Thus, when $\frac{1}{\eta\gamma}>m$, we can bound $\|\xi_t-\xi_t'\|$ as follows,

$$\begin{aligned}\|\xi_t-\xi_t'\|\le&\frac{C_1'}{\sqrt{b}}\left(L\|w_t-w_{s(t)}\|+\rho' P_t(\|w_t\|+\|w_{s(t)}\|)\right)\\ \le&\frac{C_1'}{\sqrt{b}}\left(L\cdot 5m\eta\gamma+8\log(\frac{\log(d)}{\eta\gamma})C_2\rho'\delta\right)(1+\eta\gamma)^t\|w_0\|\\ \le&\frac{C_1'}{\sqrt{b}}\left(L\cdot 5m\eta\gamma+L\cdot 8\log(\frac{\log(d)}{\eta\gamma})C_2 m\eta\gamma\right)(1+\eta\gamma)^t\|w_0\|\\ \le&\mu\cdot\eta\gamma C_1' L(1+\eta\gamma)^t\|w_0\|,\end{aligned}$$

where the second last inequality assumes $\delta\le\frac{m\eta L\gamma}{\rho'}$ and the last inequality holds as long as $\mu\ge 5+8\log(\frac{\log(d)}{\eta\gamma})C_2$. Here, we also use the fact that $P_t\le\max(\|x_{s(t)}-\widetilde{x}\|,\|x'_{s(t)}-\widetilde{x}\|,\|x_t-\widetilde{x}\|,\|x'_t-\widetilde{x}\|)\le 3\log(\frac{\log(d)}{\eta\gamma})C_2\delta$.

Overall, we know there exists large enough constant c such that the induction holds given

$$\begin{aligned}\delta&\le\min\left(\frac{\gamma}{c\log(d)\log(\frac{\log(d)}{\eta\gamma})C_2\rho},\frac{m\eta L\gamma}{\rho'}\right)\\ \eta&\le\frac{1}{c\log(d)\log(\frac{\log(d)}{\eta\gamma})C_1'C_2\cdot L}.\end{aligned}$$

Thus, we know $\|\text{Proj}_{e_1}w_t\|\ge\frac{4}{5}(1+\eta\gamma)^t\|w_0\|$ for any $t\le\frac{\log(d)}{\eta\gamma}$. Specifically, when $t=\frac{\log(d)}{\eta\gamma}$, we have

$$\begin{aligned}\|\text{Proj}_{e_1}w_t\|&\ge\frac{4}{5}(1+\eta\gamma)^t\|w_0\|\\ &\ge\frac{4}{5}(1+\eta\gamma)^{\frac{\log(d)}{\eta\gamma}}\frac{\delta}{4\sqrt{d}}\\ &>\frac{\delta}{5},\end{aligned}$$

which implies $\max(\|\text{Proj}_{e_1}x_t-\widetilde{x}\|,\|\text{Proj}_{e_1}x'_t-\widetilde{x}\|)>\frac{\delta}{10}$. This contradicts the assumption that neither sequence stops within $\frac{\log(d)}{\eta\gamma}$ steps. Thus, we know $\min(T_1,T_1')\le\frac{\log(d)}{\eta\gamma}$. Without loss of

generality, suppose $T_1 \leq \frac{\log(d)}{\eta\gamma}$, we have

$$\forall 0 \leq t \leq T_1,\ \|x_t - \widetilde{x}\| \leq 3\log(\frac{\log(d)}{\eta\gamma})C_2\delta,$$
$$\|\text{Proj}_S(x_{T_1} - \widetilde{x})\| \geq \frac{1}{10}\delta.$$

□

D.2. Proof of Lemma 19

In this section, we show that in Phase 1 the total movement is bounded by $\widetilde{O}(\delta)$ within $\frac{\log(d)}{\eta\gamma}$ steps. We recall Lemma 19 as follows.

Lemma 19 *Let T_1 be the length of Phase 1. Assume for any $0 \leq t \leq \min(T_1, \frac{\log(d)}{\eta\gamma}) - 1$, $\|\xi_t\| \leq \frac{C_1 L}{\sqrt{b}}\|x_t - x_{s(t)}\|$, where C_1 comes from Lemma 4. Then, there exists large enough constant c such that as long as*

$$\eta \leq \frac{1}{cC_1\log(nd)\log(n\frac{\log(d)}{\eta\gamma})\cdot L}, \quad \mu \geq c\log(d)\log^2(\frac{\log(d)}{\eta\gamma}), \quad \delta \leq \frac{\gamma}{\rho\mu^2},$$

we have for every $1 \leq t \leq \min(T_1, \frac{\log(d)}{\eta\gamma})$,

$$\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\delta.$$

Proof of Lemma 19.

We prove for every $1 \leq t \leq \min(T_1, \frac{\log(d)}{\eta\gamma})$, $\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\delta$ by induction. For the base case, we have $x_1 - x_0 = -\eta\nabla\hat{f}(x_0)$. Since the gradient at $\widetilde{x}$ is zero, we have

$$\begin{aligned}\|\nabla\hat{f}(x_0)\| &= \|\nabla\hat{f}(x_0) - \nabla\hat{f}(\widetilde{x})\| \\ &\leq L\|x_0 - \widetilde{x}\| \\ &\leq L\delta,\end{aligned}$$

where the first inequality holds since f ($\hat{f}$) is L-smooth. As long as $\mu \geq \eta L$, we have $\|x_1 - x_0\| \leq \mu\delta$.

Fix any $t \leq \min(T_1, \frac{\log(d)}{\eta\gamma})$, suppose for any $t' \leq t-1, \|x_{t'} - x_{t'-1}\| \leq \frac{\mu}{t'}\delta$, we will prove $\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\delta$. In order to prove $\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\delta$, we will separately bound its projections onto three orthogonal subspaces. Specifically, we consider the following three subspaces:

- S: subspace spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $[-\gamma, -\frac{\gamma}{\log(d)}]$.
- $S_-^\perp$: subspace spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $(-\frac{\gamma}{\log(d)}, 0]$.
- $S_+^\perp$: subspace spanned by the eigenvectors of $\mathcal{H}$ with eigenvalues within $(0, L]$.

Regarding the projections onto $S_-^\perp$ and $S_+^\perp$, we will use the following expansion of $x_t - x_{t-1}$,

$$\begin{aligned} & x_t - x_{t-1} \\ =& -\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0) + \eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau \\ & -\eta\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1}-x_\tau) - \eta\xi_{t-1} \end{aligned} \tag{8}$$

and bound its four terms one by one. In the expansion, we denote $\Delta_\tau := \int_0^1(\nabla^2\hat{f}(x_\tau+\theta(x_{\tau+1}-x_\tau))-\mathcal{H})d\theta$.

For the projection in subspace S, after $\frac{1}{\eta\gamma}$ steps, we cannot bound it using the above expansion since the exponential factor can be very large. Instead, we bound the projection in subspace S by the stopping condition $\|\text{Proj}_S(x_{t-1}-\widetilde{x})\| \le \frac{\delta}{10}$ using an alternative expansion,

$$\begin{aligned} x_t - x_{t-1} =& -\eta(\nabla\hat{f}(x_{t-1})+\xi_{t-1}) \\ =& -\eta\mathcal{H}(x_{t-1}-\widetilde{x}) - \eta\Delta'_{t-1}(x_{t-1}-\widetilde{x}) - \eta\xi_{t-1}, \end{aligned}$$

where $\Delta'_{t-1} = \int_0^1(\nabla^2\hat{f}(\widetilde{x}+\theta(x_{t-1}-\widetilde{x}))-\mathcal{H})d\theta$.

We will first bound the projections of $x_t - x_{t-1}$ on $S_+^\perp$ and $S_-^\perp$ by considering the four terms in Eqn. 8. For the first term, the projection in subspace $S_-^\perp$ can increase but will not increase by more than a constant factor; the projection in $S_+^\perp$ might start large but will decrease as the number of iterations increases.

Bounding $\|\text{Proj}_{S_-^\perp}\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\|$: For this term we will show that its projection on $S_-^\perp$ is small to begin with and cannot be amplified by more than a constant. Recall that $\nabla\hat{f}(x_0) = \mathcal{H}(x_0-\widetilde{x}) + \Delta(x_0-\widetilde{x})$, where $\Delta = \int_0^1(\nabla^2\hat{f}(\widetilde{x}+\theta(x_0-\widetilde{x}))-\mathcal{H})d\theta$. Due to the Hessian lipschitzness of f, we have $\|\Delta\| \le \rho\delta$. Then, we can bound $\eta\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\|$ as follows.

$$\begin{aligned} \eta\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\| =& \eta\left\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\left(\mathcal{H}(x_0-\widetilde{x})+\Delta(x_0-\widetilde{x})\right)\right\| \\ \le& \eta\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\mathcal{H}(x_0-\widetilde{x})\| \\ & + \eta\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\Delta(x_0-\widetilde{x})\| \\ \le& \eta(1+\frac{\eta\gamma}{\log(d)})^{\frac{\log(d)}{\eta\gamma}}\frac{\gamma}{\log(d)}\delta + \eta(1+\frac{\eta\gamma}{\log(d)})^{\frac{\log(d)}{\eta\gamma}}\rho\delta^2 \\ \le& \frac{e}{\log(d)}\eta\gamma\delta + e\eta\rho\delta^2 \\ \le& 2e\eta\gamma\delta, \end{aligned}$$

where the last inequality holds as long as $\delta \le \frac{\gamma}{\rho}$. Since $t \le \frac{\log(d)}{\eta\gamma}$, we have

$$\eta\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\| \le \frac{2e\log(d)}{t}\delta.$$

Bounding $\|\mathbf{Proj}_{S_+^\perp}\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\|$: The key observation here is that $\nabla\hat{f}(x_0)$ can only be large along an eigendirection if the corresponding eigenvalue λ is large; however in this case the $(I-\eta\mathcal{H})$ term will also be significantly smaller than 1 in such a direction so the contribution from this direction decreases quickly. More precisely, we have

$$\begin{aligned}\eta\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\| =&\eta\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}(\mathcal{H}(x_0-\widetilde{x})+\Delta(x_0-\widetilde{x}))\|\\ \leq&\eta\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}\mathcal{H}(x_0-\widetilde{x})\|\\ &+\eta\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}\Delta(x_0-\widetilde{x})\|\\ \leq&\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}\eta\mathcal{H}\|\delta+\eta\rho\delta^2\\ \leq&\frac{1}{t}\delta+\eta\rho\delta^2,\end{aligned}$$

where the last inequality holds since $(1-\lambda)^{t-1}\lambda\leq 1/t$ for $0<\lambda\leq 1$. Assuming $\delta\leq\frac{\gamma}{\rho}$, we can further show

$$\eta\rho\delta^2\leq\eta\gamma\delta\leq\frac{\log(d)}{t}\delta.$$

Thus, we have

$$\eta\|\text{Proj}_{S_+^\perp}(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\|\leq\frac{2\log(d)}{t}\delta.$$

Next we will bound the norm of the variance term. The main observation here is that based on induction hypothesis, we can have a good upperbound on $\|\xi_\tau\|$. Now, for subspaces $S_+^\perp$ and $S_-^\perp$, we will show that the additional matrices in front of ξ_τ will not amplify its norm by too much.

Bounding $\|\mathbf{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\|$: For each $\tau\leq t-2$, we bound variance term $\|\xi_\tau\|$ as follows,

$$\begin{aligned}\|\xi_\tau\|&=\|v_\tau-\nabla\hat{f}(x_\tau)\|\\ &\leq\frac{C_1L}{\sqrt{b}}\|x_\tau-x_{s(\tau)}\|\\ &\leq\frac{C_1L}{m}\|x_\tau-x_{s(\tau)}\|\\ &\leq\frac{C_1L}{m}\sum_{\tau'=s(\tau)+1}^{\tau}\|x_{\tau'}-x_{\tau'-1}\|\\ &\leq\frac{C_1L}{m}\sum_{\tau'=s(\tau)+1}^{\tau}\frac{\mu}{\tau'}\delta,\end{aligned}$$

where the second inequality assumes $b \geq m^2$ and the last inequality is due to the induction hypothesis. If $t \leq 2m$, we bound $\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\|$ as follows.

$$\begin{aligned}
\left\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| &\leq \eta\sum_{\tau=0}^{t-2}\|\text{Proj}_{S_+^\perp}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\|\|\xi_\tau\| \\
&\leq \eta\sum_{\tau=0}^{t-2}\frac{1}{t-1-\tau}\left(\frac{C_1L}{m}\sum_{\tau'=s(\tau)+1}^{\tau}\frac{\mu}{\tau'}\delta\right) \\
&\leq \eta\sum_{\tau=0}^{t-2}\frac{1}{t-1-\tau}\left(\frac{2C_1\log(2m)L}{m}\mu\delta\right) \\
&\leq \frac{4C_1\log^2(2m)}{m}\eta L\mu\delta \\
&\leq \frac{8C_1\log^2(2m)}{t}\eta L\mu\delta,
\end{aligned}$$

where the third inequality holds since $\sum_{\tau'=s(\tau)+1}^{\tau}\frac{1}{\tau'} \leq \log(\tau)+1 \leq \log(2m)+1 \leq 2\log(2m)$.

If $t > 2m$, we bound $\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\|$ as follows.

$$\begin{aligned}
\left\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| &\leq \eta\sum_{\tau=0}^{t-2}\|\text{Proj}_{S_+^\perp}\eta\mathcal{H}(I-\eta\mathcal{H})^{t-2-\tau}\|\|\xi_\tau\| \\
&\leq \eta\left(\sum_{\tau=0}^{m-1}\frac{1}{t-1-\tau}\|\xi_\tau\| + \eta\sum_{\tau=m}^{t-2}\frac{1}{t-1-\tau}\|\xi_\tau\|\right).
\end{aligned}$$

We bound these two terms in slightly different ways. For the first term, we have,

$$\begin{aligned}
\eta\sum_{\tau=0}^{m-1}\frac{1}{t-1-\tau}\|\xi_\tau\| &\leq \eta\sum_{\tau=0}^{m-1}\frac{1}{t-1-\tau}\left(\frac{2C_1\log(m)L}{m}\mu\delta\right) \\
&\leq \eta\sum_{\tau=0}^{m-1}\frac{2}{t}\left(\frac{2C_1\log(m)L}{m}\mu\delta\right) \\
&\leq \frac{4C_1\log(m)}{t}\eta L\mu\delta,
\end{aligned}$$

where the second inequality holds since $t - m > t/2$. For the second term, we bound it as follows.

$$\begin{aligned}
\eta \sum_{\tau=m}^{t-2} \frac{1}{t-1-\tau}\|\xi_\tau\| &\le \eta \sum_{\tau=m}^{t-2} \frac{1}{t-1-\tau}\left(\frac{C_1 L}{m}\sum_{\tau'=s(\tau)+1}^{\tau}\frac{\mu}{\tau'}\delta\right) \\
&\le \eta \sum_{\tau=m}^{t-2} \frac{1}{t-1-\tau}\left(C_1 L \frac{\mu}{s(\tau)+1}\delta\right) \\
&\le C_1 \eta L \mu \delta \sum_{\tau=m}^{t-2} \frac{1}{t-1-\tau}\cdot\frac{1}{\tau-m+1} \\
&= C_1 \eta L \mu \delta \sum_{\tau=m}^{t-2} \left(\frac{1}{t-1-\tau}+\frac{1}{\tau-m+1}\right)\frac{1}{t-m} \\
&\le \frac{8C_1 \log(\frac{\log(d)}{\eta\gamma})}{t}\eta L\mu\delta
\end{aligned}$$

where the third inequality holds because $\tau - s(\tau) \le m$. Thus, if $t > 2m$, we have

$$\left\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| \le \frac{8C_1\log(m\frac{\log(d)}{\eta\gamma})}{t}\eta L\mu\delta.$$

Thus, combining two cases when $t \le 2m$ and $t > 2m$, we know

$$\left\|\text{Proj}_{S_+^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| \le \max\left(8C_1\log^2(2m), 8C_1\log(m\frac{\log(d)}{\eta\gamma})\right)\frac{1}{t}\eta L\mu\delta.$$

Bounding $\|\text{Proj}_{S_-^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\|$: Let's now consider the projection on the $S_-^\perp$ subspace.

$$\begin{aligned}
&\left\|\text{Proj}_{S_-^\perp}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| \\
\le&\eta^2\sum_{\tau=0}^{t-2}\|\text{Proj}_{S_-^\perp}\mathcal{H}\|\|\text{Proj}_{S_-^\perp}(I-\eta\mathcal{H})^{t-2-\tau}\|\|\xi_\tau\| \\
\le&\eta^2\frac{\gamma}{\log(d)}(1+\frac{\eta\gamma}{\log(d)})^{\frac{\log(d)}{\eta\gamma}}\sum_{\tau=0}^{t-2}\|\xi_\tau\| \\
\le&\eta^2\frac{e\gamma}{\log(d)}\left(\sum_{\tau=0}^{m-1}\frac{2C_1\log(m)L}{m}\mu\delta+\sum_{\tau=m}^{t-2}C_1L\frac{1}{\tau-m+1}\mu\delta\right) \\
\le&2\log(m\frac{\log(d)}{\eta\gamma})eC_1\eta L\mu\delta\frac{\eta\gamma}{\log(d)} \\
\le&2\log(m\frac{\log(d)}{\eta\gamma})eC_1\eta L\mu\delta\frac{\log(d)}{t\log(d)} \\
=&\frac{2eC_1\log(m\frac{\log(d)}{\eta\gamma})}{t}\eta L\mu\delta.
\end{aligned}$$

Next we bound the Hessian changing term. This is easy because this term is actually of order δ^2 where δ is the radius of the initial perturbation. Therefore we can bound it as long as we make δ small.

Bounding $\|\mathbf{Proj}_{S_+^\perp \cap S_-^\perp} \eta \sum_{\tau=0}^{t-2} (I - \eta\mathcal{H})^{t-2-\tau} \Delta_\tau (x_{\tau+1} - x_\tau)\|$: First, we bound $\|\Delta_\tau\|$ for each $\tau \le t-2$.

$$\begin{aligned}
\|\Delta_\tau\| \le& \rho \max(\|x_{\tau+1} - \widetilde{x}\|, \|x_\tau - \widetilde{x}\|) \\
\le& \rho(\sum_{\tau'=1}^{\tau+1} \|x_{\tau'} - x_{\tau'-1}\| + \|x_0 - \widetilde{x}\|) \\
\le& \rho(\sum_{\tau'=1}^{\tau+1} \frac{1}{\tau'}\mu\delta + \delta) \\
\le& 3\log(\frac{\log(d)}{\eta\gamma})\rho\mu\delta,
\end{aligned}$$

where the third inequality uses the induction hypothesis. Then, for the Hessian changing term, we have

$$\begin{aligned}
&\left\| \mathrm{Proj}_{S_+^\perp \cap S_-^\perp} \eta \sum_{\tau=0}^{t-2} (I - \eta\mathcal{H})^{t-2-\tau} \Delta_\tau (x_{\tau+1} - x_\tau) \right\| \\
\le& \eta \sum_{\tau=0}^{t-2} \|\mathrm{Proj}_{S_+^\perp \cap S_-^\perp} (I - \eta\mathcal{H})^{t-2-\tau}\| \|\Delta_\tau\| \|(x_{\tau+1} - x_\tau)\| \\
\le& \eta \sum_{\tau=0}^{t-2} (1 + \frac{\eta\gamma}{\log(d)})^{\frac{\log(d)}{\eta\gamma}} \cdot 3\log(\frac{\log(d)}{\eta\gamma})\rho\mu\delta \frac{1}{\tau+1}\mu\delta \\
\le& \eta \sum_{\tau=0}^{t-2} e \cdot 3\log(\frac{\log(d)}{\eta\gamma})\rho\mu\delta \frac{1}{\tau+1}\mu\delta \\
\le& 6e\log^2(\frac{\log(d)}{\eta\gamma})\eta\rho\mu^2\delta^2 \\
\le& 6e\log^2(\frac{\log(d)}{\eta\gamma})\eta\gamma\delta \\
\le& 6e\log^2(\frac{\log(d)}{\eta\gamma})\log(d)\frac{1}{t}\delta,
\end{aligned}$$

where the second last inequality holds as long as $\delta \le \frac{\gamma}{\rho\mu^2}$.

Next, we bound the norm of the error in the last gradient estimate. This follows immediately from induction hypothesis.

Bounding $\|\eta\xi_{t-1}\|$: For the last term $\eta\xi_{t-1}$. If $t \le 2m$, we have

$$\begin{aligned}
\|\eta\xi_{t-1}\| &\le \eta\frac{2C_1\log(2m)L}{m}\mu\delta \\
&\le 4C_1\log(2m)\frac{1}{t}\eta L\mu\delta.
\end{aligned}$$

If $t > 2m$, we have

$$\begin{aligned}\|\eta\xi_{t-1}\| &\leq \eta\frac{C_1 L}{s(t-1)+1}\mu\delta \\ &\leq \eta\frac{C_1 L}{t-m}\mu\delta \\ &\leq \frac{2}{t}\eta C_1 L\mu\delta.\end{aligned}$$

Overall, we have

$$\|\eta\xi_{t-1}\| \leq 4C_1\log(2m)\frac{1}{t}\eta L\mu\delta.$$

Until now, we have already bounded the projection of $x_t - x_{t-1}$ in subspace $S_+^\perp$ and $S_-^\perp$. Finally, we bound the projection of $x_t - x_{t-1}$ on the S subspace. If $t-1 \leq \frac{1}{\eta\gamma}$, we bound it using the expansion in Eqn. 8 similar as above. If $t-1 > \frac{1}{\eta\gamma}$, we use the stopping condition to bound the projection on S.

Bounding $\|\mathbf{Proj}_S(x_t - x_{t-1})\|$ If $t-1 \leq \frac{1}{\eta\gamma}$, the exponential factor $(1+\eta\gamma)^{t-1}$ is still a constant. Similar as the analysis for the projection on subspace $S_-^\perp$, we have the following bound,

$$\begin{aligned}&\left\|\text{Proj}_S\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\right\| \leq \frac{2e\log(d)}{t}\delta, \\ &\left\|\text{Proj}_S\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| \leq \frac{2eC_1\log(m\frac{\log(d)}{\eta\gamma})\log(d)}{t}\eta L\mu\delta, \\ &\left\|\text{Proj}_S\eta\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1}-x_\tau)\right\| \leq 6e\log^2(\frac{\log(d)}{\eta\gamma})\log(d)\frac{1}{t}\delta.\end{aligned}$$

If $t-1 > \frac{1}{\eta\gamma}$, according to the stopping condition of Phase 1, we know $\|\text{Proj}_S(x_{t-1}-\widetilde{x})\| \leq \frac{\delta}{10}$. In order to better exploit this property, we express $x_t - x_{t-1}$ in the following way,

$$\begin{aligned}x_t - x_{t-1} &= -\eta(\nabla\hat{f}(x_{t-1}) + \xi_{t-1}) \\ &= -\eta\mathcal{H}(x_{t-1}-\widetilde{x}) - \eta\Delta_{t-1}(x_{t-1}-\widetilde{x}) - \eta\xi_{t-1},\end{aligned}$$

where $\Delta_{t-1} = \int_0^1(\nabla^2\hat{f}(\widetilde{x}+\theta(x_{t-1}-\widetilde{x})) - \mathcal{H})d\theta$. For the first term, we have

$$\|\text{Proj}_S\eta\mathcal{H}(x_{t-1}-\widetilde{x})\| \leq \eta\gamma\|\text{Proj}_S(x_{t-1}-\widetilde{x})\| \leq \eta\gamma\frac{\delta}{10} \leq \frac{\log(d)}{10t}\delta.$$

For the hessian changing term, we have

$$\begin{aligned}\|\text{Proj}_S\eta\Delta_{t-1}(x_{t-1}-\widetilde{x})\| &\leq \|\eta\Delta_{t-1}(x_{t-1}-\widetilde{x})\| \\ &\leq \eta\rho\|x_{t-1}-\widetilde{x}\|^2 \\ &\leq \eta\rho(3\log(\frac{\log(d)}{\eta\gamma})\mu\delta)^2 \\ &\leq 9\log^2(\frac{\log(d)}{\eta\gamma})\eta\gamma\delta \\ &\leq 9\log^2(\frac{\log(d)}{\eta\gamma})\log(d)\frac{1}{t}\delta\end{aligned}$$

where the second last inequality assumes $\delta \leq \frac{\gamma}{\rho\mu^2}$.

Combining the bound for the projections onto all three subspaces, we know there exists absolute constant c, such that

$$\|x_t - x_{t-1}\| \leq \frac{c}{2}\log(d)\log^2(\frac{\log(d)}{\eta\gamma})\frac{1}{t}\delta + \frac{c}{2}C_1\log(nd)\log(n\frac{\log(d)}{\eta\gamma})\frac{1}{t}\eta L\mu\delta,$$

assuming $\delta \leq \min(\frac{\gamma}{\rho}, \frac{\gamma}{\rho\mu^2})$. Now, we know $\|x_t - x_{t-1}\| \leq \frac{1}{t}\mu\delta$, as long as

$$\eta \leq \frac{1}{cC_1\log(nd)\log(n\frac{\log(d)}{\eta\gamma})\cdot L},$$
$$\mu \geq c\log(d)\log^2(\frac{\log(d)}{\eta\gamma}),$$
$$\delta \leq \frac{\gamma}{\rho\mu^2}.$$

□

D.3. Proofs of Phase 2

We have shown that at the end of Phase 1, $x_{T_1} - \widetilde{x}$ becomes aligned with the negative directions. Based on this property, we show the projection of $x_t - \widetilde{x}$ on S subspace grows exponentially and exceeds the threshold distance within $\widetilde{O}(\frac{1}{\eta\gamma})$ steps. We use the following expansion,

$$x_t - \widetilde{x} = (I - \eta\mathcal{H})(x_{t-1} - \widetilde{x}) - \eta\Delta_{t-1}(x_{t-1} - \widetilde{x}) - \eta\xi_{t-1},$$

where $\Delta_{t-1} = \int_0^1(\nabla^2\hat{f}(\widetilde{x} + \theta(x_{t-1} - \widetilde{x})) - \mathcal{H})d\theta$. Intuitively, if we only have the first term, it's clear that $\|\text{Proj}_S(x_t - \widetilde{x})\| \geq (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$. We show that the Hessian changing term and the variance term are negligible in the sense that $\|\eta\Delta_{t-1}(x_{t-1} - \widetilde{x}) - \eta\xi_{t-1}\| \leq \frac{\eta\gamma}{2\log(d)}\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$. The Hessian changing term can be easily bounded because the threshold distance $\mathscr{L} = \widetilde{O}(\frac{\gamma}{\rho})$. We will bound the variance by showing that $\|x_t - x_{t-1}\| \leq \widetilde{O}(1/t)\|x_{t-1} - \widetilde{x}\|$. We also need $x_{t-1} - \widetilde{x}$ to be roughly aligned with the negative directions in order to bound $\|x_{t-1} - \widetilde{x}\|$ by $\widetilde{O}(1)\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$.

There are several key differences between Phase 1 and Phase 2 . First, we use Lemma 7 to bound the variance (this is effective because the point does not move far in Phase 1), but we use Lemma 4 to bound variance in Phase 2 (this is effective because in Phase 2 the projection in the most negative eigenvalue is already large). Second, in Phase 1 we need to analyze the difference between two points, and the direction e_1 is dominating. In Phase 2 we can analyze the dynamics of a single point, and focus on the entire subspace with eigenvalues less than $-\gamma/\log d$ instead of a single e_1 direction.

Lemma 21 *Let the threshold distance $\mathscr{L} := \frac{\gamma}{C_3\rho}$. Let T be the length of the super epoch, which means $T := \inf\{t|\ \|x_t - \widetilde{x}\| \geq \mathscr{L}\}$. Assume for any $0 \leq t \leq T-1$, $\|\xi_t\| \leq \frac{C_1 L}{\sqrt{b}}\|x_t - x_{s(t)}\|$, where C_1 comes from Lemma 4. Assume Phase 1 is successful in the sense that*

$$\frac{1}{\eta\gamma} \leq T_1 \leq \frac{\log(d)}{\eta\gamma}, \qquad \forall\ 1 \leq t \leq T_1,\ \|x_t - x_{t-1}\| \leq \frac{C_2}{t}\delta,$$
$$\|\textit{Proj}_S(x_{T_1} - \widetilde{x})\| \geq \frac{\delta}{10}, \qquad \forall\ 0 \leq t \leq T_1,\ \|x_t - \widetilde{x}\| \leq C\frac{\delta}{10},$$

where C_2 comes from Lemma 19 and C comes from Lemma 20. There exists large enough absolute constant c such that as long as

$$\eta \le \frac{1}{L \cdot cCC_1 \left(\log^2(n) + \log(n \frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma}) \right)},$$
$$C_3 \ge c \left(C_2 + C \log(\frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma}) \log(d) \log(\frac{\gamma}{\rho\delta}) \right),$$
$$b \ge n^{2/3} \cdot c \left(C \log(d) \log(\frac{\gamma}{\rho\delta}) \left(C_2 + C \log(\frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma}) \log(d) \log(\frac{\gamma}{\rho\delta}) \right) \right)^{2/3},$$

we have

$$T \le T_1 + \frac{4\log(d)\log(\frac{10\gamma}{\rho\delta})}{\eta\gamma} \le \frac{\log(d) + 4\log(d)\log(\frac{10\gamma}{\rho\delta})}{\eta\gamma}.$$

Proof of Lemma 21. Let $T_{\max} = T_1 + \frac{4\log(d)\log(\frac{10\gamma}{\rho\delta})}{\eta\gamma}$. If there exists $t \le T_{\max} - 1$, $\|x_t - \widetilde{x}\| \ge \mathscr{L}$, we are done. Otherwise, we show $\|x_t - \widetilde{x}\|$ increases exponentially and will become larger than $\mathscr{L}$ after $T_{\max}$ steps.

Formally, we show the following four hypotheses hold for any $T_1 \le t \le T_{\max}$ by induction,

1. $$\|\text{Proj}_S(x_t - \widetilde{x})\| \ge (1 + \frac{\eta\gamma}{2\log(d)})^{t-T_1} \|\text{Proj}_S(x_{T_1} - \widetilde{x})\|;$$

2. $$\frac{\|\text{Proj}_{S^\perp}(x_t - \widetilde{x})\|}{\|\text{Proj}_S(x_t - \widetilde{x})\|} \le C(1 + \frac{\eta\gamma}{4\log(d)\log(\frac{10\gamma}{\rho\delta})})^{t-T_1},$$

   where $S^\perp$ denotes the orthogonal subspace of S;

3. For any $0 \le \tau \le t-1$, we have

   $$\|x_t - \widetilde{x}\| \ge \|\text{Proj}_S(x_t - \widetilde{x})\| \ge \frac{1}{eC+1} \|x_\tau - \widetilde{x}\|;$$

4. For any $1 \le \tau \le t$, we have

   $$\|x_\tau - x_{\tau-1}\| \le \frac{\mu}{\tau} \max(\|x_{\tau-1} - \widetilde{x}\|, \frac{\delta}{10}),$$

   where $\mu = \widetilde{O}(1)$.

Hypothesis 1 is our goal, which is showing the distance to the initial point increases exponentially in Phase 2. We use hypothesis 4 to bound the variance term. We also need Hypothesis 2 and 3 for some technical reason, which will only be clear in the later proof. Basically, hypothesis 2 guarantees that $x_t - \widetilde{x}$ roughly aligns with the S subspace. Hypothesis 3 guarantees that the distance to the initial point cannot shrink by too much.

Let's first check the initial case first. If $t = T_1$, the first hypothesis clearly holds. For the second hypothesis, we have

$$\frac{\|\text{Proj}_{S^\perp}(x_{T_1} - \widetilde{x})\|}{\|\text{Proj}_S(x_{T_1} - \widetilde{x})\|} \leq \frac{\|x_{T_1} - \widetilde{x}\|}{\|\text{Proj}_S(x_{T_1} - \widetilde{x})\|} \leq C.$$

The third hypothesis holds because $\|x_{T_1} - \widetilde{x}\| \geq \|\text{Proj}_S(x_{T_1} - \widetilde{x})\| \geq \delta/10$ and $\|x_t - \widetilde{x}\| \leq C\delta/10$ for any $t \leq T_1$. Since $\|x_t - x_{t-1}\| \leq \frac{C_2}{t}\delta$ for any $1 \leq t \leq T_1$, the fourth hypothesis holds as long as $\mu \geq 10C_2$.

Now, fix $T_1 < t \leq T_{\max}$, assume all four hypotheses hold for every $T_1 \leq t' \leq t-1$, we prove they still hold for t.

Proving Hypothesis 4: In order to prove Hypothesis 4, we only need to show $\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\max(\|x_{t-1} - \widetilde{x}\|, \delta/10)$. Let S^+ be the subspace spanned by all the eigenvectors of $\mathcal{H}$ with positive eigenvalues. Let S^- be the subspace spanned by all the eigenvectors of $\mathcal{H}$ with non-positive eigenvalues. We project $x_t - x_{t-1}$ into these two subspaces and bound them separately.

Bounding $\|\text{Proj}_{S^-}(x_t - x_{t-1})\|$: Consider the following expansion of $x_t - x_{t-1}$:

$$\begin{aligned} x_t - x_{t-1} =& -\eta(\nabla\hat{f}(x_{t-1}) + \xi_{t-1}) \\ =& -\eta\mathcal{H}(x_{t-1} - \widetilde{x}) - \eta\Delta_{t-1}(x_{t-1} - \widetilde{x}) - \eta\xi_{t-1}, \end{aligned}$$

where $\Delta_{t-1} = \int_0^1(\nabla^2\hat{f}(\widetilde{x} + \theta(x_{t-1} - \widetilde{x})) - \mathcal{H})d\theta$. We bound $\text{Proj}_{S^-}(x_t - x_{t-1})$ by separately considering these three terms.

The first term can be bounded because within subspace S^-, the largest singular value of $\mathcal{H}$ is just γ. Precisely, we have

$$\begin{aligned} \|\text{Proj}_{S^-}\eta\mathcal{H}(x_{t-1} - \widetilde{x})\| \leq& \eta\gamma\|x_{t-1} - \widetilde{x}\| \\ \leq& \frac{\left(\log(d) + 4\log(d)\log(\frac{10\gamma}{\rho\delta})\right)}{t}\|x_{t-1} - \widetilde{x}\|, \end{aligned}$$

where the second inequality holds because $t \leq T_{\max} \leq \frac{\left(\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})\right)}{\eta\gamma}$.

Since f is Hessian lipschitz and the total distance is upper bounded by $\frac{\gamma}{C_3\rho}$, the second term can also be well bounded. We have,

$$\begin{aligned} \|\text{Proj}_{S^-}\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})\| \leq& \|\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})\| \\ \leq& \eta\rho\|x_{t-1} - \widetilde{x}\|\|x_{t-1} - \widetilde{x}\| \\ \leq& \eta\rho\mathscr{L}\|x_{t-1} - \widetilde{x}\| \\ \leq& \eta\frac{\gamma}{C_3}\|x_{t-1} - \widetilde{x}\| \\ \leq& \frac{\log(d) + 4\log(d)\log(\frac{10\gamma}{\rho\delta})}{C_3 t}\|x_{t-1} - \widetilde{x}\|, \end{aligned}$$

where the second inequality holds due to the Hessian-lipshcitzness of f.

We can bound the variance term using Hypothesis 3 and 4. Precisely, we have

$$\begin{aligned}\|\eta\xi_{t-1}\| \leq & \eta\frac{m}{\sqrt{b}}\frac{C_1 L}{m}\sum_{\tau=s(t-1)+1}^{t-1}\|x_\tau - x_{\tau-1}\| \\ \leq & \eta\frac{m}{\sqrt{b}}\frac{C_1 L}{m}\sum_{\tau=s(t-1)+1}^{t-1}\frac{\mu}{\tau}\max(\|x_{\tau-1}-\widetilde{x}\|, \frac{\delta}{10}) \\ \leq & \eta\frac{m}{\sqrt{b}}\frac{C_1 L}{m}\sum_{\tau=s(t-1)+1}^{t-1}\frac{\mu}{\tau}(eC+1)\|x_{t-1}-\widetilde{x}\|,\end{aligned}$$

where the last inequality holds requires $\|x_{t-1}-\widetilde{x}\| \geq \frac{1}{eC+1}\max(\|x_{\tau-1}-\widetilde{x}\|, \frac{\delta}{10})$ for any $\tau \leq t-1$. According to induction hypothesis 3, we have $\|x_{t-1}-\widetilde{x}\| \geq \frac{1}{eC+1}\|x_{\tau-1}-\widetilde{x}\|$ for any $\tau \leq t-1$. By induction hypothesis 1, we have $\|x_{t-1}-\widetilde{x}\| \geq \|\text{Proj}_S(x_{t-1}-\widetilde{x})\| \geq (1+\frac{\eta\gamma}{2})^{t-1-T_1}\|\text{Proj}_S(x_{T_1}-\widetilde{x})\| \geq \frac{\delta}{10}$. Using the same analysis in Lemma 19, we further have

$$\|\eta\xi_{t-1}\| \leq \frac{m}{\sqrt{b}}4(eC+1)C_1\log(2m)\frac{1}{t}\eta L\mu\|x_{t-1}-\widetilde{x}\|.$$

Bounding $\|\textbf{Proj}_{S^+}(x_t - x_{t-1})\|$: For the projection onto S^+, we use the following expansion:

$$\begin{aligned}x_t - x_{t-1} = & -\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0) + \eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau \\ & -\eta\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1}-x_\tau) - \eta\xi_{t-1},\end{aligned}$$

Similar as the analysis in Lemma 19, we can bound the first term as follows,

$$\|\text{Proj}_{S^+}\eta(I-\eta\mathcal{H})^{t-1}\nabla\hat{f}(x_0)\| \leq \frac{2\log(d)}{t}\delta \leq \frac{20\log(d)}{t}\|x_{t-1}-\widetilde{x}\|,$$

where the second inequality holds because $\|x_{t-1}-\widetilde{x}\| \geq \delta/10$.

Using a similar analysis as in Lemma 19, we have the following bound for the second term,

$$\begin{aligned}&\left\|\text{Proj}_{S^+}\eta^2\mathcal{H}\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\xi_\tau\right\| \\ \leq & \frac{m}{\sqrt{b}}(eC+1)\max(8C_1\log^2(2m), 8C_1\log(mT_{\max}))\frac{1}{t}\eta L\mu\|x_{t-1}-\widetilde{x}\|.\end{aligned}$$

For the hessian changing term, we have

$$\begin{aligned}
&\left\|\text{Proj}_{S^+}\eta\sum_{\tau=0}^{t-2}(I-\eta\mathcal{H})^{t-2-\tau}\Delta_\tau(x_{\tau+1}-x_\tau)\right\| \\
\leq&\eta\sum_{\tau=0}^{t-2}\|\Delta_\tau\|\|x_{\tau+1}-x_\tau\| \\
\leq&\eta\sum_{\tau=0}^{t-2}\frac{\gamma}{C_3}(eC+1)\frac{\mu}{\tau+1}\|x_{t-1}-\widetilde{x}\| \\
\leq&2\log(T_{\max})\eta\gamma(eC+1)\frac{\mu}{C_3}\|x_{t-1}-\widetilde{x}\| \\
\leq&2(eC+1)\log(T_{\max})\frac{\mu}{C_3}\frac{\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})}{t}\|x_{t-1}-\widetilde{x}\| \\
\leq&2(eC+1)\log(T_{\max})\frac{\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})}{t}\|x_{t-1}-\widetilde{x}\|,
\end{aligned}$$

where the last inequality holds as long as $C_3 \geq \mu$.

Overall, we can upper bound $\|x_t - x_{t-1}\|$ as follows,

$$\begin{aligned}
&\|x_t-x_{t-1}\| \\
\leq&\Big(20\log(d)+(\frac{1}{C_3}+1+2(eC+1)\log(T_{\max}))\big(\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})\big)\Big)\frac{1}{t}\|x_{t-1}-\widetilde{x}\| \\
&+\big(4(eC+1)C_1\log(2m)+(eC+1)\max(8C_1\log^2(2m),8C_1\log(mT_{\max}))\big)\frac{1}{t}\eta L\mu\|x_{t-1}-\widetilde{x}\| \\
\leq&\Big(20\log(d)+(2+2(eC+1)\log(T_{\max}))\big(\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})\big)\Big)\frac{1}{t}\|x_{t-1}-\widetilde{x}\| \\
&+\big(4(eC+1)C_1\log(2n)+(eC+1)\max(8C_1\log^2(2n),8C_1\log(nT_{\max}))\big)\frac{1}{t}\eta L\mu\|x_{t-1}-\widetilde{x}\|,
\end{aligned}$$

assuming $C_3 \geq 1$. As long as

$$\eta \leq \frac{1}{2L\cdot\big(4(eC+1)C_1\log(2n)+(eC+1)\max(8C_1\log^2(2n),8C_1\log(nT_{\max}))\big)}$$

and

$$\mu \geq 2\left(20\log(d)+(2+2(eC+1)\log(T_{\max}))\left(\log(d)+4\log(d)\log(\frac{10\gamma}{\rho\delta})\right)\right),$$

we have $\|x_t - x_{t-1}\| \leq \frac{\mu}{t}\|x_{t-1}-\widetilde{x}\|$.

Proving Hypothesis 2: In order to prove condition 2 holds for time t, we only need to show

$$\frac{\|\text{Proj}_{S^\perp}(x_t-\widetilde{x})\|}{\|\text{Proj}_S(x_t-\widetilde{x})\|} \leq (1+\frac{\eta\gamma}{4\log(d)\log(\frac{10\gamma}{\rho\delta})})P_{t-1},$$

where $P_{t-1} := C(1+\frac{\eta\gamma}{4\log(d)\log(\frac{10\gamma}{\rho\delta})})^{t-1-T_1}$.

We can express $x_t - \widetilde{x}$ as follows,

$$x_t - \widetilde{x} = (I - \eta\mathcal{H})(x_{t-1} - \widetilde{x}) - \eta\Delta_{t-1}(x_{t-1} - \widetilde{x}) - \eta\xi_{t-1}.$$

Assuming $\|\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})\| + \|\eta\xi_{t-1}\| \leq \widetilde{C}\eta\gamma\|x_{t-1} - \widetilde{x}\|, \widetilde{C} = \widetilde{O}(1)$, we have

$$\|\text{Proj}_{S^\perp}(x_t - \widetilde{x})\| \leq (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_{S^\perp}(x_{t-1} - \widetilde{x})\| + \widetilde{C}\eta\gamma\|x_{t-1} - \widetilde{x}\|$$

and

$$\|\text{Proj}_{S}(x_t - \widetilde{x})\| \geq (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_{S}(x_{t-1} - \widetilde{x})\| - \widetilde{C}\eta\gamma\|x_{t-1} - \widetilde{x}\|.$$

Then, we have

$$\begin{aligned}
\frac{\|\text{Proj}_{S^\perp}(x_t - \widetilde{x})\|}{\|\text{Proj}_{S}(x_t - \widetilde{x})\|} \leq & \frac{P_{t-1}(1 + \frac{\eta\gamma}{\log(d)}) + (P_{t-1} + 1)\widetilde{C}\eta\gamma}{1 + \frac{\eta\gamma}{\log(d)} - (P_{t-1} + 1)\widetilde{C}\eta\gamma} \\
= & P_{t-1}\left(\frac{1 + \frac{\eta\gamma}{\log(d)} + (1 + \frac{1}{P_{t-1}})\widetilde{C}\eta\gamma}{1 + \frac{\eta\gamma}{\log(d)} - (P_{t-1} + 1)\widetilde{C}\eta\gamma}\right) \\
= & P_{t-1}\left(1 + \frac{(1 + \frac{1}{P_{t-1}})\widetilde{C}\eta\gamma + (P_{t-1} + 1)\widetilde{C}\eta\gamma}{1 + \frac{\eta\gamma}{\log(d)} - (P_{t-1} + 1)\widetilde{C}\eta\gamma}\right) \\
\leq & P_{t-1}\left(1 + (1 + \frac{1}{P_{t-1}} + P_{t-1} + 1)\widetilde{C}\eta\gamma\right) \\
\leq & P_{t-1}\left(1 + (3 + eC)\widetilde{C}\eta\gamma\right),
\end{aligned}$$

where the second last inequality holds as long as $(P_{t-1} + 1)\widetilde{C} \leq (eC + 1)\widetilde{C} \leq 1/\log(d)$ and the last inequality holds because $1 \leq P_{t-1} \leq eC$. Now, as long as $\widetilde{C} \leq \frac{1}{(3+eC)4\log(d)\log(\frac{10\gamma}{\rho\delta})}$, we have

$$\begin{aligned}
\frac{\|\text{Proj}_{S^\perp}(x_t - \widetilde{x})\|}{\|\text{Proj}_{S}(x_t - \widetilde{x})\|} \leq & (1 + \frac{\eta\gamma}{4\log(d)\log(\frac{10\gamma}{\rho\delta})})P_{t-1} \\
\leq & C(1 + \frac{\eta\gamma}{4\log(d)\log(\frac{10\gamma}{\rho\delta})})^{t-T_1}.
\end{aligned}$$

For the hessian changing term, we have $\|\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})\| \leq \frac{1}{C_3}\eta\gamma\|x_{t-1} - \widetilde{x}\|$. For the variance term, according to the previous analysis and the choosing of η, we have

$$\|\eta\xi_{t-1}\| \leq \frac{m}{2\sqrt{b}}\mu\frac{1}{t}\|x_{t-1} - \widetilde{x}\| \leq \frac{m}{2\sqrt{b}}\mu\eta\gamma\|x_{t-1} - \widetilde{x}\|$$

where the second inequality holds because $t \geq T_1 \geq \frac{1}{\eta\gamma}$. As long as $C_3 \geq \frac{2}{\widetilde{C}}$ and $b \geq (\frac{\mu}{\widetilde{C}})^{2/3}n^{2/3}$, we have $\|\eta\Delta_{t-1}(x_{t-1} - \widetilde{x})\| + \|\eta\xi_{t-1}\| \leq \widetilde{C}\eta\gamma\|x_{t-1} - \widetilde{x}\|$.

Proving Hypothesis 1. In order to prove hypothesis 1, we only need to show $\|\text{Proj}_S(x_t - \widetilde{x})\| \geq (1 + \frac{\eta\gamma}{2\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$. We know,

$$\begin{aligned}\|\text{Proj}_S(x_t - \widetilde{x})\| \geq & (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\| - \widetilde{C}\eta\gamma\|x_{t-1} - \widetilde{x}\| \\ \geq & (1 + \frac{\eta\gamma}{\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\| - (eC+1)\widetilde{C}\eta\gamma\|\text{Proj}_S(x_{t-1} - \widetilde{x})\| \\ \geq & (1 + \frac{\eta\gamma}{2\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|,\end{aligned}$$

where the last inequality holds as long as $\widetilde{C} \leq \frac{1}{2(eC+1)\log(d)}$.

Proving Hypothesis 3. For $\tau \leq t-2$, we have

$$\begin{aligned}\|x_t - \widetilde{x}\| \geq & \|\text{Proj}_S(x_t - \widetilde{x})\| \\ \geq & \|\text{Proj}_S(x_{t-1} - \widetilde{x})\| \\ \geq & \frac{1}{eC+1}\|x_\tau - \widetilde{x}\|,\end{aligned}$$

where the second inequality holds because $\|\text{Proj}_S(x_t - \widetilde{x})\| \geq (1 + \frac{\eta\gamma}{2\log(d)})\|\text{Proj}_S(x_{t-1} - \widetilde{x})\|$ and the last inequality holds due to the induction hypothesis 3.

Since $\|\text{Proj}_S(x_{t-1} - \widetilde{x})\| \geq \frac{1}{eC+1}\|x_{t-1} - \widetilde{x}\|$, we also have

$$\begin{aligned}\|x_t - \widetilde{x}\| \geq & \|\text{Proj}_S(x_t - \widetilde{x})\| \\ \geq & \|\text{Proj}_S(x_{t-1} - \widetilde{x})\| \\ \geq & \frac{1}{eC+1}\|x_{t-1} - \widetilde{x}\|.\end{aligned}$$

Thus, there exists large enough absolute constant c such that the induction holds as long as

$$\begin{aligned}&\eta \leq \frac{1}{L \cdot cCC_1\left(\log^2(n) + \log(n\frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma})\right)}, \\ &C_3 \geq c\left(C_2 + C\log(\frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma})\log(d)\log(\frac{\gamma}{\rho\delta})\right) \\ &b \geq cn^{2/3}\left(C\log(d)\log(\frac{\gamma}{\rho\delta})\left(C_2 + C\log(\frac{\log(d)\log(\frac{\gamma}{\rho\delta})}{\eta\gamma})\log(d)\log(\frac{\gamma}{\rho\delta})\right)\right)^{2/3}.\end{aligned}$$

Finally, we have

$$\begin{aligned}\|x_{T_{\max}} - \widetilde{x}\| \geq & \|\text{Proj}_S(x_{T_{\max}} - \widetilde{x})\| \\ \geq & (1 + \frac{\eta\gamma}{2\log(d)})^{T_{\max}-T_1}\|\text{Proj}_S(x_{T_1} - \widetilde{x})\| \\ \geq & (1 + \frac{\eta\gamma}{2\log(d)})^{\frac{4\log(d)\log(\frac{10\gamma}{\rho\delta})}{\eta\gamma}}\frac{\delta}{10} \\ \geq & \frac{\gamma}{\rho} \geq \frac{\gamma}{C_3\rho} := \mathscr{L}.\end{aligned}$$

□

D.4. Proof of Lemma 22

Finally, we combine the analysis for Phase 1 and Phase 2 to show that starting from a randomly perturbed point, with at least constant probability the function value decreases significantly after a super epoch.

Lemma 22 *Let $\widetilde{x}$ be the initial point with gradient $\|\nabla f(\widetilde{x})\| \le \mathscr{G}$ and $\lambda_{\min}(\mathcal{H}) = -\gamma < 0$. Define stabilized function $\hat{f}$ such that $\hat{f}(x) := f(x) - \langle \nabla f(\widetilde{x}), x - \widetilde{x}\rangle$. Let $\{x_t\}$ be the iterates of SVRG running on $\hat{f}$ starting from x_0, which is the perturbed point of $\widetilde{x}$. Let T be the length of the current super epoch. There exists $\eta = \widetilde{O}(1/L), b = \widetilde{O}(n^{2/3}), m = n/b, \delta = \widetilde{O}(\min(\frac{\gamma}{\rho}, \frac{m\gamma}{\rho'})), \mathscr{G} = \widetilde{O}(\frac{\gamma^2}{\rho}), \mathscr{L} = \widetilde{O}(\frac{\gamma}{\rho}), T_{\max} = \widetilde{O}(\frac{1}{\eta\gamma})$ such that with probability at least $1/8$,*

$$f(x_T) - f(\widetilde{x}) \le -C_5 \cdot \frac{\gamma^3}{\rho^2};$$

and with high probability,

$$f(x_T) - f(\widetilde{x}) \le \frac{C_5}{20} \cdot \frac{\gamma^3}{\rho^2};$$

where $C_5 = \widetilde{\Theta}(1)$ and $T \le T_{\max}$.

Proof of Lemma 22. Combining Lemma 20 and the coupling probabilistic argument in Lemma 17, we know from a randomly perturbed point x_0, sequence $\{x_t\}$ succeeds in Phase 1 with probability at least $1/6$. By Lemma 4, we know with high probability, there exists $C_1 = \widetilde{O}(1)$, such that $\|\xi_t\| \le \frac{C_1 L}{\sqrt{b}}\|x_t - x_{s(t)}\|$ for any $0 \le t \le T-1$, where T is the super epoch length. Then, combing with Lemma 21 and Lemma 9, with probability at least $1/8$ we know there exists $\eta = \frac{1}{C_6 L}, b = \widetilde{O}(n^{2/3}), \delta = \widetilde{O}(\min(\frac{\gamma}{\rho}, \frac{m\gamma}{\rho'})), T \le T_{\max} := \frac{C_7}{\eta\gamma}$ such that,

$$\|x_T - \widetilde{x}\| \ge \mathscr{L} := \frac{\gamma}{C_3\rho}, \quad \|x_T - x_0\|^2 \le \frac{T}{C_4 L}(\hat{f}(x_0) - \hat{f}(x_T))$$

where $C_3, C_4, C_6, C_7 = \widetilde{O}(1)$.

Since $\|x_T - x_0\|^2 \le \frac{T}{C_4 L}(\hat{f}(x_0) - \hat{f}(x_T))$, we have

$$\begin{aligned}
\hat{f}(x_0) - \hat{f}(x_T) \ge& \frac{C_4 L}{T}\|x_T - x_0\|^2 \\
\ge& \frac{C_4 L}{T}\left(\|x_T - \widetilde{x}\| - \|x_0 - \widetilde{x}\|\right)^2 \\
\ge& \frac{C_4 L}{T}\left(\frac{\gamma}{C_3\rho} - \delta\right)^2 \\
\ge& \frac{C_4 L}{T}\frac{\gamma^2}{4C_3^2\rho^2} \\
=& \frac{C_4}{4C_7 C_3^2 C_6}\frac{\gamma^3}{\rho^2},
\end{aligned}$$

where the last inequality holds as long as $\delta \le \frac{\gamma}{2C_3\rho}$.

Since $\hat{f}$ is L-smooth and $\nabla \hat{f}(\widetilde{x}) = 0$, we have

$$\hat{f}(x_0) - \hat{f}(\widetilde{x}) \leq \frac{L}{2}\|\widetilde{x} - x_0\|^2 \leq \frac{L}{2}\delta^2.$$

Let the threshold gradient $\mathscr{G} := \frac{\gamma^2}{C_8\rho}$. For the function value difference between two sequence, we have

$$f(x_T) - \hat{f}(x_T) \leq \|\nabla f(\widetilde{x})\| \cdot \|x_T - \widetilde{x}\|$$

Since T is the length of the current super epoch, we know $\|x_{T-1} - \widetilde{x}\| < \mathscr{L}$. According to the analysis in Lemma 21, we also know $\|x_T - \widetilde{x}\| \leq 3\|x_{T-1} - \widetilde{x}\| \leq 3\mathscr{L}$. Thus, we have

$$\begin{aligned} f(x_T) - \hat{f}(x_T) \leq & \mathscr{G} \cdot 3\mathscr{L} \\ \leq & \frac{\gamma^2}{C_8\rho}\frac{3\gamma}{C_3\rho} \\ = & \frac{3}{C_8C_3}\frac{\gamma^3}{\rho^2}. \end{aligned}$$

Thus, with probability at least $1/8$, we know

$$\begin{aligned} f(x_T) - f(\widetilde{x}) = & f(x_T) - \hat{f}(\widetilde{x}) \\ = & \hat{f}(x_T) - \hat{f}(x_0) + \hat{f}(x_0) - \hat{f}(\widetilde{x}) + f(x_T) - \hat{f}(x_T) \\ \leq & -\frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\rho^2} + \frac{L}{2}\delta^2 + \frac{3}{C_8C_3}\frac{\gamma^3}{\rho^2}. \end{aligned}$$

If Phase 1 is not successful, the function value may not decrease. On the other hand, we know $\hat{f}(x_T) - \hat{f}(x_0) \leq 0$ with high probability. Thus, with high probability, we know

$$f(x_T) - f(\widetilde{x}) \leq \frac{L}{2}\delta^2 + \frac{3}{C_8C_3}\frac{\gamma^3}{\rho^2}.$$

Assuming $\delta \leq \sqrt{\frac{C_4}{84C_7C_3^2C_6}\frac{\gamma^3}{\rho^2}}$ and $C_8 \geq \frac{504C_7C_3C_6}{C_4}$, we know with probability at least $1/8$,

$$f(x_T) - f(\widetilde{x}) \leq -\frac{20}{21} \cdot \frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\rho^2};$$

and with high probability,

$$f(x_T) - f(\widetilde{x}) \leq \frac{1}{21} \cdot \frac{C_4}{4C_7C_3^2C_6}\frac{\gamma^3}{\rho^2}.$$

We finish the proof by choosing $C_5 := \frac{20}{21}\frac{C_4}{4C_7C_3^2C_6}$. □

Appendix E. Proof of Theorem 3

In the previous analysis, we already showed that Algorithm 5 can decrease the function value either when the current point has a large gradient or has a large negative curvature. In this section, we combine these two cases to show Stabilized SVRG will at least once get to an ϵ-second-order stationary point within $\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2}+\frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$ time. We omit the proof for Theorem 2 since it's almost the same as the proof for Theorem 3 except for using different guarantees for negative curvature exploitation super-epoch.

Recall Theorem 3 as follows.

Theorem 3 *Assume the function $f(x)$ is ρ-Hessian Lipschitz, and each individual function $f_i(x)$ is L-smooth and ρ'-Hessian Lipschitz. Let $\Delta f := f(x_0) - f^*$, where x_0 is the initial point and f^* is the optimal value of f. There exists mini-batch size $b = \widetilde{O}(n^{2/3})$, epoch length $m = n/b$, step size $\eta = \widetilde{O}(1/L)$, perturbation radius $\delta = \widetilde{O}(\min(\frac{\sqrt{\epsilon}}{\sqrt{\rho}}, \frac{m\sqrt{\rho\epsilon}}{\rho'}))$, super epoch length $T_{\max} = \widetilde{O}(\frac{L}{\sqrt{\rho\epsilon}})$, threshold gradient $\mathscr{G} = \widetilde{O}(\epsilon)$, threshold distance $\mathscr{L} = \widetilde{O}(\frac{\sqrt{\epsilon}}{\sqrt{\rho}})$, such that Stabilized SVRG (Algorithm 5) will at least once get to an ϵ-second-order stationary point with high probability using*

$$\widetilde{O}(\frac{n^{2/3}L\Delta f}{\epsilon^2}+\frac{n\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$$

stochastic gradients.

Proof of Theorem 3. Recall that we call the steps between the beginning of perturbation and the end of perturbation a super epoch. Outside of the super epoch, we use random stopping, which is equivalent to finish the epoch first and then uniformly sample a point from this epoch. In light of Lemma 6, we divide epochs [1] into two types: if at least half of points from $\{x_\tau\}_{\tau=t+1}^{t+m}$ have gradient norm at least $\mathscr{G}$, we call it a *useful epoch*; otherwise, we call it a *wasted epoch*. For simplicity of analysis, we further define *extended epoch*, which constitutes of a useful epoch or a super epoch and all its preceding wasted epochs. With this definition, we can view the iterates of Algorithm 5 as a concatenation of extended epochs.

First, we show that within each extended epoch, the number of wasted epochs before a useful epoch or a super epoch is well bounded with high probability. Suppose $\{x_\tau\}_{\tau=t+1}^{t+m}$ is a wasted epoch, we know at least half of points from $\{x_\tau\}_{\tau=t+1}^{t+m}$ have gradient norm at most $\mathscr{G}$. Thus, uniformly sampled from $\{x_\tau\}_{\tau=t+1}^{t+m}$, point $x_{t'}$ has gradient norm $\|\nabla f(x_{t'})\| \leq \mathscr{G}$ with probability at least half. Note for different wasted epochs, returned points are independently sampled. Thus, with high probability, the number of wasted epochs in an extended epoch is $\widetilde{O}(1)$. As long as the number of "extended" epochs is polynomially many through the algorithm, by union bound the number of "wasted" epochs for every "extended" epoch is $\widetilde{O}(1)$ with high probability.

We divide the extended epochs into the following three types.

- Type-1: the extended epoch ends with a useful epoch.
- Type-2: the extended epoch ends with a super epoch whose starting point has Hessian with minimum eigenvalue less that $-\sqrt{\rho\epsilon}$.
- Type-3: the extended epoch ends with a super epoch whose starting point is an ϵ-second-order stationary point.

1. Here, we only mean the epochs outside of super epochs.

For the type-1 extended epoch, according to Lemma 6, we know with probability at least $1/5$, the function value decrease by at least $\widetilde{\Omega}(n^{1/3}\epsilon^2/L)$; and with high probability, the function value does not increase. By standard concentration bound, we know after logarithmic number of type-1 extended epochs, with high probability, at least $1/6$ fraction of them decrease the function value by $\widetilde{O}(n^{1/3}\epsilon^2/L)$.

For the type-2 extended epoch, according to Lemma 22, we know with probability at least $1/8$, the function value decreases by at least $C_5\epsilon^{1.5}/\sqrt{\rho}$; and with high probability, the function value cannot increase by more than $\frac{C_5}{20}\epsilon^{1.5}/\sqrt{\rho}$, where $C_5 = \widetilde{\Theta}(1)$. Again, by standard concentration bound, we know after logarithmic number of type-2 extended epochs, with high probability, at least $1/10$ fraction of them decreases the function value by at least $C_5\epsilon^{1.5}/\sqrt{\rho}$. Let the total number of type-2 extended epochs be N_2, we know with high probability the overall function value decrease within these type-2 extended epochs is at least $\frac{N_2C_5}{20}\epsilon^{1.5}/\sqrt{\rho}$.

Thus, after $\widetilde{O}(\frac{L\Delta f}{n^{1/3}\epsilon^2})$ number of type-1 extended epochs or $\widetilde{O}(\frac{\sqrt{\rho}\Delta f}{\epsilon^{1.5}})$ number of type-2 extended epochs, with high probability the function value decrease will be more than Δf. We also know that the time consumed within a type-1 extended epoch is $\widetilde{O}(n)$ with high probability; and that for a type-2 extended epoch is $\widetilde{O}(n + n^{2/3}L/\sqrt{\rho\epsilon})$. Therefore, after

$$\widetilde{O}\left(\frac{L\Delta f}{n^{1/3}\epsilon^2}\cdot n + \frac{\sqrt{\rho}\Delta f}{\epsilon^{1.5}}(n + \frac{n^{2/3}L}{\sqrt{\rho\epsilon}})\right)$$

stochastic gradients, we will at least once get to an ϵ-second-order stationary point with high probability. □

Appendix F. Hessian Lipschitz Parameters for Matrix Sensing

In this section we consider a simple example for non-convex optimization and show that in natural conditions the Hessian Lipschitz parameter for the average function f can be much smaller than the Hessian Lipschitz parameter for the individual functions.

The problem we consider is the symmetric matrix sensing problem. In this problem, there is an unknown low rank matrix $M^* \in \mathbb{R}^{d\times d} = U^*(U^*)^\top$ where $U^* \in \mathbb{R}^{d\times r}$. In order to find M^*, one can make observations $b_i = \langle A_i, M^*\rangle$, where A_i's are random matrices with i.i.d. standard Gaussian entries. A typical non-convex formulation of this problem is as follows:

$$\min_{U\in\mathbb{R}^{d\times r}} f(U) = \frac{1}{2n}\sum_{i=1}^{n}(\langle A_i, M\rangle - b_i)^2, \tag{9}$$

where $M := UU^\top$, $U \in \mathbb{R}^{d\times r}$. It was shown in (Bhojanapalli et al., 2016; Ge et al., 2017a) that all local minima of this objective satisfies $UU^\top = M^*$ when $n = Cd$ for a large enough constant C. We can easily view this objective as a finite sum objective by defining $f_i(U) = \frac{1}{2}(\langle A_i, M\rangle - b_i)^2$.

Without loss of generality, we will assume $\|U^*\| = 1$ (otherwise everything just scales with $\|U^*\|$). A slight complication for the objective (9) is that the function is not Hessian Lipschitz in the entire $\mathbb{R}^{d\times r}$. However, it is easy to check that if the initial U_0 satisfies $\|U_0\| \le 4$ then all the iterates U_t for gradient descent (and SVRG) will satisfy $\|U_t\| \le 4$ (with high probability for SVRG). So we will constrain our interest in the set of matrices $\mathcal{B} = \{U \in \mathbb{R}^{d\times r} : \|U\| \le 4\}$.

Theorem 23 *Assume sensing matrices A_i's are random matrices with i.i.d. standard Gaussian entries.. When $n \geq Cdr$ for some large enough universal constant C, for any U, V in $\mathcal{B} = \{U \in \mathbb{R}^{d\times r} : \|U\| \leq 4\}$, for objective f in Equation* (9)*, with high probability*

$$\|\nabla^2 f(U) - \nabla^2 f(V)\| \leq O(1)\|U - V\|_F.$$

On the other hand, for the individual function $f_i(U) = \frac{1}{2}(\langle A_i, M\rangle - b_i)^2$ with high probability, there exists U, V in $\mathcal{B}$ such that

$$\|\nabla^2 f_i(U) - \nabla^2 f_i(V)\| = \Omega(d)\|U - V\|_F.$$

Before we prove the theorem, let us first see what this implies. In a natural case when r is a constant, $n = Crd$ for large enough C, for the matrix sensing we have $\rho = O(1)$, but $\rho' = \Omega(d) = \Omega(n)$. Therefore, the guarantee for Perturbed SVRG (Theorem 2) is going to be much worse compared to the guarantee of Stabilized SVRG (Theorem 3).

Let us first adapt the notation from Ge et al. (2017a) and write out the Hessian of the objective.

Definition 24 *For matrices B, B', let $B : \mathcal{H} : B' \triangleq \frac{1}{n}\sum_{i=1}^n \langle A_i, B\rangle\langle A_i, B'\rangle$.*

Lemma 25 (Ge et al. (2017a)) *The Hessian of the objective $f(U)$ in direction $Z \in \mathbb{R}^{d\times r}$ can be computed as*

$$\nabla^2 f(U)(Z, Z) = (UZ^\top + ZU^\top) : \mathcal{H} : (UZ^\top + ZU^\top) + 2(UU^\top - M^*) : \mathcal{H} : ZZ^\top.$$

Similarly, the Hessian of an individual function $f_i(U)$ satisfies

$$\nabla^2 f_i(U)(Z, Z) = \langle UZ^\top, A_i + A_i^\top\rangle^2 + 1/2\langle UU^\top - M^*, A_i + A_i^\top\rangle\langle ZZ^\top, A_i + A_i^\top\rangle.$$

Another key property we will need is the Restrict Isometry Property (RIP) (Recht et al., 2010).

Definition 26 (Matrix RIP) *The set of sensing matrix is (r, δ)-RIP if for any matrix B of rank at most r we always have*

$$(1 - \delta)\|B\|_F^2 \leq B : \mathcal{H} : B \leq (1 + \delta)\|B\|_F^2.$$

Candes and Plan (2011) showed that random Gaussian sensing matrices satisfy RIP with high probability as long as n is sufficiently large

Theorem 27 (Candes and Plan (2011)) *Suppose $n \geq Cdr/\delta^2$, then random Gaussian sensing matrices satisfy the (r, δ)-RIP with high probability.*

Now we are ready to prove Theorem 23.

Proof of Theorem 23. We will first prove the upperbound for the average function.

For the upperbound, assume that the sensing matrices are $(2r, \delta)$-RIP for $\delta = 1/10$. By Theorem 27 we know this happens with high probability when $n \geq 200Crd$ where C was the constant in Theorem 27.

For any $\|U\|, \|V\| \leq 4$ and $Z \in \mathbb{R}^{d\times r}$, we use Lemma 25 to compute the Hessian and take the difference in the direction of Z

$$\begin{aligned}
&|\nabla^2 f(U)(Z,Z) - \nabla^2 f(V)(Z,Z)| \\
&= (UZ^\top + ZU^\top) : \mathcal{H} : (UZ^\top + ZU^\top) - (VZ^\top + ZV^\top) : \mathcal{H} : (VZ^\top + ZV^\top) \\
&\quad + 2(UU^\top - M^*) : \mathcal{H} : ZZ^\top - 2(VV^\top - M^*) : \mathcal{H} : VV^\top \\
&= (UZ^\top + ZU^\top) : \mathcal{H} : \big((U-V)Z^\top + Z(U-V)^\top\big) \\
&\quad + \big((U-V)Z^\top + Z(U-V)^\top\big) : \mathcal{H} : (VZ^\top + ZV^\top) \\
&\quad + 2(UU^\top - VV^\top) : \mathcal{H} : ZZ^\top \\
&\leq (1+\delta)\|UZ^\top + ZU^\top\|_F \|(U-V)Z^\top + Z(U-V)^\top\|_F \\
&\quad + (1+\delta)\|(U-V)Z^\top + Z(U-V)^\top\|_F \|VZ^\top + ZV^\top\|_F \\
&\quad + 2(1+\delta)\|UU^\top - VV^\top\|_F \|ZZ^\top\|_F \\
&\leq 32(1+\delta)\|Z\|_F^2\|U-V\|_F + 16(1+\delta)\|U-V\|_F\|Z\|_F^2 \\
&= 48(1+\delta)\|U-V\|_F\|Z\|_F^2,
\end{aligned}$$

where the first inequality uses the definition of RIP and Cauchy-Schwartz inequality, and the second inequality uses $\|U\|, \|V\| \leq 4$ and the fact that $\|AB\|_F \leq \|A\|\|B\|_F$. Thus, for any $U, V \in \mathcal{B}$, and any direction Z, we have

$$\frac{|\nabla^2 f(U)(Z,Z) - \nabla^2 f(V)(Z,Z)|}{\|Z\|_F^2} \leq 48(1+\delta)\|U-V\|_F.$$

This implies that $\rho \leq 48(1+\delta) = \frac{264}{5}$.

Next we prove the lowerbound for individual functions. We will consider $V = U + \epsilon\Delta$ and let ϵ go to 0. This allows us to ignore some higher order terms in ϵ. Following Lemma 25, let $A = A_i + A_i^\top$, we have

$$\nabla^2 f_i(V)(Z,Z) - \nabla^2 f_i(U)(Z,Z) = 2\epsilon\langle \Delta Z^\top, A\rangle\langle UZ^\top, A\rangle + \epsilon\langle \Delta U^\top, A\rangle\langle ZZ^\top, A\rangle + O(\epsilon^2).$$

It is easy to check that the matrix $A/\sqrt{2}$ has the same distribution as the Gaussian Orthogonal Ensemble. By standard results in random matrix theory (Bai and Yin, 1988; Tao, 2012) we know with high probability $\lambda_{max}(A) \geq \sqrt{d}$. Let $\lambda = \lambda_{max}(A)$ and v be a corresponding eigenvector. We will take $U = \Delta = Z = ve_1^\top$ where e_1 is the first basis vector. In this case, we have

$$\begin{aligned}
\nabla^2 f_i(V)(Z,Z) - \nabla^2 f_i(U)(Z,Z) &= 2\epsilon\langle \Delta Z^\top, A\rangle\langle UZ^\top, A\rangle + \epsilon\langle \Delta U^\top, A\rangle\langle ZZ^\top, A\rangle + O(\epsilon^2) \\
&= 2\epsilon\langle vv^\top, A\rangle^2 + \epsilon\langle vv^\top, A\rangle^2 + O(\epsilon^2) \\
&= 3\epsilon\lambda^2 + O(\epsilon^2) \\
&= 3\lambda^2\|U-V\|_F + o(\|U-V\|_F).
\end{aligned}$$

Note that Z satisfies $\|Z\|_F = 1$, so the calculation above implies $\rho' \geq 3\lambda^2 \geq 3d$.

□

Appendix G. Tools

Matrix concentration bounds tell us that with enough number of independent samples, the empirical mean of a random matrix can converge to the mean of this matrix.

Lemma 28 (Matrix Bernstein; Theorem 1.6 in 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*

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

Define

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

Then, for all $t \geq 0$,

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

As a corollary, we have:

Lemma 29 (Bernstein Inequality: Vector Case) *Consider a finite sequence $\{v_k\}$ of independent, random vectors with dimension d. Assume that each random vector satisfies*

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

Define

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

Then, for all $t \geq 0$,

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