Sharp Analysis for Nonconvex SGD Escaping from Saddle Points

Cong Fang FANGCONG@PKU.EDU.CN
Key Lab. of Machine Perception, School of EECS, Peking University

Zhouchen Lin[*] ZLIN@PKU.EDU.CN
Key Lab. of Machine Perception, School of EECS, Peking University

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Hong Kong University of Science and Technology

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

In this paper, we give a sharp analysis[1] for Stochastic Gradient Descent (SGD) and prove that SGD is able to efficiently escape from saddle points and find an $(\epsilon, \mathcal{O}(\epsilon^{0.5}))$-approximate second-order stationary point in $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ stochastic gradient computations for generic nonconvex optimization problems, when the objective function satisfies gradient-Lipschitz, Hessian-Lipschitz, and dispersive noise assumptions. This result subverts the classical belief that SGD requires at least $\mathcal{O}(\epsilon^{-4})$ stochastic gradient computations for obtaining an $(\epsilon, \mathcal{O}(\epsilon^{0.5}))$-approximate second-order stationary point. Such SGD rate matches, up to a polylogarithmic factor of problem-dependent parameters, the rate of most accelerated nonconvex stochastic optimization algorithms that adopt additional techniques, such as Nesterov's momentum acceleration, negative curvature search, as well as quadratic and cubic regularization tricks. Our novel analysis gives new insights into nonconvex SGD and can be potentially generalized to a broad class of stochastic optimization algorithms.

Keywords: Stochastic Gradient Descent, Non-convex Optimization, Convergence Rate, Saddle Escaping

1. Introduction

Nonconvex stochastic optimization is crucial in machine learning and have attracted tremendous attentions and unprecedented popularity. Lots of modern tasks that include low-rank matrix factorization/completion and principal component analysis (Candès and Recht, 2009; Jolliffe, 2011), dictionary learning (Sun et al., 2017), Gaussian mixture models (Reynolds et al., 2000), as well as notably deep neural networks (Hinton and Salakhutdinov, 2006) are formulated as nonconvex stochastic optimization problems. In this paper, we concentrate on finding an approximate solution to the following minimization problem:

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}} \quad f(\mathbf{x}) \equiv \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{D}}\left[F(\mathbf{x};\boldsymbol{\zeta})\right]. \tag{1}$$

Here, $F(\mathbf{x};\boldsymbol{\zeta})$ denotes a family of stochastic functions indexed by some random variable $\boldsymbol{\zeta}$ that obeys some prescribed distribution $\mathcal{D}$, and we consider the general case where $f(\mathbf{x})$ and $F(\mathbf{x};\boldsymbol{\zeta})$

[*] Corresponding author.

[1] "Sharp analysis" does not mean that our result is the tightest. It means an improved analysis.

© 2019 C. Fang, Z. Lin & T. Zhang.

Algorithm 1 SGD (Meta version)

```
for $t = 1, 2, \ldots$ do
    Draw an independent $\boldsymbol{\zeta}^t \sim \mathcal{D}$ and set $\mathbf{x}^t \leftarrow \mathbf{x}^{t-1} - \eta \nabla F(\mathbf{x}^{t-1}; \boldsymbol{\zeta}^t)$        ◇ SGD step
    if Stopping criteria is satisfied then
        break
```

have Lipschitz-continuous gradients and Hessians and might be *nonconvex*. In empirical risk minimization tasks, $\boldsymbol{\zeta}$ is an uniformly discrete distribution over the set of training sample indices, and the stochastic function $F(\mathbf{x}; \boldsymbol{\zeta})$ corresponds to the nonconvex loss associated with such a sample.

One of the classical algorithms for optimizing (1) is the **Stochastic Gradient Descent** (SGD) method, which performs descent updates iteratively via the inexpensive stochastic gradient $\nabla F(\mathbf{x}; \boldsymbol{\zeta})$ that serves as an unbiased estimator of (the inaccessible) gradient $\nabla F(\mathbf{x})$ (Robbins and Monro, 1951; Bottou and Bousquet, 2008), i.e. $\mathbb{E}_{\boldsymbol{\zeta} \sim \mathcal{D}} \left[\nabla f(\mathbf{x}; \boldsymbol{\zeta}) \right] = \nabla f(\mathbf{x})$. Let η denote the positive stepsize, then at steps $t = 1, 2, \ldots$, the iteration performs the following update:

$$\mathbf{x}^t = \mathbf{x}^{t-1} - \eta \nabla F(\mathbf{x}^{t-1}; \boldsymbol{\zeta}^t), \tag{2}$$

where $\boldsymbol{\zeta}^t$ is randomly sampled at iteration t. SGD admits perhaps the simplest update rule among stochastic first-order methods. See Algorithm 1 for a formal illustration of the meta algorithm. It has gained tremendous popularity due to its exceptional practical performance. Taking the example of training deep neural networks, the dominating algorithm at present time is SGD (Abadi et al., 2016), where the stochastic gradient is computed via one backpropagation step. Superior characteristics of SGD have been observed in many empirical studies, including but *not* limited to fast convergence, desirable solutions of low training loss, as well as its generalization ability.

Turning to the theoretical side, relatively mature and concrete analysis in existing literatures Rakhlin et al. (2012); Agarwal et al. (2009) show that SGD achieves an *optimal* rate of convergence for convex objective function under some standard regime. Specifically, the convergence rate of $\mathcal{O}(1/T)$ in term of the function optimality gap match the algorithmic lower bound for an appropriate class of strongly convex functions (Agarwal et al., 2009).

Despite the optimal convex optimization rates that SGD achieves, the provable *nonconvex* SGD convergence rate result has long stayed upon on finding an ϵ-approximate first-order stationary point $\mathbf{x}$: with high probability SGD finds an $\mathbf{x}$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$ in $\mathcal{O}(\epsilon^{-4})$ stochastic gradient computational cost under the gradient Lipschitz condition of $f(\mathbf{x})$ (Nesterov, 2004). In contrast, our goal in this paper is to find an $(\epsilon, \sqrt{\rho\epsilon})$*-approximate second-order stationary point* $\mathbf{x}$ such that $\|\nabla f(\mathbf{x})\| \leq \epsilon$ and the least eigenvalue of the Hessian matrix $\nabla^2 f(\mathbf{x})$ is $\geq -\sqrt{\rho\epsilon}$, where $\rho > 0$ denotes the so-called Hessian-Lipschitz parameter to be specified later (Nesterov and Polyak, 2006; Tripuraneni et al., 2018; Carmon et al., 2018; Agarwal et al., 2017). Putting it differently, we need to escape from all first-order stationary points that admit a strong negative Hessian eigenvalue (a.k.a. saddle points) (Dauphin et al., 2014) and lands at a point that quantitatively resembles a local minimizer in terms of the gradient norm and least Hessian eigenvalue.

Results on the convergence rate of SGD for finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point have been scarce until very recently.[2] To the best of our knowledge, Ge et al. (2015) provided the first theoretical result that SGD with *artificially injected spherical noise* can escape

[2] Some authors work with (ϵ, δ)-stationary point and we ignore such expression due to the natural choice $\delta = \sqrt{\rho\epsilon}$ in optimization literature (Nesterov and Polyak, 2006; Jin et al., 2017).

from all saddle points in polynomial time. Moreover, Ge et al. (2015) showed that SGD finds an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point at a stochastic gradient computational cost of $\tilde{\mathcal{O}}(\text{poly}(d)\epsilon^{-8})$. A recent follow-up work by Daneshmand et al. (2018) derived a convergence rate of $\tilde{\mathcal{O}}(d\epsilon^{-10})$ stochastic gradient computations. These milestone works Ge et al. (2015); Daneshmand et al. (2018) showed that SGD can always escape from saddle points and can find an approximate local solution of (1) with a stochastic gradient computational cost that is polynomially dependent on problem-specific parameters. Motivated by these recent works, the current paper tries to answer the following questions:

(i) *Is it possible to sharpen the analysis of SGD algorithm and obtain a reduced stochastic gradient computational cost for finding an $(\epsilon, \mathcal{O}(\epsilon^{0.5}))$-approximate second-order stationary point?*

(ii) *Is artificial noise injection absolutely necessary for SGD to find an approximate second-order stationary point with an almost dimension-free stochastic gradient computational cost?*

To answer aforementioned question (i), we provide a *sharp* analysis and prove that SGD with variants only on stopping criteria finds an $(\epsilon, \sqrt{\rho\epsilon})$-approximate stationary point at a remarkable $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ stochastic gradient computational cost for solving (1). This is a unexpected result because it has been conjectured by many (Xu et al., 2018; Allen-Zhu and Li, 2018; Tripuraneni et al., 2018) that an $\tilde{\mathcal{O}}(\epsilon^{-4})$ cost is required to find an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. Our result on SGD *negates* this conjecture and serves as the sharpest stochastic gradient computational cost for SGD prior to this work. To answer question (ii) above, we propose a novel *dispersive noise assumption* and prove that under such an assumption, SGD requires *no* artificial noise injection in order to achieve the aforementioned sharp stochastic gradient computational cost. Such noise assumption is satisfied in the case of infinite online samples and Gaussian sampling zeroth-order optimization, and can be satisfied automatically by injecting artificial ball-shaped, spherical uniform, or Gaussian noises.

We emphasize that the $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ stochastic gradient computational cost is, however, *not* the lower bound complexity for finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point for problem (1). Recently, Fang et al. (2018) applied a novel variance reduction technique named SPIDER *tracking* and proposed the SPIDER-SFO^{+} algorithm which achieves a stochastic gradient computational cost of $\tilde{\mathcal{O}}(\epsilon^{-3})$ for finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. It is our belief that variance reduction techniques are necessary to achieve a stochastic gradient computational cost that is strictly sharper than $\tilde{\mathcal{O}}(\epsilon^{-3.5})$. We also note that the promising $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ complexity relies on the Hessian-smooth assumption, whereas the standard $\mathcal{O}(\epsilon^{-4})$ complexity for searching an approximate first-order stationary point does not need this assumption.

1.1. Our Contributions

We study theoretically in this work the SGD algorithm for minimizing nonconvex function $\mathbb{E}[F(\mathbf{x}; \boldsymbol{\zeta})]$. Specially, this work contributes the following:

(i) We propose a sharp convergence analysis for the classical and simple SGD and prove that the total stochastic gradient computational cost to find a second-order stationary point is at most $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ under both Lipschitz-continuous gradient and Hessian assumptions of objective function. Such convergence rate matches the most accelerated nonconvex stochastic optimization results that such as Nesterov's momentum acceleration, negative curvature search, and quadratic and cubic regularization tricks.

(ii) We propose the *dispersive noise assumption* and prove that under such an assumption, SGD ensures to escape all saddles that has a strongly negative Hessian eigenvalue. Such type of noise generalizes the existing artificial ball-shaped noise and is widely applicable to many tasks.

(iii) Our novel analytic tools for proving saddle escaping and fast convergence of SGD is of independent interests, and they shed lights on developing and analyzing new stochastic optimization algorithms.

Organization The rest of the paper is organized as follows. §2 provides the SGD algorithm and the main convergence rate theorem for finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. Related Works are discussed in §3. We conclude our paper in §4 with proposed future directions. In Appendix A, we sketch the proof of our convergence rate theorem by providing and discussing three core propositions. And all the missing proofs are detailed in the Appendix rest sections.

Notation Let $\|\cdot\|$ denote the Euclidean norm of a vector or spectral norm of a square matrix. Denote $p_n = \mathcal{O}(q_n)$ for a sequence of vectors p_n and positive scalars q_n if there is a global constant C such that $|p_n| \le Cq_n$, and $p_n = \tilde{\mathcal{O}}(q_n)$ such $\tilde{C}$ hides a poly-logarithmic factor of d and ϵ. Denote $p_n = \tilde{\Omega}(q_n)$ if there is $\tilde{C}$ which hides a poly-logarithmic factor such that $|p_n| \ge \tilde{C}q_n$. We denote $p_n \asymp q_n$ if there is $\tilde{C}$ which hides a poly-logarithmic factor of d and ϵ such that $p_n = \tilde{C}q_n$. Further, we denote linear transformation of set $\mathcal{A} \subseteq \mathbb{R}^d$ as $c_1 + c_2\mathcal{A} := \{c_1 + c_2 a : a \in \mathcal{A}\}$. Let $\lambda_{\min}(\mathbf{A})$ denote the least eigenvalue of a real symmetric matrix $\mathbf{A}$. We denote $\mathcal{B}(\mathbf{x}, R)$ as the R-neighborhood of $\mathbf{x}^0$, i.e. the set $\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}^0\| \le R\}$.

2. Algorithm and Main Result

In this section, we formally state SGD and the corresponding convergence rate theorem. In §2.1, we propose the key assumptions for the objective functions and noise distributions. In §2.2, we detail SGD in Algorithm 2 and present the main convergence rate theorem.

2.1. Assumptions and Definitions

Assumption 1 (Smoothness) *We assume that the objective function satisfies some smoothness*[3] *conditions: for all* $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$*, we have*

$$\|\nabla F(\mathbf{x}; \boldsymbol{\zeta}) - \nabla F(\mathbf{x}'; \boldsymbol{\zeta})\| \le L\|\mathbf{x} - \mathbf{x}'\|, \tag{3}$$

and

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}')\| \le \rho\|\mathbf{x} - \mathbf{x}'\|. \tag{4}$$

With Hessian-Lipschitz parameter ρ prescribed in (4), we formally define the $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. To best of our knowledge, such concept firstly appeared in Nesterov and Polyak (2006):

[3] The smoothness gradient condition for $F(\mathbf{x}; \boldsymbol{\zeta})$ is only needed for searching an approximate second-order stationary point. To find a first stationary point, we only can replace (3) with a relaxed one: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \le L\|\mathbf{x} - \mathbf{x}'\|$.

Definition 1 (Second-order Stationary Point) *Call $\mathbf{x} \in \mathbb{R}^d$ an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point if*

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

Let the starting point of our SGD algorithm be $\tilde{\mathbf{x}} \in \mathbb{R}^d$. We assume the following boundedness assumption:

Assumption 2 (Boundedness) *The $\Delta := f(\tilde{\mathbf{x}}) - f^* < \infty$ where $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ is the global infimum value of $f(\mathbf{x})$.*

Turning to the assumptions on noise, we first assume the following:

Assumption 3 (Bounded Noise) *For any $\mathbf{x} \in \mathbb{R}^d$, the stochastic gradient $\nabla F(\mathbf{x}; \boldsymbol{\zeta})$ satisfies:*

$$\|\nabla F(\mathbf{x}, \boldsymbol{\zeta}) - \nabla f(\mathbf{x})\|^2 \le \sigma^2, \qquad a.s. \tag{5}$$

An alternative (slighter weaker) assumption that also works is to assume that the norm of noise satisfies subgaussian distribution, i.e. for any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\boldsymbol{\zeta}}\left[\exp(\|\nabla F(\mathbf{x}; \boldsymbol{\zeta}) - \nabla f(\mathbf{x})\|^2 / \sigma^2)\right] \le 1. \tag{6}$$

Assumptions 1, 2 and 3 are standard in nonconvex optimization literatures (Ge et al., 2015; Xu et al., 2018; Allen-Zhu and Li, 2018; Fang et al., 2018). We treat the parameters L, ρ, Δ, and σ as global constants, and focus on the dependency for stochastic gradient complexity on ϵ and d.

For the purpose of fast saddles escaping, we need an extra noise shape assumption. Let q^* be a positive real, and let $\mathbf{v}$ be a unit vector. We define a set property as follows:

Definition 2 ($(q^*, \mathbf{v})$-narrow property) *We say that a Borel set $\mathcal{A} \subseteq \mathbb{R}^d$ satisfies the $(q^*, \mathbf{v})$-**narrow property**, if for any $\mathbf{u} \in \mathcal{A}$ and $q \ge q^*$, $\mathbf{u} + q\mathbf{v} \in \mathcal{A}^c$ holds, where $\mathcal{A}^c$ denotes the complement set of $\mathcal{A}$.*

It is easy to verify that the first parameter in the narrow property is linearly scalable and translation invariant with sets, i.e. if $\mathcal{A}$ satisfies $(q^*, \mathbf{v})$-narrow property, then for any $c_1 \in \mathbb{R}^d$ and $c_2 \in \mathbb{R}$, $c_1 + c_2\mathcal{A}$ satisfies the $(|c_2|q^*, \mathbf{v})$-narrow property. Next, we introduce the $\mathbf{v}$-dispersive property as follows:

Definition 3 ($\mathbf{v}$-dispersive property) *Let $\tilde{\boldsymbol{\xi}}$ be a random vector satisfying Assumption 3. We say that $\tilde{\boldsymbol{\xi}}$ has the $\mathbf{v}$-**dispersive property**, if for an arbitrary set $\mathcal{A}$ that satisfies the $(\sigma/(4\sqrt{d}), \mathbf{v})$-narrow property (as in Definition 2) the following holds:*

$$\mathbb{P}\left(\tilde{\boldsymbol{\xi}} \in \mathcal{A}\right) \le \frac{1}{4}. \tag{7}$$

Obviously, if $\tilde{\boldsymbol{\xi}}$ satisfies $\mathbf{v}$-dispersive property, for any fixed vector $\mathbf{a}$, then $\tilde{\boldsymbol{\xi}} + \mathbf{a}$ also satisfies $\mathbf{v}$-dispersive property. We then present the dispersive noise assumption as follows:

Assumption 4 (Dispersive Noise) *For an arbitrary point $\mathbf{x} \in \mathbb{R}^d$, $\nabla f(\mathbf{x}; \boldsymbol{\zeta})$ admits the $\mathbf{v}$-dispersive property (as in Definition 3)* ***for any unit vector*** $\mathbf{v}$.

Assumption 4 is motivated from the key lemma for escaping from saddle points in Jin et al. (2017), which obtains a sharp rate for gradient descent escaping from saddle points. Such an assumption enables SGD to move out of a *stuck region* with probability $\geq 3/4$ in its first step and enables escaping from saddle points (by repeating logarithmic rounds). We would like to emphasize that the $\mathbf{v}$-dispersive noises contain many canonical examples; see the following

Examples of Dispersive Noises Here we exemplify a few noise distributions that satisfy the $\mathbf{v}$-dispersive property, that is, for an arbitrary set $\mathcal{A}$ with $(q^*, \mathbf{v})$-narrow property, where $q^* = \sigma/(4\sqrt{d})$. We have the following proposition:

Proposition 4 *For the following noise distributions,* (7) *in Definition 3 is satisfied:*

(i) Gaussian noise: $\tilde{\boldsymbol{\xi}} = \sigma/\sqrt{d} * \boldsymbol{\chi}$ where $\boldsymbol{\chi}$ is the standard Gaussian noise with covariance matrix $\mathbf{I}_d$;

(ii) Uniform ball-shaped noise or spherical noise: $\tilde{\boldsymbol{\xi}} = \sigma * \boldsymbol{\xi}_b$, where $\boldsymbol{\xi}_b$ is uniformly sampled from the unit ball centered at $\mathbf{0}$;

(iii) Artificial noise injection: $\tilde{\boldsymbol{\xi}} = \nabla f(\mathbf{x}; \boldsymbol{\zeta}) + \tilde{\gamma}$, where $\tilde{\gamma}$ is some independent artificial noise that is $\mathbf{v}$-dispersive for any $\mathbf{v}$.

The proof of Proposition 4 is shown in Appendix F.

2.2. SGD and Main Theorem

Our SGD algorithm for analysis purposes is detailed in Algorithm 2. Our SGD algorithm only differs from classical SGD algorithms on stopping criteria. Distinct from the classical ones that simply terminate in a certain number of steps and output the final iterate or a randomly drawn iterate, the SGD we consider here introduces a *ball-controlled mechanism* as the stopping criteria: if $\mathbf{x}^k$ exits a small neighborhood in K_0 iterations (Line 2 to 6), one starts over and do the next round of SGD; if exiting does *not* occur in K_0 iterations, then the algorithm simply outputs an arithmetic average of $\mathbf{x}^k$ of the last K_0 iterates within the neighborhood, which in turns is an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point with high probability. In contrast with the stopping criteria in the deterministic setting that checks the descent in function values (Jin et al., 2017), the function value in stochastic setting is reasonably costly to approximate (costs $O(\epsilon^{-2})$ stochastic gradient computations), and the error plateaus might be hard to observe theoretically.

Algorithm 2 SGD (For finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point): Input $\tilde{\mathbf{x}}$, $K_0 \asymp \epsilon^{-2}$, $\eta \asymp \epsilon^{1.5}$, and $B \asymp \epsilon^{0.5}$.

```
Set $t = 0$, $k = 0$, $\mathbf{x}^0 = \tilde{\mathbf{x}}$
while $k < K_0$ do
    Draw an independent $\boldsymbol{\zeta}^{k+1} \sim \mathcal{D}$ and set $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - \eta \nabla F(\mathbf{x}^k; \boldsymbol{\zeta}^{k+1})$    ◇ SGD step
    $t \leftarrow t + 1,\ k \leftarrow k + 1$    ◇ Counter of SGD steps
    if $\|\mathbf{x}^k - \mathbf{x}^0\| > B$ then
        $\mathbf{x}^0 \leftarrow \mathbf{x}^k$, $k \leftarrow 0$
    end if
end while
$\bar{\mathbf{x}}_{output} \leftarrow (1/K_0) \sum_{k=0}^{K_0-1} \mathbf{x}^k$    ◇ Reach this line in $t \le T_0 = \tilde{\mathcal{O}}(\epsilon^{-3.5})$ SGD steps w.h.p.
return $\bar{\mathbf{x}}_{output}$    ◇ Return an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point
```

Parameter Setting We set the hyper-parameters[4] for Algorithm 2 as follows:

$$\tilde{C}_1 = 2\left\lfloor \frac{\log(3 \cdot p^{-1})}{\log(0.7)} + 1 \right\rfloor \log\left(\frac{24\sqrt{d}}{\eta}\right) \asymp 1, \quad \delta = \sqrt{\rho\epsilon} \asymp \epsilon^{0.5},$$
$$\delta_2 = 16\delta \asymp \epsilon^{0.5}, \quad B = \frac{\delta}{\rho\tilde{C}_1} \asymp \epsilon^{0.5}, \quad K_0 = \tilde{C}_1 \eta^{-1} \delta_2^{-1} \asymp \epsilon^{-2},$$
$$\eta \le \frac{B^2\delta}{64\max(\sigma^2, 1)\tilde{C}_1 \log(48K_0/p)} \cdot \frac{1}{3 + \log(K_0)} \asymp \epsilon^{1.5}. \tag{8}$$

For brevity of analysis, we assume $B \le \min(1, \frac{\sigma}{L}, \frac{1}{L}) \asymp \mathcal{O}(1)$, and $\delta \le 1$. In other words, we assume the accuracy $\epsilon \le \mathcal{O}(1)$.

Now we are ready to present our main result of SGD theorem.

Theorem 5 (SGD Rate) *Let Assumptions 1, 2, 3, and 4 hold. Let the parameters K_0, η and B be set in (8) with $p \in (0, 1)$ being the error probability, and set $T_1 = \left\lceil \frac{7\Delta\eta K_0}{B^2} \right\rceil + 1 \asymp \epsilon^{-1.5}$, then running Algorithm 2 in $T_0 = T_1 \cdot K_0 \asymp \frac{\Delta\rho^{1/2}}{\max(\sigma^2,1)\epsilon^{3.5}} \asymp \epsilon^{-3.5}$, with probability at least $1 - (T_1 + 1) \cdot p$, SGD outputs an $\bar{\mathbf{x}}_{output}$ satisfying*

$$\|\nabla f(\bar{\mathbf{x}}_{output})\| \le 18\rho B^2 \asymp \epsilon, \qquad \lambda_{\min}(\nabla^2 f(\bar{\mathbf{x}}_{output})) \ge -17\delta \asymp -\sqrt{\rho\epsilon}. \tag{9}$$

Treating σ, L, and ρ as global constants, the stochastic gradient computational cost is $\tilde{\mathcal{O}}(\epsilon^{-3.5})$.

Strikingly, Theorem 5 indicates that SGD in Algorithm 2 achieves a stochastic gradient computation cost of $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ to find an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point[5]. Compared with existing algorithms that achieves an $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ convergence rate, SGD is comparatively simpler to implement and does *not* invoke any additional techniques or iterations such as momentum acceleration (Jin et al., 2018b), cubic regularization (Tripuraneni et al., 2018), regularization (Allen-Zhu, 2018a), or NEON-type negative curvature search (Xu et al., 2018; Allen-Zhu and Li, 2018).

[4] Set $\tilde{\eta} = \frac{B^2\delta}{512\max(\sigma^2,1)\log(48p^{-1})\left\lfloor \frac{\log(3\cdot p^{-1})}{\log(0.7)} + 1 \right\rfloor \log(d)} \asymp \epsilon^{1.5}$. Because η in (8) involves logarithmic factors on K_0 and $\tilde{C}_1$, a simple choice to set the step size is as $\eta = \tilde{\eta}\log^{-3}(\tilde{\eta}^{-1}) \asymp \epsilon^{1.5}$.

[5] For searching a more general (ϵ_g, ϵ_H)-approximate second-order stationary point, one can obtain an complexity of $\tilde{\mathcal{O}}(\epsilon_g^{-3.5} + \epsilon_H^{-7})$ using the same technique.

Algorithm 3 Noise-Scheduled SGD (For finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point): Input $\tilde{\mathbf{x}}$, $K_o = 2\log\left(\frac{24\sqrt{d}}{\eta}\right)\eta^{-1}\delta_2^{-1} \asymp \epsilon^{-2}$, $K_0 \asymp \epsilon^{-2}$, $\eta \asymp \epsilon^{1.5}$, and $B \asymp \epsilon^{0.5}$.

```
Set $t = 0$, $k = 0$, $\mathbf{x}^0 = \tilde{\mathbf{x}}$
while $k < K_0$ do
    if $\mathrm{mod}(k, K_o) = 0$ then
        Draw an independent $\boldsymbol{\zeta}^{k+1} \sim \mathcal{D}$ and Gaussian noise $\boldsymbol{\xi}_g \sim N(0, (\sigma^2/d)\mathbf{I}_d)$
            $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^{k+1} - \eta\left(\nabla F(\mathbf{x}^k; \boldsymbol{\zeta}^{k+1}) + \boldsymbol{\xi}_g\right)$    ◇ SGD step (with noise injection)
    else
        Draw an independent $\boldsymbol{\zeta}^{k+1} \sim \mathcal{D}$ and set $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - \eta\nabla F(\mathbf{x}^k; \boldsymbol{\zeta}^{k+1})$    ◇ SGD step
    end if
    $t \leftarrow t + 1$, $k \leftarrow k + 1$    ◇ Counter of SGD steps
    if $\|\mathbf{x}^k - \mathbf{x}^0\| > B$ then
        $\mathbf{x}^0 \leftarrow \mathbf{x}^k$, $k \leftarrow 0$
    end if
end while
$\bar{\mathbf{x}}_{output} \leftarrow (1/K_0)\sum_{k=0}^{K_0-1}\mathbf{x}^k$    ◇ Reach this line in $t \le T_0 = \tilde{\mathcal{O}}(\epsilon^{-3.5})$ SGD steps, w.h.p.
return $\bar{\mathbf{x}}_{output}$    ◇ Return an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point
```

Admittedly, the best-known SGD theoretical guarantee in Theorem 5 relies on a dispersive noise assumption. To remove such an assumption, we argue that only $\tilde{\mathcal{O}}(1)$ steps of each round does one need to run an SGD step of dispersive noise to enable efficient escaping. We propose a variant of SGD called *Noise-Scheduled SGD* which requires artificial noise injection but does *not* rely on a dispersive noise assumption. The algorithm is shown in Algorithm 3. One can obtain the convergence property straightforwardly.

Remark 6 *For the function class that admits the strict-saddle property (Carmon et al., 2018; Ge et al., 2015; Jin et al., 2017), an approximate second-order stationary point is guaranteed to be an approximate local minimizer. For example for optimizing a σ_*-strict-saddle function, one can first find an $(\epsilon_*, \sqrt{\rho\epsilon_*})$-approximate second-order stationary point with $\epsilon_* \le \sigma_*^2/(2\rho)$ which is guaranteed to be an approximate local minimizer due to the strict-saddle property. Our SGD convergence rate $\mathcal{O}(\epsilon_*^{-3.5}) = \mathcal{O}(\sigma_*^{-7})$ is independent of the target accuracy ϵ, and one can run a standard convex optimization theory to obtain an $\mathcal{O}(1/t)$ convergence rate in terms of the optimality gap. Limited by space we omit the details.*

3. Discussions on Related Works

Due to the recent heat of deep learning, many researchers have studied the nonconvex SGD method from various perspectives in the machine learning community. We compare our results with concurrent theoretical works on nonconvex SGD in the following discussions. For clarity, we also compare the convergence rates of some works most related to ours in Table 1.

(i) **Pioneer SGD:** The first work on SGD escaping from saddle points Ge et al. (2015) obtain a stochastic gradient computational cost of $\tilde{\mathcal{O}}(\text{poly(d)}\epsilon^{-8})$.[6] Later, Jin et al. (2017, 2018b)

[6] The analysis in (Ge et al., 2015) indicates a $poly(d)$ factor of $\mathcal{O}(d^8)$ at least.

	Algorithm		SG Comp. Cost
SGD Variants	NEON+SGD	(Xu et al., 2018)	ϵ^{-4}
	NEON2+SGD	(Allen-Zhu and Li, 2018)	
	Stochastic Cubic	(Tripuraneni et al., 2018)	$\epsilon^{-3.5}$
	RSGD5	(Allen-Zhu, 2018a)	
	Natasha2$^{\Delta}$	(Allen-Zhu, 2018b)	$\epsilon^{-3.5}$
	NEON2+SNVRG$^{\ominus}$	(Zhou et al., 2018a)	
	SPIDER	(Fang et al., 2018)	ϵ^{-3}
Original SGD	SGD	(Ge et al., 2015)	$poly(d)\epsilon^{-8}$
		(Daneshmand et al., 2018)	$d^4\epsilon^{-5}$
		(Jin et al., 2019)	ϵ^{-4}
		(this work)	$\epsilon^{-3.5}$

Table 1: Comparable results on the stochastic gradient computational cost for nonconvex optimization algorithms in finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point for problem (1) under standard assumptions. Note that each stochastic gradient computational cost may hide a poly-logarithmic factors of d, n, ϵ.

Orange-boxed: SPIDER reported in orange-boxed is the only existing variant stochastic algorithm that achieves provable faster rate by order than simple SGD.

$^{\Delta}$: Allen-Zhu (2018b) also obtains a stochastic gradient computational cost of $\tilde{\mathcal{O}}(\epsilon^{-3.25})$ for finding a relaxed $(\epsilon, \mathcal{O}(\epsilon^{0.25}))$-approximate second-order stationary point.

$^{\ominus}$: With additional third-order smoothness assumptions, SNVRG (Zhou et al., 2018a) achieves complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$.

noise-perturbed GD and AGD and achieve sharp gradient computational costs, which suggests the possibility of sharper SGD rate for escaping saddles. Our analysis in this work is partially motivated by Jin et al. (2017) for escaping from saddle points, but generalizes the noise condition and needs *no* deliberate noise injections which is *not* the original GD/SGD algorithm in a strict sense.

(ii) **Concurrent SGD:** A recent result by Daneshmand et al. (2018) obtains a stochastic computation cost of $\tilde{\mathcal{O}}(\tau^{-2}\epsilon^{-10})$ to find an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. The highlight of their work is that they need *no* injection of artificial noises. Nevertheless in their work, the Correlated Negative Curvature parameter τ^{-2} *cannot* be treated as an $\mathcal{O}(1)$-constant. Taking the case of injected spherical noise or Gaussian noise, it can be at most linearly dependent on d [Assumption 4], so the result is *not* (almost) dimension-free, and worst-case convergence rate shall be interpreted as $\tilde{\mathcal{O}}(d^4\epsilon^{-5})$. Concurrently with our work, Jin et al. (2019) also extend the technique in Jin et al. (2017) to work on SGD and prove that SGD with the injected noise can find an approximate second-order stationary point with stochastic computation cost of $\tilde{\mathcal{O}}(\epsilon^{-4})$. Besides, they further study the case when the individual function $F(\mathbf{x}; \zeta)$ does not satisfy gradient-smooth condition and obtain a complexity of $\tilde{\mathcal{O}}(d\epsilon^{-4})$.

(iii) **NC search + SGD:** The NEON+SGD (Xu et al., 2018; Allen-Zhu and Li, 2018) methods achieve a dimension-free convergence rate of $\tilde{\mathcal{O}}(\epsilon^{-4})$ for the general problem of form (1) to reach an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. Prior to this, classical nonconvex GD/SGD only achieves such a rate for finding an ϵ-approximate first-order stationary point

(Nesterov, 2004), which, with the help of NEON method, successfully escapes from saddles via a Negative Curvature (NC) search iteration (Xu et al., 2018; Allen-Zhu and Li, 2018).

(iv) **Regularization + SGD:** Very recently, Allen-Zhu (2018a) takes a quadratic regularization approach and equips it with a negative-curvature search iteration NEON2 (Allen-Zhu and Li, 2018), which successfully improves the rate to $\tilde{\mathcal{O}}(\epsilon^{-3.5})$. In comparison, our method achieves essentially the same rate without using regularization methods. Tripuraneni et al. (2018) proposed a stochastic variant of cubic regularization method (Nesterov and Polyak, 2006; Agarwal et al., 2017) and achieves the same $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ convergence rate, which is the first achieving such rate without invoking variance reduced gradient techniques.[7]

(v) **NC search + VR:** Allen-Zhu (2018b) converted a NC search method to the online stochastic setting (Carmon et al., 2018) and achieved a convergence rate of $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ for finding an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point. For finding a relaxed $(\epsilon, \mathcal{O}(\epsilon^{0.25}))$-approximate second-order stationary point, Allen-Zhu (2018b) obtains a lower stochastic gradient computational cost of $\tilde{\mathcal{O}}(\epsilon^{-3.25})$. With a recently proposed *optimal* variance reduced gradient techniques applied, SPIDER achieves the state-of-the-art $\tilde{\mathcal{O}}(\epsilon^{-3})$ stochastic gradient computational cost (Fang et al., 2018).[8] Very recently, Zhou and Gu (2019) and Shen et al. (2019) have independently designed powerful cubic algorithms using SPIDER techinique and also obtained a complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$.

3.1. More Related Works

VR Methods In the recent two years, sharper convergence rates for nonconvex stochastic optimization can be achieved using *variance reduced gradient techniques* (Schmidt et al., 2017; Johnson and Zhang, 2013; Xiao and Zhang, 2014; Defazio et al., 2014). The SVRG/SCSG (Lei et al., 2017) adopts the technique from Johnson and Zhang (2013) and novelly introduces a random stopping criteria for its inner loops and achieve a stochastic gradient costs of $\mathcal{O}(\epsilon^{-3.333})$. Very recently, two independent works, namely SPIDER (Fang et al., 2018) and SVRC (Zhou et al., 2018b), design sharper variance reduced gradient methods and obtain a stochastic gradient computational costs of $\mathcal{O}(n^{1/2}\epsilon^{-2} \wedge \epsilon^{-3})$, which is state-of-the-art and *near-optimal* in the sense that they achieve the algorithmic lower bound in the finite-sum setting.

Escaping Saddles in Single-Function Case Recently, many theoretical works care about convergence to an approximate second-order stationary point or escaping from saddles for the case of one single function (Carmon and Duchi, 2016; Jin et al., 2017; Carmon et al., 2018, 2017; Agarwal et al., 2017; Jin et al., 2018b; Lee et al., 2017; Du et al., 2017). Among them, the work Jin et al. (2017) proposed a ball-shaped-noise-perturbed variant of gradient descent which can efficiently escape saddle points and achieves a sharp stochastic gradient computational cost of ϵ^{-2}, which is also achieved by NEON+GD (Xu et al., 2018; Allen-Zhu and Li, 2018). Another line of works apply momentum acceleration techniques (Agarwal et al., 2017; Carmon et al., 2017; Jin et al., 2018b) and achieve a rate of $\epsilon^{-1.75}$ for a general optimization problem.

[7] Note in the convergence rate here, we also includes the number of stochastic Hessian-vector product evaluations, each of which takes about the same magnitude of time as per stochastic gradient evaluation.

[8] The independent work Zhou et al. (2018a) achieves a similar convergence rate for finding an ϵ-approximate second-order stationary point by imposing a third-order smoothness conditions on the objective.

Escaping Saddles in Finite-Sum Case For the finite-sum setting, many works have applied variance reduced gradient methods (Agarwal et al., 2017; Carmon et al., 2018; Fang et al., 2018; Zhou et al., 2018a) and further reduce the stochastic gradient computational cost to $\tilde{\mathcal{O}}(n\epsilon^{-1.5} + n^{3/4}\epsilon^{-1.75})$ (Agarwal et al., 2017; Allen-Zhu and Li, 2018). Reddi et al. (2018) proposed a simpler algorithm that obtains a stochastic gradient cost of $\tilde{\mathcal{O}}\left(n\epsilon^{-1.5} + n^{3/4}\epsilon^{-1.75} + n^{2/3}\epsilon^{-2}\right)$. With recursive gradient method applied (Fang et al., 2018; Zhou et al., 2018a), the stochastic gradient cost further reduces to $\tilde{\mathcal{O}}\left((n\epsilon^{-1.5} + n^{3/4}\epsilon^{-1.75}) \wedge (n + n^{1/2}\epsilon^{-2} + \epsilon^{-2.5})\right)$, which is the state-of-the-art.

Miscellaneous It is well-known that for general nonconvex optimization problem in the form of (1), finding an approximate global minimizer is in worst-case *NP-hard* (Hillar and Lim, 2013). Seeing this, many works turn to study the convergence properties based on specific models. Faster convergence rate to local or even global minimizers can be guaranteed for many statistical learning tasks such as principal component analysis (Li et al., 2018a; Jain and Kar, 2017), matrix completion (Jain et al., 2013; Ge et al., 2016; Sun and Luo, 2016), dictionary learning (Sun et al., 2015, 2017) as well as linear and nonlinear neural networks (Zhong et al., 2017; Li and Yuan, 2017; Li et al., 2018b).

In retrospect, our focus in this paper is on escaping from saddles, and we refer the readers to recent inspiring works studying how to escape from local minimizers Zhang et al. (2017); Jin et al. (2018a).

4. Conclusions and Future Direction

In this paper, we presented a sharp convergence analysis for the classical SGD algorithm. We showed that equipped with a ball-controlled stopping criterion, SGD achieves a stochastic gradient computational cost of $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ for finding an $(\epsilon, \mathcal{O}(\epsilon^{0.5}))$-approximate second-order stationary point, which improves over the best-known SGD convergence rate $\mathcal{O}\left(\min(poly(d)\epsilon^{-8}, d\epsilon^{-10})\right)$ prior to our work. While this work focuses on sharpened convergence rate, there are still some important questions left:

(i) It is still unknown whether SGD achieves a rate that is faster than $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ or $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ is exactly the lower bound for SGD to solve the general problem in the form of (1). As we mentioned in §1, it is our conjecture that variance reduction methods are necessary to achieve an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order stationary point in fewer than $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ steps.

(ii) We have *not* considered several important extensions in this work, such as the convergence rate of SGD in solving constrained optimization problems, and how one extends the analysis in this paper to the proximal case.

(iii) It will be also interesting to study the stochastic version of Nesterov's accelerated gradient descent (AGD) (Jin et al., 2018b).

Acknowledgement The authors would like to greatly thank Chris Junchi Li providing us with a proof of SGD to escape saddle points in $\tilde{\mathcal{O}}(\epsilon^{-4})$ computational costs and carefully revising our paper. The authors also would like to thank Haishan Ye for very helpful discussions and Huan Li, Zebang Shen, and Li Shen for very helpful comments.

Zhouchen Lin is supported by 973 Program of China (grant no. 2015CB352502), NSF of China (grant nos. 61625301 and 61731018), Qualcomm, and Microsoft Research Asia.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 1165–1175, 2018a.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. In *Advances in Neural Information Processing Systems*, pages 2676–2687, 2018b.

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

Peter L Bartlett, Varsha Dani, Thomas P Hayes, Sham M Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 31st Conference On Learning Theory*, 2008.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

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

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

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

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

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

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 686–696, 2018.

David A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th 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.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

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

Chi Jin, Lydia T Liu, Rong Ge, and Michael I Jordan. On the local minima of the empirical risk. In *Advances in Neural Information Processing Systems*, pages 4901–4910, 2018a.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, pages 1042–1085, 2018b.

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

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

Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

Olav Kallenberg and Rafal Sztencel. Some dimension-free features of vector-valued martingales. *Probability Theory and Related Fields*, 88(2):215–247, 1991.

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

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.

Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018a.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018b.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

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

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456, 2012.

Sashank Reddi, Manzil Zaheer, Suvrit Sra, Barnabas Poczos, Francis Bach, Ruslan Salakhutdinov, and Alex Smola. A generic approach for escaping saddle points. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1233–1242, 2018.

Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Zebang Shen, Pan Zhou, Cong Fang, and Alejandro Ribeiro. A stochastic trust region method for non-convex minimization. *arXiv preprint arXiv:1903.01540*, 2019.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

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.

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

Yi Xu, Jing Rong, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5531–5541, 2018.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, 2017.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

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

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 reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3922–3933, 2018b.

Appendix A. Proof Sketches for Theorem 5

We briefly introduce our proof techniques to prove our main Theorem 5 in this section. The rigorous proof is shown in Appendix C, D, and E. For convenience, when we study Algorithm 2 in each inner loop from Line 2 to Line 8, we override the definition of $\mathbf{x}^0$ as its initial vector. Our poof basically consists of two ingredients. The first is to prove that SGD can efficiently escape saddles: with high probability, if $\lambda_{\min}\left(\nabla^2 f(\mathbf{x}^0)\right) \leq -\delta_2 \asymp -\epsilon^{0.5}$, $\mathbf{x}^k$ moves out of $\mathcal{B}(\mathbf{x}^0, B)$ in K_0 iterations (Refer to Appendix C). The second is to show that SGD converges with a faster rate of $\tilde{\mathcal{O}}(\epsilon^{-3.5})$, rather than $\tilde{\mathcal{O}}(\epsilon^{-4})$. We further separate the second destination into two parts:

1. Throughout the execution of the algorithm, each time $\mathbf{x}^k$ moves out of $\mathcal{B}(\mathbf{x}^0, B)$, with high probability, the function value shall decrease with a magnitude at least $\tilde{\mathcal{O}}(\epsilon^{1.5})$ (Refer to Appendix D).
2. Once $\mathbf{x}^k$ does not move out of $\mathcal{B}(\mathbf{x}^0, B)$ until K_0 iteration, with high probability, we find a desired approximate second-order stationary point (Refer to Appendix E).

Let $\boldsymbol{\mathcal{F}}^k = \sigma\{\mathbf{x}^0, \boldsymbol{\zeta}^1, \cdots, \boldsymbol{\zeta}^k\}$ be the filtration involving the full information of all the previous k times iterations, where $\sigma\{\cdot\}$ denotes the sigma field. And let $\mathscr{K}_0$ be the first time (mathematically, a stopping time) that $\mathbf{x}^k$ exits the B-neighborhood of $\mathbf{x}^0$, i.e.

$$\mathscr{K}_0 = \inf_k\{k \geq 0 : \|\mathbf{x}^k - \mathbf{x}^0\| > B\}. \tag{10}$$

Both $\mathbf{x}^k$ and $\mathcal{I}_{\mathscr{K}_0 > k}$ is measurable on $\boldsymbol{\mathcal{F}}^k$, where $\mathcal{I}$ denotes the indicator function.

A.1. Part I: Escaping Saddles

Our goal is to prove the following proposition:

Proposition 7 *Assume $\lambda_{\min}\left(\nabla^2 f(\mathbf{x}^0)\right) \leq -\delta_2$, and recall the parameter set in (8). Initialized at $\mathbf{x}^0$ and running Line 2 to Line 8, with probability at least $1 - \frac{p}{3}$ we have*

$$\mathscr{K}_0 \leq K_0 = \left(\left\lfloor \frac{\log(3 \cdot p^{-1})}{\log(0.7^{-1})} \right\rfloor + 1\right) K_o, \tag{11}$$

where $K_o = 2\log\left(\frac{24\sqrt{d}}{\eta}\right)\eta^{-1}\delta_2^{-1}$.

Proposition 7 essentially says that assuming if the function has a negative Hessian eigenvalue $\leq -\delta_2$ at $\mathbf{x}^0$, the iteration exits the B-neighborhood of $\mathbf{x}^0$ in $K_0 = \tilde{\mathcal{O}}(\eta^{-1}\delta_2^{-1})$ steps with a high probability.

To prove Proposition 7, we let $\mathbf{w}^k(\mathbf{u})$, $k \geq 0$ be the iteration by SGD starting from a fixed $\mathbf{u} \in \mathbb{R}^d$ using the same stochastic samples as iteration $\mathbf{x}^k$, i.e.

$$\mathbf{w}^k(\mathbf{u}) = \mathbf{w}^{k-1}(\mathbf{u}) - \eta\nabla F(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}^k). \tag{12}$$

Obviously, we have $\mathbf{x}^k = \mathbf{w}^k(\mathbf{x}^0)$. Let $\mathcal{K}_{exit}(\mathbf{u})$ be the first step number k (a stopping time) such that $\mathbf{w}^k(\mathbf{u})$ exits the B-neighborhood of $\mathbf{x}^0$. Formally,

$$\mathcal{K}_{exit}(\mathbf{u}) := \inf\{k \geq 0 : \|\mathbf{w}^k(\mathbf{u}) - \mathbf{x}^0\| > B\}. \tag{13}$$

It is easy to see from (10) that $\mathscr{K}_0 = \mathcal{K}_{exit}(\mathbf{x}^0)$. Inspired from Jin et al. (2017), we cope with the stochasticity of gradients and define the so-called *bad initialization region* as the point $\mathbf{u}$ initialized from which iteration $\mathbf{w}^k(\mathbf{u})$ exits the B-neighborhood of $\mathbf{x}^0$ with probability ≤ 0.4:

$$\mathcal{S}^B_{K_o}(\mathbf{x}^0) := \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) < K_o\right) \leq 0.4 \right\}. \tag{14}$$

We will show that the bad initialization region $\mathcal{S}^B_{K_o}(\mathbf{x}^0)$ enjoys the $(q_0, \mathbf{e}_1)$-narrow property, where $q_0 = \frac{\sigma}{4\sqrt{d}}$. Since the first step will provide a continuous noise as supposed by Assumption 3, with the properly selected q_0, it will move the iteration out of the bad initialization region in its first step with probability $\geq 3/4$. Repeating such an argument in a logarithmic number of rounds enables escaping to occur with high probability.

The idea is to prove the following lemma:

Lemma 8 *Let the assumptions of Proposition 7 hold, and assume WLOG $\mathbf{e}_1$ be an arbitrary eigenvector of $\nabla^2 f(\mathbf{x}^0)$ corresponding to its smallest eigenvalue $-\delta_m$, which satisfies $\delta_m \geq \delta_2 > 0$. Then we have for any fixed $q \geq q_0$ and pair of points $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathcal{B}(\mathbf{x}^0, B)$ that*

$$\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) \geq K_o \text{ and } \mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1) \geq K_o\right) \leq 0.1. \tag{15}$$

Lemma 8 is inspired from Lemma 15 in Jin et al. (2017). Nevertheless due to the noise brought in at each update step, the analysis of stochastic gradient differs from that of the gradient descent in many aspects. For example, instead of showing the decrease of function value, we need to show that with a positive probability, at least one of the two iterations, $\mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1)$ or $\mathbf{w}^k(\mathbf{u})$, exits the B-neighborhood of $\mathbf{x}^0$. Our proof is also more intuitive compared with Lemma 15 in Jin et al. (2017). The core idea is to focus on analyzing the difference trajectory for $\mathbf{w}^k(\mathbf{u}+q\mathbf{e}_1)$ and $\mathbf{w}^k(\mathbf{u})$, and to show that the rotation speed for the difference trajectory is the same as the expansion speed. Detailed proof is provided in §C.1.

A.2. Part II: Faster Descent

The goal of Part II is to prove the following proposition:

Proposition 9 (Faster Descent) *For Algorithm 2 with parameter set in (8). With probability at least $1 - \frac{2}{3}p$, if $\mathbf{x}^k$ moves out of $\mathcal{B}(\mathbf{x}^0, B)$ in K_0 iteration, we have*

$$f\left(\mathbf{x}^{\mathscr{K}_0}\right) \leq f\left(\mathbf{x}^0\right) - \frac{B^2}{7\eta K_0}. \tag{16}$$

Proposition 9 is the key for SGD to achieve the reduced $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ stochastic computation costs. It shows that no matter what does the local surface of $f(\mathbf{x})$ look like, once $\mathbf{x}^k$ moves out of the ball in $K_0 \asymp \epsilon^{-2}$ iterations, the function value shall decrease with a magnitude of at least $\tilde{O}(\epsilon^{1.5})$. To put it differently, on average, the function value decreases at least $\tilde{\mathcal{O}}(\epsilon^{3.5})$ per-iteration during the execution of Algorithm 2. We will present the basic argument below.

We start with reviewing the more traditional approach for proving sufficient descent of SGD, and then we will discuss how to improve it as done in this work. The previous approaches are all based on the idea of (Nesterov, 2004), which mainly takes advantage of the gradient-smoothness condition of the objective. The proof can be briefly described below:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\zeta}} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \mathbb{E}_{\boldsymbol{\zeta}}\left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \frac{L}{2}\mathbb{E}_{\boldsymbol{\zeta}}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\
&\overset{(2)}{=} f(\mathbf{x}^k) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L}{2}\mathbb{E}_{\boldsymbol{\zeta}}\|\nabla F(\mathbf{x}^k; \boldsymbol{\zeta}^{k+1}) - \nabla f(\mathbf{x}^k)\|^2 \\
&\overset{\text{Assum.3}}{\leq} f(\mathbf{x}^k) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla f(\mathbf{x}^k)\|^2 + \frac{\eta^2 L\sigma^2}{2}. \qquad (17)
\end{aligned}$$

From the above derivation, in order to guarantee the monotone descent of function value in expectation, the step size η needs to be

$$\eta = \mathcal{O}\left(\frac{\|\nabla f(\mathbf{x}^k)\|^2}{L\sigma^2}\right) = \mathcal{O}(\epsilon^2), \qquad (18)$$

where the last equality uses $\left\|\nabla f(\mathbf{x}^k)\right\| \geq \epsilon$. Plugging (18) into (17), and using $\|\nabla f(\mathbf{x}^k)\| \geq \epsilon$, we have that the function value per-iteration would descent with a magnitude of at least $\mathcal{O}(\epsilon^4)$. Such result indicates that, in the worse case, SGD takes $\mathcal{O}(\epsilon^{-4})$ stochastic oracles to find an ϵ-approximate first-order stationary point. This simple argument is the reason why previous works conjectured that the complexity of SGD is $\mathcal{O}(\epsilon^{-4})$.

However, in this paper, we show that the above analysis can be further improved by using the Hessian-smoothness condition of the objective, and by considering the decomposition of objective function $f(\mathbf{x}) = f_+(\mathbf{x}) + f_-(\mathbf{x})$, and treating component $f_+(\mathbf{x})$ and component $f_-(\mathbf{x})$ separately as follows:

- (Case 1) The component $f_+(\mathbf{x})$ is near convex locally, in the sense that $\lambda_{\min}\left(\nabla^2 f(\mathbf{x})\right) \geq -\Omega(\epsilon^{0.5})$ for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^0, B)$. In this case, by using techniques for near convex problems, it is possible for us to take a larger stepsize $\eta = \mathcal{O}(\epsilon^{1.5})$ and prove a faster convergence rate.

- (Case 2) The component $f_-(x)$ is near concave locally, in the sense that $\lambda_{\max}\left(\nabla^2 f(\mathbf{x})\right) \leq \mathcal{O}(\epsilon^{0.5})$ for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^0, B)$. In this case, It can be shown that the last term on the right hand side of (17) can be reduced to $\mathcal{O}(\eta^2\epsilon^{0.5}\sigma^2)$. Therefore the step size can be chosen as $\eta = \mathcal{O}(\epsilon^{1.5})$, leading to a fast function value reduction.

To formalize the above observations into a rigorous proof, in this paper we introduce the quadratic approximation of $f(\mathbf{x})$ at point $\mathbf{x}^0$, defined as

$$g(\mathbf{x}) \coloneqq \left[\nabla f(\mathbf{x}^0)\right]^\top \left(\mathbf{x} - \mathbf{x}^0\right) + \frac{1}{2}\left[\mathbf{x} - \mathbf{x}^0\right]^\top \nabla^2 f(\mathbf{x}^0)\left[\mathbf{x} - \mathbf{x}^0\right]. \qquad (19)$$

We let $\mathcal{S}$ be the subspace spanned by all eigenvectors of $\nabla^2 f(\mathbf{x}^0)$ whose eigenvalue is greater than 0, and $\mathcal{S}\perp$ denotes the complement space. Also let $\boldsymbol{\mathcal{P}}_{\mathcal{S}} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp} \in \mathbb{R}^{d\times d}$ as the projection matrices onto the space of $\mathcal{S}$ and $\mathcal{S}\perp$, respectively. Also let the full SVD decomposition of $\nabla^2 f(\mathbf{x}^0)$ be $\mathbf{V}\sum\mathbf{V}^T$. We introduce $\mathbf{H}_{\mathcal{S}} = \mathbf{V}\sum_{(\lambda_i>0)}\mathbf{V}^T$ and $\mathbf{H}_{\mathcal{S}\perp} = \mathbf{V}\sum_{(\lambda_i\leq 0)}\mathbf{V}^T$ respectively, and define the following two auxiliary functions $g_{\mathcal{S}} : \mathcal{S} \to \mathbb{R}$ and $g_{\mathcal{S}\perp} : \mathcal{S}\perp \to \mathbb{R}$:

$$g_{\mathcal{S}}(\mathbf{u}) \coloneqq \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\nabla f\left(\mathbf{x}^0\right)\right]^\top \mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{H}_{\mathcal{S}}\mathbf{u}, \qquad (20)$$

and

$$g_{\mathcal{S}\perp}(\mathbf{v}) := \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\nabla f\left(\mathbf{x}^0\right)\right]^\top \mathbf{v} + \frac{1}{2}\mathbf{v}^T\mathbf{H}_{\mathcal{S}\perp}\mathbf{v}. \tag{21}$$

For the previously mentioned decomposition of $f(\mathbf{x}) = f_+(\mathbf{x}) + f_-(\mathbf{x})$, one may simply take $f_+(\mathbf{x}) = f(\boldsymbol{\mathcal{P}}_{\mathcal{S}}x)$, and let $f_-(\mathbf{x}) = f(\mathbf{x}) - f_+(\mathbf{x})$. It can be checked that $f_+(\cdot) = f_+(\mathbf{x}^0) + g_{\mathcal{S}}(\cdot) + \tilde{\mathcal{O}}(\epsilon^{1.5})$ and $f_-(\cdot) = f_-(\mathbf{x}^0) + g_{\mathcal{S}\perp}(\cdot) + \tilde{\mathcal{O}}(\epsilon^{1.5})$. It follows that we only need to separately analyze the two quadratic approximations $g_{\mathcal{S}}(\cdot)$ and $g_{\mathcal{S}\perp}(\cdot)$. We then bound the difference between $f(\mathbf{x}^{\mathscr{K}_0})$ and $g_{\mathcal{S}}(\mathbf{x}^{\mathscr{K}_0} - \mathbf{x}^0) + g_{\mathcal{S}\perp}(\mathbf{x}^{\mathscr{K}_0} - \mathbf{x}^0) + f(\mathbf{x}^0)$ as $\tilde{\mathcal{O}}(\epsilon^{1.5})$.

The analysis for $g_{\mathcal{S}\perp}(\cdot)$ can be obtained via the standard analysis informally described above in Case 2 (Refer to Lemma 18).

Our proof technique for dealing with $g_{\mathcal{S}}(\cdot)$ is to introduce an auxiliary trajectory with the following deterministic updates for $k = 0, 1, 2, \ldots,$ as:

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \eta\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right), \tag{22}$$

and $\mathbf{y}^0 = \mathbf{0}$. We then track and analyze the difference trajectory between $\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x}^{\mathscr{K}_0} - \mathbf{x}^0\right)$ and $\mathbf{y}^{\mathscr{K}_0}$ (Refer to Lemma 17). In the sense that $\mathbf{y}^k$ simply performs Gradient Descent, we can arrive our final results for $g_{\mathcal{S}}(\cdot)$ (Refer to Lemma 16), which leads to a rigorous statement of Case 1.

Finally, via the fact that $\mathbf{x}^k$ moves out of the ball in K_0 iteration throughout the execution of Algorithm 2, we prove that with high probability the sum for the norm of gradients can be lower bounded as:

$$\sum_{k=0}^{\mathscr{K}_0-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{x}^k - \mathbf{x}^0\right)\right\|^2 + \sum_{k=0}^{\mathscr{K}_0-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2 = \tilde{\Omega}(1), \tag{23}$$

which ensures sufficient descent of the function value. By putting the above arguments together, we can obtain Proposition 9.

A.3. Part III: Finding SSP

Part III proves the following proposition:

Proposition 10 *With probability of at least $1 - p$, if $\mathbf{x}^k$ has not moved out of the ball in K_0 iterations, then let $\bar{\mathbf{x}} = \sum_{k=0}^{K_0-1}\mathbf{x}^k$, we have*

$$\|\nabla f(\bar{\mathbf{x}})\| \le 18\rho B^2 \asymp \epsilon, \qquad \lambda_{\min}(\nabla^2 f(\bar{\mathbf{x}})) \ge -17\delta \asymp -\sqrt{\rho\epsilon}. \tag{24}$$

Proposition 10 can be obtained via the same idea of Part II. We first study the quadratic approximation function $g(\bar{\mathbf{x}})$ and then bound the difference between $g(\bar{\mathbf{x}})$ and $f(\bar{\mathbf{x}})$.

Finally, integrating Proposition 7, 9, and 10, and using the boundedness of the function value in Assumption 2, we know with probability at least $1 - (T_1 + 1)p$, Algorithm 2 shall stop before T_0 steps, and output an approximate second-order stationary point satisfying (9), which immediately leads to Theorem 5.

Appendix B. Concentration Inequalities

In our proofs, concentration inequalities are fundamental to obtain the high-probability result. Before we prove our results, we introduce the following two (advanced) inequalities which will be used in our proofs.

B.1. Vector-Valued Concentration Inequality

Theorem 11 (Vector-Martingale AzumaHoeffding, Theorem 3.5 in Pinelis (1994)) *Let $\boldsymbol{\epsilon}_{1:K} \in \mathbb{R}^d$ be a vector-valued martingale difference sequence with respect to $\boldsymbol{\mathcal{F}}^k$, i.e. for each $k = 1, \ldots, K$, $\mathbb{E}[\boldsymbol{\epsilon}_k \mid \boldsymbol{\mathcal{F}}^{k-1}] = 0$ and $\|\epsilon_k\|^2 \leq B_k^2$. We have*

$$\mathbb{P}\left(\left\|\sum_{k=1}^{K} \boldsymbol{\epsilon}_k\right\| \geq \lambda\right) \leq 4\exp\left(-\frac{\lambda^2}{4\sum_{k=1}^{K} B_k^2}\right), \tag{25}$$

where λ is an arbitrary real positive number.

Theorem 11 is not a straightforward derivation of one-dimensional Azuma's inequality. Because the bound on the right hand of (25) is *dimension-free*. Such result might be first found by Pinelis (1994). See also Kallenberg and Sztencel (1991), Lemma 4.4 in Zhang (2005) or Theorem 2.1 in Zhang (2005) and the references therein.

B.2. Data-Dependent Concentration Inequality

Theorem 12 (Date-Dependent Concentration Inequality, Lemma 3 in Rakhlin et al. (2012)) *Let $\epsilon_{1:K} \in \mathbb{R}$ be a martingale difference sequence with respect to $\boldsymbol{\mathcal{F}}^k$, i.e. for each $k = 1, \ldots, K$, $\mathbb{E}[\boldsymbol{\epsilon}_k \mid \boldsymbol{\mathcal{F}}^{k-1}] = 0$, and*

$$\mathbb{E}[\epsilon_k^2 \mid \boldsymbol{\mathcal{F}}^{k-1}] \leq \sigma_k^2.$$

Furthermore, assume that $\mathbb{P}(\|\epsilon_k\| \leq b \mid \boldsymbol{\mathcal{F}}^{k-1}) = 1$. Let $V_K^2 = \sum_{k=1}^{K} \sigma_k^2$, for any $\delta < 1/e$ and $K \geq 4$, we have

$$\mathbb{P}\left(\sum_{k=1}^{K} \epsilon_k > 2\max\left\{2\sqrt{V_k}, b\sqrt{\log(1/\delta)}\right\}\sqrt{\log(1/\delta)}\right) \leq \log(K)\delta. \tag{26}$$

Theorem 12 extends the standard Freedman's Inequality (Freedman, 1975) by allowing σ_k being the conditional variance. Similar results can be found in Bartlett et al. (2008) and Lemma 2 in Zhang (2005) and the references therein.

Note that Theorem 11 and 12 only list the results for the bounded martingale difference. Similar results can also be established when the martingale difference follows from a sub-gaussian distribution. In the rest of our proofs, we also only present the results for the bounded noise case, i.e. (5) in Assumption 3. Analogous analysis can be applied for sub-gaussian noise, i.e. (6) in Assumption 3.

Appendix C. Deferred Proofs of Part I: Escaping Saddles

Let the deterministic time

$$K_o = 2\log\left(\frac{24\sqrt{d}}{\eta}\right)\eta^{-1}\delta_2^{-1} \geq \left\lceil \frac{\log(6/q_0)}{\log(1+\eta(\delta_2))} \right\rceil \overset{B\leq 1}{\geq} \left\lceil \frac{\log(6B/q_0)}{\log(1+\eta(\delta_2))} \right\rceil, \tag{27}$$

where $q_0 = \frac{\sigma\eta}{4\sqrt{d}}$. We prove Proposition 7 that bound the iteration number to escape $\mathcal{B}(\mathbf{x}^0, B)$.
Proof [Proof of Proposition 7]

(i) We prove in this item that $\mathcal{S}_{K_o}^B(\mathbf{x}^0)$ satisfies the $(q_0, \mathbf{e}_1)$-narrow property, i.e. there cannot be two points $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)$ such that $q \geq q_0$. Indeed if such two points do exist, from (14) we have

$$\mathbb{P}(\mathcal{K}_{exit}(\mathbf{u}) \geq K_o) \geq 0.6 \quad \text{and} \quad \mathbb{P}(\mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1) \geq K_o) \geq 0.6,$$

and hence by inclusion-exclusion principle

$$\begin{aligned}&\mathbb{P}(\mathcal{K}_{exit}(\mathbf{u}) \geq K_o \text{ and } \mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1) \geq K_o)\\ &\geq \mathbb{P}(\mathcal{K}_{exit}(\mathbf{u}) \geq K_o) + \mathbb{P}(\mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1) \geq K_o) - 1 \geq 2(0.6) - 1 = 0.2,\end{aligned}$$

which contradicts (15) in Lemma 8.

(ii) Combining the fact that $\mathcal{S}_{K_o}^B(\mathbf{x}^0)$ satisfies the $(q_0, \mathbf{e}_1)$-narrow property (as in Definition 2) where $q_0 = \eta\sigma/4\sqrt{d}$, and Assumption 3 which allows $\nabla F(\mathbf{u}; \boldsymbol{\zeta}^1)$ to satisfy the $\mathbf{e}_1$-disperse property, we have for any $\mathbf{u} \in \mathbb{R}^d$ the following holds:

$$\begin{aligned}\mathbb{P}\left(\mathbf{w}^1(\mathbf{u}) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)\right) &= \mathbb{P}\left(\mathbf{u} - \eta\nabla F(\mathbf{u}; \boldsymbol{\zeta}^1) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)\right)\\ &= \mathbb{P}\left(\nabla F(\mathbf{u}; \boldsymbol{\zeta}^1) \in \eta^{-1}[-\mathcal{S}_{K_o}^B(\mathbf{x}^0) + \mathbf{u}]\right) \leq \frac{1}{4},\end{aligned} \tag{28}$$

where we applied (12) and that $\mathbf{w}^0(\mathbf{u}) = \mathbf{u}$. Thus

$$\begin{aligned}\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) \leq K_o\right) \geq\ & \mathbb{E}\left(\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{w}^1(\mathbf{u})) < K_o \mid \boldsymbol{\mathcal{F}}^1\right); \mathbf{w}^1(\mathbf{u}) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)\right)\\ &+ \mathbb{E}\left(\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{w}^1(\mathbf{u})) < K_o \mid \boldsymbol{\mathcal{F}}^1\right); \mathbf{w}^1(\mathbf{u}) \in [\mathcal{S}_{K_o}^B(\mathbf{x}^0)]^c\right)\\ \geq\ & \mathbb{E}\left(\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{w}^1(\mathbf{u})) < K_o \mid \boldsymbol{\mathcal{F}}^1\right); \mathbf{w}^1(\mathbf{u}) \in [\mathcal{S}_{K_o}^B(\mathbf{x}^0)]^c\right)\\ \geq\ & 0.4\mathbb{P}\left(\mathbf{w}^1(\mathbf{u}) \in [\mathcal{S}_{K_o}^B(\mathbf{x}^0)]^c\right) \geq 0.4\left(\frac{3}{4}\right) = 0.3,\end{aligned}$$

i.e. $\sup_{\mathbf{u}' \in \mathbb{R}^d} \mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}') > K_o\right) \leq 0.7$. Using (28) and Markov's property we conclude for any $N \geq 1$

$$\begin{aligned}\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) > NK_o\right) &= \mathbb{E}\left(\mathbb{P}(\mathcal{K}_{exit}(\mathbf{w}^{(N-1)K_o}(\mathbf{u})) > K_o \mid \boldsymbol{\mathcal{F}}^{K_0}); \mathcal{K}_{exit}(\mathbf{u}) > (N-1)K_o\right)\\ &\leq \sup_{\mathbf{u}' \in \mathcal{B}(\mathbf{u}, B)} \mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}') > K_o\right) \cdot \mathbb{P}(\mathcal{K}_{exit}(\mathbf{u}) > (N-1)K_o)\\ &\leq 0.7 \cdot \mathbb{P}(\mathcal{K}_{exit}(\mathbf{u}) > (N-1)K_o),\end{aligned}$$

which further leads to $\mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) > NK_o\right) \leq 0.7^N$. Letting $N = \lfloor \log(3 \cdot p^{-1}) / \log(0.7^{-1}) \rfloor + 1$ we obtain an exit probability of $\leq p/3$ which completes the proof of Proposition 7.

■

C.1. Proof of Lemma 8

This subsection denotes to the proof of Lemma 8 in the following steps:

(i) Denote for simplicity $\mathbf{w}^k \equiv \mathbf{w}^k(\mathbf{u})$, and $\bar{\mathbf{w}}^k \equiv \mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1)$. Recall from the SGD update rule we have $\mathbf{w}^0 = \mathbf{u}$, and for all $k = 1, 2, \ldots$ for a random index $\boldsymbol{\zeta}^k$ drawn from distribution $\mathcal{D}$,

$$\mathbf{w}^k = \mathbf{w}^{k-1} - \eta \nabla F(\mathbf{w}^{k-1}; \boldsymbol{\zeta}^k),$$

and

$$\bar{\mathbf{w}}^k = \bar{\mathbf{w}}^{k-1} - \eta \nabla F(\bar{\mathbf{w}}^{k-1}; \boldsymbol{\zeta}^k).$$

Recall the definition of $\mathcal{K}_{exit}(\mathbf{u})$ in (13), we let

$$\mathcal{K}_1 := \mathcal{K}_{exit}(\mathbf{u}) \wedge \mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1). \tag{29}$$

For our analysis, we define a coupled $\boldsymbol{\mathcal{F}}^k$-measurable iteration $\mathbf{z}^k$, as follows:

$$\mathbf{z}^k = \begin{cases} \bar{\mathbf{w}}^k - \mathbf{w}^k & \text{on } (k < \mathcal{K}_1) \\ \left(\mathbf{I} - \eta \nabla^2 f(\mathbf{x}^0)\right) \mathbf{z}^{k-1} & \text{on } (k \geq \mathcal{K}_1) \end{cases}, \tag{30}$$

i.e. we couple the difference iteration $\bar{\mathbf{w}}^k - \mathbf{w}^k$ on $(k < \mathcal{K}_1)$, and keep moving the iteration afterwards as if it is the difference iteration of SGD for pure quadratics (we eliminate both the Taylor remainder term and noise term after exiting). Since $\mathbf{w}^0 = \mathbf{u}$, $\bar{\mathbf{w}}^0 = \mathbf{u} + q\mathbf{e}_1$ and $(\mathcal{K}_1 > 0)$ holds, we have $\mathbf{y}^0 = q\mathbf{e}_1$. We *only* want to show for any $\mathbf{u}$ such that $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathcal{B}(\tilde{\mathbf{x}}, B)$,

$$\mathbb{P}(\mathcal{K}_1 > K_o) = \mathbb{P}\left(\mathcal{K}_{exit}(\mathbf{u}) \wedge \mathcal{K}_{exit}(\mathbf{u} + q\mathbf{e}_1) > K_o\right) \leq p. \tag{31}$$

(ii) Letting $\mathbf{H} = \nabla^2 f(\mathbf{x}^0)$, and we first conclude the following lemma to express $\mathbf{z}^k$ defined in (30):

Lemma 13 *We have for all* $k = 1, 2, \ldots$

$$\mathbf{z}^k = (\mathbf{I} - \eta \mathbf{H})\, \mathbf{z}^{k-1} + \eta \mathbf{D}^{k-1} \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k, \tag{32}$$

where

$$\|\mathbf{D}^{k-1}\| \leq \rho \max\left(\|\bar{\mathbf{w}}^{k-1} - \mathbf{x}^0\|, \|\mathbf{w}^{k-1} - \mathbf{x}^0\|\right) \leq \rho B, \tag{33}$$

$\{\boldsymbol{\xi}_d^k\}$ *forms a martingale difference sequence satisfying*

$$\|\boldsymbol{\xi}_d^k\| \leq 2L \|\mathbf{z}^{k-1}\|. \tag{34}$$

Proof [Proof of Lemma 13] By setting $\mathbf{D}^{k-1} = \mathbf{0}^{d \times d}$ and $\boldsymbol{\xi}_d^k = \mathbf{0}$, on event $(k \geq \mathcal{K}_1)$ we can easily see from (30) that all (32), (33) and (34) hold, since their left hands are zero. For its complement $(k < \mathcal{K}_1)$, we have

$$\begin{aligned}
\mathbf{z}^k = \bar{\mathbf{w}}^k - \mathbf{w}^k &= \bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1} - \eta \left(\nabla F(\bar{\mathbf{w}}^{k-1}; \boldsymbol{\zeta}^k) - \nabla F(\mathbf{w}^{k-1}; \boldsymbol{\zeta}^k)\right) \\
&= \mathbf{z}^{k-1} - \eta \left(\nabla f(\bar{\mathbf{w}}^{k-1}) - \nabla f(\mathbf{w}^{k-1})\right) \\
&\quad + \eta \left[\left(\nabla f(\bar{\mathbf{w}}^{k-1}) - \nabla f(\mathbf{w}^{k-1})\right) - \left(\nabla F(\bar{\mathbf{w}}^{k-1}; \boldsymbol{\zeta}^k) - \nabla F(\mathbf{w}^{k-1}; \boldsymbol{\zeta}^k)\right)\right] \\
&= \mathbf{z}^{k-1} - \eta \left[\int_0^1 \nabla^2 f\left(\mathbf{w}^{k-1} + \theta(\bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1})\right) d\theta\right] \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k \\
&\equiv \mathbf{z}^{k-1} - \eta \left(\mathbf{H} - \mathbf{D}^{k-1}\right) \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k,
\end{aligned}$$

where we set the following terms (35) and (36):

$$\mathbf{D}^{k-1} \equiv \nabla^2 f(\mathbf{x}^0) - \int_0^1 \nabla^2 f\left(\mathbf{w}^{k-1} + \theta(\bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1})\right) d\theta, \tag{35}$$

and the noise term $\boldsymbol{\xi}_d^k$ generated at each iteration

$$\boldsymbol{\xi}_d^k \equiv \left(\nabla f(\bar{\mathbf{w}}^{k-1}) - \nabla f(\mathbf{w}^{k-1})\right) - \left(\nabla F(\bar{\mathbf{w}}^{k-1}; \boldsymbol{\zeta}^k) - \nabla F(\mathbf{w}^{k-1}; \boldsymbol{\zeta}^k)\right), \tag{36}$$

proving (32).

It leaves us to prove (33) and (34). From (35), we have

$$\begin{aligned}
\|\mathbf{D}^{k-1}\| &\le \int_0^1 \left\|\nabla^2 f(\mathbf{x}^0) - \nabla^2 f\left(\mathbf{w}^{k-1} + \theta(\bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1})\right)\right\| d\theta \\
&\le \rho \int_0^1 \left\|\theta(\bar{\mathbf{w}}^{k-1} - \mathbf{x}^0) + (1-\theta)(\mathbf{w}^{k-1} - \mathbf{x}^0)\right\| d\theta \\
&\le \rho \max\left(\|\bar{\mathbf{w}}^{k-1} - \mathbf{x}^0\|, \|\mathbf{w}^{k-1} - \mathbf{x}^0\|\right),
\end{aligned}$$

which is bounded by ρB since $\max\left(\|\bar{\mathbf{w}}^{k-1} - \mathbf{x}^0\|, \|\mathbf{w}^{k-1} - \mathbf{x}^0\|\right) \le B$, proving (33).

The $\boldsymbol{\xi}_d^k$ defined in (36) has $\mathbb{E}[\boldsymbol{\xi}_d^k \mid \boldsymbol{\mathcal{F}}^{k-1}] = 0$ forming a Martingale Difference Sequence, and from Lipschitz continuity of the objective function, we have

$$\begin{aligned}
\|\boldsymbol{\xi}_d^k\| &\le \left\|\nabla f(\bar{\mathbf{w}}^{k-1}) - \nabla f(\mathbf{w}^{k-1})\right\| + \left\|\nabla F(\bar{\mathbf{w}}^{k-1}; \boldsymbol{\zeta}^k) - \nabla F(\mathbf{w}^{k-1}; \boldsymbol{\zeta}^k)\right\| \\
&\le L\left\|\bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1}\right\| + L\left\|\bar{\mathbf{w}}^{k-1} - \mathbf{w}^{k-1}\right\| = 2L\|\mathbf{z}^{k-1}\|.
\end{aligned}$$

This completes the proof of (34), and hence the lemma.

■

(iii) We observe from (32) that if $\nabla^2 f(\mathbf{z})$ does *not* rotate in the sense that each pair of Hessian matrices $\nabla^2 f(\mathbf{w}_1)$ and $\nabla^2 f(\mathbf{w}_2)$ can be spectrally decomposed via the same orthogonal matrix, one can analyze the iteration coordinate-wisely. Here, the rotation effect of Hessian matrix *cannot* be ignored. Hence, we analyze the difference iteration $\mathbf{z}^k$ in two aspects: (i) $\mathbf{z}^k$ has a *rotation effect* after standardization, and (ii) its norm $\|\mathbf{z}^k\|$ has an *expansion effect*.

To decouple these two effect, we define a rescaled iteration as follows. Let δ_m denote the negated least eigenvalue $\lambda_{\min}(\nabla^2 f(\mathbf{x}^0))$ of Hessian so $\delta_m \ge \delta_2$. Let for each $k = 0, 1, \ldots$

$$\boldsymbol{\psi}^k \equiv q^{-1}(1+\eta\delta_m)^{-k}\mathbf{z}^k. \tag{37}$$

We state the following lemma for the update rule of $\boldsymbol{\psi}^k$.

Lemma 14 *Let* $\hat{\mathbf{D}}^k \equiv (1+\eta\delta_m)^{-1}\mathbf{D}^k$, *and* $\boldsymbol{\zeta}_d^k \equiv q^{-1}(1+\eta\delta_m)^{-k}\boldsymbol{\xi}_d^k$. *We have* $\boldsymbol{\psi}^0 = \mathbf{e}_1$ *and*

$$\boldsymbol{\psi}^k = \frac{(\mathbf{I} - \eta\mathbf{H})}{1+\eta\delta_m}\boldsymbol{\psi}^{k-1} + \eta\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \eta\boldsymbol{\zeta}_d^k, \tag{38}$$

where

$$\|\hat{\mathbf{D}}^{k-1}\| \le \rho B, \tag{39}$$

and the rescaled noise iteration $\boldsymbol{\zeta}_d^k$ has

$$\|\boldsymbol{\zeta}_d^k\| \le 2L\|\boldsymbol{\psi}^{k-1}\|, \quad k \ge 1. \tag{40}$$

Then with the step size set in (8)[9]*, we have on the event $\boldsymbol{\mathcal{H}}_o$ ((49) happens), the norm of $\boldsymbol{\psi}^k$ satisfies*

$$\|\boldsymbol{\psi}^k\|^2 \le 4, \tag{41}$$

and for the projection of $\boldsymbol{\psi}^k$ onto the first coordinate,

$$\mathbf{e}_1^\top \boldsymbol{\psi}^k > \frac{1}{2}. \tag{42}$$

Proof [Proof of Lemma 14] We have from the definition of $\boldsymbol{\zeta}_d^k$

$$\begin{aligned}\|\boldsymbol{\zeta}_d^k\| &\le q^{-1}(1+\eta\delta_m)^{-k}\|\boldsymbol{\xi}_d^k\| \\ &\le 2Lq^{-1}\frac{(1+\eta\delta_m)^{-(k-1)}}{1+\eta\delta_m}\|\mathbf{z}^{k-1}\| \le 2L\|\boldsymbol{\psi}^{k-1}\|,\end{aligned}$$

establishing (40), and hence

$$\begin{aligned}\boldsymbol{\psi}^k &= q^{-1}(1+\eta\delta_m)^{-k}\mathbf{z}^k \\ &= \frac{(\mathbf{I}-\eta\mathbf{H})}{1+\eta\delta_m}q^{-1}(1+\eta\delta_m)^{-(k-1)}\mathbf{z}^{k-1} \\ &\qquad + \eta\frac{\mathbf{D}^{k-1}}{1+\eta\delta_m}q^{-1}(1+\eta\delta_m)^{-(k-1)}\mathbf{z}^{k-1} + \eta q^{-1}(1+\eta\delta_m)^{-k}\boldsymbol{\xi}_d^k \\ &= \frac{(\mathbf{I}-\eta\mathbf{H})}{1+\eta\delta_m}\boldsymbol{\psi}^{k-1} + \eta\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \eta\boldsymbol{\zeta}_d^k,\end{aligned}$$

proving (38) and (39).

To handle the term involving the $\boldsymbol{\zeta}_d^k$ terms on the right hands of (42) and (41), we first set

$$\hat{\boldsymbol{\psi}}^{k-1} = \frac{[\mathbf{I}-\eta\mathbf{H}]}{1+\eta\delta_m}\boldsymbol{\psi}^{k-1}. \tag{43}$$

Since $\eta L \le 1$ we simply have $[\mathbf{I}-\eta\mathbf{H}]$ is symmetric and has all eigenvalues in $[0, 1+\eta\delta_m]$, so $\|\mathbf{I}-\eta\mathbf{H}\| \le 1+\eta\delta_m$. This implies $\|\hat{\boldsymbol{\psi}}^{k-1}\| \le \|\boldsymbol{\psi}^{k-1}\|$.

On the other hand, for all $k \ge 1$, we have

$$\mathbb{E}\left[\hat{\boldsymbol{\psi}}^{k-1\top}\boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2} \mid \boldsymbol{\mathcal{F}}^{k-1}\right] \overset{a}{=} \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2} \cdot \mathbb{E}[\hat{\boldsymbol{\psi}}^{k-1\top}\boldsymbol{\zeta}_d^k \mid \boldsymbol{\mathcal{F}}^{k-1}] = 0, \tag{44}$$

and

$$\mathbb{E}\left[\left|\hat{\boldsymbol{\psi}}^{k-1\top}\boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right|^2 \mid \boldsymbol{\mathcal{F}}^{k-1}\right] \overset{a\ \&\ (40)}{=} \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2} \cdot 2L\|\boldsymbol{\psi}^{k-1}\|^2 \le 8L, \tag{45}$$

[9] We actually only need $\eta \le \tilde{\mathcal{O}}(\epsilon^{0.5})$ to obtain Lemma 14.

where $\mathcal{I}$ denotes the indicator function, $\overset{a}{=}$ uses $\boldsymbol{\psi}^{k-1}$ and $\hat{\boldsymbol{\psi}}^{k-1}$ are measurable on $\boldsymbol{\mathcal{F}}^{k-1}$. By the standard Azuma's inequality, with probability $1 - 0.1/(2K_0)$, for any l from 1 to K_0,

$$\left|\sum_{k=1}^{l} \hat{\boldsymbol{\psi}}^{k-1\,\top} \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right| \le 4\sqrt{Ll\log(40K_0)} \le 4\sqrt{LK_0\log(40K_0)} \overset{(8)}{\le} \frac{1}{\eta}. \tag{46}$$

Analogously, we also have

$$\mathbb{E}\left[\mathbf{e}_1^\top \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2} \mid \boldsymbol{\mathcal{F}}^{k-1}\right] = \mathbf{0}, \quad \mathbb{E}\left[\left|\mathbf{e}_1^\top \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right|^2 \mid \boldsymbol{\mathcal{F}}^{k-1}\right] \le 4L. \tag{47}$$

Thus with standard Azuma's inequality,

$$\left|\sum_{k=1}^{l} \mathbf{e}_1^\top \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right| \le \sqrt{8Ll\log(40k_0)} \le \sqrt{8LK_0\log(12K_0/p)} \overset{(8)}{\le} \frac{1}{4\eta} \tag{48}$$

happens with probability at least $1 - 0.1/(2K_0)$.

So by union bound, there exists a high-probability event $\boldsymbol{\mathcal{H}}_o$ happening with probability at least 0.9 such that the following inequalities hold for each $l = 1, 2, \ldots, K_0$,

$$\left|\sum_{k=1}^{l} \hat{\boldsymbol{\psi}}^{k-1\,\top} \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right| \le \frac{1}{\eta}, \quad \left|\sum_{k=1}^{l} \mathbf{e}_1^\top \boldsymbol{\zeta}_d^k \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{k-1}\|\le 2}\right| \le \frac{1}{4\eta}. \tag{49}$$

On the other hand, we have from (38) and (43) that for all $k \ge 1$,

$$\begin{aligned}
\|\boldsymbol{\psi}^k\|^2 &= \left\|\frac{[\mathbf{I} - \eta\mathbf{H}]}{1 + \eta\delta_m}\boldsymbol{\psi}^{k-1} + \eta\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \eta\boldsymbol{\zeta}_d^k\right\|^2 \\
&= \|\hat{\boldsymbol{\psi}}^{k-1}\|^2 + 2\eta\hat{\boldsymbol{\psi}}^{k-1\,\top}\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \eta^2\left\|\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \boldsymbol{\zeta}_d^k\right\|^2 + 2\eta\hat{\boldsymbol{\psi}}^{k-1\,\top}\boldsymbol{\zeta}_d^k \\
&\le \|\boldsymbol{\psi}^{k-1}\|^2 + Q_{1,k} + Q_{2,k} + Q_{3,k}
\end{aligned}$$

Hence from (39),

$$Q_{1,k} = 2\eta\hat{\boldsymbol{\psi}}^{k-1\,\top}\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} \le 2\eta \cdot \rho B\|\boldsymbol{\psi}^{k-1}\|^2$$

and

$$\begin{aligned}
Q_{2,k} &= \eta^2\left\|\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1} + \boldsymbol{\zeta}_d^{k-1}\right\|^2 \\
&\le 2\eta^2\|\hat{\mathbf{D}}^{k-1}\boldsymbol{\psi}^{k-1}\|^2 + 2\eta^2\|\boldsymbol{\zeta}_d^{k-1}\|^2 \\
&\le 2\eta^2 \cdot \rho^2 B^2\|\boldsymbol{\psi}^{k-1}\|^2 + 8\eta^2 L^2\|\boldsymbol{\psi}^{k-1}\|^2 \\
&\le 16\eta^2 \cdot L^2\|\boldsymbol{\psi}^{k-1}\|^2,
\end{aligned}$$

and

$$Q_{3,k} = 2\eta\hat{\boldsymbol{\psi}}^{k-1\,\top}\boldsymbol{\zeta}_d^{k-1}.$$

Under the event $\boldsymbol{\mathcal{H}}_0$ happens, by induction, when $k = 0$, $\|\boldsymbol{\psi}^0\| = \|\mathbf{e}_1\| \le 2$, suppose $\|\boldsymbol{\psi}^l\| \le 2$ holds for all $l = 0$ to $k-1$, we have for the step k,

$$\begin{aligned}
\|\boldsymbol{\psi}^k\|^2 &\le \|\boldsymbol{\psi}^0\|^2 + \sum_{s=1}^{k} Q_{1,s} + \sum_{s=1}^{k} Q_{2,s} + \sum_{s=1}^{k} Q_{3,s} \\
&\le 1 + 2\eta \sum_{s=1}^{k} \rho B \|\boldsymbol{\psi}^{s-1}\|^2 + 16\eta^2 \cdot L^2 \sum_{s=1}^{k} \|\boldsymbol{\psi}^s\|^2 + 2\eta \sum_{s=1}^{k} \hat{\boldsymbol{\psi}}^{s-1\,\top} \boldsymbol{\zeta}_d^s \\
&\le 1 + 2\rho B \cdot 4 \cdot \eta k + 16\eta^2 \cdot L^2 \cdot 4 \cdot k + 2\eta \sum_{s=1}^{k} \hat{\boldsymbol{\psi}}^{s-1\,\top} \boldsymbol{\zeta}_d^s \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{s-1}\| \le 2} \\
&\overset{a}{\le} 1 + 16\rho B \cdot \eta k + 2\eta \sum_{s=1}^{k} \hat{\boldsymbol{\psi}}^{s-1\,\top} \boldsymbol{\zeta}_d^s \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{s-1}\| \le 2} \le 1 + 1 + 2 = 4,
\end{aligned}$$

where $\overset{a}{\le}$ uses $\eta \le \frac{\rho B}{8L^2}$ (because (8) and $B \le \frac{1}{L}$). This conclude the proof of (41). For $\mathbf{e}_1^\top \boldsymbol{\psi}^k$, we have

$$\begin{aligned}
\mathbf{e}_1^\top \boldsymbol{\psi}^k &= \mathbf{e}_1^\top \boldsymbol{\psi}^0 + \sum_{s=0}^{k-1} \eta \mathbf{e}_1^\top \hat{\mathbf{D}}_s \psi^1 + \sum_{s=0}^{k-1} \eta \mathbf{e}_1^\top \boldsymbol{\zeta}_d^s \\
&\ge 1 - \eta \sum_{s=0}^{k-1} 2\rho B \cdot \|\boldsymbol{\psi}^{s-1}\| + \eta \sum_{s=1}^{k} \mathbf{e}_1^\top \boldsymbol{\zeta}_d^s \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{s-1}\| \le 2} \\
&\ge 1 - \eta \cdot k \cdot 2\rho B \cdot 2 + \eta \sum_{s=0}^{k-1} \mathbf{e}_1^\top \boldsymbol{\zeta}_d^s \cdot \mathcal{I}_{\|\boldsymbol{\psi}^{s-1}\| \le 2} \ge 1 - \frac{1}{8} - \frac{2}{8} > \frac{1}{2},
\end{aligned}$$

concluding (42), and hence the lemma. ■

(iv) Now, we have all the ingredients necessary to prove our final lemma.

Proof [Proof of Lemma 8] Recall that the deterministic time K_o was defined in (27), we have on the event $(\mathcal{K}_1 > K_o)$ that $\mathbf{z}^{K_o} = \bar{\mathbf{w}}^{K_o} - \mathbf{w}^{K_o}$ and hence $\|\mathbf{z}^{K_o}\| \le \|\bar{\mathbf{w}}^{K_o}\| + \|\mathbf{w}^{K_o}\| \le 2B$, which concludes

$$(\mathcal{K}_1 > K_o) \subseteq (\|\mathbf{z}^{K_o}\| \le 2B). \tag{50}$$

In the mean time, from (30) we know that on the event $(\mathcal{K}_1 > K_o)$, $\mathbf{z}^k = \bar{\mathbf{w}}^k - \mathbf{w}^k$ for all $k \le K_o$, from (42)

$$\mathbf{e}_1^\top \boldsymbol{\psi}^{K_o} > \frac{1}{2}.$$

So on the event $(\mathcal{K}_1 > K_o) \cap \boldsymbol{\mathcal{H}}_0$

$$\begin{aligned}
\|\mathbf{z}^{K_o}\| &= q\,(1 + \eta\delta_m)^{K_o} \|\boldsymbol{\psi}^{K_o}\| \ge q_0\,(1 + \eta(\delta_2))^{K_o} |\mathbf{e}_1^\top \boldsymbol{\psi}^{K_o}| \\
&> q_0 \cdot \frac{6B}{q_0} \cdot \frac{1}{2} = 3B,
\end{aligned}$$

so

$$(\mathcal{K}_1 > K_o) \cap \boldsymbol{\mathcal{H}}_0 \subseteq (\|\mathbf{z}^{K_o}\| > 3B) \tag{51}$$

Combining (50), (51) and the fact that $B > 0$ gives

$$(\mathcal{K}_1 > K_o) \cap \boldsymbol{\mathcal{H}}_0 \subseteq (\|\mathbf{z}^{K_o}\| \leq 2B) \cap \boldsymbol{\mathcal{H}}_0 \cap (\|\mathbf{z}^{K_o}\| > 3B) = \varnothing,$$

and hence $(\mathcal{K}_1 > K_o) \subseteq \boldsymbol{\mathcal{H}}_0^c$ which leads to

$$\mathbb{P}(\mathcal{K}_1 > K_o) \leq \mathbb{P}(\boldsymbol{\mathcal{H}}_0^c) \leq 0.1,$$

proving (31). Hence (15) and Lemma 8 hold.

■

Appendix D. Deferred Proofs of Part II: Faster Descent

D.1. Definition and Preliminary

In Part II, we still use $\mathbf{H}$ to denote $\nabla^2 f(\mathbf{x}^0)$ and let

$$\boldsymbol{\xi}^{k+1} = \nabla F(\mathbf{x}^k, \boldsymbol{\zeta}^{k+1}) - \nabla f(\mathbf{x}^k), \quad k \geq 0.$$

Recall the definition of $\mathcal{S}$, $\boldsymbol{\mathcal{P}}_{\mathcal{S}}$, $\mathcal{S}\perp$, and $\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}$ in Appendix A.2. Let $\mathbf{u}^k = \boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x}^k - \mathbf{x}^0\right)$, and $\mathbf{v}^k = \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x}^k - \mathbf{x}^0\right)$. We can decompose the update equation of SGD as:

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \eta\boldsymbol{\mathcal{P}}_{\mathcal{S}}\nabla f\left(\mathbf{x}^k\right) - \eta\boldsymbol{\mathcal{P}}_{\mathcal{S}}\boldsymbol{\xi}^{k+1}. \tag{52}$$

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \eta\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\nabla f\left(\mathbf{x}^k\right) - \eta\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\boldsymbol{\xi}^{k+1}, \tag{53}$$

with $k \geq 0$. And $\mathbf{u}^0 = \mathbf{0}$, $\mathbf{v}^0 = \mathbf{0}$. From the definition of $g(\mathbf{x})$ in Appendix A.2, we have

$$\begin{aligned}
g(\mathbf{x}) &= \left[\nabla f(\mathbf{x}^0)\right]^\top \left(\mathbf{x} - \mathbf{x}^0\right) + \frac{1}{2}\left[\mathbf{x} - \mathbf{x}^0\right]^\top \mathbf{H}\left[\mathbf{x} - \mathbf{x}^0\right] \qquad (54)\\
&= \left[\nabla f(\mathbf{x}^0)\right]^\top \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right) + \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right] + \frac{1}{2}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right)\right]^\top \mathbf{H}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right)\right]\\
&\quad + \frac{1}{2}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right]^\top \mathbf{H}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right]\\
&= \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\nabla f(\mathbf{x}^0)\right]^\top \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right)\right] + \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\nabla f(\mathbf{x}^0)\right]^\top \left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right]\\
&\quad + \frac{1}{2}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right)\right]^\top \mathbf{H}_{\mathcal{S}}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}}\left(\mathbf{x} - \mathbf{x}^0\right)\right] + \frac{1}{2}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right]^\top \mathbf{H}_{\mathcal{S}\perp}\left[\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\mathbf{x} - \mathbf{x}^0\right)\right],
\end{aligned}$$

where in the last equality we use $\boldsymbol{\mathcal{P}}_{\mathcal{S}}^2 = \boldsymbol{\mathcal{P}}_{\mathcal{S}}$ and $\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}^2 = \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}$, because $\boldsymbol{\mathcal{P}}_{\mathcal{S}}$ and $\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}$ are projection matrices. Thus if $\mathbf{u} = \boldsymbol{\mathcal{P}}_{\mathcal{S}}(\mathbf{x} - \mathbf{x}^0)$ and $\mathbf{v} = \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}(\mathbf{x} - \mathbf{x}^0)$, we have

$$g(\mathbf{x}) = g_{\mathcal{S}}(\mathbf{u}) + g_{\mathcal{S}\perp}(\mathbf{v}). \tag{55}$$

For clarify, we denote $\nabla_{\mathbf{u}} f(\mathbf{x}^k) = \boldsymbol{\mathcal{P}}_{\mathcal{S}}\nabla f(\mathbf{x}^k)$, and $\nabla_{\mathbf{v}} f(\mathbf{x}^k) = \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\nabla f(\mathbf{x}^k)$, respectively. Similarly, let $\boldsymbol{\xi}_{\mathbf{u}}^k = \boldsymbol{\mathcal{P}}_{\mathcal{S}}\boldsymbol{\xi}^k$, and $\boldsymbol{\xi}_{\mathbf{v}}^k = \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\boldsymbol{\xi}^k$. In the following, we denote $\mathscr{K} = \mathscr{K}_0 \wedge K_0$ which is also a stopping time. The Lemma below is basic to obtain our result.

Lemma 15 *Given* $\mathbf{x}^0$*, for any* $\mathbf{x}$*, if* $\left\|\mathbf{x}-\mathbf{x}^0\right\| \le B$*, then*

$$\|\nabla f(\mathbf{x}) - \nabla g(\mathbf{x})\| \le \rho B^2/2. \tag{56}$$

For any symmetric matrix $\mathbf{A}$*, with* $0 < a \le \frac{1}{\|\mathbf{A}\|_2}$*, for any* $i = 0, 1, \dots,$ *and* $j = 0, 1, \dots,$ *we have*

$$\left\|(\mathbf{I} - a\mathbf{A})^i \mathbf{A} (\mathbf{I} - a\mathbf{A})^j\right\|_2 \le \frac{1}{a(i+j+1)}. \tag{57}$$

Proof For (56), we have

$$\begin{aligned}
& \|\nabla f(\mathbf{x}) - \nabla g(\mathbf{x})\| \\
= & \left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^0) - \mathbf{H}(\mathbf{x} - \mathbf{x}^0)\right\| \\
= & \left\|\left(\int_0^1 \nabla^2 f(\mathbf{x}^0 + \theta(\mathbf{x} - \mathbf{x}^0)) d_\theta - \mathbf{H}\right)(\mathbf{x} - \mathbf{x}^0)\right\| \\
\overset{a}{\le} & \left\|\left(\int_0^1 \rho\theta \left\|\mathbf{x} - \mathbf{x}^0\right\| d_\theta\right)\right\| \left\|\mathbf{x} - \mathbf{x}^0\right\| \\
\le & \rho B^2/2
\end{aligned} \tag{58}$$

where in $\overset{a}{\le}$, we use (4) that $f(\mathbf{x})$ has ρ-Lipschitz continuous Hessian.

(57) is from Jin et al. (2017). To prove it, suppose the eigenvalue of $\{\mathbf{A}\}$ is $\{\lambda_l\}$, thus the eigenvalue of $(\mathbf{I} - a\mathbf{A})^i \mathbf{H} (\mathbf{I} - a\mathbf{A})^j$ is $\{\lambda_l(1 - a\lambda_l)^{i+j}\}$. For the function of $\lambda(1 - a\lambda)^{i+j}$, we can compute out its derivative as $(1 - a\lambda)^{i+j} - (i + j)a\lambda(1 - a\lambda)^{i+j-1}$. Then with simple analysis, we can find that the maximal point is obtained only at $\frac{1}{(1+i+j)a}$. If $i = 0$ and $j = 0$, (57) clearly holds. Otherwise, we have

$$\left\|(\mathbf{I} - a\mathbf{A})^i \mathbf{A} (\mathbf{I} - a\mathbf{A})^j\right\|_2 \le \frac{1}{(1+j+i)a}\left(1 - \frac{1}{1+j+i}\right)^{j+i} \le \frac{1}{(1+j+i)a}. \tag{59}$$

■

D.2. Analysis on Quadratic Approximation

As have been introduced before, our main technique to obtain a faster $\tilde{\mathcal{O}}(\epsilon^{-3.5})$ convergence rate is by separately analyzing the two quadratic approximations: $g_{\mathcal{S}}(\cdot)$ and $g_{\mathcal{S}\perp}(\cdot)$. We will show in this section that the noise effect can be upper bounded by $\tilde{\mathcal{O}}(\epsilon^{1.5})$ instead of $\mathcal{O}(\epsilon)$ via our tool.

We first summarize our result for $g_{\mathcal{S}}(\cdot)$ in the following lemma:

Lemma 16 *Set hyper-parameters in* (8) *for Algorithm* 2*. With probability at least* $1 - p/4$*, we have*

$$\begin{aligned}
g_{\mathcal{S}}(\mathbf{u}^{\mathscr{K}}) - g_{\mathcal{S}}(\mathbf{u}^0) \quad \le & \quad -\frac{25\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{32} + 4\eta\sigma^2\left(\log(K_0) + 3\right)\log(48K_0/p) + \eta\rho^2 B^4 K_0 \\
= & \quad -\frac{25\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{32} + \tilde{\mathcal{O}}\left(\epsilon^{1.5}\right).
\end{aligned} \tag{60}$$

Proofs of Lemma 16
Our novel technique to analyze $g_{\mathcal{S}}(\cdot)$ is by first considering an auxiliary Gradient Descent trajectory, which performs update as:

$$\mathbf{y}^{k+1} = \mathbf{y}^{k} - \eta \nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right), \quad k \geq 0. \tag{61}$$

and $\mathbf{y}^0 = \mathbf{u}^0$. $\mathbf{y}^k$ preforms Gradient Decent on $g_{\mathcal{S}}(\cdot)$, which is deterministic given $\mathbf{x}^0$. We study the property of $\mathbf{y}^{\mathscr{K}}$ and obtain the following standard results:

- Because $g_{\mathcal{S}}(\cdot)$ has L-Lipschitz continuous gradient ($\|\mathbf{H}_{\mathcal{S}}\|_2 \leq L$), we have

$$\begin{aligned} & g_{\mathcal{S}}\left(\mathbf{y}^{k+1}\right) \\ \leq\ & g_{\mathcal{S}}\left(\mathbf{y}^{k}\right) + \left\langle \nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right), \mathbf{y}^{k+1} - \mathbf{y}^{k} \right\rangle + \frac{L}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^{k}\right\|^2 \\ \overset{(61)}{=}\ & g_{\mathcal{S}}\left(\mathbf{y}^{k}\right) - \eta\left(1 - \frac{L\eta}{2}\right)\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right)\right\|^2. \end{aligned} \tag{62}$$

- By telescoping (62) from 0 to $\mathscr{K} - 1$, we have

$$\begin{aligned} g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right) \quad & \leq \quad g_{\mathcal{S}}\left(\mathbf{y}^{0}\right) - \eta\left(1 - \frac{L\eta}{2}\right)\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right)\right\|^2 \\ & \overset{L\eta \leq \frac{1}{16}}{\leq} \quad g_{\mathcal{S}}\left(\mathbf{y}^{0}\right) - \frac{31\eta}{32}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right)\right\|^2. \end{aligned} \tag{63}$$

To obtain Lemma 16, we bound the difference between $\mathbf{u}^{\mathscr{K}}$ and $\mathbf{y}^{\mathscr{K}}$. Define

$$\mathbf{z}^k := \mathbf{u}^k - \mathbf{y}^k.$$

The remaining is to conclude the properties of $\mathbf{z}^{\mathscr{K}}$, stated as follows:

Lemma 17 *With probability at least $1 - p/6$, we have*

$$\left\|\mathbf{z}^{\mathscr{K}}\right\| \leq \frac{3B}{32} \asymp \epsilon^{0.5}. \tag{64}$$

and

$$\left(\mathbf{z}^{\mathscr{K}}\right)^{\top}\mathbf{H}_{\mathcal{S}}\left(\mathbf{z}^{\mathscr{K}}\right) \leq 8\sigma^2\eta\left(\log(K_0) + 1\right)\log(48K_0/p) + \eta\rho^2 B^4 K_0 \asymp \epsilon^{1.5}. \tag{65}$$

Proof [Proofs of Lemma 17] With $\mathbf{z}^k = \mathbf{u}^k - \mathbf{y}^k$ being the difference iteration, we have

$$\begin{aligned} \mathbf{z}^{k+1} \quad & = \quad \mathbf{z}^{k} - \eta\left(\nabla g_{\mathcal{S}}\left(\mathbf{u}^{k}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{y}^{k}\right)\right) - \eta\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^{k}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^{k}\right)\right) - \eta\boldsymbol{\xi}_{\mathbf{u}}^{k+1} \\ & = \quad (\mathbf{I} - \eta\mathbf{H}_{\mathcal{S}})\mathbf{z}^{k} - \eta\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^{k}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^{k}\right)\right) - \eta\boldsymbol{\xi}_{\mathbf{u}}^{k+1}, \quad k \geq 0. \end{aligned} \tag{66}$$

And $\mathbf{z}^0 = \mathbf{0}$. Thus we can obtain the general solution of (66) as

$$\mathbf{z}^k = -\sum_{j=1}^{k} \eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{k-j}\boldsymbol{\xi}_{\mathbf{u}}^j - \eta\sum_{j=0}^{k-1}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{k-1-j}\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right), \quad k \geq 0. \tag{67}$$

Setting $k = \mathscr{K}$, by triangle inequality, we have

$$\left\|\mathbf{z}^{\mathscr{K}}\right\| \leq \left\|\sum_{j=1}^{\mathscr{K}} \eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j}\boldsymbol{\xi}_{\mathbf{u}}^j\right\| + \left\|\eta\sum_{j=0}^{\mathscr{K}-1}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j}\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)\right\|. \tag{68}$$

We separately bound the two terms in the right hand sides of (68). For the first term, for any fixed l from 1 to K_0, and any j from 1 to l, we have

$$\mathbb{E}\left[\eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j}\boldsymbol{\xi}_{\mathbf{u}}^j \mid \boldsymbol{\mathcal{F}}^{j-1}\right] \overset{a}{=} \mathbf{0}, \qquad \left\|\eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j}\boldsymbol{\xi}_{\mathbf{u}}^j\right\| \overset{b}{\leq} \eta\sigma, \tag{69}$$

where $\overset{a}{=}$ uses that $\|\boldsymbol{\xi}_{\mathbf{u}}^j\| = \|\boldsymbol{\mathcal{P}}_{\mathcal{S}}\boldsymbol{\xi}^j\|$, $\overset{b}{\leq}$ further uses $\|\boldsymbol{\xi}_{\mathbf{u}}^j\| \leq \|\boldsymbol{\xi}^j\| \leq \sigma$, because $\boldsymbol{\mathcal{P}}$ is projection matrix, and the the bounded noise assumption in (5) and $\|(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j}\|_2 \leq 1$ for all j from 1 to l. Thus by the Vector-Martingale Concentration Inequality in Theorem 11, we have with probability $1-p/(12K_0)$,

$$\left\|\sum_{j=1}^{l} \eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j}\boldsymbol{\xi}_{\mathbf{u}}^j\right\| \leq 2\eta\sigma\sqrt{(l)\log(48K_0/p)} \leq 2\eta\sigma\sqrt{K_0\log(48K_0/p)} \overset{(8)}{\leq} \frac{B}{16}. \tag{70}$$

By union bound, with probability at least $1-p/12$, (70) holds for all l from 1 to K_0. Because $1 \leq \mathscr{K} \leq K_0$, with probability at least $1-p/12$,

$$\left\|\sum_{j=1}^{\mathscr{K}} \eta(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j}\boldsymbol{\xi}_{\mathbf{u}}^j\right\| \leq \frac{B}{16}. \tag{71}$$

For the second term in the right hand side of (68), we have

$$\begin{aligned}
\left\|\eta\sum_{j=0}^{\mathscr{K}-1}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j}\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)\right\| &\overset{a}{\leq} \eta\sum_{j=0}^{\mathscr{K}-1}\left\|\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)\right\| \\
&\overset{b}{\leq} \eta\sum_{j=0}^{\mathscr{K}-1}\left\|\nabla f\left(\mathbf{x}^j\right) - \nabla g\left(\mathbf{x}^j\right)\right\| \\
&\overset{(56)}{\leq} \frac{\rho\eta B^2 K_0}{2} \overset{(8)}{\leq} \frac{B}{32},
\end{aligned} \tag{72}$$

where in $\overset{a}{\leq}$, we use triangle inequality, and $\|\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}}\|_2^{\mathscr{K}-1-j} \leq 1$ with j from 0 to $\mathscr{K}-1$; $\overset{b}{\leq}$ uses $\|\boldsymbol{\mathcal{P}}_{\mathcal{S}}(\nabla f(\mathbf{x}) - \nabla g(\mathbf{x}))\| \leq \|\nabla f(\mathbf{x}) - \nabla g(\mathbf{x})\|$ becuase $\boldsymbol{\mathcal{P}}_{\mathcal{S}}$ is projected matrix. Substituting (71) and (72) into (68), we obtain (64).

To prove (65), using the fact that $(\mathbf{a}+\mathbf{b})^\top \mathbf{A}(\mathbf{a}+\mathbf{b}) \le 2\mathbf{a}^\top \mathbf{A}\mathbf{a} + 2\mathbf{b}^\top \mathbf{A}\mathbf{b}$ holds for any symmetry positive definite matrix $\mathbf{A}$, we have

$$
\begin{aligned}
&\left(\mathbf{z}^{\mathscr{K}}\right)^\top \mathbf{H}_{\mathcal{S}} \left(\mathbf{z}^{\mathscr{K}}\right) \\
\le\ & 2\eta^2 \left(\sum_{j=1}^{\mathscr{K}} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)^\top \mathbf{H}_{\mathcal{S}} \left(\sum_{j=1}^{\mathscr{K}} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right) \\
& +2\eta^2 \left(\sum_{j=0}^{\mathscr{K}-1} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j} \left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)\right)^\top \mathbf{H}_{\mathcal{S}} \left(\sum_{j=0}^{\mathscr{K}-1} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j} \left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)\right) \\
=\ & 2\left\|\eta\left(\sum_{j=1}^{\mathscr{K}} \mathbf{H}_{\mathcal{S}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)\right\|^2 \\
& +2\eta^2 \sum_{j=0}^{\mathscr{K}-1} \sum_{l=0}^{\mathscr{K}-1} \left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right)^\top (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j} \mathbf{H}_{\mathcal{S}} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-l} \left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^j\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^j\right)\right) \\
\overset{(56)}{\le}\ & 2\left\|\eta\left(\sum_{j=1}^{\mathscr{K}} \mathbf{H}_{\mathcal{S}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)\right\|^2 + 2\eta^2 \frac{\rho^2 B^4}{4} \sum_{j=0}^{\mathscr{K}-1} \sum_{l=0}^{\mathscr{K}-1} \left\|(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j} \mathbf{H}_{\mathcal{S}} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-l}\right\|_2.
\end{aligned}
\tag{73}
$$

For the first term in the right hand side of (73), for any fixed l from 1 to K_0, and any j from 1 to l, we have

$$
\mathbb{E}\left[\eta\left(\mathbf{H}_{\mathcal{S}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right) \mid \boldsymbol{\mathcal{F}}^{j-1}\right] = \mathbf{0},
$$

and

$$
\begin{aligned}
&\left\|\eta\left(\mathbf{H}_{\mathcal{S}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)\right\|^2 \\
\le\ & \eta^2 \|\boldsymbol{\xi}_{\mathbf{u}}^j\| \left\|(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j} \mathbf{H}_{\mathcal{S}} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j}\right\|_2 \|\boldsymbol{\xi}_{\mathbf{u}}^j\| \overset{(57)\ \&\ \|\boldsymbol{\xi}_{\mathbf{u}}^j\|\le\sigma}{\le} \frac{\eta\sigma^2}{1+2(l-j)},
\end{aligned}
\tag{74}
$$

by the Vector-Martingale Concentration Inequality in §11, we have with probability $1-p/(12K_0)$

$$
\begin{aligned}
&\left\|\eta\left(\sum_{j=1}^{l} \boldsymbol{\mathcal{H}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)\right\|^2 \le 4\eta\sigma^2 \log(48K_0/p) \sum_{j=1}^{l} \frac{1}{1+2(l-j)} \\
\le\ & 4\sigma^2\eta \log(48K_0/p) \sum_{j=0}^{K_0-1} \frac{1}{1+j} \le 4\sigma^2\eta\left(\log(K_0)+1\right)\log(48K_0/p).
\end{aligned}
\tag{75}
$$

By union bound, with probability at least $1-p/12$, (75) holds for all l from 1 to K_0. Because $1 \le \mathscr{K} \le K_0$, with probability at least $1-p/12$,

$$
\left\|\eta\left(\sum_{j=1}^{\mathscr{K}} \boldsymbol{\mathcal{H}}^{1/2} (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j\right)\right\|^2 \le 4\sigma^2\eta\left(\log(K_0)+1\right)\log(48K_0/p).
\tag{76}
$$

For the second term in the right hand side of (73), we have

$$\begin{aligned}
& \eta^2 \frac{\rho^2 B^4}{4} \sum_{j=0}^{\mathscr{K}-1} \sum_{l=0}^{\mathscr{K}-1} \left\| (\mathbf{I}-\eta \mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-j} \mathbf{H} (\mathbf{I}-\eta \mathbf{H}_{\mathcal{S}})^{\mathscr{K}-1-l} \right\|_2 \\
\overset{(56)}{\leq} \quad & \eta \frac{\rho^2 B^4}{4} \sum_{j=0}^{\mathscr{K}-1} \sum_{l=0}^{\mathscr{K}-1} \frac{1}{1+(\mathscr{K}-1-j)+(\mathscr{K}-1-l)} \overset{\mathscr{K}\leq K_0}{\leq} \eta \frac{\rho^2 B^4}{4} \sum_{j=0}^{K_0-1} \sum_{l=0}^{K_0-1} \frac{1}{1+j+l} \\
= \quad & \eta \frac{\rho^2 B^4}{4} \sum_{j=0}^{2(K_0-1)} \frac{\min(1+j, 2K_0-1-j)}{1+j} \leq \frac{\eta \rho^2 B^4 K_0}{2}. \qquad (77)
\end{aligned}$$

Substituting (76) and (77) into (73), we obtain (65). ■

Proof [Proofs of Lemma 16] Let $\mathbf{y}^* = \operatorname{argmin}_{\mathbf{y}} g_{\mathcal{S}}(\mathbf{y})$. By the optimal condition of $\mathbf{y}^*$, we have

$$\nabla_{\mathcal{S}} f\left(\mathbf{x}^0\right) = -\mathbf{H}_{\mathcal{S}} \mathbf{y}^*.$$

Define $\tilde{\mathbf{y}}^k = \mathbf{y}^k - \mathbf{y}^*$. From the update rule of $\mathbf{y}^k$ in (61), we have

$$\begin{aligned}
\mathbf{H}_{\mathcal{S}} \tilde{\mathbf{y}}^k &= \nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right), \qquad (78) \\
\tilde{\mathbf{y}}^{k+1} &= \tilde{\mathbf{y}}^k - \eta \mathbf{H}_{\mathcal{S}} \tilde{\mathbf{y}}^k. \qquad (79)
\end{aligned}$$

We can bound the first-order difference between $g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}}\right)$ and $g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right)$ as:

$$\begin{aligned}
& \left\langle \nabla g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right), \mathbf{u}^{\mathscr{K}} - \mathbf{y}^{\mathscr{K}} \right\rangle \overset{(78)}{=} \left\langle \tilde{\mathbf{y}}^{\mathscr{K}}, \mathbf{z}^{\mathscr{K}} \right\rangle_{\mathbf{H}_{\mathcal{S}}} \qquad (80) \\
\overset{(79)\,(66)}{=} \quad & \left\langle (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})\tilde{\mathbf{y}}^{\mathscr{K}-1}, (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})\mathbf{z}^{\mathscr{K}-1} - \eta \boldsymbol{\xi}_{\mathbf{u}}^{\mathscr{K}} - \eta\left(\nabla_{\mathbf{u}} f\left(\mathbf{x}^{\mathscr{K}-1}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}-1}\right)\right)\right\rangle_{\mathbf{H}_{\mathcal{S}}} \\
\overset{a}{=} \quad & \left\langle \tilde{\mathbf{y}}^{\mathscr{K}-1}, \mathbf{z}^{\mathscr{K}-1} \right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^2} - \eta \left\langle \tilde{\mathbf{y}}^{\mathscr{K}-1}, \boldsymbol{\xi}_{\mathbf{u}}^{\mathscr{K}} \right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})} \\
& -\eta \left\langle \tilde{\mathbf{y}}^{\mathscr{K}-1}, \nabla_{\mathbf{u}} f\left(\mathbf{x}^{\mathscr{K}-1}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}-1}\right)\right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})} \\
\overset{b}{=} & \eta \sum_{k=1}^{\mathscr{K}} \left\langle \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_{\mathbf{u}}^{k} \right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k+1}} - \eta \sum_{k=0}^{\mathscr{K}-1} \left\langle \tilde{\mathbf{y}}^{k}, \nabla_{\mathbf{u}} f\left(\mathbf{x}^{k}\right) - \nabla g_{\mathcal{S}}\left(\mathbf{u}^{k}\right)\right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k}},
\end{aligned}$$

where in $\overset{a}{=}$, we use $(\mathbf{I}-\mathbf{H}_{\mathcal{S}})\mathbf{H}_{\mathcal{S}} = \mathbf{H}_{\mathcal{S}}(\mathbf{I}-\mathbf{H}_{\mathcal{S}})$, and in $\overset{b}{=}$, we use $\mathbf{z}^0 = \mathbf{0}$.

We also bound the two terms in the right hand side of (80). For any fixed l from 1 to K_0, and any j from 1 to l, we have

$$\begin{aligned}
& \left| \left\langle \tilde{\mathbf{y}}^{j-1}, \boldsymbol{\xi}_{\mathbf{u}}^{l} \right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j+1}} \right|^2 = \left| \left\langle \mathbf{H}_{\mathcal{S}} \tilde{\mathbf{y}}^{j-1}, \boldsymbol{\xi}_{\mathbf{u}}^{j} \right\rangle_{(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j+1}} \right|^2 \\
\overset{(78)}{=} \quad & \left| \left\langle \nabla g_{\mathcal{S}}\left(\mathbf{y}^{j-1}\right), \boldsymbol{\xi}_{\mathbf{u}}^{j} \right\rangle_{(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j+1}} \right|^2 \\
\overset{a}{\leq} \quad & \sigma^2 \left\| \nabla g_{\mathcal{S}}\left(\mathbf{y}^{j-1}\right) \right\|^2 \left\| (\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j+1} \right\|_2^2 \overset{b}{\leq} \sigma^2 \left\| \nabla g_{\mathcal{S}}\left(\mathbf{y}^{j-1}\right) \right\|^2, \qquad (81)
\end{aligned}$$

where $\overset{a}{\leq}$ uses $\|\boldsymbol{\xi}_{\mathbf{u}}^j\| \leq \|\boldsymbol{\xi}^j\| \leq \sigma$, $\overset{b}{\leq}$ uses $\|(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-j+1}\|_2 \leq 1$ for all j from 1 to l. So for any l from $1 \leq k \leq K_0$, by standard AzumaHoeffding inequality, using $\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)$ is measurable on $\boldsymbol{\mathcal{F}}^0$, with probability at least $1-p/(12K_0)$, we have

$$\left|\eta\sum_{k=1}^{l}\left\langle\tilde{\mathbf{y}}^{k-1},\boldsymbol{\xi}_{\mathbf{u}}^{k}\right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{l-k+1}}\right| \quad\leq\quad \sqrt{2\eta^2\sigma^2\log(24K_0/p)\sum_{k=0}^{l-1}\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\|^2}. \tag{82}$$

By union bound, with probability at least $1-p/12$, (82) holds for all l from 1 to K_0. we have with probability at least $1-p/12$

$$\begin{aligned}\left|\eta\sum_{k=1}^{\mathscr{K}}\left\langle\tilde{\mathbf{y}}^{k-1},\boldsymbol{\xi}_{\mathbf{u}}^{k}\right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k+1}}\right| &\leq \sqrt{2\eta^2\sigma^2\log(24K_0/p)\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}(\mathbf{y}^k)\right\|^2} \qquad (83)\\ &\overset{a}{\leq} \frac{\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}(\mathbf{y}^k)\right\|^2}{16}+8\eta\sigma^2\log(48K_0/p),\end{aligned}$$

where in $\overset{a}{\leq}$, we use $\sqrt{ab}\leq\frac{a+b}{2}$ with $a\geq 0$ and $b\geq 0$.

For the second term in the right hand side of (80), we have

$$\begin{aligned}&\eta\sum_{k=0}^{\mathscr{K}-1}\left\langle\tilde{\mathbf{y}}^k,\nabla_{\mathbf{u}}f\left(\mathbf{x}^k\right)-\nabla g_{\mathcal{S}}\left(\mathbf{u}^k\right)\right\rangle_{\mathbf{H}_{\mathcal{S}}(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k}}\\ \overset{(78)}{=}\;&\eta\sum_{k=0}^{\mathscr{K}-1}\left\langle\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right),\nabla_{\mathbf{u}}f\left(\mathbf{x}^k\right)-\nabla g_{\mathcal{S}}\left(\mathbf{u}^k\right)\right\rangle_{(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k}}\\ \overset{a}{\leq}\;&\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|\cdot\left\|\nabla_{\mathbf{u}}f\left(\mathbf{x}^k\right)-\nabla g_{\mathcal{S}}\left(\mathbf{u}^k\right)\right\|\\ \overset{b}{\leq}\;&\frac{\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{8}+2\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla_{\mathbf{u}}f\left(\mathbf{x}^k\right)-\nabla g_{\mathcal{S}}\left(\mathbf{u}^k\right)\right\|^2\\ \overset{(56)}{\leq}\;&\frac{\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{8}+\eta\rho^2B^4K_0/2,\end{aligned}\tag{84}$$

where $\overset{a}{\leq}$ uses $\left\|(\mathbf{I}-\eta\mathbf{H}_{\mathcal{S}})^{\mathscr{K}-k}\right\|_2\leq 1$ with $k<\mathscr{K}$, $\overset{b}{\leq}$ uses $ab\leq\frac{a^2+b^2}{2}$.

Substituting (83) and (84) into (80), and using (65), we have

$$\begin{aligned}&g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}}\right) \qquad (85)\\ =\;&g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right)+\left\langle\nabla g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right),\mathbf{u}^{\mathscr{K}}-\mathbf{y}^{\mathscr{K}}\right\rangle+\frac{1}{2}\left(\mathbf{u}^{\mathscr{K}}-\mathbf{y}^{\mathscr{K}}\right)^{\top}\mathbf{H}\left(\mathbf{u}^{\mathscr{K}}-\mathbf{y}^{\mathscr{K}}\right)\\ \overset{(80)\,(65)}{\leq}\;&g_{\mathcal{S}}\left(\mathbf{y}^{\mathscr{K}}\right)+\frac{3\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}(\mathbf{y}^k)\right\|^2}{16}+4\eta\sigma^2(\log(K_0)+3)\log(48K_0/p)+\rho^2\eta B^4K_0.\end{aligned}$$

Then by adding (85) and (63), we have

$$\begin{aligned}&g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}}\right)-g_{\mathcal{S}}\left(\mathbf{u}^0\right)\\ \leq\;&-\frac{25\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{32}+4\eta\sigma^2(3+\log(K_0))\log(48K_0/p)+\rho^2\eta B^4K_0,\end{aligned}\tag{86}$$

implying Lemma 16. ■

We then investigate $g_{\mathcal{S}\perp}(\cdot)$ and summarize its property as follows:

Lemma 18 *With hyper-parameters set in* (8) *for Algorithm* 2*, we have*

$$\begin{aligned} g_{\mathcal{S}\perp}(\mathbf{v}^{\mathscr{K}}) &\leq g_{\mathcal{S}\perp}(\mathbf{v}^0) - \sum_{k=1}^{\mathscr{K}} \eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k \right\rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathscr{K}-1} \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{x}^k\right) \right\|^2 + \frac{\rho^2 B^4 \eta K_0}{2} \\ &= g_{\mathcal{S}\perp}(\mathbf{v}^0) - \sum_{k=1}^{\mathscr{K}} \eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k \right\rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathscr{K}-1} \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{x}^k\right) \right\|^2 + \tilde{\mathcal{O}}(\epsilon^{1.5}). \end{aligned} \tag{87}$$

Proof [Proofs of Lemma 18] Lemma 18 can be obtained via the standard analysis. Specifically, from the definition of $g_{\mathcal{S}\perp}(\cdot)$, we have

$$\begin{aligned} g_{\mathcal{S}\perp}\left(\mathbf{v}^{k+1}\right) &= g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \mathbf{v}^{k+1} - \mathbf{v}^k \right\rangle + \left[\mathbf{v}^{k+1} - \mathbf{v}^k\right]^\top \frac{\mathbf{H}_{\mathcal{S}\perp}}{2} \left[\mathbf{v}^{k+1} - \mathbf{v}^k\right] \\ &\overset{\mathbf{H}_{\mathcal{S}\perp} \preceq \mathbf{0}}{\leq} g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \mathbf{v}^{k+1} - \mathbf{v}^k \right\rangle \\ &\overset{(53)}{=} g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) - \eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) + \boldsymbol{\xi}_{\mathbf{v}}^{k+1} \right\rangle. \end{aligned} \tag{88}$$

We can further bound the right hand side of (88) as follows:

$$\begin{aligned} & -\left\langle \eta \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) \right\rangle \\ = & -\eta \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\|^2 - \left\langle \eta \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \nabla_{\mathbf{v}} f(\mathbf{x}^k) - \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\rangle \\ \leq & -\frac{7\eta}{8} \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\|^2 + 2\eta \left\| \nabla_{\mathbf{v}} f(\mathbf{x}^k) - \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\|^2, \end{aligned} \tag{89}$$

where in the last inequality, we apply:

$$\left\langle \eta \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right), \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) - \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) \right\rangle \leq \frac{\eta \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\|^2}{8} + 2\eta \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) - \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) \right\|^2. \tag{90}$$

Substituting (89) into (88), and telescoping the results with k from 0 to $\mathscr{K} - 1$, we have

$$\begin{aligned} & g_{\mathcal{S}\perp}\left(\mathbf{v}^{\mathscr{K}}\right) \\ \leq & g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right) - \sum_{k=1}^{\mathscr{K}} \eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k \right\rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathscr{K}-1} \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{x}^k\right) \right\|^2 + 2\eta \sum_{k=0}^{\mathscr{K}-1} \left\| \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) - \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) \right\|^2 \\ \overset{a}{\leq} & g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right) - \sum_{k=1}^{\mathscr{K}} \eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k \right\rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathscr{K}-1} \left\| \nabla g_{\mathcal{S}\perp}\left(\mathbf{x}^k\right) \right\|^2 + \frac{\rho^2 B^4 \eta K_0}{2}, \end{aligned} \tag{91}$$

where in $\overset{a}{\leq}$, we use

$$\|\nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right) - \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\| = \left\| \boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}\left(\nabla f\left(\mathbf{x}^k\right) - \nabla g\left(\mathbf{x}^k\right)\right) \right\| \leq \left\| \nabla f\left(\mathbf{x}^k\right) - \nabla g\left(\mathbf{x}^k\right) \right\| \overset{(56)}{\leq} \rho B^2/2,$$

holds for all $k \leq \mathscr{K} - 1$.

■

D.3. Proofs of Proposition 9

With Lemma 16 and 18 in hand, the mainly rest to do is to prove

$$\sum_{k=0}^{\mathscr{K}_0-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2+\sum_{k=0}^{\mathscr{K}_0-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2=\tilde{\Omega}(1)$$

and bound the noise term $-\sum_{k=1}^{\mathscr{K}}\eta\left\langle\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right),\boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle$. We separately consider two cases:

1. $\left\|\nabla f\left(\mathbf{x}^0\right)\right\|>5\sigma\asymp 1$,
2. $\left\|\nabla f\left(\mathbf{x}^0\right)\right\|\le 5\sigma\asymp 1$.

(i) **Case 1:** in the sense that the gradient is large, we show that function value is guaranteed to decrease monotonously.

Proof [Proofs of Proposition 9 in Case 1] Because $\left\|\nabla f\left(\mathbf{x}^0\right)\right\|>5\sigma$, we have, for all $0\le k\le\mathscr{K}-1$,

$$\left\|\nabla f\left(\mathbf{x}^k\right)\right\|\ge\left\|\nabla f\left(\mathbf{x}^0\right)\right\|-\left\|\nabla f\left(\mathbf{x}^k\right)-\nabla f\left(\mathbf{x}^0\right)\right\|\overset{a}{\ge}5\sigma-LB\overset{LB\le\sigma}{\ge}\frac{9}{2}\sigma,\tag{92}$$

where $\overset{a}{\ge}$ uses $\|\mathbf{x}^k-\mathbf{x}^0\|\le B$ for all $k\le\mathscr{K}_1$, the L-Lipschitz continuous of the gradient. Furthermore, we also have

$$\begin{aligned}
&f\left(\mathbf{x}^{k+1}\right)-f\left(\mathbf{x}^k\right)\\
\le\;&\left\langle\nabla f\left(\mathbf{x}^k\right),\mathbf{x}^{k+1}-\mathbf{x}^k\right\rangle+\frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\\
\overset{a}{=}\;&-\eta\left\langle\nabla f\left(\mathbf{x}^k\right),\nabla f(\mathbf{x}^k)+\boldsymbol{\xi}^{k+1}\right\rangle+\frac{L\eta^2}{2}\|\nabla f(\mathbf{x}^k)+\boldsymbol{\xi}^{k+1}\|^2\\
\overset{b}{\le}\;&-\eta\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2-\eta\left\langle\nabla f\left(\mathbf{x}^k\right),\boldsymbol{\xi}^{k+1}\right\rangle+L\eta^2\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2+L\eta^2\left\|\boldsymbol{\xi}^{k+1}\right\|^2\\
\overset{c}{\le}\;&-\frac{15\eta}{16}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2+\frac{5\eta}{32}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2+\frac{8}{5}\eta\sigma^2+L\eta^2\sigma^2\\
\le\;&-\frac{25\eta}{32}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2+2\eta\sigma^2\overset{(92)}{\le}-\eta\left(\frac{25}{32}-\frac{8}{81}\right)\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2,
\end{aligned}\tag{93}$$

where $\overset{a}{=}$ uses the update rule of SGD: $\mathbf{x}^{k+1}=\mathbf{x}^k-\eta\nabla f\left(\mathbf{x}^k\right)-\eta\boldsymbol{\xi}^{k+1}$, $\overset{b}{\le}$ uses $\|\mathbf{a}+\mathbf{b}\|^2\le 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$, in $\overset{c}{\le}$, we use $L\eta\le\frac{1}{16}$ from (8), $-\left\langle\nabla f\left(\mathbf{x}^k\right),\boldsymbol{\xi}^{k+1}\right\rangle\le\frac{5}{32}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2+\frac{8}{5}\left\|\boldsymbol{\xi}^{k+1}\right\|^2$, and $\left\|\boldsymbol{\xi}^{k+1}\right\|\le\sigma$. By telescoping (93) with k from 0 to $\mathscr{K}-1$, we have

$$f\left(\mathbf{x}^{\mathcal{K}}\right)-f\left(\mathbf{x}^0\right)\le-\eta\left(\frac{25}{32}-\frac{8}{81}\right)\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2.\tag{94}$$

On the other hand, again by the update rule of SGD, we have

$$\begin{aligned}
&\left\|\eta\sum_{k=0}^{\mathscr{K}-1}\nabla f\left(\mathbf{x}^k\right)\right\| = \left\|-\eta\sum_{k=0}^{\mathscr{K}-1}\nabla f\left(\mathbf{x}^k\right)\right\| \\
= \quad & \left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0 + \eta\sum_{k=1}^{\mathscr{K}}\boldsymbol{\xi}^k\right\| \\
\geq \quad & \left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\right\| - \left\|\eta\sum_{k=1}^{\mathscr{K}}\boldsymbol{\xi}^k\right\|. \qquad (95)
\end{aligned}$$

By the Vector-Martingale Concentration Inequality in Theorem 11, we have with probability $1-p/12$,

$$\left\|\sum_{k=1}^{\mathscr{K}}\boldsymbol{\xi}^k\right\| \leq \left\|\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\cdot\mathcal{I}_{\mathscr{K}\geq k}\right\| \overset{a}{\leq} 2\eta\sigma\sqrt{K_0\log(48/p)} \leq \frac{B}{16}, \qquad (96)$$

where $\overset{a}{\leq}$ uses $\mathcal{I}_{k\leq\mathscr{K}}$ is measurable on $\boldsymbol{\mathcal{F}}^{k-1}$ and $\|\boldsymbol{\xi}^k\|\leq\sigma$. So if (96) happens, and $\mathbf{x}^k$ exits $\mathcal{B}\left(\mathbf{x}^0, B\right)$ in K_0 iterations, we have

$$\begin{aligned}
&\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla f\left(\mathbf{x}^k\right)\right\|^2 \\
\overset{a}{\geq} \quad & \frac{1}{\eta\mathscr{K}}\left\|\eta\sum_{k=0}^{\mathscr{K}-1}\nabla f\left(\mathbf{x}^k\right)\right\|^2 \overset{(95)}{\geq} \frac{1}{\eta\mathscr{K}}\left(B-\frac{1}{16}B\right)^2 \\
\geq \quad & \frac{15^2B^2}{16^2\eta\mathscr{K}} \overset{\mathscr{K}\leq K_0}{\geq} \frac{15^2B^2}{16^2\eta K_0}, \qquad (97)
\end{aligned}$$

where in $\overset{a}{\geq}$, we use the inequality that

$$\left\|\sum_{i=1}^{l}\mathbf{a}_i\right\|^2 \leq l\sum_{i=1}^{l}\|\mathbf{a}_i\|^2,$$

holds for all $l\geq 1$. Plugging (97) into (94), with probability at least $1-p/12$ ((96) happens), we have

$$f\left(\mathbf{x}^{\mathscr{K}_0}\right) \leq f\left(\mathbf{x}^0\right) - \left(\frac{25}{32}-\frac{8}{81}\right)\frac{15^2B^2}{16^2\eta K_0} \leq f\left(\mathbf{x}^0\right) - \frac{B^2}{7\eta K_0}. \qquad (98)$$

■

Case 2: To obtain the result, we first prepare the following lemmas:

(ii) We fuse Lemma 16 and 18 and obtain the lemma shown below:

Lemma 19 *With the parameters set in (8), and if $\|\nabla f(\mathbf{x}^0)\| \leq 5\sigma$, with probability $1 - p/4$ ((70), (75) and (83) happen), we have*

$$\begin{aligned} f\left(\mathbf{x}^{\mathscr{K}}\right) \quad &\leq \quad f\left(\mathbf{x}^0\right) - \eta\sum_{k=1}^{\mathscr{K}}\left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle + \left(\frac{3}{256}+\frac{1}{80}\right)\frac{B^2}{\eta K_0} \qquad (99)\\ &\quad -\frac{7\eta}{8}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 - \frac{25\eta}{32}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2. \end{aligned}$$

Proof [Proofs of Lemma 19] Because $\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right)\right\| = \left\|\nabla_{\mathbf{v}} f\left(\mathbf{x}^0\right)\right\| \leq \left\|\nabla f\left(\mathbf{x}^0\right)\right\| \leq 5\sigma$, for all $0 \leq k \leq \mathscr{K} - 1$, we have

$$\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\| \leq \left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right)\right\| + \left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) - \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right)\right\| \overset{a}{\leq} 5\sigma + LB \overset{LB\leq\frac{1}{2}}{\leq} \frac{11}{2}\sigma, \qquad (100)$$

where in $\overset{a}{\leq}$, we use $\left\|\mathbf{v}^k - \mathbf{v}^0\right\| = \left\|\boldsymbol{\mathcal{P}}_{\mathcal{S}\perp}(\mathbf{x}^k - \mathbf{x}^0)\right\| \leq B$ and L-Lipschitz continuous gradient for $g_{\mathcal{S}\perp}(\cdot)$. In the same way, for all $0 \leq k \leq \mathscr{K} - 1$, we have

$$\left\|\nabla f\left(\mathbf{x}^k\right)\right\| \leq \frac{11\sigma}{2}, \qquad (101)$$

which indicates that

$$\left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\right\| \leq \left\|\mathbf{x}^0 - \mathbf{x}^{\mathscr{K}-1}\right\| + \left\|\nabla f\left(\mathbf{x}^{\mathscr{K}-1}\right) + \boldsymbol{\xi}^{\mathscr{K}}\right\| \leq B + \frac{13}{2}\eta\sigma \overset{(8)}{\leq} B + \frac{B}{100}. \qquad (102)$$

We then bound the difference between $f(\mathbf{x}^{\mathscr{K}})$ and $g(\mathbf{x}^{\mathscr{K}})$: using ρ-smoothness of Hessian, we have

$$f\left(\mathbf{x}^{\mathscr{K}}\right) - f\left(\mathbf{x}^0\right) - g_{\mathcal{S}}\left(\mathbf{u}^{\mathscr{K}}\right) - g_{\mathcal{S}\perp}\left(\mathbf{v}^{\mathscr{K}}\right) \leq \frac{\rho}{6}\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\|^3 \overset{(102)}{\leq} \frac{\rho B^3}{5}, \qquad (103)$$

Then by adding (86) and (91), using (103), and $g_{\mathcal{S}}\left(\mathbf{u}^0\right) + g_{\mathcal{S}\perp}\left(\mathbf{v}^0\right) = 0$, we have, with probability at least $1 - p/4$, ((70), (75) and (83) happen)

$$\begin{aligned} &f\left(\mathbf{x}^{\mathscr{K}}\right) \qquad (104)\\ \leq\ & f\left(\mathbf{x}^0\right) - \eta\sum_{k=1}^{\mathscr{K}}\left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle + 4\eta\sigma^2\left(1 + 3\log(K_0)\right)\log(48/p)\\ &-\frac{7\eta}{8}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 - \frac{25\eta}{32}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2 + \frac{3\rho^2 B^4\eta K_0}{2} + \frac{\rho B^3}{5}. \end{aligned}$$

With the parameter set in (8),

$$4\eta\sigma^2\left(1 + 3\log(K_0)\right)\log(48/p) \leq \frac{B^2}{256\eta K_0}, \qquad (105)$$

$$\frac{3\rho^2 B^4 \eta K_0}{2} \leq \frac{B^2}{128\eta K_0}, \tag{106}$$

and

$$\frac{\rho B^3}{5} \leq \frac{B^2}{80\eta K_0}, \tag{107}$$

we obtain with probability at least $1-p/4$, ((70), (75) and (83) happen)

$$\begin{aligned} f\left(\mathbf{x}^{\mathscr{K}}\right) \quad &\leq \quad f\left(\mathbf{x}^0\right) - \eta \sum_{k=1}^{\mathscr{K}} \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k \right\rangle + \left(\frac{3}{256} + \frac{1}{80}\right) \frac{B^2}{\eta K_0} \\ &\quad - \frac{7\eta}{8} \sum_{k=0}^{\mathscr{K}-1} \left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 - \frac{25\eta \sum_{k=0}^{\mathscr{K}-1} \left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2}{32}, \end{aligned}$$

implying (99). ■

(iii) Furthermore, the following lemma ensures the function value sufficient descent:

Lemma 20 *With probability* $1-\frac{p}{6}$ *((70) and (110) happen), if* $\mathbf{x}^k$ *exits* $\mathcal{B}\left(\mathbf{x}^{\mathcal{K}}, B\right)$ *in* K_0 *iterations, we have*

$$\eta \sum_{k=0}^{\mathscr{K}-1} \left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 + \eta \sum_{k=0}^{\mathscr{K}-1} \left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2 \geq \frac{169B^2}{512\eta K_0}. \tag{108}$$

Proof [Proofs of Lemma 20] From the update rule of SGD, we have

$$\begin{aligned} &\left\|\eta \sum_{k=0}^{\mathscr{K}-1} \left(\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right)\right\| = \left\|-\eta \sum_{k=0}^{\mathscr{K}-1} \left(\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right)\right\| \\ \overset{a}{=} \quad & \left\|\mathbf{v}^{\mathscr{K}} - \mathbf{v}^0 + \eta \sum_{k=0}^{\mathscr{K}-1} \left(\boldsymbol{\xi}_{\mathbf{v}}^{k+1} - \nabla g_{\mathcal{S}\perp}(\mathbf{v}^k) + \nabla_{\mathbf{v}} f(\mathbf{x}^k)\right) + \mathbf{y}^{\mathscr{K}} - \mathbf{y}^0\right\| \\ \overset{b}{\geq} \quad & \left\|\mathbf{v}^{\mathscr{K}} - \mathbf{v}^0 + \eta \sum_{k=0}^{\mathscr{K}-1} \boldsymbol{\xi}_{\mathbf{v}}^{k+1} + \left(\mathbf{u}^{\mathscr{K}} - \mathbf{u}^0\right) - \left(\mathbf{z}^{\mathscr{K}} - \mathbf{z}^0\right)\right\| \\ & - \left\|\eta \sum_{k=0}^{\mathscr{K}-1} \left(\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) - \nabla_{\mathbf{v}} f\left(\mathbf{x}^k\right)\right)\right\| \\ \overset{c}{\geq} \quad & \left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\right\| - \left\|\mathbf{z}^{\mathscr{K}} - \mathbf{z}^0\right\| - \eta \left\|\sum_{k=1}^{\mathscr{K}} \boldsymbol{\xi}_{\mathbf{v}}^k\right\| - \frac{\eta K_0 \rho B^2}{2} \\ \overset{(8)}{\geq} \quad & \left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\right\| - \left\|\mathbf{z}^{\mathscr{K}} - \mathbf{z}^0\right\| - \frac{B}{32} - \eta \left\|\sum_{k=1}^{\mathscr{K}} \boldsymbol{\xi}_{\mathbf{v}}^k\right\|., \end{aligned} \tag{109}$$

where $\overset{a}{=}$ uses $\mathbf{v}^k = \mathbf{v}^{k-1} - \eta\boldsymbol{\xi}_{\mathbf{v}}^{k+1} - \eta\nabla_{\mathbf{v}} f(\mathbf{x}^k)$ and $\mathbf{y}^k = \mathbf{y}^{k-1} - \eta\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)$, $\overset{b}{\geq}$ uses $\mathbf{z}^{\mathscr{K}} = \mathbf{u}^{\mathscr{K}} - \mathbf{y}^{\mathscr{K}}$, $\mathbf{z}^0 = \mathbf{u}^0 = \mathbf{y}^0 = \mathbf{0}$, and triangle inequality, $\overset{c}{\geq}$ uses (56).

From (64), with probability at least $1 - \frac{1}{12}p$, we have $\left\|\mathbf{z}^{\mathscr{K}} - \mathbf{z}^0\right\| \leq \frac{3B}{32}$. By the Vector-Martingale Concentration Inequality in Theorem 11, we have with probability $1 - p/12$,

$$\left\|\eta\sum_{k=1}^{\mathscr{K}}\boldsymbol{\xi}_{\mathbf{v}}^k\right\| = \left\|\eta\sum_{k=1}^{K_0}\left(\boldsymbol{\xi}_{\mathbf{v}}^k \cdot \mathcal{I}_{k\leq\mathscr{K}}\right)\right\| \overset{a}{\leq} 2\eta\sigma\sqrt{K_0\log(48/p)} \overset{(8)}{\leq} \frac{B}{16}, \tag{110}$$

where $\overset{a}{\leq}$ uses $\mathcal{I}_{k\leq\mathscr{K}}$ is measurable on $\boldsymbol{\mathcal{F}}^{k-1}$ and $\|\boldsymbol{\xi}_{\mathbf{v}}^k\| \leq \sigma$. We obtain

$$\left\|\eta\sum_{k=0}^{\mathscr{K}-1}\left(\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right)\right\| \geq \left\|\mathbf{x}^{\mathscr{K}} - \mathbf{x}^0\right\| - \frac{3B}{16}. \tag{111}$$

So with probability $1 - p/6$, if $\mathbf{x}^k$ exits $\mathcal{B}\left(\mathbf{x}^{\mathcal{K}}, B\right)$ in K_0 iterations, we have

$$\begin{aligned}
&\eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 + \eta\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2\\
\overset{a}{\geq}\;& \frac{1}{2\eta\mathscr{K}}\left\|\eta\sum_{k=0}^{\mathscr{K}-1}\left(\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right) + \nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right)\right\|^2\\
\overset{(111)}{\geq}\;& \frac{169B^2}{512\eta\mathscr{K}} \overset{\mathscr{K}\leq K_0}{\geq} \frac{169B^2}{512\eta K_0},
\end{aligned} \tag{112}$$

where $\overset{a}{\geq}$ uses

$$\left\|\sum_{i=1}^{l}\mathbf{a}_i\right\|^2 \leq l\sum_{i=1}^{l}\|\mathbf{a}_i\|^2,$$

holds for all $l \geq 1$. ■

(iv) Now, we have all the ingredients necessary to prove Proposition 9:

Proof [Proofs of Proposition 9 in Case 2] We first bound the noise term $\sum_{k=1}^{\mathscr{K}}\left\langle\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle$. We have for all k from 1 to K_0

$$\mathbb{E}\left[\eta\left\langle\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle \cdot \mathcal{I}_{k\leq\mathscr{K}} \mid \boldsymbol{\mathcal{F}}^{k-1}\right] = \mathbf{0}, \tag{113}$$

From (100), and $\|\boldsymbol{\xi}_{\mathbf{v}}^k\| \leq \sigma$, we have

$$\left|-\eta\left\langle\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle \cdot \mathcal{I}_{k\leq\mathscr{K}}\right| \leq \frac{11\eta\sigma^2}{2}, \tag{114}$$

and

$$\mathbb{E}\left|\eta\left\langle\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^k\right\rangle \cdot \mathcal{I}_{k\leq\mathscr{K}} \mid \boldsymbol{\mathcal{F}}^{k-1}\right|^2 \leq \eta^2\sigma^2\mathcal{I}_{k\leq\mathscr{K}}\left\|g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right)\right\|^2 \tag{115}$$

by Data-Dependent Berinstein inequality in Theorem 12 with $\delta = \frac{p}{3\log(K_0)}$, we have with probability at least $1 - \frac{p}{3}$,

$$\sum_{k=1}^{K_0} \left\{ -\eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^{k} \right\rangle \cdot \mathcal{I}_{k \leq \mathscr{K}} \right\} \tag{116}$$

$$\leq \quad \max\left\{ 11\eta\sigma^2 \cdot \log\left(\frac{3\log(K_0)}{p}\right), 4\sqrt{\eta^2\sigma^2 \sum_{k=0}^{\mathscr{K}-1} \|\nabla g_{\mathcal{S}\perp}(\mathbf{v}^k)\|^2} \cdot \sqrt{\log\left(\frac{3\log(K_0)}{p}\right)} \right\}.$$

With the parameter set in (8), we have

$$11\eta\sigma^2 \cdot \log\left(\frac{3\log(K_0)}{p}\right) \leq \frac{B^2}{100\eta K_0}, \tag{117}$$

and

$$\begin{aligned} & 4\eta\sqrt{\sigma^2 \sum_{k=0}^{\mathscr{K}-1} \|\nabla g_{\mathcal{S}\perp}(\mathbf{v}^{k-1})\|^2} \cdot \sqrt{\log\left(\frac{3\log(K_0)}{p}\right)} \\ \overset{a}{\leq} \quad & 32\log\left(\frac{3\log(K_0)}{p}\right)\eta\sigma^2 + \frac{\eta}{8}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 \\ \overset{(8)}{\leq} \quad & \frac{B^2}{32\eta K_0} + \frac{\eta}{8}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2, \end{aligned} \tag{118}$$

where $\overset{a}{\leq}$ uses $\sqrt{ab} \leq \frac{a+b}{2}$ for $a \geq 0$. Substituting (117) and (118) into (116), with probability at least $1 - 3/p$, we have

$$\sum_{k=1}^{\mathscr{K}} \left\{ -\eta \left\langle \nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^{k-1}\right), \boldsymbol{\xi}_{\mathbf{v}}^{k} \right\rangle \right\} \leq \frac{B^2}{32\eta K_0} + \frac{\eta}{8}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2. \tag{119}$$

Fusing (119) with (99) in Lemma 19, using $\frac{7}{8} - \frac{1}{8} \leq \frac{25}{32}$, we have with probability at least $1 - \frac{7}{12}p$ ((70), (75), (83), and (116) happen),

$$f\left(\mathbf{x}^{\mathscr{K}}\right) \quad \leq \quad f\left(\mathbf{x}^0\right) - \left(\frac{3}{256} + \frac{1}{80} + \frac{1}{32}\right)\frac{B^2}{\eta K_0} - \frac{3\eta}{4}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}\perp}\left(\mathbf{v}^k\right)\right\|^2 - \frac{3\eta}{4}\sum_{k=0}^{\mathscr{K}-1}\left\|\nabla g_{\mathcal{S}}\left(\mathbf{y}^k\right)\right\|^2.$$

Finally, applying Lemma 20, if $\mathbf{x}^k$ moves out of the ball in K_0 iteration, with probability at least $1 - \frac{2}{3}p$ ((70), (75), (83), (110), and (116) happen), we have

$$f\left(\mathbf{x}^{\mathscr{K}_0}\right) - f\left(\mathbf{x}^0\right) \leq -\left(\frac{3}{4}\cdot\frac{169}{512} - \frac{3}{256} - \frac{1}{80} - \frac{1}{32}\right)\frac{B^2}{\eta K_0} \leq -\frac{B^2}{7\eta K_0}. \tag{120}$$

Combining Case 1 and Case 2, we obtain Proposition 9. ■

Appendix E. Deferred Proofs of Part III: Finding SSP

Proof [Proofs of Proposition 10] Clearly, under the random event $\boldsymbol{\mathcal{H}}_0$ in Part I happens, we know that if $\lambda_{\min}\nabla f(\mathbf{x}^0) \le -\delta_2$, $\mathbf{x}^k$ must gone out of the ball. Thus with probability at least $1-p/3$ (the random events $\boldsymbol{\mathcal{H}}_0$ in Part I happens), if $\mathbf{x}^k$ does not move out the ball in K_0 steps, we have $\lambda_{\min}\left(\nabla f(\mathbf{x}^0)\right) \ge -\delta_2$. Using that $f(\mathbf{x})$ has continuous Hessian, we have

$$\lambda_{\min}\left(\nabla f(\bar{\mathbf{x}})\right) \ge \lambda_{\min}\left(\nabla f(\mathbf{x}^0)\right) - \rho\left\|\bar{\mathbf{x}}-\mathbf{x}^0\right\|_2 \ge -\delta_2 - \frac{\rho}{K_0}\sum_{k=0}^{K_0-1}\left\|\bar{\mathbf{x}}-\mathbf{x}^0\right\| \ge -\frac{17}{16}\delta_2 = -17\delta. \tag{121}$$

To a give upper bound on the $\|\nabla f(\bar{\mathbf{x}})\|^2$, we follow the idea by considering quadratic approximations in Part II. We have

$$\begin{aligned}
\|\nabla g(\bar{\mathbf{x}})\| &\overset{a}{=} \left\|\frac{1}{K_0}\sum_{k=0}^{K_0-1}\nabla g(\mathbf{x}^k)\right\| \\
&\le \left\|\frac{1}{K_0}\sum_{k=0}^{K_0-1}\nabla f(\mathbf{x}^k)\right\| + \frac{1}{K_0}\sum_{k=0}^{K_0-1}\left\|\nabla f(\mathbf{x}^k)-\nabla g(\mathbf{x}^k)\right\| \\
&= \frac{1}{K_0\eta}\left\|\mathbf{x}^{K_0-1}-\mathbf{x}^0-\eta\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\right\| + \frac{1}{K_0}\sum_{k=0}^{K_0-1}\left\|\nabla f(\mathbf{x}^k)-\nabla g(\mathbf{x}^k)\right\| \\
&\overset{(56)}{\le} \frac{1}{K_0\eta}\left\|\mathbf{x}^{K_0-1}-\mathbf{x}^0\right\| + \frac{1}{K_0}\left\|\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\right\| + \frac{\rho}{2}B^2 \\
&\le \frac{B}{K_0\eta} + \frac{\rho B^2}{2} + \frac{1}{K_0}\left\|\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\right\| \\
&\le \left(\frac{16}{\tilde{C}_1}+\frac{1}{2}\right)\rho B^2 + \frac{1}{K_0}\left\|\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\right\|,
\end{aligned} \tag{122}$$

where in $\overset{a}{\le}$, we use the gradient of the quadratic function $g(\cdot)$ is a linear mapping.

By the Vector-Martingale Concentration Inequality, we have with probability $1-2p/3$,

$$\frac{1}{K_0}\left\|\sum_{k=1}^{K_0}\boldsymbol{\xi}^k\right\| \le 2\sigma\sqrt{K_0\log(6/p)}/K_0 \le \rho B^2. \tag{123}$$

Using $\|\bar{\mathbf{x}}-\mathbf{x}^0\| \le B$, we have $\|\nabla f(\bar{\mathbf{x}})\| \le \|\nabla g(\bar{\mathbf{x}})\| + \frac{\rho B^2}{2} \le 18\rho B^2$.

In all, we have with probability at least $1-p$, $\lambda_{\min}\left(\nabla f(\bar{\mathbf{x}})\right) \ge -17\delta$, and $\|\nabla f(\bar{\mathbf{x}})\| \le 18\rho B^2$.

■

Proof [Proofs of Theorem 5] By union bound, with probability at least $1-T_1\cdot p$, if at step $T_0 = T_1\cdot K_0$, Algorithm 2 has not stopped, $\mathbf{x}^k$ must have moved out of the ball at least T_1 times, then from Proposition 9, the function values shall decrease at least

$$T_1\cdot\frac{B^2}{7\eta K_0} \ge \Delta + \frac{B^2}{7\eta K_0} > \Delta.$$

Contradiction with Assumption 2. Thus with probability at least $1 - T_1 \cdot p$, Algorithm 2 shall stop before T_0 steps. Further, fusing with Proposition 10, we have with probability at least $1-(T_1+1)\cdot p$, Algorithm 2 outputs a second-order stationary point satisfying (9) in T_0 steps.

■

Appendix F. Proof of Proposition 4

Proof [Proof of Proposition 4]

(i) Recall the multivariate gaussian noise $\tilde{\boldsymbol{\xi}} = \sigma/\sqrt{d} * \boldsymbol{\chi}$ where $\boldsymbol{\chi} \sim N(0, \mathbf{I}_d)$. We show that it satisfies (7). Clearly, it satisfies (6).

Let $\mathbf{v}$ be an arbitrary unit vector, and due to symmetry in below we assume WLOG $\mathbf{v} = \mathbf{e}_1$. Recall we have set $\mathcal{A}$ satisfying the $(q^*, \mathbf{v})$-narrow property in Definition 2. Then

$$\{\mathbf{u} + q\mathbf{e}_1 : \mathbf{u} \in \mathcal{A},\ q \in [q^*, \infty)\} \subseteq \mathcal{A}^c.$$

If set $\mathcal{A}$ contains no points of $\mathbf{u}, \mathbf{u}+q\mathbf{e}_1$ for each $q \geq q^*$, then $\mathcal{A}[\bullet, \mathbf{a}_{\setminus 1}] := \{a_1 : (a_1, \mathbf{a}_{\setminus 1})^\top \in \mathcal{A}\}$ is a subset of $\mathbb{R}$ and has Lebesgue measure $\leq 1.1q^*$. This is because that for any given $\mathbf{a}_{\setminus 1} = (a_2, \ldots, a_d)$ there exists an a_1^* such that $(a_1^*, \mathbf{a}_{\setminus 1})^\top \in \mathcal{A}$ and we pick a_1^* to be the infimum of such. Then it is easy to conclude that $(a_1^* + q, \mathbf{a}_{\setminus 1})^\top \in \mathcal{A}^c$ for any $q > 1.1q^*$, and that

$$\mathcal{A}[\bullet, \mathbf{a}_{\setminus 1}] \subseteq [a_1^*, a_1^* + 1.1q^*].$$

Therefore we have for any $\mathcal{A}$ admitting $(q^*, \mathbf{v})$-narrow property where $q^* = (\sigma/4\sqrt{d})$, that for any given $\boldsymbol{\chi}_{\setminus 1}$,

$$\begin{aligned}\mathbb{P}(\sigma/\sqrt{d} * \boldsymbol{\chi}_1 \in \mathcal{A}[\bullet, \boldsymbol{\chi}_{\setminus 1}] \mid \boldsymbol{\chi}_{\setminus 1}) &\leq \frac{1}{\sqrt{2\pi}} \int_{(4q^*)^{-1}\mathcal{A}[\bullet, \boldsymbol{\chi}_{\setminus 1}]} \exp(-z^2/2) dz \\ &\leq \frac{1.1q^*}{4q^*} \cdot \frac{1}{\sqrt{2\pi}} < \frac{1}{4},\end{aligned}$$

where $\mathcal{A}[\bullet, \boldsymbol{\chi}_{\setminus 1}]$ is of Lebesgue measure $\leq 1.1q^*$. Taking expectation again gives

$$\mathbb{P}(\sigma/\sqrt{d} * \boldsymbol{\chi} \in \mathcal{A}) = \mathbb{E}\left[\mathbb{P}(\sigma/\sqrt{d} * \boldsymbol{\chi}_1 \in \mathcal{A}[\bullet, \boldsymbol{\chi}_{\setminus 1}] \mid \boldsymbol{\chi}_{\setminus 1})\right] \leq \frac{1}{4},$$

and we complete the proof that $\boldsymbol{\xi} = \sigma/\sqrt{d} * \boldsymbol{\chi}$ is $\mathbf{v}$-disperse for any $\mathbf{v}$.

(ii) For example, recall the uniform ball-shaped noise $\tilde{\boldsymbol{\xi}} = \sigma * \boldsymbol{\xi}_b$, where $\boldsymbol{\xi}_b$ is uniformly sampled from $\mathcal{B}^d$, the unit ball centered at $\mathbf{0}$. We prove that (7) holds in this case. Assume once again that $\mathbf{v} = \mathbf{e}_1$ because of symmetry. Using classical results in Multivariate Calculus (or see Jin et al. (2017)) and $(q^*, \mathbf{v})$-narrow property property in Definition 2 of set $\mathcal{A}^*$ we have

$$\mathbb{P}(\sigma * \boldsymbol{\xi}_b \in \mathcal{A}) = \frac{Vol_d((\sigma^{-1}\mathcal{A}) \cap \mathcal{B}^d)}{Vol_d(\mathcal{B}^d)} \leq \frac{q^*}{\sigma} \cdot \frac{Vol_{d-1}(\mathcal{B}^{d-1})}{Vol_d(\mathcal{B}^d)}. \quad (124)$$

It is well known that the d-dimensional unit ball $\mathcal{B}^d$ of $\mathbb{R}^d$ has volume being

$$Vol_d(\mathcal{B}^d) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)},$$

and analogously for $\mathcal{B}^{d-1}$. We have

$$\frac{Vol_{d-1}(\mathcal{B}^{d-1})}{Vol_d(\mathcal{B}^d)} = \frac{\pi^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}+1\right)} \cdot \frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{d/2}} = \frac{\Gamma\left(\frac{d+1}{2}+\frac{1}{2}\right)}{\pi^{1/2}\Gamma\left(\frac{d+1}{2}\right)} \le \sqrt{\frac{d+1}{2\pi}} \le \sqrt{d},$$

where we applied a well-known fact that $\Gamma(x+1/2) \le \Gamma(x)\sqrt{x}$ for all $x > 0$. Plugging in the definition $q^* := \sigma/4\sqrt{d}$ in (124), we have proved (7) that $\tilde{\boldsymbol{\xi}}$ is $\mathbf{v}$-disperse for any $\mathbf{v}$.

(iii) For stochastic gradients injected by artificial, dispersive noise, we prove that the $\mathbf{v}$-disperse property still holds. Let $\tilde{\boldsymbol{\gamma}}$ be some artificial noise that has the $\mathbf{v}$-dispersive property, that is, for an arbitrary set $\mathcal{A}$ with $(q^*, \mathbf{v})$-narrow property, where $q^* = \sigma/4\sqrt{d}$. Then as in Definition 2 one has, by the linearly scalable property after Definition 2, that $\mathbb{P}\left(\tilde{\boldsymbol{\gamma}} \in \mathcal{A} - \mathbf{g}\right) \le 1/4$ for any fixed vector $\mathbf{g} \in \mathbb{R}^d$. Then we have by injecting such independent noise to the stochastic gradient $\nabla f(\mathbf{w}; \boldsymbol{\zeta})$ that

$$\mathbb{P}\left(\nabla f(\mathbf{w}; \boldsymbol{\zeta}) + \tilde{\boldsymbol{\gamma}} \in \mathcal{A} \mid \nabla f(\mathbf{w}; \boldsymbol{\zeta})\right) = \mathbb{P}\left(\tilde{\boldsymbol{\gamma}} \in \mathcal{A} - \nabla f(\mathbf{w}; \boldsymbol{\zeta}) \,\middle|\, \nabla f(\mathbf{w}; \boldsymbol{\zeta})\right) \le \frac{1}{4},$$

where in the last step we used the independence of $\tilde{\boldsymbol{\gamma}} \in \mathcal{A}$ and $\nabla f(\mathbf{w}; \boldsymbol{\zeta})$. Taking expectation in the last line gives

$$\mathbb{P}\left(\nabla f(\mathbf{w}; \boldsymbol{\zeta}) + \tilde{\boldsymbol{\gamma}} \in \mathcal{A}\right) \le \frac{1}{4}, \tag{125}$$

so (7) is satisfied for this noise-injected stochastic gradient $\nabla f(\mathbf{w}; \boldsymbol{\zeta}) + \tilde{\boldsymbol{\gamma}}$.

■