# Towards Better Generalization Bounds of Stochastic Optimization for Nonconvex Learning

Yunwen Lei

***Abstract*—Stochastic optimization is the workhorse behind the success of many machine learning algorithms. The existing theoretical analysis of stochastic optimization mainly focuses on the behavior on the training dataset or requires a convexity assumption. In this paper, we provide a comprehensive analysis on the generalization behavior of stochastic optimization with nonconvex problems. We first present both upper and lower bounds on the uniform convergence of gradients. Our analysis outperforms existing results by incorporating the 2nd moment of the gradient at a single model into the upper bound. Based on this uniform convergence, we provide a high-probability bound on the gradient norm of population risks for stochastic gradient descent (SGD), which significantly improves the existing results. We show that better bounds can be achieved under further assumptions such as quasi-convexity or Polyak-Łojasiewicz condition. Our analysis shows the computation cost can be further decreased by taking the variance-reduction trick. Finally, we study the utility guarantee of SGD under a privacy constraint. Our results show a linear speed up with respect to the batch size, which shows the benefit of computing gradients in a distributed manner.**

***Index Terms*—Learning Theory, Generalization Analysis, Stochastic Optimization, Stochastic Gradient Descent**

## I. Introduction

STOCHASTIC optimization such as stochastic gradient descent (SGD) has found wide applications in training complex models in the big-data era [1]. A basic idea of stochastic optimization is to introduce randomness into the optimization process to speed up the optimization by using the sum structure of objective functions in machine learning (ML). For example, SGD builds an unbiased estimate of gradients by drawing a single example or a minibatch of examples. Variance reduction techniques were introduced to further decrease the variance of stochastic gradients [2–4].

The popularity of stochastic optimization motivates a lot of theoretical studies to understand the convergence of algorithms under different assumptions such as the Lipschitz continuity, smoothness and convexity. For example, convergence rates on the suboptimality of function values were developed for convex problems [1, 5, 6], while convergence rates on the gradient of objectives were derived for nonconvex problems [3, 7–11].

Most of the theoretical analysis focuses on the behavior of empirical risks (training errors) of models on the training dataset. However, the ultimate goal of ML is to train a model with a good behavior on the testing dataset [12]. Based on this consideration, researchers have studied an important issue called the generalization gap to understand the difference between training and testing [13]. Two popular approaches to study the generalization are the algorithmic stability approach [14] and the uniform convergence approach [12]. The former shows that generalization is closely related to the sensitivity of an algorithm up to a perturbation of the training dataset, while the latter uses concentration inequalities for empirical process to bound the generalization gap. Algorithmic stability considers only the property of the output model, and can imply capacity-independent bounds [14]. As a comparison, uniform convergence considers the uniform deviation between training and testing over a function space, and yields capacity-dependent bounds [12, 15–17]. However, stability analysis often requires a convexity assumption to get meaningful stability bounds [13], and therefore have limited applications to nonconvex problems.

To study the generalization of SGD with nonconvex problems, recent studies took a uniform convergence approach and gave a high-probability bound $\|\nabla F(\text{out})\|_2^2 = \widetilde{O}(\sqrt{d/n})$ (we use $\widetilde{O}$ to ignore logarithmic factors) [18, 19], where "out" denotes the output model, $F$ denotes the population risk, $d$ is the dimensionality of the model and $n$ is the sample size. In the literature, it was shown that $\sup_{\mathbf{w}} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2^2 = \widetilde{O}(d/n)$ [16], where $F_S$ denotes the empirical risk. Therefore there is a gap between the high-probability generalization bound of order $\widetilde{O}(\sqrt{d/n})$ [18, 19] and the uniform convergence rate of gradients of the order $\widetilde{O}(d/n)$. This observation motivates the following question: can we further improve the existing high-probability analysis on the generalization of stochastic optimization?

In this paper, we work toward a tighter generalization analysis of stochastic optimization with nonconvex problems. Our major contributions are summarized as follows.

- We develop both upper and lower bounds for the uniform convergence of empirical gradients to the expectation over a ball. Under a Lipschitzness assumption, we develop uniform convergence rates with a *logarithmic* dependency on the radius of the ball. We also develop an upper bound with a linear dependency without the Lipschitzness assumption, which still outperforms existing uniform convergence rates [18] since the radius has a multiplicative factor of $d/n$ instead of the multiplicative factor of $\sqrt{d}/\sqrt{n}$ in [18]. Our upper bounds involve the 2nd moment of the gradient at an output model, which is significantly smaller than the uniform Lipschitz constant. Furthermore, we also develop dimension-free uniform convergence of gradients for functions with a structure.
- We apply our uniform convergence of gradients to study the generalization behavior of SGD. We develop a high-probability bound of the order $\widetilde{O}(d/n)$ for squared-

Yunwen Lei is with the Department of Mathematics, The University of Hong Kong, Hong Kong, China (e-mail: leiyw@hku.hk).

norm of population gradients without a Lipschitzness assumption, which substantially improves the existing bound $\widetilde{O}(\sqrt{d/n})$ [18, 19]. Under a further assumption such as the quasi-convexity or Polyak-Łojasiewicz (PL) condition, we get high-probability bounds on the excess population risks.
- We extend the analysis to study variance reduction optimization algorithms, and show that similar convergence of population gradients can be achieved with less computation than the vanilla SGD.
- Finally, we extend our discussions to differentially private SGD to deal with sensitive data, which is often encountered in application domains in finance and health care. We give high-probability bounds on the utility guarantees. Our results show a linear speed up with respect to (w.r.t.) the batch size, meaning the iteration number decays by a factor of the batch size. This is effective for large-scale optimization since the computation of gradients with minibatch can be implemented in a distributed manner.

The remaining parts of the paper are structured as follows. We discuss the related work in Section II and introduce the problem formulation in Section III. We discuss the uniform convergence of gradients in Section IV and present our results on generalization in Section V. We present the proof on SGD in Section VI-A and leave other proofs to the appendix. We conclude the paper in Section VII.

## II. Related Work

In this section, we discuss related work on the generalization analysis of stochastic optimization methods. We divide our discussions into two parts: generalization via algorithmic stability and generalization via uniform convergence.

Algorithmic stability is a fundamental concept to study the generalization issues of learning algorithms, which measures the sensitivity of the output model up to a perturbation of a dataset [14]. A most widely used stability concept is the uniform stability, which was used to study regularized learning algorithms [14], SGD [13, 20] and differentially private SGD [21]. Several other stability concepts have been introduced to study generalization under different assumptions, including hypothesis stability [14], Bayesian stability [22, 23] and on-average stability [24–26]. For example, argument stability can yield generalization bounds for SGD with nonsmooth problems [26, 27], while on-average stability can incorporate the empirical risks into the generalization bounds, which implies fast rates under a low noise condition [24, 26]. A downside of stability analysis is that it often requires a convexity or weak convexity assumption [28] to get meaningful stability bounds. For general nonconvex problems, the stability parameter grows as an exponential function of the summation of step sizes [13]. Therefore, one needs to choose fast-decaying step sizes of order $\eta_t = O(1/t)$ to get the summation of step sizes controlled [13]. While this step size choice yields good stability and generalization, it leads to a very slow decay of the optimization error. The recent stability analysis only implies sub-optimal bounds for nonconvex problems [29].

Another popular approach to study generalization of stochastic optimization algorithms is the uniform convergence, which considers the uniform deviation of the empirical process indexed by a function class. For convex problems, one often uses the uniform convergence of empirical risks to population risks [15, 30, 31], while for nonconvex problems one often resorts to the uniform convergence of empirical gradients to population gradients [16, 32–34]. The underlying reason is that we can derive convergence rates of excess empirical risks for convex problems [5, 35], and convergence rates of empirical gradients for nonconvex problems [7]. These convergence rates can be combined with the uniform convergence to yield meaningful bounds for quantities related to testing. For nonsmooth problems, the gradients are not well-defined and one resorts to the uniform convergence of gradients of Moreau envelops [36]. The most related work is the recent study of SGD with nonconvex objectives [18, 37], where high-probability bounds of the order $\widetilde{O}(\sqrt{d/n})$ were developed for squared-norm of population gradients. These discussions were extended to SGD with heavy tails [19]. The uniform convergence of gradients was also used to study the utility guarantee of gradient descent with nonconvex problems under a Lipschitz continuity of the Hessian of loss functions [38]. The above discussions consider SGD for solving general problems under some smoothness assumptions.

## III. Problem Formulation

Let $\rho$ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y} \subset \mathbb{R}$ is an output space. Let $S = \{z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)\}$ be a dataset drawn independently from $\rho$, where $n$ is the sample size. Based on $S$ we wish to build a function $g : \mathcal{X} \mapsto \mathcal{Y}$ to learn the relationship between an input and an output. We consider parametric models where a model is determined by a parameter $\mathbf{w}$ in a parameter space $\mathcal{W} = \mathbb{R}^d$, where $d \in \mathbb{N}$ is the dimension. The performance of a model $\mathbf{w}$ on an example $z$ can be quantified by $f(\mathbf{w}; z)$, where $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ is a loss function. The average behavior of $\mathbf{w}$ on a test example can be quantified by the population risk $F(\mathbf{w}) := \mathbb{E}_z[f(\mathbf{w}; z)]$, where $\mathbb{E}_z[\cdot]$ denotes the expectation w.r.t. $z$. As a comparison, the empirical behavior on $S$ is measured by the empirical risk $F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}; z_i)$. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ be a model with the minimal population risk.

In ML, we often apply an algorithm $A$ to minimize the empirical risk to get a model for future prediction. We use $A(S)$ to denote the output of $A$ when applied to the dataset $S$. SGD is a very popular algorithm due to its simplicity and efficiency. Let $\mathbf{w}_1$ be the zero vector and $\{\eta_t\}_t$ be a sequence of positive step sizes. At the $t$-th iteration, we first randomly draw $i_t$ from the uniform distribution over $[n] := \{1, 2, \ldots, n\}$ and then update $\mathbf{w}_{t+1}$ as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}), \tag{1}$$

where $\nabla f$ denotes the gradient of $f$ w.r.t. the first argument. There is a lot of work on the convergence analysis of SGD in terms of empirical behavior. For example, if $F_S$ is nonconvex and smooth, a classical result shows that

$\min_{t\in[T]} \mathbb{E}[\|\nabla F_S(\mathbf{w}_t)\|_2] = O(1/T^{\frac{1}{4}})$ [7]. In this paper, we consider a more challenging problem on the convergence of SGD in terms of the behavior on testing examples, which is the quantity that matters in ML. We consider two performance measures: one is the population gradient $\|\nabla F(A(S))\|_2$ and the other is the excess population risk $F(A(S)) - F(\mathbf{w}^*)$, where $\|\cdot\|_2$ denotes the Euclidean norm. To this aim, we introduce several assumptions.

Our first assumption is the smoothness of loss functions, which is a standard and popular assumption in the literature of nonconvex optimization [7, 8].

**Assumption 1.** We assume that for any $z$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is $L$-smooth, i.e., for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, z \in \mathcal{Z}$

$$\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}'; z)\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2.$$

Our second assumption is the bounded variance of stochastic gradients, which is widely used to study either optimization errors [7] or stability of SGD [25] for nonconvex problems.

**Assumption 2.** We assume the existence of $\sigma > 0$ such that

$$\mathbb{E}_{i_t}\big[\big\|\nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\big\|_2^2\big] \leq \sigma^2, \quad \forall t \in \mathbb{N},$$

where $\mathbb{E}_{i_t}$ denotes the expectation w.r.t. $i_t$.

Our third assumption is on the Lipschitz continuity of loss functions, which is widely considered in the literature for stochastic optimization with both convex [39, 40] and nonconvex problems [8, 41, 42], especially for high-probability analysis [39, 40] and differential privacy analysis [42]. This assumption holds for robust regression, generalized linear models and learning with shallow neural networks. In Section G (Supplementary Material), we provide more nonconvex problems for which Assumption 3 holds with a universal $G$.

**Assumption 3.** We assume there exists some $G > 0$ such that

$$\|\nabla f(\mathbf{w}_t; z)\|_2 \leq G, \quad \forall t \in \mathbb{N}, z \in \mathcal{Z}.$$

It should be mentioned that Assumption 3 is stronger than Assumption 2, i.e., we can always choose $\sigma = G$ in Assumption 2 if Assumption 3 holds. However, as we will also consider generalization bounds (Theorem 7) for SGD without Assumption 3, we keep Assumption 2 here. Furthermore, $\sigma$ can be much smaller than $G$ and therefore we include Assumption 2 to get a better dependency on $G$.

## IV. Uniform Convergence of Gradients

In this paper we are particularly interested in the generalization behavior of stochastic optimization algorithms measured by the decay of $\|\nabla F(\mathbf{w}_t)\|_2$. To this aim, we require a quantitative connection between population gradients and empirical gradients at the output of the algorithm. Since the output depends on the training sample, we turn to the uniform deviation between population and empirical gradients over the whole function class.

### A. Upper Bounds

The following theorem gives upper bounds for the uniform convergence over a ball of finite radius. Let $B_R = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq R\}$. For any $\mathbf{w}$, define $L_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i)\|_2^2$.

We say $A \lesssim B$ if there exists some universal constant $C > 0$ such that $A \leq BC$. We use the notation $A \asymp B$ if $A \lesssim B$ and $B \lesssim A$. For simplicity of presentation, we assume $\log\log n \lesssim d\log(LR/G)$ and $L \lesssim Gd \lesssim Gn$. The proof is given in Section A (Supplementary Material).

**Theorem 1.** *Let $\delta \in (0,1)$ and $S = \{z_1, \ldots, z_n\}$ be drawn independently from $\rho$. Suppose Assumption 1 and Assumption 3 hold. Then with probability at least $1-\delta$ the following inequality holds simultaneously for all $\mathbf{w} \in B_R$*

$$\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim \frac{G\big(\log(1/\delta) + d\log(Rn)\big)}{n} + \Big(\frac{L_S(\mathbf{w})(\log(1/\delta) + d\log(Rn))}{n}\Big)^{\frac{1}{2}}. \quad (2)$$

**Remark 1** (Explanation)**.** We ignore logarithmic factors in this remark. Eq. (2) involves a slow-decaying term $\frac{\sqrt{dL_S(\mathbf{w})}}{\sqrt{n}}$ and a fast-decaying term $\frac{Gd}{n}$ (we assume $d \ll n$ here). Note that the Lipschitz constant $G$ appears only in the fast-decaying term and therefore can be ignored if $n$ is sufficiently large. For example, if $G^2 d \lesssim n$, the dominating term is $\frac{\sqrt{dL_S(\mathbf{w})}}{\sqrt{n}}$, and Eq. (2) becomes

$$\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 = \widetilde{O}\Big(\frac{L_S^{\frac{1}{2}}(\mathbf{w})(\log^{\frac{1}{2}}(1/\delta) + \sqrt{d})}{\sqrt{n}}\Big), \quad (3)$$

where we absorb the logarithmic factors in the notation $\widetilde{O}$. A notable property of Eq. (3) is that the Lipschitz constant $G$ is replaced by $L_S^{\frac{1}{2}}(\mathbf{w})$, which is the 2nd-moment of $\|\nabla f(\mathbf{w}; z)\|_2$ on $S$. For the generalization analysis, we are only interested in $\big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2$ at the particular model $\mathbf{w} = A(S)$ instead of the uniform convergence. Then, an application of Theorem 1 implies a generalization bound depending on $L_S^{\frac{1}{2}}(A(S))$, which is significantly smaller than the uniform Lipschitz constant $G$. Note $G = \sup_{\mathbf{w}\in B_R}\sup_z \|\nabla f(\mathbf{w}; z)\|_2$ involves two supremums. One supremum over $\mathbf{w}$ is replaced by $A(S)$, and the other supremum over $z$ is replaced by an average over $S$. By the self-bounding property $\|\nabla f(\mathbf{w}; z)\|_2^2 \leq 2Lf(\mathbf{w}; z)$ [43], we know

$$L_S(\mathbf{w}) \leq \frac{2L}{n}\sum_{i=1}^n f(\mathbf{w}; z_i) = 2LF_S(\mathbf{w}).$$

Then, we apply Eq. (3) to the particular model $A(S)$ to derive

$$\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2 = \widetilde{O}\Big(\frac{\sqrt{LF_S(A(S))}(\log^{\frac{1}{2}}(1/\delta) + \sqrt{d})}{\sqrt{n}}\Big).$$

This bound is of the order $o(\sqrt{Ld/n})$ in an interpolation setting where $F_S(A(S)) = o(1)$.

**Remark 2** (Comparison)**.** Under an assumption that $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex, $L$-smooth and nonnegative for all $z$, Lemma 1 with $\epsilon = 1/n$ and Lemma 2 in [44] show the following inequality with probability $1-\delta$ for all $\mathbf{w} \in B_R$

$$\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\|_2 \lesssim \frac{L\|\mathbf{w}-\mathbf{w}^*\|_2(\log(1/\delta)+d\log(nR))}{n} + \Big(\frac{L(F(\mathbf{w})-F(\mathbf{w}^*))(\log(1/\delta)+d\log(nR))}{n}\Big)^{\frac{1}{2}} + \Big(\frac{LF(\mathbf{w}^*)\log(1/\delta)}{n}\Big)^{\frac{1}{2}}. \quad (4)$$

A key difference is that the analysis in [44] requires a convexity assumption. Indeed, the idea of their analysis is to control the uniform convergence for excess gradient, i.e., $\sup_{\mathbf{w}} \|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}^*) - (\nabla \hat{F}_S(\mathbf{w}) - \nabla \hat{F}_S(\mathbf{w}^*))\|_2$. To apply Bernstein's inequality, Zhang et al. [44] first gave an estimate on the variance as follows

$$\mathbb{E}_z\big[\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}^*; z)\|_2^2\big] \leq L(F(\mathbf{w}) - F(\mathbf{w}^*)). \quad (5)$$

This variance estimate requires a convexity assumption. As a comparison, our analysis applies to nonconvex loss functions since we directly consider the loss $\mathbf{w} \mapsto f(\mathbf{w}; z)$ instead of the excess loss $\mathbf{w} \mapsto f(\mathbf{w}; z) - f(\mathbf{w}^*; z)$. Then, we use the self-bounding property of smooth functions to show that $\|\nabla f(\mathbf{w}; z)\|_2^2 \leq 2Lf(\mathbf{w}; z)$ [43]. By this strategy, we remove the convexity assumption in [44] to control the variance in Eq. (5). Intuitively, self-bounding property controls gradients by function values, which is widely used to derive optimistic rates for smooth problems [26, 43, 45].

The second difference is that Eq. (4) involves the excess population risk $F(\mathbf{w}) - F(\mathbf{w}^*)$, which cannot be computed from the dataset since $\rho$ is unknown. As a comparison, our upper bound in Theorem 1 involves $L_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i)\|_2^2$, which is a data-dependent quantity. As compared to the excess risk, $L_S(\mathbf{w})$ is easier to estimate. For example, in the proof of Theorem 7, we show $\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t L_S(\mathbf{w}_t) \lesssim \sigma^2 \log(1/\delta)$ with probability at least $1-\delta$. This estimation of $L_S(\mathbf{w}_t)$ simplifies the application of Theorem 1.

Theorem 1 gives a bound with a *logarithmic dependency* on $R$ under a Lipschitzness assumption. We can get a bound with a *linear dependency* without this assumption by noting $\|\nabla f(\mathbf{w}; z) - \nabla f(0; z)\| \leq L\|\mathbf{w}\|_2$. Then we put $G = LR + b$ in Theorem 1 to immediately get the following theorem, where $b = \sup_z \|\nabla f(0; z)\|_2$. We omit the proof for simplicity.

**Theorem 2.** *Let $\delta \in (0,1)$ and $S = \{z_1, \ldots, z_n\}$ be drawn independently from $\rho$. Suppose Assumption 1 holds and $b = \sup_z \|\nabla f(0; z)\|_2 < \infty$. Then with probability at least $1-\delta$ the following inequality holds simultaneously for all $\mathbf{w} \in B_R$*

$$\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim \frac{(LR+b)\big(\log(1/\delta)+d\log(Rn)\big)}{n} + \Big(\frac{L_S(\mathbf{w})(\log(1/\delta)+d\log(Rn))}{n}\Big)^{\frac{1}{2}}. \quad (6)$$

**Remark 3** (Comparison)**.** The uniform convergence analysis of gradients was initialized by a seminal paper [16] under assumptions on gradient statistical noise, Hessian statistical noise and Hessian regularity. Similar bounds were developed by Rademacher complexities or covering numbers [18, 44]

$$\sup_{\mathbf{w}\in B_R} \big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\big\|_2 \lesssim \frac{(LR+b)\sqrt{d}}{\sqrt{n}}. \quad (7)$$

Note that both Eq. (6) and Eq. (7) have a linear dependency on $R$. However, there is a multiplicative factor of $d/n$ in front of $R$ in Eq. (6), while the multiplicative factor in Eq. (7) is $\sqrt{d}/\sqrt{n}$. Therefore, our analysis outperforms the existing uniform convergence rates under the same condition. As we will show, this improved uniform convergence allows us to get improved bounds for SGD under the same assumption in [18].

### *B. Dimension-independent Bounds*

The uniform convergence rates in the previous subsection have an explicit dependency on the dimensionality $d$, which are not appealing for high-dimensional problems [46]. This is especially the case for overparameterized models in modern ML. In this subsection, we aim to relax this issue by developing dimension-independent uniform convergence for problems with a structure when the weight vector $\mathbf{w}$ and the data $\phi(x)$ are appropriately controlled by norms. We consider loss functions of the form

$$f(\mathbf{w}; z) = \ell(y, \langle \mathbf{w}, \phi(x)\rangle), \quad (8)$$

where $\phi : \mathcal{X} \mapsto \mathcal{W}$ is a feature map and $\ell : \mathbb{R}^2 \mapsto \mathbb{R}_+$. This include many problems such as generalized linear models, robust regression models [16, 32] and shallow neural networks. We denote $a \vee b = \max\{a, b\}$. Let $(\lambda_i)_i$ be the eigenvalue of the operator $\mathbf{v} \mapsto \mathbb{E}_X[\langle \mathbf{v}, \phi(X)\rangle \phi(X)]$ arranged in a nonincreasing order.

**Theorem 3.** *Suppose $f$ takes the form in Eq. (8). Assume $a \mapsto \ell(y, a)$ is $L_\ell$-smooth for all $y$. Let $B_\ell = (\mathbb{E}_Y(\ell'(Y, 0))^2)^{\frac{1}{2}}, \sup_x \|\phi(x)\|_2 \leq B_\phi$ and $\delta \in (0,1)$. With probability at least $1-\delta$ the following inequality holds uniformly for any $\mathbf{w} \in B_R$*

$$\Big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\Big\|_2 = \widetilde{O}\Big(\frac{L_\ell B_\phi^2 R\log(1/\delta)}{n} + B_\phi L_\ell\Big(\frac{1}{n}\min_{h\in\mathbb{N}_+}\Big(1 + \tilde{r}(\mathbf{w})h + R^2\sum_{j=h+1}^\infty \lambda_j\Big)\Big)^{\frac{1}{2}}\Big),$$

*where $\tilde{r}(\mathbf{w}) := V(\mathbf{w}) \vee \frac{1}{n}$ and $V(\mathbf{w}) := \mathbb{E}_X[\langle \mathbf{w}, \phi(X)\rangle^2]$.*

For functions of form (8), the uniform convergence of gradients were developed in [32]

$$\sup_{\mathbf{w}\in B_R} \Big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\Big\|_2 \lesssim \frac{L_\ell R B_\phi^2 \log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}}. \quad (9)$$

Theorem 3 improves it by replacing $RB_\phi$ with $\min_{h\in\mathbb{N}_+}\big(1 + \tilde{r}(\mathbf{w})h + R^2\sum_{j=h+1}^\infty \lambda_j\big)^{\frac{1}{2}}$, which is always smaller since $\sum_{i=1}^\infty \lambda_i \leq B_\phi^2$. Furthermore, as we will show, Theorem 3 implies a fast rate if there is a fast decay of eigenvalues. We achieve this improvement by taking a localization analysis and

using several techniques such as local Rademacher complexity, peeling trick and structural results on covering numbers [47].

To understand the benefit of Theorem 3, we impose some assumptions on the decay of eigenvalues.

**Assumption 4** (Polynomial decay)**.** We assume the eigenvalues $\{\lambda_j\}_j$ of $K(x,x') := \langle \phi(x), \phi(x')\rangle$ admit a polynomial decay of degree $p > 1$, i.e., there exists a $\beta > 0$ such that $\lambda_j \leq \beta j^{-p}, \forall j \in \mathbb{N}$.

Assumption 4 is widely used in deriving fast rates of kernel learning methods [48–50]. Below we present a lemma on the polynomial decay of eigenvalues.

**Lemma 4** ([49])**.** *Let $\mathcal{X} = [0,1]^d$ and $K \in \mathcal{C}^\alpha(\mathcal{X}\times\mathcal{X})$, where $\mathcal{C}^\alpha(\mathcal{X}\times\mathcal{X})$ is the space of functions $f : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$ whose sth partial derivatives $D^s f = \frac{\partial^s f}{\partial x_1^{s_1}\cdots\partial x_d^{s_d}}$ are continuous if $\|s\|_1 \leq \alpha$ for any $s = (s_1,\ldots,s_d) \in \mathbb{N}^d$. If the marginal distribution of $\rho$ on $\mathcal{X}$ is a Borel measure, then the corresponding eigenvalues have a polynomial decay with $p = \frac{\alpha}{d} + \frac{1}{2}$.*

**Theorem 5.** *Let assumptions in Theorem 3 and Assumption 4 hold. Then, we have*

$$\min_{h\in\mathbb{N}}\Big\{\tilde{r}(\mathbf{w})h + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big\} \lesssim \frac{p\tilde{r}(\mathbf{w})^{1-\frac{1}{p}}\beta^{\frac{1}{p}}R^{\frac{2}{p}}}{p-1}.$$

*Furthermore, with probability at least $1-\delta$ the following inequality holds uniformly for any $\mathbf{w} \in B_R$*

$$\Big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\Big\|_2 = \widetilde{O}\Big(\frac{L_\ell B_\phi^2 R\log\frac{1}{\delta}}{n} + \frac{c_p B_\phi L_\ell \tilde{r}(\mathbf{w})^{\frac{1}{2}-\frac{1}{2p}}R^{\frac{1}{p}}}{n^{\frac{1}{2}}}\Big)$$

*where $c_p = \beta^{\frac{1}{2p}}p^{\frac{1}{2}}/(p-1)^{\frac{1}{2}} + 1$.*

**Remark 4.** While the uniform convergence rate in Theorem 5 grows as a linear function of $R$, the linear term only appears in $\frac{L_\ell B_\phi^2 R\log\frac{1}{\delta}}{n}$, which is not a dominating term due to the factor of $1/n$. The second term in the upper bound enjoys a sublinear dependency as $R^{\frac{1}{p}}$, which is sharper than the linear dependency in Eq. (9). The difference between $R^{\frac{1}{p}}$ and $R$ is significant if $p$ is large (e.g., $\alpha$ is large in Lemma 4), which is the case for Gaussian kernels $K_\sigma(x,x') := \exp(-\|x-x'\|^2/(2\sigma^2)), \sigma > 0$. Indeed, Gaussian kernels belong to $C^\alpha(\mathcal{X}\times\mathcal{X})$ for any $\alpha \in \mathbb{N}$. This shows the strength of localization analysis in our discussions.

### C. Lower Bounds

In this subsection, we present specific examples to develop lower bounds for the uniform convergence of gradients. Note that $f$ defined in Eq. (10) is smooth since $\sigma$ is smooth, and not convex since $-\sigma(w_d x)$ is not a convex function of $w_d$.

**Proposition 6.** *Let $\mathcal{X} = \{-1,+1\}$. Consider $f : \mathcal{W}\times\mathcal{X} \mapsto \mathbb{R}$:*

$$f(\mathbf{w};x) = \sum_{j=1}^{d-1}\sigma(w_j x) - \sigma(w_d x),\ \mathbf{w} = (w_1,\ldots,w_d)^\top, \quad (10)$$

*where $\sigma(t) = t_+^2/2$ and $t_+ := \max\{t,0\}$. Let $R = (d-1)^{\frac{1}{2}}$ and assume $x_1,\ldots,x_n$ are drawn independently from the uniform distribution over $\{-1,+1\}$. With probability at least $1-\exp(-1/16)$ we know*

$$\sup_{\mathbf{w}\in B_R}\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \geq \frac{(d-1)^{\frac{1}{2}}}{4\sqrt{2n}}. \quad (11)$$

**Remark 5.** Proposition 6 shows that the square-root dependency is necessary for the uniform convergence of gradients for general nonconvex and smooth learning problems. Furthermore, the lower bound in Eq. (11) matches the upper bound in Eq. (3) up to a logarithmic factor. This shows the tightness of our analysis.

## V. Generalization Analysis

In this section, we present the generalization analysis for stochastic optimization algorithms. We first consider SGD under various assumptions, and then study its extension to privacy-preserving and variance-reduced variants.

### A. Stochastic Gradient Descent

We first consider error bounds of SGD on testing datasets for three problem classes: general nonconvex problems, quasi-weakly convex problems and gradient-dominated problems.

**General Nonconvex Problems**. We first consider general smooth problems, for which we measure the performance via population gradients. The underlying reason is that SGD generally only guarantees a local minimizer if the problem is nonconvex. Our idea to prove Theorem 7 is to decompose $\|\nabla F(\mathbf{w}_t)\|_2^2$ into two terms as follows

$$\|\nabla F(\mathbf{w}_t)\|_2^2 \leq 2\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2^2 + 2\|\nabla F_S(\mathbf{w}_t)\|_2^2.$$

We call the first term the generalization error, which shows how the empirical behavior on $S$ would generalize to the test set. We call the second term the optimization error, which shows how the empirical behavior would improve along the optimization process. We will use the uniform convergence of gradients to control the generalization error, and apply tools in optimization theory to control the optimization error. Eq. (12) is also imposed in [18], which is milder than the Lipschitzness assumption since $\eta_t$ is often small (e.g., of order $1/\sqrt{t}$). Theorem 7 focuses on the underparamterized setting, i.e., $d \ll n$. The detailed proof is given in Section VI-A.

**Theorem 7.** *Let $\{\mathbf{w}_t\}$ be produced by Eq. (1) with $\eta_t \asymp 1/\sqrt{t}$. Let Assumptions 1 and 2 hold. Assume there is $G_0$ with*

$$\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t;z)\|_2 \leq G_0, \quad \forall t\in[T], z\in\mathcal{Z}. \quad (12)$$

*For any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\frac{1}{\sum_{t=1}^T\eta_t}\sum_{t=1}^T\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{d\sigma^2\log 1/\delta}{n} + L(\sigma^2+G_0^2)\log(1/\delta)\Big(\frac{1}{\sqrt{T}} + \frac{L^2\sqrt{T}d^2}{n^2}\Big)\Big). \quad (13)$$

**Remark 6** (Comparison)**.** Under the same assumptions, the following bound for SGD was recently developed if taking $\eta_t \asymp 1/\sqrt{t}$ [18, 19]

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(L(G_0^2+\sigma^2)\log(1/\delta)\Big(\frac{1}{\sqrt{T}}+\frac{dL^2\sqrt{T}}{n}\Big)\Big). \quad (14)$$

Both this bound and Theorem 7 involve the term $L(G_0^2+\sigma^2)\frac{\log(1/\delta)}{\sqrt{T}}$, which corresponds to the optimization error bound. The remaining terms are due to the generalization bounds. It is clear that our risk bound is always better than Eq. (14) since (we assume $d \lesssim n$ here)

$$\frac{d\sigma^2}{n}+\frac{L^3(\sigma^2+G_0^2)\sqrt{T}d^2}{n^2} \ll (G_0^2+\sigma^2)\frac{dL^3\sqrt{T}}{n}. \quad (15)$$

If the optimization error bound dominates the generalization (this happens if $T$ is small), then both Theorem 7 and Eq. (14) yield the same risk bound. If the generalization dominates optimization, our risk bound is significantly sharper. For example, if we choose $T \asymp \frac{n^2}{d^2L^2}$, then Theorem 7 implies risk bounds of order $\widetilde{O}(d/n)$

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{L^2(\sigma^2+G_0^2)d\log(1/\delta)}{n}\Big).$$

As a comparison, Eq. (14) implies vacuous bounds of order $\widetilde{O}(1)$ for this $T$. Indeed, Eq. (14) requires to stop at a much earlier iteration to balance the optimization and generalization. By Eq. (14), the optimal choice is $T \asymp \frac{n}{L^2 d}$ and in this case Eq. (14) implies risk bounds

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{\sqrt{d}L^2(\sigma^2+G_0^2)\log(1/\delta)}{\sqrt{n}}\Big). \quad (16)$$

For $T \asymp \frac{n}{L^2 d}$, both Eq. (13) and Eq. (14) yield risk bounds of the same order. However, as $T$ increases from this point, the generalization part dominates and Eq. (15) shows that the existing bound in [18] is larger by a factor of $\max\{\sqrt{T}, n/d\}$ than ours.

The discussion in [18] is based on the uniform convergence rate in Eq. (7), which is a linear function of $R$. A major step in [18, 19] is to show with high probability for $\eta_t \asymp 1/\sqrt{t}$

$$\|\mathbf{w}_t\|_2^2 \le R_T^2 = \widetilde{O}\big(L(\sigma^2+G_0^2)\sqrt{T}\log(1/\delta)\big),\ t\in[T], \quad (17)$$

which, according to Eq. (7), implies

$$\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{L^3(\sigma^2+G_0^2)d\sqrt{T}\log^2(1/\delta)}{n}\Big). \quad (18)$$

As a comparison, we use our improved uniform convergence in Theorem 2 with $R = R_T$ defined in Eq. (17), and get

$$\big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\big\|_2^2 = \widetilde{O}\Big(\frac{dL_S(\mathbf{w}_t)}{n} + \frac{(\sigma^2+G_0^2)L^3\sqrt{T}d^2\log^2(1/\delta)}{n^2}\Big),$$

which is much better than (18) since $d/n$ there is replaced by $d^2/n^2$ in the last term.

The above risk bound involves a linear dependency on $d$. We can remove this dependency by imposing constraints on the norm of data and weight $\mathbf{w}$ for problems with a structure.

**Theorem 8.** *Let $f$ take the structure in Eq.* (8)*. Let assumptions in Theorem 3, Eq.* (12) *and Assumption 2 hold. Let $\{\mathbf{w}_t\}$ be produced by Eq.* (1) *with $\eta_t \asymp 1/\sqrt{t}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{B_\phi^2 L_\ell^2}{n\sum_{t=1}^T\eta_t}\sum_{t=1}^T \eta_t \min_{h\in\mathbb{N}_+}\Big(1 + \tilde{r}(\mathbf{w}_t)h + \big((\sigma^2+G_0^2)\sqrt{T}L_\ell B_\phi^2\log(1/\delta)\big)\sum_{j=h+1}^\infty \lambda_j\Big) + L_\ell B_\phi^2(\sigma^2+G_0^2)\Big(\frac{\log(1/\delta)}{\sqrt{T}}+\frac{\sqrt{T}L_\ell^2 B_\phi^4\log(1/\delta)}{n^2}\Big)\Big),$$

*where $\tilde{r}(\cdot)$ is defined in Theorem 3.*

**Remark 7.** Theorem 8 uses localization arguments to derive risk bounds depending on the eigenvalues of the operator $\mathbf{v} \mapsto \mathbb{E}_X[\langle \mathbf{v}, \phi(X)\rangle\phi(X)]$. To understand how this localization improves the analysis, we impose Assumption 4 with $p > 1$ and assume $\tilde{r}(\mathbf{w}_t) = \widetilde{O}(1)$. If we focus only on the dependency on $n$, then we can choose $T \asymp n^{\frac{2p}{p+1}}$ to get $\frac{1}{\sum_{t=1}^T\eta_t}\sum_{t=1}^T \eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\big(n^{-\frac{p}{p+1}}\big)$, which becomes $O(1/n)$ if $p \to \infty$. As a comparison, the analysis based on Eq. (9) can only imply rates of order $\widetilde{O}(1/\sqrt{n})$.

**Quasi-weakly Convex Problems**. Now we consider quasi-weakly convex problems, which means that the suboptimality $F(\mathbf{w})-F(\mathbf{w}^*)$ can be bounded by the inner product of $\mathbf{w}-\mathbf{w}^*$ and $\nabla F(\mathbf{w})$. This shows $\mathbf{w}-\mathbf{w}^*$ is positively correlated to $\nabla F(\mathbf{w})$, and therefore $-\nabla F(\mathbf{w})$ is a direction towards $\mathbf{w}^*$.

**Assumption 5.** Let $\alpha > 0$. We assume $F$ is $\alpha$-quasi-weakly convex in the sense that for all $\mathbf{w} \in \mathcal{W}$

$$\langle \mathbf{w}-\mathbf{w}^*, \nabla F(\mathbf{w})\rangle \ge \alpha(F(\mathbf{w})-F(\mathbf{w}^*)). \quad (19)$$

The class of quasi-weakly convex functions is characterized by a parameter $\alpha \in [0,1]$ [51]. If $\alpha = 1$, then Eq. (19) is known as star convexity [52]. As $\alpha$ becomes smaller, the function becomes "more nonconvex". In the following theorem to be proved in Section D (supplementary material), we show that the assumption on quasi-weak convexity allows us to derive bounds for excess population risks. This implies that SGD is able to identify an approximate global minimizer of the population risk. In the remainder of the paper, we always impose Assumption 3 and **ignore** the term $\frac{G}{n}\big(\log(1/\delta)+d\log(Rn)\big)$ in the uniform convergence rates, i.e., we only use Eq. (3) for simplicity of presentation. We also only assume $d \lesssim n$ for brevity.

**Theorem 9.** *Let Assumptions 1, 2, 3 and 5 hold. Let $\delta \in (0,1)$ and $\{\mathbf{w}_t\}$ be produced by Eq.* (1)*. With probability at least*

$1-\delta$ *we have*

$$\alpha\sum_{t=1}^{T}\eta_t(F(\mathbf{w}_t)-F(\mathbf{w}^*)) \lesssim \|\mathbf{w}^*\|_2^2 + G^2\sum_{t=1}^{T}\eta_t^2\log^2\frac{1}{\delta} + \frac{G^2(d\log(R_T'n)+\log(1/\delta))}{n}\Big(\sum_{t=1}^{T}\eta_t\Big)^2. \quad (20)$$

*In particular, if* $\eta_t \asymp 1/\sqrt{t}$ *and* $T \asymp n/d$*, we get*

$$\alpha\Big(\sum_{t=1}^{T}\eta_t\Big)^{-1}\sum_{t=1}^{T}\eta_t(F(\mathbf{w}_t)-F(\mathbf{w}^*)) = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2^2\sqrt{d}}{\sqrt{n}} + \frac{G^2\sqrt{d}\log(1/\delta)}{\sqrt{n}} + \frac{G^2\log(1/\delta)\sqrt{d}}{\sqrt{n}}\Big).$$

**Remark 8.** We now discuss the related work on the generalization analysis of SGD to study the decay of $F(\mathbf{w}_t)-F(\mathbf{w}^*)$. Generalization bounds of SGD were established for convex problems based on algorithmic stability [13, 25–27]. High-probability generalization bounds of SGD were also studied for convex problems based on the uniform convergence of empirical risks to population risks [31]. A key assumption for these discussions is the convexity of loss functions, which is relaxed to a quasi-weak convexity assumption in Theorem 9.

**Remark 9.** To prove Theorem 9, we need to give a high-probability bound of $\sum_{t=1}^{T}\tilde{\xi}_t$, where

$$\tilde{\xi}_t = \eta_t\langle \mathbf{w}^*-\mathbf{w}_t, \nabla f(\mathbf{w}_t;z_{i_t}) - \nabla F_S(\mathbf{w}_t)\rangle.$$

To apply a concentration inequality to handle it, we need to bound the magnitude of $\tilde{\xi}_t$ as follows

$$|\tilde{\xi}_t| \le \eta_t\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla f(\mathbf{w}_t;z_{i_t}) - \nabla F_S(\mathbf{w}_t)\|_2.$$

Therefore, to get an informative bound we need to give a good estimate of $\|\mathbf{w}_t\|_2$. Indeed, the most essential part in proving Theorem 9 is to control $\|\mathbf{w}_t\|_2$. This is totally different from the proof of Theorem 7, where a crude bound of $\|\mathbf{w}_t\|_2$ is sufficient due to the logarithmic dependency of the uniform convergence on the radius under the Lipschitzness assumption.

**Gradient-dominated Problems**. Finally, we consider gradient-dominated problems, which are common in non-convex optimization [8, 53, 54], and are shown to hold true for deep (linear) and shallow neural networks [20, 55]. Roughly speaking, gradient dominance means that the suboptimality in terms of function values can be bounded by gradients.

**Assumption 6** (PL Condition)**.** We assume $F$ satisfies PL or gradient-dominated condition with parameter $\mu > 0$, i.e.,

$$F(\mathbf{w})-F(\mathbf{w}^*) \le \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|_2^2, \quad \forall \mathbf{w}\in\mathcal{W}. \quad (21)$$

Under the PL condition, we can derive excess population risk bounds of the order $O(d/(n\mu))$ by stopping SGD after an appropriate number of iterations. As a comparison, most of existing studies of stochastic optimization for gradient-dominated problems consider excess empirical risk bounds, i.e., how $F_S(\mathbf{w}_T)-\inf_{\mathbf{w}}F_S(\mathbf{w})$ would decay as a function of $T$. The proof is given in Section VI-B.

**Theorem 10.** *Let Assumptions 1, 2, 3 and 6 hold. Let* $\delta\in(0,1)$ *and* $\{\mathbf{w}_t\}$ *be produced by Eq.* (1) *with* $\eta_t = 2/(\mu(t+1))$*. Denote* $\widetilde{L}_T := T^{-2}\sum_{t=1}^{T}tL_S(\mathbf{w}_t)$*. With probability* $1-\delta$

$$F(\mathbf{w}_T)-F(\mathbf{w}^*) = \widetilde{O}\Big(\frac{G^2\log^{\frac{1}{2}}(1/\delta)}{\sqrt{T}\mu} + \frac{d\widetilde{L}_T\log(1/\delta)}{n\mu} + \frac{LG^2}{T\mu^2}\Big).$$

If $T \gtrsim n^2G^4/(d^2\widetilde{L}_T^2)$ and $T \gtrsim LnG^2/(\mu d\widetilde{L}_T)$, then Theorem 10 implies bounds of order $\widetilde{O}\big(\frac{d\widetilde{L}_T\log(1/\delta)}{n\mu}\big)$.

**Algorithm 1** Differentially Private SGD

**Input:** $\mathbf{w}_1=0$, learning rates $\{\eta_t\}_t$, parameter $\beta,\epsilon,\delta>0$ and dataset $S=\{z_1,\dots,z_n\}$

Set noise variance $\sigma_T^2 := \frac{8TG^2\log(1/\delta)}{n^2\epsilon^2}$

Set batch size $m := \max\{1, n\sqrt{\epsilon/(4T)}\}$

**for** $t=1,2,\dots,T$ **do**

  sample a batch $B_t=\{z_{i_{t,1}},\dots,z_{i_{t,m}}\}$ with replacement uniformly from $S$

  update $\mathbf{w}_{t+1}$ according to Eq. (22)

**Output:** $\{\mathbf{w}_t\}$

### *B. Differentially Private SGD*

In this subsection, we use our previous generalization analysis to develop differentially private algorithms to handle sensitive data. Differential privacy measures how the perturbation of a training dataset would change the distribution of output models [56]. We say $S$ and $S'$ are two neighboring datasets if they differ by a single example.

**Definition 1** (Differential Privacy)**.** Let $\epsilon>0$ and $\delta\in(0,1)$. A randomized mechanism $\mathcal{A}$ provides $(\epsilon,\delta)$-differential privacy (DP) if for any two neighboring datasets $S$ and $S'$, and any set $E$ in the range of $\mathcal{A}$ there holds

$$\mathbb{P}\{\mathcal{A}(S)\in E\} \le e^{\epsilon}\mathbb{P}\{\mathcal{A}(S')\in E\} + \delta.$$

A basic idea to develop differentially private algorithms is to inject noises in the learning process according to the sensitivity of the algorithm. We consider a differentially-private SGD introduced in [21] (Algorithm 1), where the stochastic gradient is estimated based on a minibatch of samples. At each iteration, we first sample a batch $B_t=\{z_{i_{t,1}},\dots,z_{i_{t,m}}\}$ with replacement from the uniform distribution of $S$. Then we update the model as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\Big(\frac{1}{m}\sum_{j=1}^{m}\nabla f(\mathbf{w}_t;z_{i_{t,j}}) + G_t\Big), \quad (22)$$

where $G_t\sim N(0,\sigma_T^2\mathbb{I}_d)$ (Gaussian distribution) and

$$\sigma_T^2 := \frac{8TG^2\log(1/\delta)}{n^2\epsilon^2}. \quad (23)$$

We present the pseudo-code in Algorithm 1 whose privacy guarantee is given in the following lemma.

**Lemma 11** (Privacy guarantee [21])**.** *If Assumption 3 holds, then Algorithm 1 is* $(\epsilon,\delta)$*-differentially private.*

In the following two theorems, we establish the utility guarantees of Algorithm 1. Theorem 12 considers the utility

guarantee as measured by the gradient norm of empirical risks, while Theorem 13 considers the utility guarantee as measured by the gradient norm of population risks.

**Theorem 12.** *Suppose Assumptions 1, 2 and 3 hold. Let $\{\mathbf{w}_t\}$ be produced by Eq. (22). Then for any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$*

$$\eta\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 \lesssim 1+\eta\log(1/\delta)(G^2+\sigma_T^2)+ G^2L\eta^2\big(T\log(1/\delta)\big)^{\frac{1}{2}}+\frac{\sigma^2L\eta^2T}{m}+L\eta^2\sigma_T^2Td.$$

**Theorem 13.** *Under the same assumptions of Theorem 12, we have*

$$\eta\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2=\widetilde{O}\Big(\frac{\eta(d+\log(1/\delta))}{n}\sum_{t=1}^{T}L_S(\mathbf{w}_t) +1+\eta\log(1/\delta)(G^2+\sigma_T^2)+G^2L\eta^2\big(T\log(1/\delta)\big)^{\frac{1}{2}} +\frac{\sigma^2L\eta^2T}{m}+L\eta^2\sigma_T^2Td\Big).$$

We specify parameters in the utility guarantees and derive the following corollary on the utility guarantee of DP-SGD.

**Corollary 14.** *Let assumptions in Theorem 12 hold. Suppose we choose $m$ according to Algorithm 1 and $\eta=\min\big\{\frac{\sqrt{m}}{\sigma\sqrt{LT}},\frac{1}{\sqrt{TLd}\sigma_T}\big\}$. If $m\lesssim\sqrt{T}\sigma^2/G^2$ and $T\geq\frac{n^2\epsilon^2\sigma^2}{mdG^2}$, then with probability at least $1-\delta$ we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2\lesssim\frac{G^2\log(1/\delta)}{n^2\epsilon^2}+\frac{G\sqrt{Ld\log(1/\delta)}}{n\epsilon} \quad (24)$$

*and*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2=\widetilde{O}\Big(\frac{d+\log(1/\delta)}{n}\frac{1}{T}\sum_{t=1}^{T}L_S(\mathbf{w}_t)+ \frac{G^2\log(1/\delta)}{n^2\epsilon^2}+\frac{G\sqrt{Ld\log(1/\delta)}}{n\epsilon}\Big). \quad (25)$$

According to the choice of $m$ in Algorithm 1, the requirement $m\lesssim\sqrt{T}\sigma^2/G^2$ corresponds to $1\lesssim\sigma^2\sqrt{T}/G^2$ and $n\sqrt{\epsilon/(4T)}\lesssim\sqrt{T}\sigma^2/G^2$, which further corresponds to $T\gtrsim\max\{G^4/\sigma^4,G^2n\sqrt{\epsilon}/\sigma^2\}$. Furthermore, the condition $T\geq\frac{n^2\epsilon^2\sigma^2}{mdG^2}$ means either $T\geq\frac{n^2\epsilon^2\sigma^2}{dG^2}$ or $T\geq 4\epsilon^3\sigma^4n^2/(G^4d^2)$.

**Remark 10.** Under the assumption $\|\mathbf{w}_t\|_2\leq D$ for some $D>0$, utility guarantees of the order $\mathbb{E}[\|\nabla F_S(A(S))\|_2^2]\lesssim\frac{LGD\sqrt{d\log(n/\delta)\log(1/\delta)}}{n\epsilon}$ have been derived for a *Random Round Private Stochastic Gradient Descent* in [57], which requires $O(n^2)$ gradient evaluations. This discussion was extended to other private algorithms with utility guarantee bounds on population gradient squares [38, 58]. However, these discussions require a computation of full gradient per iteration and is therefore not computationally efficient to handle large-scale data. For example, the algorithm in [58] requires $O(n\epsilon\sqrt{L}/(G\sqrt{d\log(1/\delta)}))$ iterations with $O(n)$ gradient evaluations per iteration, leading to a total gradient computation $O(n^2\epsilon\sqrt{L}/(G\sqrt{d\log(1/\delta)}))$. As a comparison, our algorithm requires $O\big(\frac{n^2\epsilon^2\sigma^2}{mdG^2}\big)$ iterations and therefore the gradient computation complexity is $O\big(\frac{n^2\epsilon^2\sigma^2}{dG^2}\big)$. Better utility guarantees of order $\mathbb{E}[\|\nabla F_S(A(S))\|_2]\lesssim\big(\frac{GLd\log(1/\delta)}{n\epsilon}\big)^{\frac{2}{3}}$ were developed for a differentially-private stochastic recursive variance reduced descent method proposed in [42][1], which enjoys a smaller total gradient computational complexity of the order $O((n\epsilon)^2/(d\log(1/\delta)))$. Their utility guarantee is measured in terms of $\|\nabla F_S(A(S))\|_2$, while we also provide guarantee as measured by $\|\nabla F(A(S))\|_2$. Furthermore, the discussions in [38, 42, 58] developed utility guarantees in expectation, while our analysis gives high-probability bounds. Finally, our analysis in Corollary 14 shows that if $m\lesssim\sqrt{T}$, our algorithm achieves a linear speed up on the iteration number w.r.t. the batch size, i.e., the number $T$ of iterations decays by a factor of $m$. The gradient computation per iteration can be performed in a distributed manner.

### C. Stochastic Variance Reduced Optimization

In this subsection, we consider a class of stochastic variance reduced optimization algorithms [2, 3, 8, 9, 54], which are implemented in epochs. Let $\tilde{\mathbf{w}}_0$ be an initialization point. For the $s$-th epoch, we first set a reference point $\mathbf{w}_0=\tilde{\mathbf{w}}_{s-1}$, draw a batch $\tilde{I}_s\subseteq[n]$ and compute $\mathbf{v}_0=\nabla f_{\tilde{I}_s}(\mathbf{w}_0)$. Here we use the notation $f_I(\mathbf{w})=\frac{1}{|I|}\sum_{i\in I}f(\mathbf{w};z_i)$ for $I\subseteq[n]$ with $|I|$ being the cardinality of $I$. We can set $\tilde{I}_s=[n]$ [2, 8, 54] or build $\tilde{I}_s$ by drawing with replacement from the uniform distribution over $[n]$ [9, 10]. Then we run $m_s$ inner iterations. At the $t$-th inner iteration, we first draw a batch $I_t\subseteq[n]$ from the uniform distribution over $[n]$ and update models with gradient estimators of decreased variance. The original SVRG [2, 8] takes the following gradient estimator (we omit the dependency on $s$ for brevity)

$$\mathbf{v}_t=\nabla f_{I_t}(\mathbf{w}_t)-\nabla f_{I_t}(\mathbf{w}_0)+\mathbf{v}_0, \quad (26)$$

while the recent discussions propose the following gradient estimator [3, 9]

$$\mathbf{v}_t=\nabla f_{I_t}(\mathbf{w}_t)-\nabla f_{I_t}(\mathbf{w}_{t-1})+\mathbf{v}_{t-1}. \quad (27)$$

The variance of these $\mathbf{v}_t$ diminishes to zero as we run more and more iterations, which allows us to update iterates with a constant step size $\mathbf{w}_{t+1}=\mathbf{w}_t-\eta\mathbf{v}_t$ while still enjoying convergence [2]. The framework of stochastic variance-reduced optimization is described in Algorithm 2.

We now give the population gradient bounds for stochastic variance reduced optimization algorithms. Theorem 15 considers the SVRG in [8], while Theorem 16 considers the SARAH in [3] and the Spider in [9]. The proofs of these two theorems are given in Section E (supplementary material). We present convergence rates in expectation here since the existing optimization error bounds are stated in expectation.

**Theorem 15.** *Let Assumptions 1, 2 and 3 hold. Let $A$ be the SVRG in [8]. We can take $O(n+Ln^{\frac{2}{3}}/\epsilon^2)$ stochastic gradient evaluations to get a model $A(S)$ with*

$$\mathbb{E}[\|\nabla F(A(S))\|_2]=\widetilde{O}\Big(\epsilon+\frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big).$$

[1] the measure is slightly different here: they consider the gradient norm while we consider the gradient norm square.

**Algorithm 2** Stochastic Variance Reduced Optimization

**Input:** step size $\eta$, initialization $\tilde{\mathbf{w}}_0$, $\{m_s\}$

**for** $s = 1, 2, \ldots$ **do**
  set $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$
  draw a batch $\tilde{I}_s \subseteq [n]$
  compute $\mathbf{v}_0 = \nabla f_{\tilde{I}_s}(\mathbf{w}_0)$
  update $\mathbf{w}_1 = \mathbf{w}_0 - \eta \mathbf{v}_0$
  **for** $t = 1, \ldots, m_s - 1$ **do**
    draw a batch $I_t \subseteq [n]$
    compute $\mathbf{v}_t$ by either (26) or (27)
    update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{v}_t$
  set $\tilde{\mathbf{w}}_s$ as $\mathbf{w}_{i_s}$, where $i_s$ is drawn according to a distribution on $[m_s]$
choose the output from $\{\tilde{\mathbf{w}}_s\}$ according to some strategy

**Theorem 16.** *Let Assumptions 1, 2 and 3 hold. Let $A$ be either the SARAH in [3] or the Spider in [9]. We can take $\widetilde{O}\big(\min\{\sigma^3/\epsilon^3, n + \sqrt{n}L/\epsilon^2\}\big)$ stochastic gradient evaluations to get a model $A(S)$ with*

$$\mathbb{E}[\|\nabla F(A(S))\|_2] = \widetilde{O}\Big(\epsilon + \frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big).$$

**Remark 11.** We now consider the iteration complexity of these algorithms to achieve $\mathbb{E}[\|\nabla F(A(S))\|_2] = \widetilde{O}(\sqrt{d/n})$. For simplicity, we assume $\mathbb{E}[L_S^{\frac{1}{2}}(A(S))] \lesssim 1$. Taking $\epsilon = O(\sqrt{d/n})$, Theorem 15 shows SVRG requires $O(n + Ln^{\frac{2}{3}}/\epsilon^2) = O(n + Ln^{\frac{5}{3}}/d)$ stochastic gradient evaluations. By comparison, Theorem 16 shows SARAH/Spider requires

$$\widetilde{O}\Big(\min\Big\{\frac{\sigma^3}{\epsilon^3}, n + \frac{\sqrt{n}L}{\epsilon^2}\Big\}\Big) = \widetilde{O}\Big(\min\Big\{\frac{n^{\frac{3}{2}}\sigma^3}{d^{\frac{3}{2}}}, n + n^{\frac{3}{2}}L/d\Big\}\Big)$$

stochastic gradient evaluations, which are less than that of SVRG. Furthermore, Theorem 7 shows that SGD requires $O(n^2/(L^2d^2))$ stochastic gradient evaluations to achieve $\mathbb{E}[\|\nabla F(A(S))\|_2] = \widetilde{O}(\sqrt{d/n})$, which is also larger than that of SARAH/Spider. It should be mentioned that for SGD we derive population gradient bounds with high probability, while for stochastic variance reduced optimization we derive bounds in expectation. Other than SARAH/Spider, a stochastic nested variance-reduced gradient descent (SNVRG) was developed for nonconvex optimization, which uses $K+1$ nested reference points to build semi-stochastic gradients for further variance reduction [10]. This algorithm also achieves the computational complexity of order $\widetilde{O}\big(\min\{\sigma^3/\epsilon^3, n + \sqrt{n}L/\epsilon^2\}\big)$ to achieve the accuracy $\epsilon$. Therefore, our analysis also implies similar generalization bounds for SNVRG.

## VI. Proof on Stochastic Gradient Descent

In this section, we present the theoretical analysis of SGD. We first consider general nonconvex problems in Section VI-A and gradient-dominated problems in Section VI-B. Finally, we consider SGD under a privacy constraint in Section VI-C.

### A. Proof of Theorem 7

Below we present the proof of Theorem 7. By the elementary inequality $(a+b)^2 \leq 2(a^2+b^2)$, we know

$$\begin{aligned}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 &= \sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t) + \nabla F_S(\mathbf{w}_t)\|_2^2 \\ &\leq 2\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2^2 + 2\sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2.\end{aligned} \tag{28}$$

*Proof of Theorem 7.* It was shown in [18] that with probability at least $1 - \delta/3$

$$\|\mathbf{w}_t\|_2^2 = \widetilde{O}\Big((\sigma^2 + G_0^2)L\sqrt{T}\log(1/\delta)\Big),\ \forall t \in [T]. \tag{29}$$

Then, the $L$-smoothness of $f$ implies

$$\|\nabla f(\mathbf{w}_t; z)\|_2 = \widetilde{O}((\sigma + G_0)L^{\frac{3}{2}}T^{\frac{1}{4}}\log^{\frac{1}{2}}(1/\delta)) := \widetilde{G}. \tag{30}$$

Then, we can apply Theorem 1 with $G = \widetilde{G}$ to derive the following inequality with probability at least $1 - 2\delta/3$ simultaneously for all $t \in [T]$

$$\begin{aligned}\big\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\big\|_2 = \widetilde{O}\Big(&\frac{\sqrt{dL_S(\mathbf{w}_t)}}{\sqrt{n}} + \\ &\frac{(\sigma + G_0)L^{\frac{3}{2}}T^{\frac{1}{4}}(d + \log 1/\delta)\log^{\frac{1}{2}}(1/\delta)}{n}\Big).\end{aligned}$$

The following optimization error bound was shown in [18] with probability at least $1 - \delta/3$

$$\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{L(\sigma^2 + G_0^2)\log(1/\delta)}{\sqrt{T}}\Big). \tag{31}$$

We combine the above two inequalities and Eq. (28) together, and derive

$$\begin{aligned}\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(&\frac{L(\sigma^2 + G_0^2)\log(1/\delta)}{\sqrt{T}} + \\ &\frac{d}{n\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t L_S(\mathbf{w}_t) + \frac{(\sigma^2 + G_0^2)L^3\sqrt{T}d^2\log^3(1/\delta)}{n^2}\Big).\end{aligned} \tag{32}$$

By Assumption 2, we know

$$\begin{aligned}L_S(\mathbf{w}_t) &= \mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\|_2^2] + \|\nabla F_S(\mathbf{w}_t)\|_2^2 \\ &\leq \sigma^2 + \|\nabla F_S(\mathbf{w}_t)\|_2^2.\end{aligned}$$

It then follows from Eq. (31) that

$$\begin{aligned}&\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t L_S(\mathbf{w}_t) \leq \sigma^2 + \frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|_2^2 \\ &= \widetilde{O}\Big(\sigma^2 + L(\sigma^2 + G_0^2)\log(1/\delta)/\sqrt{T}\Big) = \widetilde{O}\Big(\sigma^2\log(1/\delta)\Big).\end{aligned}$$

We plug the above inequality back into Eq. (32) and get the stated bound. The proof is completed. □

*Proof of Theorem 8.* By the structure, we know that $f$ is $L = (L_\ell B_\phi^2)$-smooth. By Eq. (29) and Theorem 3 with

$R = \widetilde{O}\big((\sigma^2+G_0^2)^{\frac{1}{2}}\sqrt{L_\ell}B_\phi T^{\frac{1}{4}}\log^{\frac{1}{2}}(1/\delta)\big)$, we derive the following inequality with probability $1-2\delta/3$ simultaneously for all $t\in[T]$

$$\Big\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\Big\|_2 = \widetilde{O}\Big(\frac{L_\ell B_\phi^2 R\log(1/\delta)}{n} + B_\phi L_\ell\Big(\frac{1}{n}\min_{h\in\mathbb{N}_+}\Big(1+\tilde{r}(\mathbf{w}_t)h + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big)^{\frac{1}{2}}\Big).$$

We combine it with Eq. (31) and derive

$$\frac{1}{\sum_{t=1}^T\eta_t}\sum_{t=1}^T\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{L_\ell B_\phi^2(\sigma^2+G_0^2)\log(1/\delta)}{\sqrt{T}} + \frac{L_\ell^2 B_\phi^4 R^2\log^2(1/\delta)}{n^2} + \frac{B_\phi^2 L_\ell^2}{n\sum_{t=1}^T\eta_t}\sum_{t=1}^T\eta_t\min_{h\in\mathbb{N}_+}\Big(1+\tilde{r}(\mathbf{w}_t)h + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big).$$

The proof is completed. □

### B. Proof of Theorem 10

In this section, we prove Theorem 10 on excess population risk bounds for SGD under the PL condition.

*Proof of Theorem 10.* According to the update (1), we know

$$\|\mathbf{w}_t\|_2 \le G\sum_{k=1}^T\eta_k = G\sum_{k=1}^T\frac{2}{\mu(k+1)} := R_T. \tag{33}$$

According to Assumption 1, we know

$$F(\mathbf{w}_{t+1}) \le F(\mathbf{w}_t) + \langle\mathbf{w}_{t+1}-\mathbf{w}_t,\nabla F(\mathbf{w}_t)\rangle + \frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 = F(\mathbf{w}_t) - \eta_t\langle\nabla f(\mathbf{w}_t;z_{i_t}),\nabla F(\mathbf{w}_t)\rangle + \frac{L\eta_t^2\|\nabla f(\mathbf{w}_t;z_{i_t})\|_2^2}{2}.$$

According to Assumption 3, we further get

$$F(\mathbf{w}_{t+1}) \le F(\mathbf{w}_t) - \eta_t\langle\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle - \eta_t\langle\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle - \eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{L\eta_t^2G^2}{2}.$$

By the Schwarz's inequality, we further get

$$F(\mathbf{w}_{t+1}) \le F(\mathbf{w}_t) - \eta_t\langle\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle + \frac{\eta_t}{2}\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2^2 + \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{L\eta_t^2G^2}{2}.$$

It then follows from Assumption 6 that

$$F(\mathbf{w}_{t+1}) \le F(\mathbf{w}_t) - \eta_t\langle\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle + \frac{\eta_t}{2}\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2^2 - \mu\eta_t\big(F(\mathbf{w}_t)-F(\mathbf{w}^*)\big) + \frac{L\eta_t^2G^2}{2}.$$

Denote $\triangle_t := F(\mathbf{w}_t)-F(\mathbf{w}^*)$. The above inequality can be reformulated as $(1-\mu\eta_t = 1-2/(t+1))$

$$\triangle_{t+1} \le \Big(1-\frac{2}{t+1}\Big)\triangle_t - \frac{2\langle\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle}{\mu(t+1)} + \frac{1}{\mu(t+1)}\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2^2 + \frac{2LG^2}{\mu^2(t+1)^2}.$$

We multiply both sides by $t(t+1)$ and get

$$t(t+1)\triangle_{t+1} \le (t-1)t\triangle_t - \frac{2t\langle\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t),\nabla F(\mathbf{w}_t)\rangle}{\mu} + \frac{t}{\mu}\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2^2 + \frac{2LG^2}{\mu^2}.$$

Taking a summation of this inequality from $t=1$ to $T$ shows

$$T(T+1)\triangle_{T+1} \le \sum_{t=1}^T\frac{2t}{\mu}\langle\nabla F_S(\mathbf{w}_t)-\nabla f(\mathbf{w}_t;z_{i_t}),\nabla F(\mathbf{w}_t)\rangle + \sum_{t=1}^T\frac{t}{\mu}\|\nabla F(\mathbf{w}_t)-\nabla F_S(\mathbf{w}_t)\|_2^2 + \frac{2LG^2T}{\mu^2}. \tag{34}$$

Introduce the following martingale difference sequences

$$\xi_t := \frac{t}{\mu}\langle\nabla F_S(\mathbf{w}_t)-\nabla f(\mathbf{w}_t;z_{i_t}),\nabla F(\mathbf{w}_t)\rangle,\quad \forall t\in[T].$$

According to Assumption 3, we know $|\xi_t| \le \frac{2G^2t}{\mu}$ and therefore one can apply Lemma D.1 to derive the following inequality with probability at least $1-\delta/2$

$$\sum_{t=1}^T\xi_t \le \frac{2G^2}{\mu}\Big(2\sum_{t=1}^T t^2\log(2/\delta)\Big)^{\frac{1}{2}} \lesssim \frac{G^2T^{\frac{3}{2}}\log^{1/2}(1/\delta)}{\mu}.$$

According to Eq. (3), we have the following inequality with probability at least $1-\delta/2$

$$\sum_{t=1}^T t\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\|_2^2 = \widetilde{O}\Big(\frac{T^2\widetilde{L}_T(d+\log(1/\delta))}{n}\Big).$$

We plug the above two inequalities back into Eq. (34) and derive the following inequality with probability at least $1-\delta$

$$T(T+1)\triangle_{T+1} = O\Big(\frac{G^2T^{\frac{3}{2}}\log^{1/2}(1/\delta)}{\mu}\Big) + \widetilde{O}\Big(\frac{T^2\widetilde{L}_T(d+\log(1/\delta))}{n\mu}\Big) + \frac{2LG^2T}{\mu^2}.$$

This gives the stated bound. □

### C. Proofs on Differentially Private SGD

In this section, we present the proof on utility guarantees of Algorithm 1. To this aim, we first introduce the following lemma on the concentration behavior of elementary random variables [59]. Let $\chi^2(n)$ denote the chi-square distribution with the degree of freedom being $n$.

**Lemma 17** ([59]). *Let $X\sim\chi^2(n)$. Then*

$$\mathbb{P}\big\{X \ge n+2\sqrt{nt}+2t\big\} \le \exp(-t).$$

*Proof of Theorem 12.* According to Assumption 2 and the i.i.d. property of $i_{t,j}$, we know

$$\mathbb{E}_{B_t}\Big\|\frac{1}{m}\sum_{j=1}^m\nabla f(\mathbf{w}_t;z_{i_{t,j}})-\nabla F_S(\mathbf{w}_t)\Big\|_2^2 \le \frac{\sigma^2}{m}. \tag{35}$$

By the $L$-smoothness of $F_S$ we know

$$F_S(\mathbf{w}_{t+1}) \leq F_S(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2$$
$$= F_S(\mathbf{w}_t) - \eta\Big\langle \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) + G_t, \nabla F_S(\mathbf{w}_t)\Big\rangle$$
$$+ \frac{L\eta^2}{2}\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) + G_t\Big\|_2^2.$$

By the standard inequality $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$ we know

$$\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) + G_t\Big\|_2^2$$
$$\leq 2\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t)\Big\|_2^2 + 2\|\nabla F_S(\mathbf{w}_t) + G_t\|_2^2$$
$$\leq 2\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t)\Big\|_2^2 + 4\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 4\|G_t\|_2^2$$

We combine the above two inequalities together and derive

$$F_S(\mathbf{w}_{t+1})$$
$$\leq F_S(\mathbf{w}_t) - \eta\Big\langle \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\Big\rangle$$
$$- \eta\langle G_t, \nabla F_S(\mathbf{w}_t)\rangle - \eta\|\nabla F_S(\mathbf{w}_t)\|_2^2 + L\eta^2\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}})$$
$$- \nabla F_S(\mathbf{w}_t)\Big\|_2^2 + 2L\eta^2\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2L\eta^2\|G_t\|_2^2. \quad (36)$$

For any $t \in [T]$, we define

$$\xi_t = \Big\langle \nabla F_S(\mathbf{w}_t) - \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}), \nabla F_S(\mathbf{w}_t)\Big\rangle,$$

$$\xi_t' = \Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t)\Big\|_2^2$$
$$- \mathbb{E}_{B_t}\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t)\Big\|_2^2.$$

It is clear that $\{\xi_t\}, \{\xi_t'\}$ are two martingale difference sequences. Taking a summation of Eq. (36) from $t = 1$ to $t = T$ and using Eq. (35), we get

$$F_S(\mathbf{w}_{T+1}) \leq F_S(\mathbf{w}_1) + \eta\sum_{t=1}^{T}\xi_t - \eta\sum_{t=1}^{T}\langle G_t, \nabla F_S(\mathbf{w}_t)\rangle + L\eta^2\sum_{t=1}^{T}\xi_t'$$
$$+ \frac{\sigma^2 L\eta^2 T}{m} + (2L\eta^2 - \eta)\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2L\eta^2\sum_{t=1}^{T}\|G_t\|_2^2. \quad (37)$$

According to Assumption 3, we know

$$\xi_t - \mathbb{E}_{B_t}\xi_t = \xi_t \leq 2G^2 \quad \text{and} \quad |\xi_t' - \mathbb{E}_{B_t}\xi_t'| \leq 4G^2.$$

It then follows from the Schwarz's inequality and Eq. (35) that

$$\mathbb{E}_{B_t}[(\xi_t - \mathbb{E}_{B_t}\xi_t)^2] = \mathbb{E}_{B_t}[\xi_t^2] \leq \|\nabla F_S(\mathbf{w}_t)\|_2^2 \times$$
$$\mathbb{E}_{B_t}\Big\| \frac{1}{m}\sum_{j=1}^{m} \nabla f(\mathbf{w}_t; z_{i_{t,j}}) - \nabla F_S(\mathbf{w}_t)\Big\|_2^2 \leq \frac{\sigma^2\|\nabla F_S(\mathbf{w}_t)\|_2^2}{m}. \quad (38)$$

By Part (a) of Lemma D.1 we get the following inequality with probability at least $1 - \delta/4$

$$\sum_{t=1}^{T}\xi_t' \leq 4G^2\Big(2T\log(4/\delta)\Big)^{\frac{1}{2}}. \quad (39)$$

According to Part (b) of Lemma D.1 and Eq. (38), we get the following inequality with probability at least $1 - \delta/4$

$$\sum_{t=1}^{T}\xi_t \leq \frac{\rho\sigma^2\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2}{2G^2 m} + \frac{2G^2\log(4/\delta)}{\rho}.$$

Choosing $\rho = \min\{G^2 m/(3\sigma^2), 1\}$ implies the following inequality with probability at least $1 - \delta/4$

$$\sum_{t=1}^{T}\xi_t \leq \frac{1}{6}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 2G^2\log\frac{4}{\delta}\max\Big\{1, \frac{3\sigma^2}{G^2 m}\Big\}. \quad (40)$$

Let $\tilde{\xi}_t = -\langle G_t, \nabla F_S(\mathbf{w}_t)\rangle$. Then, it is clear that $\tilde{\xi}_t$ (conditioned on $\nabla F_S(\mathbf{w}_t)$) is a Gaussian random variable with mean 0 and variance $\sigma_T^2\|\nabla F_S(\mathbf{w}_t)\|_2^2$. Therefore, we have

$$\log \mathbb{E}_{G_t}\exp(\rho\tilde{\xi}_t) \leq \frac{\rho^2\sigma_T^2\|\nabla F_S(\mathbf{w}_t)\|_2^2}{2}.$$

We then apply Lemma D.1 (Part (c)) to derive the following inequality with probability at least $1 - \delta/4$

$$-\sum_{t=1}^{T}\langle G_t, \nabla F_S(\mathbf{w}_t)\rangle \leq \frac{\rho\sigma_T^2}{2}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 + \frac{\log(4/\delta)}{\rho}.$$

Taking $\rho = 1/(3\sigma_T^2)$ gives the following inequality

$$-\eta\sum_{t=1}^{T}\langle G_t, \nabla F_S(\mathbf{w}_t)\rangle \leq \frac{\eta}{6}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 + 3\sigma_T^2\eta\log(4/\delta). \quad (41)$$

We know that $\sum_{t=1}^{T}\|G_t\|_2^2/\sigma_T^2 \sim \chi^2(Td)$, where $\chi^2(Td)$ means the chi-square distribution with the degree of freedom $Td$. By Lemma 17, with probability at least $1 - \delta/4$ we have

$$\sum_{t=1}^{T}\|G_t\|_2^2/\sigma_T^2 \leq Td + 2\sqrt{Td\log(4/\delta)} + 2\log(4/\delta). \quad (42)$$

We plug Eq. (39), (40), (41), (42) back into Eq. (37), and derive the following inequality with probability at least $1 - \delta$

$$F_S(\mathbf{w}_{T+1}) \leq F_S(\mathbf{w}_1) + 2\eta G^2\log(4/\delta)\max\{1, 3\sigma^2/(G^2 m)\}$$
$$+ 3\eta\sigma_T^2\log(4/\delta)) + 4G^2L\eta^2\Big(2\log(4/\delta)T\Big)^{\frac{1}{2}}$$
$$+ \frac{\sigma^2 L\eta^2 T}{m} + (\eta/6 + \eta/6 + 2L\eta^2 - \eta)\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2$$
$$+ 2L\eta^2\sigma_T^2\Big(Td + 2\sqrt{Td\log(4/\delta)} + 2\log(4/\delta)\Big).$$

Since $\eta \leq 1/(12L)$, we further get the following inequality with probability at least $1-\delta$

$$\frac{\eta}{2}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 \leq F_S(\mathbf{w}_1) + 4G^2L\eta^2\Big(2\log(4/\delta)T\Big)^{\frac{1}{2}} + 2\eta G^2\log(4/\delta)\max\{1, 3\sigma^2/(G^2m)\} + 3\eta\sigma_T^2\log(4/\delta)) + \frac{\sigma^2L\eta^2T}{m} + 2L\eta^2\sigma_T^2\Big(Td + 2\sqrt{Td\log(1/\delta)} + 2\log(1/\delta)\Big).$$

The proof is completed. □

*Proof of Theorem 13.* According to Eq. (22), we know

$$\mathbf{w}_{t+1} = -\frac{\eta}{m}\sum_{k=1}^{t}\sum_{j=1}^{m}\nabla f(\mathbf{w}_k; z_{i_{k,j}}) - \eta\sum_{k=1}^{t}G_k.$$

It is clear that $\sum_{k=1}^{t}G_k$ follows the Gaussian distribution $N(0, t\sigma_T^2\mathbb{I}_d)$ and therefore

$$\|\mathbf{w}_{t+1}\|_2 \leq \frac{\eta}{m}\sum_{k=1}^{t}\sum_{j=1}^{m}\|\nabla f(\mathbf{w}_k; z_{i_{k,j}})\|_2 + \eta\|N(0, t\sigma_T^2\mathbb{I}_d)\|_2 \leq \eta tG + \eta\|N(0, t\sigma_T^2\mathbb{I}_d)\|_2.$$

Let $X \sim N(0, t\sigma_T^2\mathbb{I}_d)$. Then $\|X\|^2/\sigma_T^2 \sim \chi^2(td)$. According to Lemma 17, we know

$$\mathbb{P}\{\|X\|_2 \geq \sigma_T(\sqrt{td} + \sqrt{2a})\} \leq \mathbb{P}\{\|X\|_2^2 \geq \sigma_T^2(td + 2\sqrt{tda} + 2a)\} \leq \exp(-a).$$

That is, with probability at least $1-\delta/(2T)$ we have

$$\|N(0, t\sigma_T^2\mathbb{I}_d)\|_2 \leq \sigma_T\big(\sqrt{td} + \sqrt{2\log(2T/\delta)}\big).$$

By the union bounds of probability, we know with probability at least $1-\delta/2$ the following inequality holds for all $t \in [T]$

$$\|\mathbf{w}_{t+1}\|_2 \leq \eta TG + \eta\sigma_T\big(\sqrt{Td} + \sqrt{2\log(2T/\delta)}\big) := R_T. \quad (43)$$

According to Eq. (3), we get the following inequality with probability at least $1-\delta/4$

$$\eta\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{\eta(d+\log(1/\delta))}{n}\sum_{t=1}^{T}L_S(\mathbf{w}_t)\Big). \quad (44)$$

By Theorem 12, the following inequality holds with probability at least $1-\delta/4$

$$\eta\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 = O\Big(1 + \eta\log(1/\delta)(G^2+\sigma_T^2) + G^2L\eta^2\big(T\log(1/\delta)\big)^{\frac{1}{2}} + \frac{\sigma^2L\eta^2T}{m} + L\eta^2\sigma_T^2Td\Big). \quad (45)$$

Let $A$ be the event that Eq. (43), (44), (45) hold. Then we know $\mathbb{P}\{A\} \geq 1-\delta$, under which we use Eq. (28) and derive

$$\eta\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{\eta(d+\log(1/\delta))}{n}\sum_{t=1}^{T}L_S(\mathbf{w}_t) + 1 + \eta\log(1/\delta)(G^2+\sigma_T^2) + G^2L\eta^2\big(T\log(1/\delta)\big)^{\frac{1}{2}} + \frac{\sigma^2L\eta^2T}{m} + L\eta^2\sigma_T^2Td\Big).$$

The proof is completed. □

*Proof of Corollary 14.* According to Theorem 12, the following inequality holds with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 \lesssim \frac{1}{T\eta} + \frac{\log(1/\delta)(G^2+\sigma_T^2)}{T} + \frac{G^2L\eta\log^{\frac{1}{2}}(1/\delta)}{\sqrt{T}} + \frac{\sigma^2L\eta}{m} + L\eta\sigma_T^2d.$$

Since $m \lesssim \sqrt{T}\sigma^2/G^2$ and $\eta = \min\big\{\frac{\sqrt{m}}{\sigma\sqrt{LT}}, \frac{1}{\sqrt{TLd}\sigma_T}\big\}$, we further get

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|_2^2 \\ &\lesssim \frac{1}{T\eta} + \frac{\log(1/\delta)(G^2+\sigma_T^2)}{T} + \frac{\sigma^2L\eta\log^{\frac{1}{2}}(1/\delta)}{m} + L\eta\sigma_T^2d \\ &\lesssim \frac{G^2\log(1/\delta)}{n^2\epsilon^2} + \frac{\sigma\sqrt{L}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{Tm}} + \frac{\sqrt{Ld}\sigma_T}{\sqrt{T}} \\ &\lesssim \frac{G^2\log(1/\delta)}{n^2\epsilon^2} + \frac{\sigma\sqrt{L}\log^{\frac{1}{2}}(1/\delta)}{\sqrt{Tm}} + \frac{G\sqrt{Ld\log(1/\delta)}}{n\epsilon}.\end{aligned}$$

Since $T \geq \frac{n^2\epsilon^2\sigma^2}{mdG^2}$, we know $\frac{\sigma\sqrt{L}}{\sqrt{Tm}} \leq \frac{G\sqrt{Ld}}{n\epsilon}$ and therefore we get the stated bound Eq. (24). We now prove Eq. (25). According to Theorem 13 and the above deductions, the following inequality holds with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2 = \widetilde{O}\Big(\frac{(d+\log(1/\delta))}{n}\frac{1}{T}\sum_{t=1}^{T}L_S(\mathbf{w}_t) + \frac{G^2\log(1/\delta)}{n^2\epsilon^2} + \frac{G\sqrt{Ld\log(1/\delta)}}{n\epsilon}\Big).$$

The proof is completed. □

## VII. Conclusions

We study the generalization of stochastic optimization algorithms for nonconvex problems via the uniform convergence of gradients. Our uniform convergence rate incorporates the 2nd moment of the stochastic gradients of a particular model, and is significantly better than the existing rates. We develop high-probability bounds of the order $\widetilde{O}(d/n)$ for SGD with nonconvex problems, which significantly improves the existing bounds [18, 19]. We further remove the dependency on $d$ for problems with a structure. We show that improved bounds are possible under further assumptions such as quasi-convexity or PL condition. Finally, we extend our discussions to variance-reduced variants and SGD under privacy constraints. Our results show a linear speed up w.r.t. the batch size by exploiting the smoothness assumption.

There remain several questions worthy of further discussion. For example, our discussion requires a smoothness assumption. It would be interesting to extend our analysis to relaxed smoothness assumptions such as quasi-smoothness [60] and Hölder continuity of gradients [26].

## ACKNOWLEDGEMENTS

We are grateful to the editor and referees for the constructive comments, which are very helpful for us to improve the paper. The work is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723, 17302624].

## REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[2] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, 2013, pp. 315–323.

[3] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *International Conference on Machine Learning*. JMLR. org, 2017, pp. 2613–2621.

[4] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[5] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *International Conference on Machine Learning*, 2004, pp. 919–926.

[6] W. Zhang, L. Zhang, Z. Jin, R. Jin, D. Cai, X. Li, R. Liang, and X. He, "Sparse learning with stochastic composite optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1223–1236, 2017.

[7] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[8] S. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International Conference on Machine Learning*, 2016, pp. 314–323.

[9] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems*, 2018, pp. 689–699.

[10] D. Zhou, P. Xu, and Q. Gu, "Stochastic nested variance reduction for nonconvex optimization," *Journal of Machine Learning Research*, vol. 21, no. 103, pp. 1–63, 2020.

[11] K. Huang, X. Li, and S. Pu, "Distributed stochastic optimization under a general variance condition," *IEEE Transactions on Automatic Control*, 2024.

[12] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[13] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*, 2016, pp. 1225–1234.

[14] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[15] P. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[16] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[17] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7404–7413.

[18] Y. Lei and K. Tang, "Learning rates for stochastic gradient descent with nonconvex objectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4505–4511, 2021.

[19] S. Li and Y. Liu, "High probability guarantees for nonconvex stochastic gradient descent with heavy tails," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 931–12 963.

[20] Z. Charles and D. Papailiopoulos, "Stability and generalization of learning algorithms that converge to global optima," in *International Conference on Machine Learning*, 2018, pp. 744–753.

[21] R. Bassily, V. Feldman, K. Talwar, and A. G. Thakurta, "Private stochastic convex optimization with optimal rates," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 279–11 288.

[22] J. Li, X. Luo, and M. Qiao, "On generalization error bounds of noisy gradient methods for non-convex learning," in *International Conference on Learning Representations*, 2020.

[23] W. Mou, L. Wang, X. Zhai, and K. Zheng, "Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints," in *Conference on Learning Theory*, 2018, pp. 605–638.

[24] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2635–2670, 2010.

[25] I. Kuzborskij and C. Lampert, "Data-dependent stability of stochastic gradient descent," in *International Conference on Machine Learning*, 2018, pp. 2820–2829.

[26] Y. Lei and Y. Ying, "Fine-grained analysis of stability and generalization for stochastic gradient descent," in *International Conference on Machine Learning*, 2020, pp. 5809–5819.

[27] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar, "Stability of stochastic gradient descent on nonsmooth convex losses," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[28] D. Richards and I. Kuzborskij, "Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[29] Y. Lei, "Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems," in *Annual Conference on Learning Theory*, 2023.

[30] J. Lin, L. Rosasco, and D.-X. Zhou, "Iterative regularization for learning with convex loss functions," *Journal of Machine Learning Research*, vol. 17, no. 77, pp. 1–38, 2016.

[31] Y. Lei, T. Hu, and K. Tang, "Generalization performance of multi-pass stochastic gradient descent with convex loss functions," *Journal of Machine Learning Research*, vol. 22, pp. 1–41, 2021.

[32] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 8759–8770.

[33] S. Zhang, Y. Hu, L. Zhang, and N. He, "Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 694–702.

[34] Y. Xu and A. Zeevi, "Towards optimal problem dependent generalization error bounds in statistical learning theory," *Mathematics of Operations Research*, 2024.

[35] F. Orabona, "A modern introduction to online learning," *arXiv preprint arXiv:1912.13213*, 2019.

[36] D. Davis and D. Drusvyatskiy, "Graphical convergence of subgradients in nonconvex optimization and learning," *Mathematics of Operations Research*, vol. 47, no. 1, pp. 209–231, 2022.

[37] S. Li and Y. Liu, "Learning rates for nonconvex pairwise learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9996–10 011, 2023.

[38] D. Wang and J. Xu, "Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1182–1189.

[39] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent," in *Conference on Learning Theory*, 2010, pp. 14–26.

[40] N. J. Harvey, C. Liaw, and S. Randhawa, "Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent," *arXiv preprint arXiv:1909.00843*, 2019.

[41] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes," in *International Conference on Machine Learning*, 2013, pp. 71–79.

[42] L. Wang, B. Jayaraman, D. Evans, and Q. Gu, "Efficient privacy-preserving stochastic nonconvex optimization," in *Uncertainty in Artificial Intelligence*. PMLR, 2023, pp. 2203–2213.

[43] N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in *Advances in Neural Information Processing Systems*, 2010, pp. 2199–2207.

[44] L. Zhang, T. Yang, and R. Jin, "Empirical risk minimization for stochastic convex optimization: $\mathrm{O}(1/n)$-and $\mathrm{O}(1/n^2)$-type of risk bounds," in *Conference on Learning Theory*, 2017, pp. 1954–1979.

[45] P. Zhao, Y.-J. Zhang, L. Zhang, and Z.-H. Zhou, "Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization," *Journal of Machine Learning Research*, vol. 25, no. 98, pp. 1–52, 2024.

[46] V. Feldman, "Generalization of erm in stochastic convex optimization: The dimension strikes back," in *Advances in Neural Information Processing Systems*, 2016, pp. 3576–3584.

[47] P. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[48] I. Steinwart, D. R. Hush, C. Scovel *et al.*, "Optimal rates for regularized least squares regression." in *Conference on Learning Theory*, 2009, pp. 79–93.

[49] S. Mendelson and J. Neeman, "Regularization in kernel learning," *Annals of Statistics*, vol. 38, no. 1, pp. 526–565, 2010.

[50] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Conference on Learning Theory*, 2013, pp. 185–209.

[51] R. Gower, O. Sebbouh, and N. Loizou, "Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1315–1323.

[52] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[53] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *European Conference on Machine Learning*, 2016, pp. 795–811.

[54] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, "Spiderboost and momentum: Faster variance reduction algorithms," in *Advances in Neural Information Processing Systems*, 2019, pp. 2403–2413.

[55] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in *Advances in Neural Information Processing Systems*, 2017, pp. 597–607.

[56] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[57] J. Zhang, K. Zheng, W. Mou, and L. Wang, "Efficient private ERM for smooth objectives," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 3922–3928.

[58] Y. Zhou, X. Chen, M. Hong, Z. S. Wu, and A. Banerjee, "Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds," *arXiv preprint arXiv:2006.13501*, 2020.

[59] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[60] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representations*, 2019.

**Yunwen Lei** received the Ph.D. degree from the Wuhan University, Wuhan, China, in 2014. He is currently an Assistant Professor at the Department of Mathematics, The University of Hong Kong. His research interests include machine learning, learning theory and stochastic optimization. He has published papers in prestigious journals and conference proceedings, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Information Theory, Journal of Machine Learning Research, COLT, ICLR, ICML and NeurIPS. He is an associate editor for Machine Learning, Transactions on Machine Learning Research, IEEE Transactions on Neural Networks and Learning Systems, and an area chair for NeurIPS, ICLR and AISTATS.

# Supplemental Material for "Towards Better Generalization Bounds of Stochastic Optimization for Nonconvex Learning"

Yunwen Lei

**Abstract**

In the Supplemental Material, we prove some theoretical results stated in the main text. The Supplemental Material consists of six parts: the proofs on the upper bounds of uniform convergence, the proofs on dimension-free bounds, the proofs on the lower bounds of the uniform convergence, the proofs of SGD for quasi-weak convex problems, the proofs for stochastic variance reduced optimization, and the proofs of a concentration inequality for an empirical process.

## A. Proof on Upper Bounds of Uniform Convergence

### A. Definitions and Classical Lemmas

The uniform convergence depends on the complexity of functions. Two classical complexity measures of function classes are Rademacher complexities and covering numbers.

**Definition 1** (Rademacher complexity)**.** Let $\mathcal{F} = \{f : \mathcal{Z} \mapsto \mathbb{R}\}$ be a function class. Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher variables with $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$. We define the empirical Rademacher complexity as follows

$$\mathfrak{R}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i f(z_i).$$

**Definition 2** (Covering number)**.** Let $(\mathcal{G}, d)$ be a metric space and set $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon > 0$, a set $\mathcal{F}^\triangle \subset \mathcal{F}$ is called an $\epsilon$-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$ we can find an element $g \in \mathcal{F}^\triangle$ satisfying $d(f, g) \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the cardinality of the minimal $\epsilon$-cover of $\mathcal{F}$:

$$\mathcal{N}(\epsilon, \mathcal{F}, d) := \min\Big\{|\mathcal{F}^\triangle| : \mathcal{F}^\triangle \text{ is an } \epsilon\text{-cover of } \mathcal{F}\Big\}.$$

Dudley's entropy integral provides a connection between these two complexity measures. We consider the refined entropy integral in Srebro et al. (2010).

**Lemma A.1** (Dudley's entropy integral)**.** *Let $\mathcal{F} = \{f : \mathcal{Z} \mapsto \mathbb{R}\}$ be a function class with $\sup_{f\in\mathcal{F}} d_S(f, 0) \leq D$ and $S = \{z_1, \ldots, z_n\}$, where $d_S$ is a pseudometric on $\mathcal{F}$ defined as follows*

$$d_S(f, g) := \Big(\frac{1}{n}\sum_{i=1}^n (f(z_i) - g(z_i))^2\Big)^{\frac{1}{2}}.$$

*Then, there holds*

$$\mathfrak{R}_S(\mathcal{F}) \leq \inf_{\alpha\geq 0}\Big\{4\alpha + 12\int_\alpha^D \sqrt{\frac{\log \mathcal{N}(r, \mathcal{F}, d_S)}{n}}\mathrm{d}r\Big\}.$$

The following classical result gives estimates on the covering numbers of a Euclidean ball.

**Lemma A.2** (Pisier 1999)**.** *Define the metric $d_2(\mathbf{w}, \mathbf{w}') := \|\mathbf{w} - \mathbf{w}'\|_2$ over ball $B_R$. Then* $\log \mathcal{N}(r, B_R, d_2) \leq d \log(3R/r)$.

The following lemma gives a Bernstein inequality for random variables taking values in a Hilbert space.

**Lemma A.3** (Smale and Zhou 2007)**.** *Let $H$ be a Hilbert space with the norm $\|\cdot\|$ and let $\xi$ be a random variables with values in $H$. Assume $\|\xi\| \leq \widetilde{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}[\|\xi\|^2]$. Let $\{\xi_i\}_{i=1}^n$ be $n$ independent draws of $\xi$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have*

$$\Big\|\frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbb{E}[\xi_i])\Big\| \leq \frac{2\widetilde{M}\log(2/\delta)}{n} + \Big(\frac{2\sigma^2(\xi)\log(2/\delta)}{n}\Big)^{\frac{1}{2}}.$$

The following lemma gives the self-bounding property of a smooth and nonnegative function.

**Lemma A.4** (Srebro et al. 2010)**.** *If for all $z$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and $L$-smooth, then $\|\nabla f(\mathbf{w}; z)\|_2^2 \leq 2Lf(\mathbf{w}; z)$.*

Our analysis requires to estimate some integrals, and the following two lemmas are useful.

**Lemma A.5.** *Let $a > 0$. Then*

$$\int_a^\infty \exp(-x^2)\mathrm{d}x \leq \frac{\sqrt{\pi}\exp(-a^2)}{2}.$$

*Proof.* A standard result about Gaussian integral shows that

$$\int_0^a \exp(-x^2)\mathrm{d}x \geq \frac{\sqrt{\pi}}{2}\big(1-\exp(-a^2)\big)^{\frac{1}{2}}.$$

It then follows from $\int_0^\infty \exp(-x^2)\mathrm{d}x = \sqrt{\pi}/2$ that

$$\begin{aligned}\int_a^\infty \exp(-x^2)\mathrm{d}x &= \frac{\sqrt{\pi}}{2} - \int_0^a \exp(-x^2)\mathrm{d}x \leq \frac{\sqrt{\pi}}{2}\Big(1-\big(1-\exp(-a^2)\big)^{\frac{1}{2}}\Big)\\ &= \frac{\sqrt{\pi}}{2}\frac{1-\big(1-\exp(-a^2)\big)}{1+\big(1-\exp(-a^2)\big)^{\frac{1}{2}}} \leq \frac{\sqrt{\pi}\exp(-a^2)}{2}.\end{aligned}$$

The proof is completed. □

**Lemma A.6.** *Let $a, b > 0$. Then*

$$\int_0^b \log^{\frac{1}{2}}(a/\epsilon)\mathrm{d}\epsilon \leq b\log^{\frac{1}{2}}(a/b) + 2^{-1}b\sqrt{\pi}.$$

*Proof.* Let $x = \log^{\frac{1}{2}}(a/\epsilon)$, we know $\epsilon = a\exp(-x^2)$. It then follows from integrating by parts that

$$\begin{aligned}\int_0^b \log^{\frac{1}{2}}(a/\epsilon)\mathrm{d}\epsilon &= a\int_\infty^{\log^{\frac{1}{2}}(a/b)} x\mathrm{d}\exp(-x^2) = ax\cdot\exp(-x^2)\Big|_\infty^{\log^{\frac{1}{2}}(a/b)} - a\int_\infty^{\log^{\frac{1}{2}}(a/b)}\exp(-x^2)\mathrm{d}x\\ &= a\log^{\frac{1}{2}}(a/b)\exp(-\log(a/b)) + a\int_{\log^{\frac{1}{2}}(a/b)}^\infty \exp(-x^2)\mathrm{d}x \leq b\log^{\frac{1}{2}}(a/b) + \frac{a\sqrt{\pi}}{2}\exp(-\log(a/b)),\end{aligned}$$

where we have used Lemma A.5. The proof is completed. □

### *B. Bounds on Square Norm of Gradients*

In this subsection, we present results on relating $L_S(\mathbf{w})$ to $L(\mathbf{w})$, which are defined as follows

$$L_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i)\|_2^2, \quad L(\mathbf{w}) := \mathbb{E}_z[\|\nabla f(\mathbf{w}; z)\|_2^2]. \tag{A.1}$$

This connection is illustrated below.

**Lemma A.7.** *Let Assumptions 1, 3 hold. For any $x > 0$, with probability at least $1 - 2\exp(-x)$ the following inequalities hold simultaneously for any $\mathbf{w} \in B_R$*

$$\begin{aligned}L_S(\mathbf{w}) - 2L(\mathbf{w}) &\lesssim \frac{G^2 d\log(LR/G)}{n} + \frac{G^2(x+\log\log n)}{n},\\ L(\mathbf{w}) - 2L_S(\mathbf{w}) &\lesssim \frac{G^2 d\log(LR/G)}{n} + \frac{G^2(x+\log\log n)}{n}.\end{aligned} \tag{A.2}$$

To prove Lemma A.7, we introduce the following lemma to be proved in Section F, which is a variant of Theorem 6.1 in Bousquet (2002).

**Definition 3** (Sub-root function)**.** We say $\phi : [0,\infty) \mapsto [0,\infty]$ is a sub-root function if it is non-decreasing, not identically zero and $r \mapsto \phi(r)/\sqrt{r}$ is non-increasing. The fixed point of $\phi$ is the unique point $r^*$ such that $\phi(r^*) = r^*$.

**Lemma A.8.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ to $[0, b]$. Let $\phi_n$ be a sub-root function such that*

$$\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^n f(z_i)\leq r}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\Big] \leq \phi_n(r),$$

*where $\epsilon_i$ are Rademacher variables. Define $r_n^*$ as the solution of the equation $\phi_n(r) = r$, and denote $r_0 = b(x+6\log\log n)/n$. Then for any $x > 0$, with probability at least $1 - 2\exp(-x)$, the following inequalities hold simultaneously for all $f \in \mathcal{F}$*

$$\frac{1}{n}\sum_{i=1}^n f(z_i) \le 2\mathbb{E}_z[f(z)] + 108r_n^* + 20r_0 + 38\sqrt{r_0 r_n^*}, \tag{A.3}$$

$$\mathbb{E}_z[f(z)] \le \frac{2}{n}\sum_{i=1}^n f(z_i) + 106r_n^* + 48r_0. \tag{A.4}$$

*Proof of Lemma A.7.* For any $\mathbf{w}$, we introduce $\tilde{f}_{\mathbf{w}}(z) := \|\nabla f(\mathbf{w}; z)\|_2^2$, and

$$\widetilde{\mathcal{F}}_R = \big\{z \mapsto \tilde{f}_{\mathbf{w}}(z) : \mathbf{w} \in B_R\big\}, \quad \widetilde{\mathcal{F}}_{R,r} = \big\{z \mapsto \tilde{f}_{\mathbf{w}}(z) : \mathbf{w} \in B_R, L_S(\mathbf{w}) \le r\big\}.$$

We define a metric $\tilde{d}_S$ over $\widetilde{\mathcal{F}}_R$ by

$$\tilde{d}_S(\tilde{f}_{\mathbf{w}}, \tilde{f}_{\mathbf{w}'}) = \Big(\frac{1}{n}\sum_{i=1}^n \big(\|\nabla f(\mathbf{w}; z_i)\|_2^2 - \|\nabla f(\mathbf{w}'; z_i)\|_2^2\big)^2\Big)^{\frac{1}{2}}.$$

For any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ with $L_S(\mathbf{w}) \le r, L_S(\mathbf{w}') \le r$, we know

$$\begin{aligned}
\tilde{d}_S^2(\tilde{f}_{\mathbf{w}}, \tilde{f}_{\mathbf{w}'}) &= \frac{1}{n}\sum_{i=1}^n \big(\|\nabla f(\mathbf{w}; z_i)\|_2 - \|\nabla f(\mathbf{w}'; z_i)\|_2\big)^2\big(\|\nabla f(\mathbf{w}; z_i)\|_2 + \|\nabla f(\mathbf{w}'; z_i)\|_2\big)^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i) - \nabla f(\mathbf{w}'; z_i)\|_2^2\big(\|\nabla f(\mathbf{w}; z_i)\|_2^2 + \|\nabla f(\mathbf{w}'; z_i)\|_2^2\big) \\
&\le \frac{2L^2\|\mathbf{w} - \mathbf{w}'\|_2^2}{n}\sum_{i=1}^n \big(\|\nabla f(\mathbf{w}; z_i)\|_2^2 + \|\nabla f(\mathbf{w}'; z_i)\|_2^2\big) \\
&= 2L^2\|\mathbf{w} - \mathbf{w}'\|_2^2(L_S(\mathbf{w}) + L_S(\mathbf{w}')) \le 4L^2\|\mathbf{w} - \mathbf{w}'\|_2^2 r,
\end{aligned}$$

where we have used $(a+b)^2 \le 2a^2 + 2b^2$. That is, $\tilde{d}_S(\tilde{f}_{\mathbf{w}}, \tilde{f}_{\mathbf{w}'}) \le 2Lr^{\frac{1}{2}}\|\mathbf{w} - \mathbf{w}'\|_2$. It then follows from Lemma A.2 that

$$\log\mathcal{N}\big(\epsilon, \widetilde{\mathcal{F}}_{R,r}, \tilde{d}_S\big) \le \log\mathcal{N}\big(\epsilon/(2Lr^{\frac{1}{2}}), B_R, d_2\big) \le d\log\big(6LRr^{\frac{1}{2}}\epsilon^{-1}\big).$$

Note that for any $\mathbf{w}$ with $L_S(\mathbf{w}) \le r$, we have

$$\tilde{d}_S^2(\tilde{f}_{\mathbf{w}}, 0) = \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i)\|_2^4 \le \frac{G^2}{n}\sum_{i=1}^n \|\nabla f(\mathbf{w}; z_i)\|_2^2 = G^2 L_S(\mathbf{w}) \le G^2 r.$$

It then follows from Lemma A.1 and Lemma A.6 that

$$\mathfrak{R}_S(\widetilde{\mathcal{F}}_{R,r}) \le \frac{12\sqrt{d}}{\sqrt{n}}\int_0^{G\sqrt{r}} \log^{\frac{1}{2}}\big(6LRr^{\frac{1}{2}}\epsilon^{-1}\big)\mathrm{d}\epsilon \le \frac{12\sqrt{d}}{\sqrt{n}}\Big(G\sqrt{r}\log^{\frac{1}{2}}(6LR/G) + 2^{-1}G\sqrt{r\pi}\Big) := \phi_n(r). \tag{A.5}$$

It is clear that $\phi_n$ is a sub-root function and the fixed-point $r_n^*$ is

$$r_n^* = \frac{12^2G^2d}{n}\Big(\log^{\frac{1}{2}}(6LR/G) + 2^{-1}\sqrt{\pi}\Big)^2 \lesssim \frac{G^2 d\log(LR/G)}{n}.$$

The stated bounds then follow directly from Lemma A.8 with $\mathcal{F} = \widetilde{\mathcal{F}}_R$ and $b = G^2$. The proof is completed. □

### C. Proof of Theorem 1

*Proof of Theorem 1.* Let $\epsilon > 0$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be an $\epsilon$-cover of $B_R$ under the metric $d_2(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|_2$. Then, Lemma A.2 shows that

$$\log m \le d\log(3R/\epsilon). \tag{A.6}$$

We now consider any $j \in [n]$ and define $\xi_i = \nabla f(\mathbf{w}_j; z_i)$. Then, it is clear that $\|\xi_i\|_2 \le G$. Furthermore, by the definition of $L$ we know

$$\mathbb{E}_{z_i}[\|\xi_i\|_2^2] = \mathbb{E}_{z_i}[\|\nabla f(\mathbf{w}_j; z_i)\|_2^2] = L(\mathbf{w}_j).$$

We then apply Lemma A.3 and get the following inequality with probability at least $1 - \delta/m$

$$\begin{aligned}
\big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 &= \Big\|\frac{1}{n}\big(\sum_{i=1}^n \xi_i - \mathbb{E}_{z_i}[\xi_i]\big)\Big\|_2 \le \frac{2G\log(2m/\delta)}{n} + \Big(\frac{2L(\mathbf{w}_j)\log(2m/\delta)}{n}\Big)^{\frac{1}{2}} \\
&\le \frac{2G\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n} + \Big(\frac{2L(\mathbf{w}_j)\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n}\Big)^{\frac{1}{2}}.
\end{aligned}$$

By the union bounds of probability, with probability at least $1-\delta$ the following inequality holds simultaneously for all $j \in [m]$

$$\big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 \leq \frac{2G\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n} + \Big(\frac{2L(\mathbf{w}_j)\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n}\Big)^{\frac{1}{2}}. \tag{A.7}$$

For any $\mathbf{w} \in B_R$, by the construction of $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$, we know there exists an $j \in [m]$ such that $\|\mathbf{w} - \mathbf{w}_j\|_2 \leq \epsilon$ and therefore

$$\begin{aligned}
\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 &\leq \big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 + \big\|\big(\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big) - \big(\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big)\big\|_2 \\
&\leq \big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 + \|\nabla F_S(\mathbf{w}_j) - \nabla F_S(\mathbf{w})\|_2 + \|\nabla F(\mathbf{w}_j) - \nabla F(\mathbf{w})\|_2 \\
&\leq \big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 + 2L\|\mathbf{w} - \mathbf{w}_j\| \leq \big\|\nabla F_S(\mathbf{w}_j) - \nabla F(\mathbf{w}_j)\big\|_2 + 2L\epsilon \\
&\leq \frac{2G\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n} + \Big(\frac{2L(\mathbf{w}_j)\big(\log(2/\delta) + d\log(3R/\epsilon)\big)}{n}\Big)^{\frac{1}{2}} + 2L\epsilon,
\end{aligned} \tag{A.8}$$

where we have used Eq. (A.7) in the last step. By the standard inequality $(a+b)^2 \leq 2(a^2+b^2)$, we know

$$\begin{aligned}
L(\mathbf{w}_j) = \mathbb{E}_z[\|\nabla f(\mathbf{w}_j; z)\|_2^2] &\leq 2\mathbb{E}_z\big[\|\nabla f(\mathbf{w}_j; z) - \nabla f(\mathbf{w}; z)\|_2^2\big] + 2\mathbb{E}_z[\|\nabla f(\mathbf{w}; z)\|_2^2] \\
&\leq 2L^2\|\mathbf{w}_j - \mathbf{w}\|_2^2 + 2L(\mathbf{w}) \leq 2L^2\epsilon^2 + 2L(\mathbf{w}).
\end{aligned}$$

By Lemma A.7, we further get the following inequality with probability at least $1-\delta$

$$L(\mathbf{w}_j) \lesssim L^2\epsilon^2 + L_S(\mathbf{w}) + \frac{G^2 d\log(LR/G)}{n} + \frac{G^2(\log 1/\delta + \log\log n)}{n}.$$

We plug the above inequality back into Eq. (A.8) and derive the following inequality with probability at least $1-2\delta$ simultaneously for all $\mathbf{w} \in B_R$ (we use the assumption $\log\log n \lesssim d\log(LR/G)$)

$$\begin{aligned}
\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim{}& \frac{G\big(\log(1/\delta) + d\log(R/\epsilon)\big)}{n} + \\
&\Big(L\epsilon + \sqrt{L_S(\mathbf{w})} + \frac{G\sqrt{d}\log^{\frac{1}{2}}(LR/G) + G\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}}\Big)\Big(\frac{\log(1/\delta) + d\log(R/\epsilon)}{n}\Big)^{\frac{1}{2}} + L\epsilon.
\end{aligned}$$

We choose $\epsilon = 1/n$ and derive

$$\begin{aligned}
\big\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\big\|_2 \lesssim{}& \frac{G\big(\log(1/\delta) + d\log(Rn)\big)}{n} + \frac{L}{n} \\
&+ \Big(L/n + \sqrt{L_S(\mathbf{w})} + \frac{G\sqrt{d}\log^{\frac{1}{2}}(LR/G) + G\log^{\frac{1}{2}}(1/\delta)}{\sqrt{n}}\Big)\Big(\frac{\log(1/\delta) + d\log(Rn)}{n}\Big)^{\frac{1}{2}}.
\end{aligned}$$

Since $L \lesssim Gd \lesssim Gn$, we know $d\log(LR/G) + \log(1/\delta) \lesssim \log(1/\delta) + d\log(Rn)$ and therefore get the stated bound. The proof is completed. □

**Remark A.1.** In the proof, we conduct the localization analysis to study the convergence of $\nabla f(\mathbf{w}; z)$ to its expectation. In the literature, some works considered the excess loss $f(\mathbf{w}; z) - f(\mathbf{w}^*; z)$ or the excess gradient $\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}^*; z)$ in their localization analysis (Bartlett et al., 2005; Zhang et al., 2017). Their motivation is to use a Bernstein-type condition. Indeed, the analysis in Bartlett et al. (2005) considers the following Bernstein condition

$$\mathbb{E}_z[(f(\mathbf{w}; z) - f(\mathbf{w}^*; z))^2] \lesssim F(\mathbf{w}) - F(\mathbf{w}^*), \tag{A.9}$$

while the analysis in Zhang et al. (2017) considers the Bernstein condition in Eq. (5). Typically, one requires a convexity assumption to satisfy these Bernstein conditions. For example, the paper (Bartlett et al., 2006) introduces the modulus of convexity to show Eq. (A.9), while the paper (Zhang et al., 2017) uses the convexity of $\mathbf{w} \mapsto f(\mathbf{w}; z)$ to show Eq. (5). As we consider nonconvex problems, we do not have these Bernstein conditions. This explains why we conduct the analysis directly on $f$ instead of the excess gradient $z \mapsto \nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}^*; z)$.

## B. Proof of Theorem 3

The following lemma is the Talagrand's inequality to control the uniform deviation between expectation and empirical average by incorporating the variance information (Bartlett et al., 2005). Let $\mathrm{Var}[X]$ denote the variance of a random variable $X$.

**Lemma B.1** (Bartlett et al. 2005)**.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ into $[-b, b]$. Assume there is some $r > 0$ such that $\mathrm{Var}[f(Z)] \leq r$ for any $f \in \mathcal{F}$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$\sup_{f\in\mathcal{F}}\Big(\mathbb{E}_Z[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\Big) \leq 4\mathbb{E}\big[\mathfrak{R}_S(\mathcal{F})\big] + \Big(\frac{2r\log(2/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{8b\log(2/\delta)}{3n}.$$

The following lemma shows that the Gaussian complexity can be bounded by covering numbers from below.

**Lemma B.2** (Sudakov minoration inequality (Ledoux and Talagrand, 1991)). *Let $\mathcal{F}$ be a class of real-valued functions, $S = \{z_1, \ldots, z_n\}$ and $g_i$ be a sequence of $N(0,1)$ Gaussian random variables. Then*

$$\mathbb{E}_{\mathbf{g}}\Big[\sup_{f\in\mathcal{F}}\sum_{i\in[n]} g_i f(z_i)\Big] \geq \sqrt{n}\sup_{\epsilon>0}\epsilon\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F},d_S),$$

*where $d_S(f,g) = \big(\frac{1}{n}\sum_{i=1}^n (f(z_i)-g(z_i))^2\big)^{\frac{1}{2}}$.*

The following lemma provides covering number estimates for linear function classes.

**Lemma B.3** (Zhang 2002). *If $\|\phi(x)\|_2 \leq B_\phi$, then for any $\epsilon > 0$, we have*

$$\log\mathcal{N}\big(\epsilon, \{x \leftarrow \langle\phi(x),\mathbf{w}\rangle : \|\mathbf{w}\|_2 \leq R\}, d_{S,\infty}\big) \leq \frac{36R^2B_\phi^2}{\epsilon^2}\log_2\big(6n + 8B_\phi R/\epsilon\big),$$

*where $d_{S,\infty}(f,g) = \max_{i\in[n]}|f(x_i)-g(x_i)|$.*

The following lemma provides a contraction property for Rademacher complexities. It also holds for Gaussian complexities, i.e., with $\epsilon_i$ replaced by standard normal random variables $g_i$.

**Lemma B.4** (Contraction Lemma (Bartlett and Mendelson, 2002)). *Suppose $\tau : \mathbb{R} \mapsto \mathbb{R}$ is $G$-Lipschitz in the sense that $|\tau(t)-\tau(\tilde{t})| \leq G|t-\tilde{t}|$. Then the following inequality holds for any $\mathcal{F}$*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i\tau\big(f(x_i)\big) \leq G\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\epsilon_i f(x_i).$$

The following lemma provides estimates on vector Rademacher complexities for functions with a structure.

**Lemma B.5.** *Suppose $f$ takes the form $f(\mathbf{w};z) = \ell(y,\langle\mathbf{w},\phi(x)\rangle)$, where $\phi : \mathcal{X}\mapsto\mathcal{W}$ is a feature map and $\ell : \mathbb{R}^2 \mapsto \mathbb{R}_+$. Assume $a \mapsto \ell(y,a)$ is $G_\ell$-Lipschitz continuous and $L_\ell$-smooth for all $y$. Then*

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\epsilon\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n\epsilon_i\nabla f(\mathbf{w};z_i)\Big\|_2 &\leq \frac{12}{\sqrt{n}}\\ &+\Big(\frac{24\sqrt{2}B_\phi L_\ell}{n}\mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}}\sum_{i=1}^n g_i\langle\mathbf{w},\phi(x_i)\rangle\Big] + \frac{144\sqrt{2}B_\phi\widehat{V}_{\mathcal{W}}\log_2^{\frac{1}{2}}\big(6n+8\sqrt{n}B_\phi G_\ell\big)}{\sqrt{n}}\Big)\log(G_\ell B_\phi\sqrt{n}/3),\end{aligned}$$

*where $\widehat{V}_{\mathcal{W}} = \sup_{\mathbf{w}\in\mathcal{W}}\Big(\frac{1}{n}\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)\big)^2\Big)^{\frac{1}{2}}$.*

*Proof.* For any $\mathbf{w}\in\mathcal{W}, \mathbf{v}\in B_1$, define

$$h_{\mathbf{w},\mathbf{v}}(z) = \langle\nabla f(\mathbf{w};z),\mathbf{v}\rangle = \ell'(y,\langle\mathbf{w},\phi(x)\rangle)\langle\phi(x),\mathbf{v}\rangle.$$

The following inequality was developed in Lei and Tang (2021)

$$\begin{aligned}&\sum_{i=1}^n\big(h_{\mathbf{w},\mathbf{v}}(z_i)-h_{\mathbf{w}',\mathbf{v}'}(z_i)\big)^2\\ &\leq 2\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)-\ell'(y_i,\langle\mathbf{w}',\phi(x_i)\rangle)\big)^2\langle\phi(x_i),\mathbf{v}\rangle^2 + 2\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w}',\phi(x_i)\rangle)\big)^2\big(\langle\phi(x_i),\mathbf{v}\rangle-\langle\phi(x_i),\mathbf{v}'\rangle\big)^2\\ &\leq 2B_\phi^2\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)-\ell'(y_i,\langle\mathbf{w}',\phi(x_i)\rangle)\big)^2 + 2\Big(\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w}',\phi(x_i)\rangle)\big)^2\Big)\Big(\max_{i\in[n]}\big(\langle\phi(x_i),\mathbf{v}\rangle-\langle\phi(x_i),\mathbf{v}'\rangle\big)^2\Big),\end{aligned}$$

where we used $|\langle\phi(x),\mathbf{v}\rangle| \leq B_\phi$. For any $f,g$, define the following metrics

$$d_S(f,g) := \Big(\frac{1}{n}\sum_{i=1}^n(f(z_i)-g(z_i))^2\Big)^{\frac{1}{2}}, \quad d_{S,\infty}(f,g) := \max_{i\in[n]}\big|f(z_i)-g(z_i)\big|. \tag{B.1}$$

Then, we have

$$d_S\big(h_{\mathbf{w},\mathbf{v}},h_{\mathbf{w}',\mathbf{v}'}\big) \leq \sqrt{2}B_\phi d_S(h^{(1)}_{\mathbf{w}},h^{(1)}_{\mathbf{w}'}) + \sqrt{2}\Big(\frac{1}{n}\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w}',\phi(x_i)\rangle)\big)^2\Big)^{\frac{1}{2}}d_{S,\infty}(h^{(2)}_{\mathbf{v}},h^{(2)}_{\mathbf{v}'}),$$

where we introduce

$$h^{(1)}_{\mathbf{w}}(z) := \ell'(y, \langle \mathbf{w}, \phi(x)\rangle) \quad \text{and} \quad h^{(2)}_{\mathbf{v}}(x) = \langle \phi(x), \mathbf{v}\rangle. \tag{B.2}$$

That is, to build an $\epsilon$-cover of $\big\{z \mapsto h_{\mathbf{w},\mathbf{v}}(z) : \mathbf{w} \in \mathcal{W}, \mathbf{v} \in B_1\big\}$ w.r.t. the metric $d_S$, it suffices to build an $\epsilon/(2\sqrt{2}B_\phi)$-cover of $\big\{z \mapsto \ell'(y, \langle \mathbf{w}, \phi(x)\rangle)\big\}$ w.r.t. the metric $d_S$, and also an $\epsilon/(2\sqrt{2}\widehat{V}_{\mathcal{W}})$-cover of $\big\{x \mapsto \langle \phi(x), \mathbf{v}\rangle : \mathbf{v} \in B_1\big\}$ w.r.t. the metric $d_{S,\infty}$. Note the latter two function classes are indexed by $\mathbf{w}$ and $\mathbf{v}$, respectively. Therefore, we have

$$\begin{aligned}
&\log \mathcal{N}\Big(\epsilon, \big\{z \mapsto h_{\mathbf{w},\mathbf{v}}(z) : \mathbf{w} \in \mathcal{W}, \mathbf{v} \in B_1\big\}, d_S\Big) \leq \\
&\log \mathcal{N}\Big(\epsilon/(2\sqrt{2}B_\phi), \big\{z \mapsto \ell'(y, \langle \mathbf{w}, \phi(x)\rangle) : \mathbf{w} \in \mathcal{W}\big\}, d_S\Big) + \log \mathcal{N}\Big(\epsilon/(2\sqrt{2}\widehat{V}_{\mathcal{W}}), \big\{x \mapsto \langle \phi(x), \mathbf{v}\rangle : \mathbf{v} \in B_1\big\}, d_{S,\infty}\Big).
\end{aligned} \tag{B.3}$$

By Lemma B.2, we know

$$\log^{\frac{1}{2}} \mathcal{N}\Big(\epsilon, \big\{z \mapsto \ell'(y, \langle \mathbf{w}, \phi(x)\rangle) : \mathbf{w} \in \mathcal{W}\big\}, d_S\Big) \leq \frac{1}{\sqrt{n}\epsilon} \mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n g_i \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle)\Big].$$

By Lemma B.3, we know

$$\log^{\frac{1}{2}} \mathcal{N}\Big(\epsilon, \big\{x \mapsto \langle \phi(x), \mathbf{v}\rangle : \mathbf{v} \in B_1\big\}, d_{S,\infty}\Big) \leq \frac{6B_\phi \log_2^{\frac{1}{2}}\big(6n + 8B_\phi/\epsilon\big)}{\epsilon}.$$

We plug the above two inequalities back into Eq. (B.3) and derive

$$\begin{aligned}
&\log^{\frac{1}{2}} \mathcal{N}\Big(\epsilon, \big\{z \mapsto h_{\mathbf{w},\mathbf{v}}(z) : \mathbf{w} \in \mathcal{W}, \mathbf{v} \in B_1\big\}, d_S\Big) \\
&\leq \frac{2\sqrt{2}B_\phi}{\sqrt{n}\epsilon} \mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n g_i \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle)\Big] + \frac{12\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}} \log_2^{\frac{1}{2}}\big(6n + 16\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}}/\epsilon\big)}{\epsilon} \\
&\leq \frac{2\sqrt{2}B_\phi L_\ell}{\sqrt{n}\epsilon} \mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n g_i \langle \mathbf{w}, \phi(x_i)\rangle\Big] + \frac{12\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}} \log_2^{\frac{1}{2}}\big(6n + 16\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}}/\epsilon\big)}{\epsilon},
\end{aligned}$$

where the last inequality follows from the contraction principle for Gaussian complexity together with the $L_\ell$-Lipschitz continuity of $\ell'$ (Lemma B.4). Note that $|h_{\mathbf{w},\mathbf{v}}(z)| \leq G_\ell \|\phi(x)\|_2 \|\mathbf{v}\|_2 \leq G_\ell B_\phi$. Therefore, we can apply Lemma A.1 with $\alpha = 3/\sqrt{n}$ to derive the following inequality

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon \sup_{\mathbf{w}\in\mathcal{W}, \mathbf{v}\in B_1} \sum_{i=1}^n \epsilon_i h_{\mathbf{w},\mathbf{v}}(z_i) \\
&\leq \frac{12}{\sqrt{n}} + \frac{12}{\sqrt{n}} \int_{3/\sqrt{n}}^{G_\ell B_\phi} \Big(\frac{2\sqrt{2}B_\phi L_\ell}{\sqrt{n}\epsilon} \mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n g_i \langle \mathbf{w}, \phi(x_i)\rangle\Big] + \frac{12\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}} \log_2^{\frac{1}{2}}\big(6n + 8\sqrt{n}B_\phi \widehat{V}_{\mathcal{W}}\big)}{\epsilon}\Big) d\epsilon \\
&\leq \frac{12}{\sqrt{n}} + \Big(\frac{24\sqrt{2}B_\phi L_\ell}{n} \mathbb{E}_g\Big[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^n g_i \langle \mathbf{w}, \phi(x_i)\rangle\Big] + \frac{144\sqrt{2}B_\phi \widehat{V}_{\mathcal{W}} \log_2^{\frac{1}{2}}\big(6n + 8\sqrt{n}B_\phi \widehat{V}_{\mathcal{W}}\big)}{\sqrt{n}}\Big) \int_{3/\sqrt{n}}^{G_\ell B_\phi} \frac{1}{\epsilon} d\epsilon.
\end{aligned}$$

The proof is completed by noting that $\int_{3/\sqrt{n}}^{G_\ell B_\phi} \frac{1}{\epsilon} d\epsilon = \log(G_\ell B_\phi \sqrt{n}/3)$ and $\widehat{V}_{\mathcal{W}} \leq G_\ell$. □

The following lemma provides estimates on the local Gaussian complexity for linear function classes (Bartlett et al., 2005). The original estimate considers Rademacher variables. It can be directly checked from the proof that it also holds for Gaussian variables.

**Lemma B.6** (Bartlett et al. 2005)**.** *Let $V(\mathbf{w}) = \mathbb{E}_Z[\langle \mathbf{w}, \phi(x)\rangle^2]$. Let $(\lambda_i)_i$ be the eigenvalue of the operator $\mathbf{v} \mapsto \mathbb{E}_X[\langle \mathbf{v}, \phi(X)\rangle \phi(X)]$ arranged in a nonincreasing order. Then*

$$\frac{1}{n}\mathbb{E}\big[\sup_{\mathbf{w}: V(\mathbf{w})\leq r, \|\mathbf{w}\|\leq R} \sum_{i=1}^n g_i \langle \mathbf{w}, \phi(x_i)\rangle\big] \leq \Big(\frac{2}{n} \min_{h\in\mathbb{N}} \Big(rh + R^2 \sum_{j=h+1}^\infty \lambda_j\Big)\Big)^{\frac{1}{2}},$$

*where $g_i$ are standard Gaussian random variables.*

The following lemma is standard and we omit the proof for brevity.

**Lemma B.7.** *Let $a, b \geq 0$. If $x^2 \leq ax + b$, then $x^2 \leq a^2 + 2b$. On the other hand, if $x = a + \sqrt{b}$ for $a, b \geq 0$, then $ax + b \leq x^2$.*

**Theorem B.8.** *Suppose $f$ takes the form $f(\mathbf{w};z)=\ell(y,\langle\mathbf{w},\phi(x)\rangle)$, where $\phi:\mathcal{X}\mapsto\mathcal{W}$ is a feature map and $\ell:\mathbb{R}^2\mapsto\mathbb{R}_+$. Assume $a\mapsto\ell(y,a)$ is $L_\ell$-smooth for all $y$. Let $B_\ell=(\mathbb{E}_Y(\ell'(Y,0))^2)^{\frac{1}{2}},\sup_x\|\phi(x)\|_2\leq B_\phi$ and $\delta\in(0,1)$. Let $V(\mathbf{w})=\mathbb{E}_X[\langle\mathbf{w},\phi(X)\rangle^2]$. With probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\Big\|_2=\widetilde{O}\Big(B_\phi L_\ell\Big(\frac{1}{n}\min_{h\in\mathbb{N}}\Big(rh+R^2\sum_{j=h+1}^\infty\lambda_j\Big)\Big)^{\frac{1}{2}}+\frac{L_\ell B_\phi^2R\log(1/\delta)}{n}+B_\phi(B_\ell+L_\ell\sqrt{r})\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{1}{2}}\Big).$$

*Proof.* Due to the $L_\ell$-smoothness of $\ell$, we know the following inequality for any $\mathbf{w}\in B_R$

$$|\ell'(y,\langle\mathbf{w},\phi(x)\rangle)|\leq|\ell'(y,0)|+L_\ell|\langle\mathbf{w},\phi(x)\rangle|\leq|\ell'(y,0)|+L_\ell\|\mathbf{w}\|_2\|\phi(x)\|\leq|\ell'(y,0)|+L_\ell RB_\phi.$$

Therefore, $\ell$ is $G_\ell$-Lipschitz continuous in $B_R$ with $G_\ell\lesssim L_\ell B_\phi R$. For any $\mathbf{w},\mathbf{v}$ with $V(\mathbf{w})\leq r,\|\mathbf{v}\|_2\leq1$, we know the variance of $z\mapsto\langle\nabla f(\mathbf{w};z),\mathbf{v}\rangle$ satisfies

$$\begin{aligned}\mathrm{Var}_Z\big[\langle\nabla f(\mathbf{w};Z),\mathbf{v}\rangle\big]&\leq\mathbb{E}_Z[\langle\nabla f(\mathbf{w};Z),\mathbf{v}\rangle^2]\leq\mathbb{E}_Z[\|\nabla f(\mathbf{w};Z)\|_2^2\|\mathbf{v}\|_2^2]\leq\mathbb{E}_Z[\|\nabla f(\mathbf{w};Z)\|_2^2]\\&=\mathbb{E}_Z\big[|\ell'(Y,\langle\mathbf{w},\phi(X)\rangle)|^2\|\phi(X)\|_2^2\big]\leq B_\phi^2\mathbb{E}_Z\big[|\ell'(Y,\langle\mathbf{w},\phi(X)\rangle)|^2\big]\\&\leq2B_\phi^2(\mathbb{E}_Y[(\ell'(Y,0))^2]+L_\ell^2\mathbb{E}_Z[\langle\mathbf{w},\phi(X)\rangle^2])\leq2B_\phi^2(B_\ell^2+L_\ell^2r),\end{aligned}\tag{B.4}$$

where we have used the inequality $|\ell'(y,a)|\leq|\ell'(y,0)|+L_\ell|a-0|$ due to the smoothness of $\ell$ and the standard inequality $(a+b)^2\leq2a^2+2b^2$. Furthermore, there holds

$$|\langle\nabla f(\mathbf{w};Z),\mathbf{v}\rangle|\leq\|\nabla f(\mathbf{w};Z)\|_2\|\mathbf{v}\|_2\leq|\ell'(Y,\langle\mathbf{w},\phi(X)\rangle)|\|\phi(X)\|_2\leq G_\ell B_\phi.$$

We can apply Lemma B.1 to show the following inequality with probability at least $1-\delta$

$$\begin{aligned}&\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big\|\nabla F(\mathbf{w})-\nabla F_S(\mathbf{w})\Big\|_2=\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r,\|\mathbf{v}\|_2\leq1}\Big\langle\mathbb{E}_Z[\nabla f(\mathbf{w};Z)]-\frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{w};z_i),\mathbf{v}\Big\rangle\\&\leq4\mathbb{E}\Big[\mathfrak{R}_S\Big(\big\{z\mapsto\langle\nabla f(\mathbf{w};z),\mathbf{v}\rangle:\mathbf{w}\in B_R,V(\mathbf{w})\leq r,\|\mathbf{v}\|_2\leq1\big\}\Big)\Big]+2B_\phi(B_\ell+L_\ell\sqrt{r})\Big(\frac{\log(2/\delta)}{n}\Big)^{\frac{1}{2}}+\frac{8G_\ell B_\phi\log(2/\delta)}{3n}\\&=\frac{4}{n}\mathbb{E}\sup_{\mathbf{w}\in B_R,V(\mathbf{w})\leq r,\|\mathbf{v}\|_2\leq1}\sum_{i=1}^n\epsilon_i\langle\nabla f(\mathbf{w};z_i),\mathbf{v}\rangle+2B_\phi(B_\ell+L_\ell\sqrt{r})\Big(\frac{\log(2/\delta)}{n}\Big)^{\frac{1}{2}}+\frac{8G_\ell B_\phi\log(2/\delta)}{3n}\\&=\frac{4}{n}\mathbb{E}\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big\|\sum_{i=1}^n\epsilon_i\nabla f(\mathbf{w};z_i)\Big\|_2+2B_\phi(B_\ell+L_\ell\sqrt{r})\Big(\frac{\log(2/\delta)}{n}\Big)^{\frac{1}{2}}+\frac{8G_\ell B_\phi\log(2/\delta)}{3n}.\end{aligned}\tag{B.5}$$

By Lemma B.5 and Lemma B.6, we know

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big\|\sum_{i=1}^n\epsilon_i\nabla f(\mathbf{w};z_i)\Big\|_2\\&\leq\frac{24\sqrt{2}B_\phi L_\ell\log(G_\ell B_\phi\sqrt{n}/3)\mathbb{E}\big[\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\sum_{i=1}^n g_i\langle\mathbf{w},\phi(x_i)\rangle\big]}{n}+\frac{12+144\sqrt{2}B_\phi\mathbb{E}[\widehat{V}_r]\log_2^{\frac{1}{2}}\big(6n+8\sqrt{n}B_\phi G_\ell\big)\log(G_\ell B_\phi\sqrt{n}/3)}{\sqrt{n}}\\&\leq48B_\phi L_\ell\log(G_\ell B_\phi\sqrt{n}/3)\Big(\frac{1}{n}\min_{h\in\mathbb{N}}\Big(rh+R^2\sum_{j=h+1}^\infty\lambda_j\Big)\Big)^{\frac{1}{2}}+\frac{12+144\sqrt{2}B_\phi\mathbb{E}[\widehat{V}_r]\log_2^{\frac{1}{2}}\big(6n+8\sqrt{n}B_\phi G_\ell\big)\log(G_\ell B_\phi\sqrt{n}/3)}{\sqrt{n}},\end{aligned}\tag{B.6}$$

where we introduce

$$\widehat{V}_r:=\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big(\frac{1}{n}\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)\big)^2\Big)^{\frac{1}{2}}.$$

We now estimate $\mathbb{E}[\widehat{V}_r]$. By the standard symmetrization trick (Bartlett et al., 2005), we can relate the uniform deviation by Rademacher complexity as follows

$$\mathbb{E}\Big[\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\Big(\frac{1}{n}\sum_{i=1}^n\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)\big)^2-\mathbb{E}_Z\big[\big(\ell'(y,\langle\mathbf{w},\phi(x)\rangle)\big)^2\big]\Big)\Big]\leq2\mathbb{E}\Big[\sup_{\mathbf{w}\in B_R:V(\mathbf{w})\leq r}\frac{1}{n}\sum_{i=1}^n\epsilon_i\big(\ell'(y_i,\langle\mathbf{w},\phi(x_i)\rangle)\big)^2\Big].\tag{B.7}$$

For any $\mathbf{w}, \mathbf{w}' \in B_R$ with $V(\mathbf{w}) \le r, V(\mathbf{w}') \le r$, we know

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \Big( \big(\ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle)\big)^2 - \big(\ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle)\big)^2 \Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big( \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle) + \ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle) \Big)^2 \Big( \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle) - \ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle) \Big)^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \Big( \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle) + \ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle) \Big)^2 \max_{i\in[n]} \Big( \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle) - \ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle) \Big)^2 \\
&\le 4\widehat{V}_r^2 \max_{i\in[n]} \Big( \ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle) - \ell'(y_i, \langle \mathbf{w}', \phi(x_i)\rangle) \Big)^2 \le 4\widehat{V}_r^2 L_\ell^2 \max_{i\in[n]} |\langle \mathbf{w}, \phi(x_i)\rangle - \langle \mathbf{w}', \phi(x_i)\rangle|^2,
\end{aligned}$$

where the last second inequality follows from the definition of $\widehat{V}_r$ and the inequality $(a+b)^2 \le 2a^2 + 2b^2$. It then follows that (recall the definition of $h_{\mathbf{w}}^{(1)}, h_{\mathbf{w}}^{(2)}$ in Eq. (B.2) and the definition of $d_{S,\infty}$ in Eq. (B.1))

$$d_S((h_{\mathbf{w}}^{(1)})^2, (h_{\mathbf{w}'}^{(1)})^2) \le 2\widehat{V}_r L_\ell d_{S,\infty}(h_{\mathbf{w}}^{(2)}, h_{\mathbf{w}'}^{(2)})$$

and therefore

$$\begin{aligned}
\log \mathcal{N}\Big(\epsilon, \{(h_{\mathbf{w}}^{(1)})^2 : \mathbf{w} \in B_R : V(\mathbf{w}) \le r\}, d_S\Big) &\le \mathcal{N}\Big(\epsilon/(2\widehat{V}_r L_\ell), \{h_{\mathbf{w}}^{(2)} : \mathbf{w} \in B_R : V(\mathbf{w}) \le r\}, d_{S,\infty}\Big) \\
&\le \frac{144\widehat{V}_r^2 L_\ell^2 R^2 B_\phi^2}{\epsilon^2} \log_2\big(6n + 16\widehat{V}_r L_\ell R B_\phi/\epsilon\big),
\end{aligned}$$

where we have used Lemma B.3 on covering numbers for linear function classes. Lemma A.1 with $\alpha = 3/\sqrt{n}$ and $|h_{\mathbf{w}}^{(1)}(z)| \le G_\ell$ then imply that

$$\begin{aligned}
\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in B_R: V(\mathbf{w})\le r} \frac{1}{n}\sum_{i=1}^n \epsilon_i \big(\ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle)\big)^2\Big] &\le \frac{12}{\sqrt{n}} + \frac{144\widehat{V}_r L_\ell R B_\phi \log_2^{\frac{1}{2}}\big(6n + 6\sqrt{n}\widehat{V}_r L_\ell R B_\phi\big)}{\sqrt{n}} \int_{3/\sqrt{n}}^{G_\ell^2} \frac{1}{\epsilon} d\epsilon \\
&\le \frac{12}{\sqrt{n}} + \frac{144\widehat{V}_r L_\ell R B_\phi \log_2^{\frac{1}{2}}\big(6n + 6\sqrt{n} G_\ell L_\ell R B_\phi\big) \log(G_\ell^2\sqrt{n}/3)}{\sqrt{n}}.
\end{aligned}$$

We combine this inequality and Eq. (B.7) and derive

$$\begin{aligned}
\mathbb{E}[\widehat{V}_r^2] &\le \sup_{\mathbf{w}\in B_R: V(\mathbf{w})\le r} \mathbb{E}_Z\big[\big(\ell'(Y, \langle \mathbf{w}, \phi(X)\rangle)\big)^2\big] + 2\mathbb{E}\Big[\sup_{\mathbf{w}\in B_R: V(\mathbf{w})\le r} \frac{1}{n}\sum_{i=1}^n \epsilon_i \big(\ell'(y_i, \langle \mathbf{w}, \phi(x_i)\rangle)\big)^2\Big] \\
&\le 2(B_\ell^2 + L_\ell^2 r) + \frac{24}{\sqrt{n}} + \frac{288(\mathbb{E}[\widehat{V}_r^2])^{\frac{1}{2}} L_\ell R B_\phi \log_2^{\frac{1}{2}}\big(6n + 6\sqrt{n} G_\ell L_\ell R B_\phi\big) \log(G_\ell^2\sqrt{n}/3)}{\sqrt{n}},
\end{aligned}$$

where the last step is due to Eq. (B.4). The above inequality is a quadratic inequality of $\big(\mathbb{E}[\widehat{V}_r^2]\big)^{\frac{1}{2}}$. We can apply Lemma B.7 to show that

$$\mathbb{E}[\widehat{V}_r^2] \lesssim B_\ell^2 + L_\ell^2 r + \frac{1}{\sqrt{n}} + \frac{L_\ell^2 R^2 B_\phi^2 \log_2\big(n + \sqrt{n} G_\ell L_\ell R B_\phi\big) \log_2^2(G_\ell^2\sqrt{n})}{n}.$$

We plug the above inequality back into Eq. (B.6), and derive that

$$\frac{1}{n}\mathbb{E} \sup_{\mathbf{w}\in B_R: V(\mathbf{w})\le r} \Big\| \sum_{i=1}^n \epsilon_i \nabla f(\mathbf{w}; z_i)\Big\|_2 = \widetilde{O}\Big(B_\phi L_\ell \Big(\frac{1}{n}\min_{h\in\mathbb{N}} \Big(rh + R^2 \sum_{j=h+1}^\infty \lambda_j\Big)\Big)^{\frac{1}{2}} + \frac{B_\phi}{\sqrt{n}}\Big(B_\ell + L_\ell\sqrt{r} + \frac{L_\ell B_\phi R}{\sqrt{n}}\Big)\Big).$$

We plug the above inequality back into Eq. (B.5) and derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}
&\sup_{\mathbf{w}\in B_R: V(\mathbf{w})\le r} \Big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\Big\|_2 = \\
&\qquad \widetilde{O}\Big(B_\phi L_\ell \Big(\frac{1}{n}\min_{h\in\mathbb{N}} \Big(rh + R^2 \sum_{j=h+1}^\infty \lambda_j\Big)\Big)^{\frac{1}{2}} + \frac{L_\ell B_\phi^2 R}{n} + B_\phi(B_\ell + L_\ell\sqrt{r})\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{G_\ell B_\phi \log(1/\delta)}{n}\Big).
\end{aligned}$$

The proof is completed by noting that $G_\ell \lesssim L_\ell R B_\phi$. □

To conduct a localization analysis, we require the following uniform localized convergence argument developed in Xu and Zeevi (2024) based on the peeling trick. While the original statement holds for the uniform convergence of function values, it is direct to extend their argument to the uniform convergence of gradients. Recall that $a \vee b = \max\{a, b\}$.

**Lemma B.9** (Uniform localized convergence argument (Xu and Zeevi, 2024)). *For a function class $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ and the functional $\widetilde{V} : \mathcal{W} \mapsto [0, \widetilde{R}]$, assume there is a function $\psi(r;\delta)$, which is non-decreasing w.r.t. $r$ and satisfies that $\forall \delta \in (0,1), \forall r \in [0, \widetilde{R}]$, with probability at least $1-\delta$*

$$\sup_{\mathbf{w}\in\mathcal{W}:\widetilde{V}(\mathbf{w})\le r} \big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\mathbf{w}}(z_i) - \mathbb{E}_Z[\nabla f_{\mathbf{w}}(Z)]\big\|_2 \le \psi(r;\delta).$$

*Then, given any $\delta \in (0,1)$ and $r_0 \in (0, \widetilde{R}]$, with probability at least $1-\delta$, for all $\mathbf{w} \in \mathcal{W}$*

$$\big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\mathbf{w}}(z_i) - \mathbb{E}_Z[\nabla f_{\mathbf{w}}(Z)]\big\|_2 \le \psi\Big(2\widetilde{V}(\mathbf{w}) \vee r_0; \delta/\log_2(2\widetilde{R}/r_0)\Big). \tag{B.8}$$

*Proof of Theorem 3.* We define

$$\psi(r,\delta) := \widetilde{C}\Big(B_\phi L_\ell\Big(\frac{1}{n}\min_{h\in\mathbb{N}}\Big(rh + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big)^{\frac{1}{2}} + \frac{L_\ell B_\phi^2 R\log(1/\delta)}{n} + B_\phi(B_\ell + L_\ell\sqrt{r})\Big(\frac{\log(1/\delta)}{n}\Big)^{\frac{1}{2}}\Big),$$

where $\widetilde{C}$ is a constant which only has a logarithmic dependency on $R, n$ and other parameters. Then, by Theorem B.8, $\psi$ satisfies the condition in Eq. (B.8) with $\widetilde{V} = V$. Therefore, we can apply Lemma B.9 with $\widetilde{V} = V, \widetilde{R} = \sup_{\mathbf{w}\in B_R} V(\mathbf{w})$ and $r_0 = 1/n$ to derive the following inequality uniformly for all $\mathbf{w} \in B_R$

$$\Big\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\Big\|_2 \le \widetilde{C}\Big(B_\phi L_\ell\Big(\frac{1}{n}\min_{h\in\mathbb{N}}\Big(2(V(\mathbf{w})\vee 1/n)h + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big)^{\frac{1}{2}} + \frac{L_\ell B_\phi^2 R\log(\log_2(2\widetilde{R}n)/\delta)}{n} + B_\phi(B_\ell + L_\ell\sqrt{2(V(\mathbf{w})\vee 1/n)})\Big(\frac{\log(\log_2(2\widetilde{R}n)/\delta)}{n}\Big)^{\frac{1}{2}}\Big).$$

The stated bound then follows since $B_\phi(B_\ell + L_\ell\sqrt{2(V(\mathbf{w})\vee 1/n)})n^{-\frac{1}{2}}$ is not the dominating term if we restrict $h \ge 1$. □

*Proof of Theorem 5.* By the polynomial decay, we know

$$\begin{aligned}\tilde{r}(\mathbf{w})h + R^2\sum_{j=h+1}^{\infty}\lambda_j &\le \tilde{r}(\mathbf{w})h + \beta R^2\sum_{j=h+1}^{\infty} j^{-p}\\ &\le \tilde{r}(\mathbf{w})h + \beta R^2\int_h^{\infty} x^{-p}dx = \tilde{r}(\mathbf{w})h + \beta R^2 h^{1-p}(p-1)^{-1}.\end{aligned}$$

We can choose $h = \lceil\big(\frac{\beta R^2}{\tilde{r}(\mathbf{w})}\big)^{\frac{1}{p}}\rceil$. Then, we have

$$\begin{aligned}\tilde{r}(\mathbf{w})h &\lesssim \tilde{r}(\mathbf{w})\Big(\frac{\beta R^2}{\tilde{r}(\mathbf{w})}\Big)^{\frac{1}{p}} = \tilde{r}(\mathbf{w})^{1-\frac{1}{p}}\beta^{\frac{1}{p}}R^{\frac{2}{p}}\\ \frac{\beta R^2 h^{1-p}}{p-1} &\le \frac{\beta R^2}{p-1}\Big(\frac{\beta R^2}{\tilde{r}(\mathbf{w})}\Big)^{\frac{1-p}{p}} = \frac{\tilde{r}(\mathbf{w})^{1-\frac{1}{p}}R^{\frac{2}{p}}\beta^{\frac{1}{p}}}{p-1}.\end{aligned}$$

It then follows that

$$\min_{h\in\mathbb{N}}\Big\{\tilde{r}(\mathbf{w})h + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big\} \lesssim \frac{p\tilde{r}(\mathbf{w})^{1-\frac{1}{p}}\beta^{\frac{1}{p}}R^{\frac{2}{p}}}{p-1}.$$

We then plug this bound into Theorem 3 to get the stated bound. □

## C. Lower Bounds of Uniform Convergence

### A. Proof of Proposition 6

In this subsection, we prove Proposition 6 on lower bounds on the uniform convergence of gradients. To this aim, we first introduce a concentration inequality called the McDiarmid's inequality.

**Lemma C.1** (McDiarmid's inequality (McDiarmid, 1989)). *Let $X_1, \ldots, X_n$ be independent random variables and $g : \mathcal{X}^n \mapsto \mathbb{R}$. Assume for any index $i$ and $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ we have*

$$\big|g(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\big| \le c_i, \tag{C.1}$$

*where $c_i \ge 0, i \in [n]$. Then, for any $a > 0$ we have*

$$\Pr\big\{g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)] \le -a\big\} \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big).$$

*Proof of Proposition 6.* Since $\sigma'(t) = t_+$, we know

$$\nabla f(\mathbf{w}; x) = \begin{pmatrix} \sigma'(w_1 x)x \\ \vdots \\ \sigma'(w_{d-1}x)x \\ -\sigma'(w_d x)x \end{pmatrix} = \begin{pmatrix} (w_1 x)_+ x \\ \vdots \\ (w_{d-1}x)_+ x \\ -(w_d x)_+ x \end{pmatrix}$$

Then, we have

$$\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (w_1 x_i)_+ x_i - \mathbb{E}_X[(w_1 X)_+ X] \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n (w_{d-1} x_i)_+ x_i - \mathbb{E}_X[(w_{d-1} X)_+ X] \\ -\frac{1}{n}\sum_{i=1}^n (w_d x_i)_+ x_i + \mathbb{E}_X[(w_d X)_+ X] \end{pmatrix}.$$

We choose $w_1 = \ldots = w_{d-1} = v \in [-1, 1]$ and $w_d = 0$. It is clear that $\mathbf{w} \in B_R$ and $(w_d x)_+ = 0$ for any $x$. It then follows that

$$\|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 = (d-1)^{\frac{1}{2}} \Big| \frac{1}{n}\sum_{i=1}^n (v x_i)_+ x_i - \mathbb{E}_X[(vX)_+ X] \Big|$$

and therefore

$$\begin{aligned} \sup_{\mathbf{w} \in B_R} \|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 &\geq (d-1)^{\frac{1}{2}} \sup_{v:|v|\leq 1} \Big| \frac{1}{n}\sum_{i=1}^n (v x_i)_+ x_i - \mathbb{E}_X[(vX)_+ X] \Big| \\ &= (d-1)^{\frac{1}{2}} \sup_{v:|v|\leq 1} \Big| \frac{1}{n}\Big( \sum_{i \in I_+} (v x_i)_+ x_i + \sum_{i \in I_-} (v x_i)_+ x_i \Big) - \mathbb{E}_X[(vX)_+ X] \Big|, \end{aligned}$$

where $I_+ = \{i \in [n] : x_i = 1\}$ and $I_- = \{i \in [n] : x_i = -1\}$. For any $i \in I_+$ we know $(v x_i)_+ x_i = v_+$, and for any $i \in I_-$ we know $(v x_i)_+ x_i = -(-v)_+$. Furthermore, we know

$$\mathbb{E}_X[(vX)_+ X] = \frac{1}{2}\big(v_+ - (-v)_+\big) = \frac{1}{2} v.$$

It then follows that ($|I|$ denotes the cardinality of a set $I$)

$$\begin{aligned} \sup_{\mathbf{w} \in B_R} \|\nabla F_S(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 &\geq (d-1)^{\frac{1}{2}} \sup_{v:|v|\leq 1} \Big| \frac{1}{n}\Big( |I_+| v_+ - |I_-|(-v)_+ \Big) - \frac{v}{2} \Big| \\ &\geq (d-1)^{\frac{1}{2}} \Big| \frac{1}{n}\Big( |I_+| 1_+ - |I_-|(-1)_+ \Big) - \frac{1}{2} \Big| \\ &= (d-1)^{\frac{1}{2}} \Big| \frac{|I_+|}{n} - \frac{1}{2} \Big|, \end{aligned} \tag{C.2}$$

where we have taken $v = 1$ in the second inequality. We now give bounds on $K_n := \big|\frac{|I_+|}{n} - \frac{1}{2}\big|$. Let $\epsilon_i = 2\mathbb{I}_{[x_i = 1]} - 1$, where $\mathbb{I}_{[\cdot]}$ denotes the indicator function, i.e., taking values 1 if the argument is true, and 0 otherwise. Then it is clear that $\epsilon_i$ is a Rademache variable, i.e., taking values in $\{\pm 1\}$ with the same probability. We know

$$\sum_{i=1}^n \epsilon_i = 2\sum_{i=1}^n \mathbb{I}_{[x_i = 1]} - n = 2|I_+| - n.$$

It then follows that $K_n = \frac{1}{2n}\big|\sum_{i=1}^n \epsilon_i\big|$ and

$$\mathbb{E}[K_n] = \frac{1}{2n}\mathbb{E}\Big[\Big|\sum_{i=1}^n \epsilon_i\Big|\Big] \geq \frac{1}{2\sqrt{2n}}, \tag{C.3}$$

where we have used the Khitchine-Kahane inequality $\mathbb{E}_\epsilon|\sum_{i=1}^n \epsilon_i| \geq 2^{-\frac{1}{2}} n^{\frac{1}{2}}$ (Haagerup, 1981). Define $g(\epsilon_1, \ldots, \epsilon_n) = \frac{1}{2n}\big|\sum_{i=1}^n \epsilon_i\big|$. It is clear that for any $\epsilon_1, \ldots, \epsilon_n, \epsilon_i'$, we have

$$\Big| \Big|\sum_{j=1}^n \epsilon_j\Big| - \Big|\sum_{j \in [n]: j \neq i} \epsilon_j + \epsilon_i' \Big| \Big| \leq \Big| \sum_{j \in [n]: j \neq i} \epsilon_j + \epsilon_i - \sum_{j \in [n]: j \neq i} \epsilon_j - \epsilon_i' \Big| \leq 2.$$

Therefore, $g$ satisfies the bounded increment condition in Eq. (C.1) with $c_i = \frac{1}{n}$. By Lemma C.1 with $a = \frac{1}{4\sqrt{2n}}$ and $c_i = \frac{1}{n}$, we know

$$\Pr\Big\{ g(\epsilon_1, \ldots, \epsilon_n) - \mathbb{E}[g(\epsilon_1, \ldots, \epsilon_n)] \leq -\frac{1}{4\sqrt{2n}} \Big\} \leq \exp\Big( -\frac{2}{(4\sqrt{2n})^2 \sum_{i=1}^n n^{-2}} \Big) = \exp\big( -\frac{1}{16} \big).$$

Therefore, with probability at least $1-\exp(-1/16)$, we know $K_n - \mathbb{E}[K_n] > -\frac{1}{4\sqrt{2n}}$, which implies

$$K_n = \mathbb{E}[K_n] + \big(K_n - \mathbb{E}[K_n]\big) \geq \frac{1}{2\sqrt{2n}} - \frac{1}{4\sqrt{2n}} = \frac{1}{4\sqrt{2n}},$$

where we have used Eq. (C.3). This together with Eq. (C.2) implies the stated bound with probability at least $1-\exp(-1/16)$. □

### *B. Lower Bounds in Expectation*

In this subsection, we give lower bounds for the uniform convergence in expectation. We first consider a general $f$, and present lower bounds in terms of covering numbers. For any $f$ we define $\tilde{f}(\mathbf{w};z) = f(\mathbf{w};z) - \mathbb{E}_z[f(\mathbf{w};z)]$. We consider covering numbers w.r.t. the following distance metric

$$\hat{d}_S((\mathbf{w},\mathbf{v}),(\mathbf{w}',\mathbf{v}')) = \Big(\frac{1}{n}\sum_{i\in[n]}(\langle\nabla\tilde{f}(\mathbf{w};z_i),\mathbf{v}\rangle - \langle\nabla\tilde{f}(\mathbf{w}';z_i),\mathbf{v}'\rangle)^2\Big)^{\frac{1}{2}}$$

over the function class

$$\mathcal{F}_{\mathcal{W},B_1} = \big\{z\mapsto\langle\nabla\tilde{f}(\mathbf{w};z),\mathbf{v}\rangle : \mathbf{w}\in\mathcal{W},\mathbf{v}\in B_1\big\}.$$

Eq. (C.4) below shows that the uniform convergence of gradients can be bounded by $\mathbb{E}\big[\sup_{\mathbf{w}\in\mathcal{W}}\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\|_2\big]$ from below, which, according to Eq. (C.5) can be further bounded by covering numbers from below.

**Lemma C.2** (Lower bounds in expectation)**.** *Let $S=\{z_i : i\in[n]\}\subset\mathcal{Z}$. Let $\{\epsilon_i, i\in[n]\}$ be independent Rademacher variables (i.e., $\epsilon_i$ take values in $\{\pm1\}$ with the same probability). Then*

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\Big\|_2\Big] \geq \frac{1}{2n}\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\|_2\Big]. \tag{C.4}$$

*Furthermore, there holds*

$$\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}}\big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\big\|_2\Big] \geq \frac{\sqrt{n}\sup_{\epsilon>0}\big\{\epsilon\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F}_{\mathcal{W},B_1},\hat{d}_S)\big\}}{2\log^{\frac{1}{2}}(n)}. \tag{C.5}$$

**Remark C.1** (Upper bounds in expectation)**.** Since $\tilde{f}(\mathbf{w};z) = f(\mathbf{w};z) - \mathbb{E}_Z[f(\mathbf{w};Z)]$, we know

$$\begin{aligned}\mathbb{E}_Z[\tilde{f}(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n\tilde{f}(\mathbf{w};z_i) &= \mathbb{E}_{Z'}\Big[f(\mathbf{w};Z') - \mathbb{E}_Z[f(\mathbf{w};Z)]\Big] - \frac{1}{n}\sum_{i=1}^n\Big(f(\mathbf{w};z_i) - \mathbb{E}_Z[f(\mathbf{w};Z)]\Big)\\ &= \frac{1}{n}\sum_{i=1}^n\Big(\mathbb{E}_Z[f(\mathbf{w};Z)] - f(\mathbf{w};z_i)\Big) = \mathbb{E}_Z[f(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n f(\mathbf{w};z_i).\end{aligned}$$

It then follows that (by taking gradients)

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\mathbb{E}_Z[\nabla f(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{w};z_i)\Big\|_2\Big] = \mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\mathbb{E}_Z[\nabla\tilde{f}(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big].$$

By the standard symmetrization trick (Bartlett et al., 2005), we know

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\mathbb{E}_Z[\nabla\tilde{f}(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big] \leq \frac{2}{n}\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big].$$

We combine the above two inequalities together and derive

$$\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\mathbb{E}_Z[\nabla f(\mathbf{w};Z)] - \frac{1}{n}\sum_{i=1}^n\nabla f(\mathbf{w};z_i)\Big\|_2\Big] \leq \frac{2}{n}\mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big]. \tag{C.6}$$

Furthermore, by Lemma A.1 we get

$$\begin{aligned}\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}}\big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\big\|_2\Big] &= \mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in B_1}\Big\langle\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i),\mathbf{v}\Big\rangle\Big]\\ &\leq \inf_\alpha\Big\{4n\alpha + 12\sqrt{n}\int_\alpha^{\widetilde{D}}\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F}_{\mathcal{W},B_1},\hat{d}_S)\mathrm{d}\epsilon\Big\},\end{aligned} \tag{C.7}$$

where $\widetilde{D} = \sup_{\mathbf{w}\in\mathcal{W}} \left(\frac{1}{n}\sum_{i=1}^n \|\nabla \tilde{f}(\mathbf{w}; z_i)\|_2^2\right)^{\frac{1}{2}}$.

**Remark C.2** (Tightness of bounds in expectation)**.** Eq. (C.6) shows $\mathbb{E}\big[\sup_{\mathbf{w}} \big\|\nabla F_S(\mathbf{w})–\nabla F(\mathbf{w})\big\|_2\big] \leq \frac{2}{n}\mathbb{E}\big[\sup_{\mathbf{w}}\big\|\sum_{i=1}^n \epsilon_i \nabla \tilde{f}(\mathbf{w}; z_i)\big\|_2\big]$. This shows that the lower bound in Eq. (C.4) is tight up to a constant factor of 4. Furthermore, by Eq. (C.7), we have the following inequality for any $\alpha > 0$

$$\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}} \big\|\sum_{i=1}^n \epsilon_i \nabla \tilde{f}(\mathbf{w}; z_i)\big\|_2\Big] \leq 4n\alpha + 12\sqrt{n}\int_\alpha^{\widetilde{D}} \epsilon^{-1}\epsilon \log^{\frac{1}{2}} \mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\mathrm{d}\epsilon \tag{C.8}$$

$$\begin{aligned}
&\leq 4n\alpha + 12\sqrt{n}\int_\alpha^{\widetilde{D}} \epsilon^{-1} \sup_{\epsilon'>0}\big\{\epsilon' \log^{\frac{1}{2}} \mathcal{N}(\epsilon', \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\big\}\mathrm{d}\epsilon \\
&\leq 4n\alpha + 12\sqrt{n}\sup_{\epsilon'>0}\big\{\epsilon' \log^{\frac{1}{2}} \mathcal{N}(\epsilon', \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\big\}\int_\alpha^{\widetilde{D}} \epsilon^{-1}\mathrm{d}\epsilon \\
&= 4n\alpha + 12\sqrt{n}\sup_{\epsilon>0}\big\{\epsilon \log^{\frac{1}{2}} \mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\big\}\log(\widetilde{D}/\alpha).
\end{aligned} \tag{C.9}$$

We take $\alpha = n^{-\frac{1}{2}}\sup_{\epsilon>0}\big\{\epsilon \log^{\frac{1}{2}} \mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\big\}$ and derive that

$$\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}} \big\|\sum_{i=1}^n \epsilon_i \nabla \tilde{f}(\mathbf{w}; z_i)\big\|_2\Big] = \widetilde{O}\Big(\sqrt{n}\sup_{\epsilon>0}\big\{\epsilon \log^{\frac{1}{2}} \mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{W},B_1}, \hat{d}_S)\big\}\Big), \tag{C.10}$$

which matches the lower bound in Eq. (C.5) up to a logarithmic factor and justifies the tightness of our lower bound.

We now prove Lemma C.2. To this aim, we introduce the following lemma as an extension of the contraction principle for Rademacher complexities.

**Lemma C.3.** *Let $a_1, \ldots, a_n \in \mathbb{R}$ and $\mathcal{F}$ be a class of real-valued functions. Let $S = \{z_1, \ldots, z_n\}$ and $\epsilon_i$ be a sequence of Rademacher random variables. Then*

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\Big\|\sum_{i\in[n]} a_i\epsilon_i \nabla f(z_i)\Big\|_2 \leq \max_{i\in[n]}|a_i|\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\Big\|\sum_{i\in[n]} \epsilon_i \nabla f(z_i)\Big\|_2.$$

*Proof.* We know

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\Big\|\sum_{i\in[n]} a_i\epsilon_i \nabla f(z_i)\Big\|_2 &= \mathbb{E}_\epsilon \sup_{f\in\mathcal{F},\mathbf{v}\in B_1}\Big\langle\sum_{i\in[n]} a_i\epsilon_i \nabla f(z_i), \mathbf{v}\Big\rangle \leq \max_{i\in[n]}|a_i|\mathbb{E}_\epsilon \sup_{f\in\mathcal{F},\mathbf{v}\in B_1}\Big\langle\sum_{i\in[n]} \epsilon_i \nabla f(z_i), \mathbf{v}\Big\rangle \\
&= \max_{i\in[n]}|a_i|\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\Big\|\sum_{i\in[n]} \epsilon_i \nabla f(z_i)\Big\|_2,
\end{aligned}$$

where we have used the contraction principle of Rademacher complexities (Bartlett and Mendelson, 2002)

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\sum_{i\in[n]} a_i\epsilon_i g(z_i) \leq \max_{i\in[n]}|a_i|\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\sum_{i\in[n]} \epsilon_i g(z_i).$$

The proof is completed. □

*Proof of Lemma C.2.* Let $S' = \{z_1', \ldots, z_n'\}$ be independently drawn from $\rho$. By the Jensen's inequality, we know

$$\begin{aligned}
\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \|\sum_{i\in[n]} \epsilon_i \nabla \tilde{f}(\mathbf{w}; z_i)\|_2\Big] &= \mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]} \epsilon_i\big(\nabla f(\mathbf{w}; z_i) - \mathbb{E}_z[\nabla f(\mathbf{w}; z)]\big)\Big\|_2\Big] \\
&= \mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]} \epsilon_i \nabla f(\mathbf{w}; z_i) - \sum_{i\in[n]} \epsilon_i \mathbb{E}_{z_i'}[\nabla f(\mathbf{w}; z_i')]\Big\|_2\Big] \\
&\leq \mathbb{E}_{S,S',\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]} \epsilon_i \nabla f(\mathbf{w}; z_i) - \sum_{i\in[n]} \epsilon_i \nabla f(\mathbf{w}; z_i')\Big\|_2\Big] \\
&= \mathbb{E}_{S,S'}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]} \nabla f(\mathbf{w}; z_i) - \sum_{i\in[n]} \nabla f(\mathbf{w}; z_i')\Big\|_2\Big],
\end{aligned}$$

where we have used the symmetry between $S$ and $S'$ in the last identity. By the triangle inequality, we further know

$$\begin{aligned}
&\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\|_2\Big]\\
&\le \mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]}\nabla f(\mathbf{w};z_i)-n\mathbb{E}_z\nabla f(\mathbf{w};z)\Big\|_2\Big]+\mathbb{E}_{S'}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]}\nabla f(\mathbf{w};z_i')-n\mathbb{E}_z\nabla f(\mathbf{w};z)\Big\|_2\Big]\\
&= 2\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i\in[n]}\nabla f(\mathbf{w};z_i)-n\mathbb{E}_z\nabla f(\mathbf{w};z)\Big\|_2\Big].
\end{aligned}$$

This shows the first inequality. We now prove the second inequality. Let $g_1,\ldots,g_n$ be independent $N(0,1)$ random variables. Furthermore, by the Jensen's inequality we have (note $|g_i|\epsilon_i$ has the same distribution of $g_i$)

$$\mathbb{E}_{\mathbf{g}}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n g_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big]=\mathbb{E}_{\epsilon,\mathbf{g}}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n|g_i|\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big]\le\mathbb{E}_{\mathbf{g}}\max_{i\in[n]}|g_i|\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big],$$

where we have used Lemma C.3. Since $\mathbb{E}_{\mathbf{g}}\max_{i\in[n]}|g_i|\le 2\log^{\frac{1}{2}}(n)$, we get

$$\begin{aligned}
\mathbb{E}_\epsilon\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n\epsilon_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big]&\ge\frac{1}{2\log^{\frac{1}{2}}(n)}\mathbb{E}_{\mathbf{g}}\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\sum_{i=1}^n g_i\nabla\tilde{f}(\mathbf{w};z_i)\Big\|_2\Big]\\
&=\frac{1}{2\log^{\frac{1}{2}}(n)}\mathbb{E}_{\mathbf{g}}\Big[\sup_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in B_1}\Big\langle\sum_{i=1}^n g_i\nabla\tilde{f}(\mathbf{w};z_i),\mathbf{v}\Big\rangle\Big]\\
&=\frac{1}{2\log^{\frac{1}{2}}(n)}\mathbb{E}_{\mathbf{g}}\Big[\sup_{\mathbf{w}\in\mathcal{W},\mathbf{v}\in B_1}\sum_{i=1}^n g_i\big\langle\nabla\tilde{f}(\mathbf{w};z_i),\mathbf{v}\big\rangle\Big]\\
&\ge\frac{\sqrt{n}}{2\log^{\frac{1}{2}}(n)}\sup_{\epsilon>0}\big\{\epsilon\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F}_{\mathcal{W},B_1},\hat{d}_S)\big\},
\end{aligned}$$

where in the last step we have used Lemma B.2. The proof is completed. □

We now present explicit lower and upper bounds on the uniform convergence of gradients for a specific function class of the form in Eq. (8), where we can give dimension-independent bounds. Note Eq. (C.11) holds if $R$ is sufficiently large. The above upper and lower bounds in Eq. (C.12) match up to constant factors.

**Proposition C.4.** *Let* $f(\mathbf{w};z)=\frac{1}{2}(\mathbf{w}^\top x-y)^2$. *If*

$$R\mathbb{E}\Big\|\sum_{i\in[n]}\epsilon_i(x_ix_i^\top-\mathbb{E}[XX^\top])\Big\|_{op}\ge 2\Big\|\sum_{i=1}^n\epsilon_i\big(\mathbb{E}_Z[YX-y_ix_i]\big)\Big\|_2,\tag{C.11}$$

*then*

$$\frac{R}{\sqrt{n}}\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n(x_ix_i^\top-\mathbb{E}[XX^\top])^2\Big\|_{op}^{\frac{1}{2}}\Big]\lesssim\mathbb{E}_S\Big[\sup_{\mathbf{w}\in\mathcal{W}}\Big\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\Big\|_2\Big]\lesssim\frac{R\sup_x\|x\|_2^2}{\sqrt{n}}.\tag{C.12}$$

To prove Proposition C.4, we require the following lemma on lower bounds for operator norm of random matrices.

**Lemma C.5** (Tropp 2016)**.** *Let* $A_1,\ldots,A_n\in\mathbb{R}^{d_1\times d_2}$ *be* $n$ *symmetric matrices. Then*

$$\mathbb{E}\Big[\Big\|\sum_{i\in[n]}\epsilon_iA_i\Big\|_{op}\Big]\ge\Big(\frac{1}{8}\Big\|\sum_{i=1}^nA_i^2\Big\|_{op}\Big)^{\frac{1}{2}}+\frac{\max_i\|A_i\|_{op}}{8},$$

*where* $\|\cdot\|_{op}$ *denotes the operator norm of a matrix.*

*Proof of Proposition C.4.* For $f(\mathbf{w};z)=\frac{1}{2}(\mathbf{w}^\top x-y)^2$, we know

$$\nabla\tilde{f}(\mathbf{w};z)=xx^\top\mathbf{w}-yx-\mathbb{E}_Z[XX^\top\mathbf{w}-YX]=(xx^\top-\mathbb{E}[XX^\top])\mathbf{w}+\mathbb{E}_Z[YX-yx].$$

Therefore, there holds

$$\begin{aligned}\mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z)\Big\|_2 &= \mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i(x_ix_i^\top-\mathbb{E}[XX^\top])\mathbf{w}+\sum_{i\in[n]}\epsilon_i\mathbb{E}_Z[YX-y_ix_i]\Big\|_2\\ &\geq \mathbb{E}_\epsilon \sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i(x_ix_i^\top-\mathbb{E}[XX^\top])\mathbf{w}\Big\|_2-\mathbb{E}_\epsilon\Big\|\sum_{i=1}^n\epsilon_i\big(\mathbb{E}_Z[YX-y_ix_i]\big)\Big\|_2\\ &= R\mathbb{E}_\epsilon\Big\|\sum_{i\in[n]}\epsilon_i(x_ix_i^\top-\mathbb{E}[XX^\top])\Big\|_{\text{op}}-\mathbb{E}_\epsilon\Big\|\sum_{i=1}^n\epsilon_i\big(\mathbb{E}_Z[YX-y_ix_i]\big)\Big\|_2.\end{aligned}$$

By Eq. (C.11) and Lemma C.5, we get

$$\mathbb{E}\sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z)\Big\|_2\geq 2^{-1}R\mathbb{E}\Big\|\sum_{i\in[n]}\epsilon_i(x_ix_i^\top-\mathbb{E}[XX^\top])\Big\|_{\text{op}}\geq R\mathbb{E}\Big[\Big(\frac{1}{32}\Big\|\sum_{i=1}^n(x_ix_i^\top-\mathbb{E}[XX^\top])^2\Big\|_{\text{op}}\Big)^{\frac{1}{2}}\Big].$$

Furthermore, the analysis in Lei and Tang (2021) shows that

$$\frac{1}{n}\mathbb{E}\sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z)\Big\|_2\lesssim\frac{R\sup_x\|x\|_2^2}{\sqrt{n}}.$$

We combine the above two inequalities together, and derive

$$\frac{R}{\sqrt{n}}\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n(x_ix_i^\top-\mathbb{E}[XX^\top])^2\Big\|_{\text{op}}^{\frac{1}{2}}\Big]\lesssim\frac{1}{n}\mathbb{E}\sup_{\mathbf{w}\in B_R}\Big\|\sum_{i\in[n]}\epsilon_i\nabla\tilde{f}(\mathbf{w};z)\Big\|_2\lesssim\frac{R\sup_x\|x\|_2^2}{\sqrt{n}}.$$

The proof is completed by noting Eq. (C.6) and Eq. (C.4) together. □

In this subsection, we consider bounds in expectation. As a comparison, we develop bounds with high probability in the main text. The following lemma shows that one can transfer bounds with high probability to bounds in expectation.

**Lemma C.6.** *Let $X$ be a non-negative random variable and $a>0$. Suppose that for any $\delta\in(0,1)$, with probability at least $1-\delta$, $X<a\log(1/\delta)$. Then, $\mathbb{E}[X]\leq a$.*

*Proof.* Let $t=a\log(1/\delta)$. Then, we have $\delta=\exp(-t/a)$. The assumption then shows that $\Pr\{X\geq t\}\leq\exp(-t/a)$. It then follows that

$$\begin{aligned}\mathbb{E}[X]&=\int_0^\infty\Pr\{X\geq t\}\mathrm{d}t\leq\int_0^\infty\exp(-t/a)\mathrm{d}t=a\int_0^\infty\exp(-t/a)\mathrm{d}(t/a)\\ &=a\int_0^\infty\exp(-t)\mathrm{d}t=-a\int_0^\infty\mathrm{d}\exp(-t)=a.\end{aligned}$$

The proof is completed. □

Based on Lemma C.6, we can directly transfer the high-probability bounds to bounds in expectation. For example, in Theorem 1, we show with probability at least $1-\delta$ that

$$\sup_{\mathbf{w}\in B_R}\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\|_2\lesssim\log(1/\delta)\Big(\frac{Gd\log(Rn)}{n}+\Big(\frac{L_S(\mathbf{w})d\log(Rn))}{n}\Big)^{\frac{1}{2}}\Big).$$

Then, Lemma C.6 shows that

$$\mathbb{E}\Big[\sup_{\mathbf{w}\in B_R}\|\nabla F_S(\mathbf{w})-\nabla F(\mathbf{w})\|_2\Big]\lesssim\frac{Gd\log(Rn)}{n}+\Big(\frac{L_S(\mathbf{w})d\log(Rn))}{n}\Big)^{\frac{1}{2}}.$$

## D. Proof of Theorem 9

In this section, we prove Theorem 9. In our proof, we require to apply concentration inequalities for martingale difference sequences. In Lemma D.1, we present concentration inequalities for real-valued martingales. Part (a) is the classical Azuma-Hoeffding inequality for martingales with bounded increments (Hoeffding, 1963), while Part (b) and Part (c) are Bernstein-type inequalities where the concentration behavior is better quantified in terms of the variance (Zhang, 2005).

**Lemma D.1.** *Let $z_1,\ldots,z_n$ be a sequence of independent random variables. Consider a sequence of functionals $\xi_k(z_1,\ldots,z_k),k=1,\ldots,n$.*

*(a) Assume that $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b_k$ for each $k$. Let $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

$$\sum_{k=1}^{n} \xi_k - \sum_{k=1}^{n} \mathbb{E}_{z_k}[\xi_k] \leq \Big(2\sum_{k=1}^{n} b_k^2 \log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{D.1}$$

*(b) Let $\sigma_n^2 = \sum_{k=1}^{n} \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big]$ be the conditional variance. Assume that $\xi_k - \mathbb{E}_{z_k}[\xi_k] \leq b$ for each $k$. Let $\rho \in (0,1]$ and $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

$$\sum_{k=1}^{n} \xi_k - \sum_{k=1}^{n} \mathbb{E}_{z_k}[\xi_k] \leq \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}. \tag{D.2}$$

*(c) Let $\rho > 0$. With probability at least $1-\delta$ we have*

$$\sum_{k=1}^{n} \xi_k \leq \frac{1}{\rho}\sum_{k=1}^{n} \log \mathbb{E}_{z_k} \exp(\rho\xi_k) + \frac{\log(1/\delta)}{\rho}.$$

Lemma D.2 is the Pinelis-Bernstein inequality for martingale difference sequences in a Hilbert space (Tarres and Yao, 2014).

**Lemma D.2.** *Let $\{\xi_k\}_{k\in\mathbb{N}}$ be a martingale difference sequence in $\mathbb{R}^d$. Suppose that almost surely $\|\xi_k\|_2 \leq B$ and*

$$\sum_{k=1}^{t} \mathbb{E}[\|\xi_k\|_2^2 | \xi_1, \ldots, \xi_{k-1}] \leq \sigma_t^2.$$

*Then, for any $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta$*

$$\max_{1\leq j\leq t} \Big\| \sum_{k=1}^{j} \xi_k \Big\|_2 \leq 2\Big(\frac{B}{3} + \sigma_t\Big)\log(2/\delta).$$

We now present the proof of Theorem 9.

*Proof of Theorem 9.* By Eq. (1) and Assumption 3, we know

$$\|\mathbf{w}_t\|_2 = \Big\| \sum_{k=1}^{t-1} \eta_k \nabla f(\mathbf{w}_k; z_{i_k}) \Big\|_2 \leq \sum_{k=1}^{t-1} \eta_k G. \tag{D.3}$$

Define

$$R_T \;=\; \Big(4\|\mathbf{w}^*\|_2^2 + 2G^2\sum_{t=1}^{T}\eta_t^2\Big)^{\frac{1}{2}} + 16\Big(\frac{2G\eta_1}{3} + \sigma\big(\sum_{t=1}^{T}\eta_t^2\big)^{\frac{1}{2}}\Big)\log(4/\delta) + 8G\widetilde{C}\sqrt{\frac{d\log(R_T' n) + \log(1/\delta)}{n}}\sum_{t=1}^{T}\eta_t \tag{D.4}$$

and $R_T' = G\sum_{t=1}^{T}\eta_t$, where $\widetilde{C}$ is a constant independent of $n, T, L, G$ and $d$ (to be defined later). According to Eq. (1) and Assumption 3, we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2 + 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t})\rangle \\ &\leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2 G^2 + 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t})\rangle.\end{aligned}$$

It then follows that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2 G^2 &+ 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\rangle \\ &+ 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\rangle + 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F(\mathbf{w}_t)\rangle.\end{aligned}$$

According to Assumption 5, we further get

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2 G^2 &+ 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\rangle \\ &+ 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\rangle + 2\eta_t\alpha(F(\mathbf{w}^*) - F(\mathbf{w}_t)).\end{aligned}$$

It then follows that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2 G^2 &+ 2\eta_t\langle \mathbf{w}^* - \mathbf{w}_t, \nabla f(\mathbf{w}_t; z_{i_t}) - \nabla F_S(\mathbf{w}_t)\rangle \\ &+ 2\eta_t\|\mathbf{w}^* - \mathbf{w}_t\|_2\|\nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|_2 + 2\eta_t\alpha(F(\mathbf{w}^*) - F(\mathbf{w}_t)).\end{aligned}$$

We take a summation of the above inequality and get

$$\|\mathbf{w}_{T+1}-\mathbf{w}^*\|_2^2+2\alpha\sum_{t=1}^{T}\eta_t(F(\mathbf{w}_t)-F(\mathbf{w}^*))\le\|\mathbf{w}^*\|_2^2+G^2\sum_{t=1}^{T}\eta_t^2 \\ +2\sum_{t=1}^{T}\eta_t\langle\mathbf{w}^*-\mathbf{w}_t,\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\rangle+2\sum_{t=1}^{T}\eta_t\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2. \quad \text{(D.5)}$$

We define two random variable sequences

$$\xi_t=\eta_t\langle\mathbf{w}^*-\mathbf{w}_t,\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T},$$
$$\xi_t'=\eta_t\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla F_S(\mathbf{w}_t)-\nabla F(\mathbf{w}_t)\|_2\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T},\quad t=1,2,\ldots,T,$$

where $\mathbb{I}_A$ is the indicator function for an event $A$, i.e., $\mathbb{I}_A=1$ if $A$ holds and 0 otherwise. It is clear that $\{\xi_t\}$ is a martingale difference sequence

$$\mathbb{E}_{i_t}[\xi_t]=\eta_t\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T}\mathbb{E}_{i_t}\big[\langle\mathbf{w}^*-\mathbf{w}_t,\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\rangle\big]=0.$$

Assumption 2 implies

$$\begin{aligned}\mathbb{E}_{i_t}[(\xi_t-\mathbb{E}_{i_t}[\xi_t])^2]&=\eta_t^2\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T}\mathbb{E}_{i_t}\Big[\langle\mathbf{w}^*-\mathbf{w}_t,\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\rangle^2\Big]\\&\le\eta_t^2\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T}\|\mathbf{w}^*-\mathbf{w}_t\|_2^2\mathbb{E}_{i_t}\big[\|\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\|_2^2\big]\\&\le\eta_t^2\big(R_T+\|\mathbf{w}^*\|_2\big)^2\sigma^2.\end{aligned}$$

Furthermore, Assumption 3 implies

$$\begin{aligned}|\xi_t|&\le\eta_t\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla f(\mathbf{w}_t;z_{i_t})-\nabla F_S(\mathbf{w}_t)\|_2\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T}\\&\le 2G\eta_t\|\mathbf{w}^*-\mathbf{w}_t\|_2\mathbb{I}_{\|\mathbf{w}_t\|_2\le R_T}\le 2G\eta_t(\|\mathbf{w}^*\|_2+R_T).\end{aligned}$$

According to Lemma D.2, we derive the following inequality with probability at least $1-\delta/2$ ($\eta_t$ is nonincreasing)

$$\max_{t\in[T]}\sum_{k=1}^{t}\xi_k\le 2\Big(\frac{2G\eta_1(\|\mathbf{w}^*\|_2+R_T)}{3}+\big(R_T+\|\mathbf{w}^*\|_2\big)\sigma\big(\sum_{t=1}^{T}\eta_t^2\big)^{\frac{1}{2}}\Big)\log(4/\delta). \quad \text{(D.6)}$$

Furthermore, Theorem 1 and Eq. (3) imply the following inequality with probability at least $1-\delta/2$ simultaneously for all $t\in[T]$ (note $\|\mathbf{w}_t\|_2\le R_T'$ for $t\in[T]$ by Eq. (D.3))

$$\sum_{k=1}^{t}\xi_t'\le\widetilde{C}G\sqrt{\frac{d\log(R_T'n)+\log(1/\delta)}{n}}\sum_{k=1}^{t}\eta_k\|\mathbf{w}^*-\mathbf{w}_k\|_2\mathbb{I}_{\|\mathbf{w}_k\|_2\le R_T}\le\widetilde{C}G(\|\mathbf{w}^*\|_2+R_T)\sqrt{\frac{d\log(R_T'n)+\log(1/\delta)}{n}}\sum_{k=1}^{t}\eta_k, \quad \text{(D.7)}$$

where $\widetilde{C}$ is a universal constant independent of $n,d,L,G$ and $T$. Let $A$ be the event that both Eq. (D.6) and (D.7) hold for $t\in[T]$. The above discussions imply that $\mathbb{P}(A)\ge 1-\delta$. We now assume the event $A$ happens and use mathematical induction to prove $\|\mathbf{w}_t\|_2\le R_T$ under this condition. The case $t=1$ is clear since $\mathbf{w}_1$ is the zero vector. We now assume $\max_{k\in[t]}\|\mathbf{w}_k\|_2\le R_T$ and we want to prove $\|\mathbf{w}_{t+1}\|_2\le R_T$. Since $\mathbb{I}_{\|\mathbf{w}_k\|_2\le R_T}=1$ for $k\in[t]$ we know

$$\xi_k=\eta_k\langle\mathbf{w}^*-\mathbf{w}_k,\nabla f(\mathbf{w}_k;z_{i_k})-\nabla F_S(\mathbf{w}_k)\rangle,$$
$$\xi_k'=\eta_k\|\mathbf{w}^*-\mathbf{w}_k\|_2\|\nabla F_S(\mathbf{w}_k)-\nabla F(\mathbf{w}_k)\|_2,\quad k\in[t].$$

It then follows from Eq. (D.5) that ($T$ replaced by $t$)

$$\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\le\|\mathbf{w}^*\|_2^2+G^2\sum_{k=1}^{t}\eta_k^2+2\sum_{k=1}^{t}\xi_k+2\sum_{k=1}^{t}\xi_k'.$$

It then follows from $\|\mathbf{w}_{t+1}\|_2^2\le 2\|\mathbf{w}^*\|_2^2+2\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2$ that

$$\|\mathbf{w}_{t+1}\|_2^2\le 4\|\mathbf{w}^*\|_2^2+2G^2\sum_{k=1}^{t}\eta_k^2+4\sum_{k=1}^{t}\xi_k+4\sum_{k=1}^{t}\xi_k'.$$

We can plug Eq. (D.6) and Eq. (D.7) back into the above inequality and get

$$\|\mathbf{w}_{t+1}\|_2^2 \leq 4\|\mathbf{w}^*\|_2^2 + 2G^2\sum_{k=1}^t \eta_k^2 + 8\Big(\frac{2G\eta_1(\|\mathbf{w}^*\|_2 + R_T)}{3} + \big(R_T + \|\mathbf{w}^*\|_2\big)\sigma\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big)\log(4/\delta) \\ + 4\widetilde{C}G(\|\mathbf{w}^*\|_2 + R_T)\sqrt{\frac{d\log(R_T' n) + \log(1/\delta)}{n}}\sum_{k=1}^t \eta_k.$$

Since $\|\mathbf{w}^*\|_2 \leq R_T$, we further get

$$\|\mathbf{w}_{t+1}\|_2^2 \leq 4\|\mathbf{w}^*\|_2^2 + 2G^2\sum_{k=1}^t \eta_k^2 + 16R_T\Big(\frac{2G\eta_1}{3} + \sigma\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big)\log(4/\delta) + 8\widetilde{C}GR_T\sqrt{\frac{d\log(R_T' n) + \log(1/\delta)}{n}}\sum_{t=1}^T \eta_t.$$

Note the right hand side of the above inequality is a linear function of $R_T$. For the $R_T$ defined in Eq. (D.4), we know the right hand side of the above inequality is smaller than $R_T^2$ (by Lemma B.7). This proves the case for $k = t+1$ and finishes the proof of showing $\max_{t\in[T]}\|\mathbf{w}_t\|_2 \leq R_T$.

Under the event of $A$, we have $\|\mathbf{w}_t\|_2 \leq R_T$. It follows from (D.5), (D.6) and (D.7) that

$$2\alpha\sum_{t=1}^T \eta_t(F(\mathbf{w}_t) - F(\mathbf{w}^*)) \leq \|\mathbf{w}^*\|_2^2 + G^2\sum_{t=1}^T \eta_t^2 + 2\sum_{t=1}^T \xi_t + 2\sum_{t=1}^T \xi_t' \\ \leq \|\mathbf{w}^*\|_2^2 + G^2\sum_{t=1}^T \eta_t^2 + 4\Big(\frac{2G\eta_1(\|\mathbf{w}^*\|_2 + R_T)}{3} + \big(R_T + \|\mathbf{w}^*\|_2\big)\sigma\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big)\log(4/\delta) \\ + 2\widetilde{C}G(\|\mathbf{w}^*\|_2 + R_T)\sqrt{\frac{d\log(R_T' n) + \log(1/\delta)}{n}}\sum_{t=1}^T \eta_t \leq R_T^2.$$

According to our definition of $R_T$, we know

$$R_T^2 \lesssim \|\mathbf{w}^*\|_2^2 + G^2\sum_{t=1}^T \eta_t^2\log^2(1/\delta) + \frac{G^2(d\log(R_T' n) + \log(1/\delta))}{n}\Big(\sum_{t=1}^T \eta_t\Big)^2.$$

This gives Eq. (20).

For $\eta_t \asymp 1/\sqrt{T}$ we have $\sum_{t=1}^T \eta_t^2 \lesssim 1$ and $\sum_{t=1}^T \eta_t \asymp \sqrt{T}$. It then follows that

$$\alpha\Big(\sum_{t=1}^T \eta_t\Big)^{-1}\sum_{t=1}^T \eta_t(F(\mathbf{w}_t) - F(\mathbf{w}^*)) \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\sum_{t=1}^T \eta_t} + \frac{G^2\sum_{t=1}^T \eta_t^2\log^2(1/\delta)}{\sum_{t=1}^T \eta_t} + \frac{G^2(d\log(R_T' n) + \log(1/\delta))\sum_{t=1}^T \eta_t}{n} \\ \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\sqrt{T}} + \frac{G^2\log^2(1/\delta)}{\sqrt{T}} + \frac{G^2(d\log(R_T' n) + \log(1/\delta))\sqrt{T}}{n}.$$

This gives the stated bound and finishes the proof. □

## E. Proof on Stochastic Variance Reduced Optimization

In this section, we prove Theorem 15 and Theorem 16 on bounds of stochastic variance reduced optimization algorithms.

*Proof of Theorem 15.* To achieve the empirical accuracy $\mathbb{E}[\|\nabla F_S(A(S))\|_2] = O(\epsilon)$, it was shown that SVRG requires $\widetilde{T} = O(n + Ln^{\frac{2}{3}}/\epsilon^2)$ stochastic gradient evaluations (Reddi et al., 2016). According to Assumption 3, we know $\|\mathbf{v}_t\|_2 \leq 3G$ and therefore $\|A(S)\|_2 \leq 3G\widetilde{T} \lesssim Gn + GLn^{\frac{2}{3}}/\epsilon^2$. Eq. (3) then implies

$$\mathbb{E}\Big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\Big] = \widetilde{O}\Big(\frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big)$$

It then follows that

$$\mathbb{E}\Big[\big\|\nabla F(A(S))\big\|_2\Big] \leq \mathbb{E}\Big[\big\|\nabla F_S(A(S))\big\|_2\Big] + \mathbb{E}\Big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\Big] = \epsilon + \widetilde{O}\Big(\frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big).$$

The proof is completed. □

*Proof of Theorem 16.* To achieve the empirical accuracy $\mathbb{E}[\|\nabla F_S(A(S))\|_2] = O(\epsilon)$, it was shown that Spider/SARAH requires $\widetilde{T} = \widetilde{O}\big(\min\{\sigma^3/\epsilon^3, n + \sqrt{n}L/\epsilon^2\}\big)$ stochastic gradient evaluations (Nguyen et al., 2017; Fang et al., 2018). According to Assumption 3, we know $\|\mathbf{v}_t\|_2 \leq 3G$ and therefore $\|A(S)\|_2 \leq 3G\widetilde{T} \lesssim Gn + \sqrt{n}GL/\epsilon^2$. Eq. (3) then implies

$$\mathbb{E}\Big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\Big] = \widetilde{O}\Big(\frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big).$$

It then follows that

$$\mathbb{E}\Big[\big\|\nabla F(A(S))\big\|_2\Big] \leq \mathbb{E}\Big[\big\|\nabla F_S(A(S))\big\|_2\Big] + \mathbb{E}\Big[\big\|\nabla F(A(S)) - \nabla F_S(A(S))\big\|_2\Big] = \epsilon + \widetilde{O}\Big(\frac{\sqrt{d}\mathbb{E}[L_S^{\frac{1}{2}}(A(S))]}{\sqrt{n}}\Big).$$

The proof is completed. □

## F. Proof of Lemma A.8

In this section, we prove Lemma A.8. To this aim, we introduce several necessary lemmas. The following two lemmas show the uniform deviation of an empirical process in terms of local Rademacher complexities. For simplicity, for a function $f : \mathcal{Z} \mapsto \mathbb{R}$ and $S = \{z_1, \ldots, z_n\}$, we denote

$$P_n f = \frac{1}{n}\sum_{i=1}^n f(z_i), \quad Pf = \mathbb{E}_z[f(z)].$$

**Lemma F.1** (Lemma 6.1 in Bousquet (2002))**.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ to $[a, b]$. Assume there exists some $r > 0$ such that for every $f \in \mathcal{F}$, we have $Pf^2 \leq r$. Then, for every $x > 0$, with probability at least $1 - 3\exp(-x)$*

$$\sup_{f\in\mathcal{F}} \big|Pf - P_n f| \leq 6\mathfrak{R}_S(\mathcal{F}) + \sqrt{2rx/n} + \frac{16(b-a)x}{3n}.$$

**Lemma F.2** (Lemma 6.2 in Bousquet (2002))**.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ to $[a, b]$. Let $\mathcal{F}_k$ be a sequence of subsets of $\mathcal{F}$ such that $\sup_{f\in\mathcal{F}_k} Pf^2 \leq K\delta_k$, where $\delta_k = (b-a)2^{-k}$ and $K > 0$. For any $\delta > 0$, denote*

$$x(\delta) = 2\log\Big(\frac{\pi}{\sqrt{2}}\log_2\frac{2(b-a)}{\delta}\Big). \tag{F.1}$$

*Then for all $x > 0$, with probability at least $1 - \exp(-x)$, simultaneously for all $k \geq 0$ and all $f \in \mathcal{F}_k$*

$$\big|Pf - P_n f\big| \leq 6\mathfrak{R}_S(\mathcal{F}_k) + \sqrt{\frac{2K\delta_k(x + x(\delta_k))}{n}} + \frac{16(b-a)(x + x(\delta_k))}{3n}.$$

We now present the proof of Lemma A.8, which follows closely from Section 6.2 in Bousquet (2002). For a sub-root function $\phi$ with the fixed point $r^*$, we have that $\phi(rr^*) \leq \sqrt{rr^*}$ for any $r \geq r^*$. We will use this property several times in the following proof.

*Proof of Lemma A.8.* Eq. (A.4) was proved in Bousquet (2002). We only prove Eq. (A.3). For any $k \geq 0$, denote $\delta_k = b2^{-k}$ and introduce

$$\mathcal{F}_k = \big\{f \in \mathcal{F} : \delta_{k+1} < Pf \leq \delta_k\big\}.$$

It is clear that $Pf^2 \leq bPf \leq b\delta_k$ for any $f \in \mathcal{F}_k, k \geq 0$. According to Lemma F.2, with probability at least $1 - \exp(-x)$, simultaneously for all $k \geq 0$ and for all $f \in \mathcal{F}_k$

$$\big|Pf - P_n f\big| \leq 6\mathfrak{R}_S(\mathcal{F}_k) + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{16b(x + x(\delta_k))}{3n}, \tag{F.2}$$

where $x(\delta)$ is defined in Eq. (F.1). Now we always assume that Eq. (F.2) holds, which happens with probability $1 - \exp(-x)$. Introduce the notation

$$U_k = \delta_k + 6\mathfrak{R}_S(\mathcal{F}_k) + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{16b(x + x(\delta_k))}{3n}.$$

Then for any $f \in \mathcal{F}_k$, we have $P_n f \leq U_k$, which implies $\mathcal{F}_k \subseteq \{f \in \mathcal{F} : P_n f \leq U_k\}$ and therefore

$$\mathbb{E}_\epsilon\Big[\sup_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\Big] \leq \mathbb{E}_\epsilon\Big[\sup_{f:P_n f\leq U_k}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\Big].$$

This together with the definition of $U_k$ implies

$$U_k \leq \delta_k + 6\phi_n(U_k) + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{16b(x + x(\delta_k))}{3n}.$$

Let $k_0$ be the largest number such that $\delta_{k_0} > b/n$, from which we know $\delta_{k_0} \leq 2b/n$. We now consider two cases. In the **first** case, we consider any $k < k_0$. For these $k$, we know $\delta_k \geq 2\delta_{k_0} > 2b/n$ and therefore ($n \geq 5$) (Bousquet, 2002)

$$x(\delta_k) \leq 2\log(\pi/\sqrt{2}\log_2 n) \leq 6\log\log n.$$

If $U_k > r_n^*$, the property of sub-root function then implies $\phi_n(U_k) \leq \sqrt{U_k r_n^*}$. It then follows that

$$U_k \leq \delta_k + 6\phi_n(U_k) + \sqrt{2\delta_k r_0} + 16r_0/3 \leq \delta_k + 6\sqrt{U_k r_n^*} + \sqrt{2\delta_k r_0} + 16r_0/3.$$

Solving this quadratic inequality of $\sqrt{U_k}$ then gives (Lemma B.7 with $x = \sqrt{U_k}$)

$$U_k \leq 36r_n^* + 2\delta_k + \sqrt{8\delta_k r_0} + 32r_0/3 := r_n(\delta_k).$$

In the other case that $U_k \leq r_n^*$, the above inequality still holds. Therefore, we can apply Eq. (F.2) to derive that for all $k \leq k_0$ and $f \in \mathcal{F}_k$

$$\begin{aligned}
&|Pf - P_n f| \leq 6\mathbb{E}_\epsilon\Big[\sup_{f:P_n f \leq r_n(\delta_k)} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\Big] + \sqrt{2\delta_k r_0} + 16r_0/3 \\
&\leq 6\phi_n(r_n(\delta_k)) + \sqrt{2\delta_k r_0} + 16r_0/3 \leq 6\phi_n(r_n(2Pf)) + \sqrt{2\delta_k r_0} + 16r_0/3 \\
&\leq 6\sqrt{r_n^* r_n(2Pf)} + \sqrt{2\delta_k r_0} + 16r_0/3 \leq 6\sqrt{r_n^*}\Big(36r_n^* + 4Pf + 4\sqrt{Pfr_0} + 32r_0/3\Big)^{\frac{1}{2}} + \sqrt{2\delta_k r_0} + 16r_0/3 \\
&\leq 6\sqrt{r_n^*}\Big(36r_n^* + 8Pf + 35r_0/3\Big)^{\frac{1}{2}} + 2\sqrt{Pfr_0} + 16r_0/3 \leq 2\sqrt{Pf}\big(6\sqrt{2r_n^*} + \sqrt{r_0}\big) + 6\sqrt{r_n^*}\Big(36r_n^* + 35r_0/3\Big)^{\frac{1}{2}} + 16r_0/3,
\end{aligned}$$

where we have used the fact that $\delta_k \leq 2Pf$ for $f \in \mathcal{F}_k$ in the second inequality and the fact $\phi_n(r_n(2Pf)) \leq \sqrt{r_n^* r_n(2Pf)}$ due to the sub-root property. It then follows that

$$P_n f \leq Pf + 2\sqrt{Pf}\big(6\sqrt{2r_n^*} + \sqrt{r_0}\big) + 6\sqrt{r_n^*}\Big(36r_n^* + 35r_0/3\Big)^{\frac{1}{2}} + 16r_0/3,$$

which further implies (note $12\sqrt{2} + 6\sqrt{12} \leq 38$)

$$\begin{aligned}
P_n f &\leq Pf + Pf + \big(6\sqrt{2r_n^*} + \sqrt{r_0}\big)^2 + 36r_n^* + 6\sqrt{12r_0 r_n^*} + 16r_0/3 \\
&\leq 2Pf + 72r_n^* + r_0 + 12\sqrt{2r_n^* r_0} + 36r_n^* + 6\sqrt{12r_0 r_n^*} + 7r_0 \leq 2Pf + 108r_n^* + 10r_0 + 38\sqrt{r_0 r_n^*}.
\end{aligned}$$

In the **second** case, we consider $k \geq k_0$. Then $Pf \leq 2b/n$ for any $f \in \mathcal{F}_k$. Then we introduce

$$\widetilde{\mathcal{F}} = \big\{f \in \mathcal{F} : Pf \leq 2b/n\big\}.$$

For any $f \in \widetilde{\mathcal{F}}$, we have $Pf^2 \leq bPf \leq 2b^2/n$. We then apply Lemma F.1 to derive the following inequality with probability at least $1 - 3\exp(-x')$

$$\sup_{f\in\widetilde{\mathcal{F}}} \big|Pf - P_n f| \leq 6\mathfrak{R}_S(\widetilde{\mathcal{F}}) + \sqrt{4b^2x'/n^2} + \frac{16bx'}{3n}. \tag{F.3}$$

We always assume Eq. (F.3) holds, which happens with probability at least $1 - 3\exp(-x')$. It then follows that

$$P_n f \leq \frac{2b}{n} + 6\mathfrak{R}_S(\widetilde{\mathcal{F}}) + \sqrt{4b^2x'/n^2} + \frac{16bx'}{3n} := \widetilde{U}.$$

That is, $\widetilde{\mathcal{F}} \subseteq \{f \in \mathcal{F} : P_n f \leq \widetilde{U}\}$ and therefore ($x' \geq 1$)

$$\widetilde{U} \leq \frac{2b}{n} + 6\mathbb{E}_\epsilon\Big[\sup_{f:P_n f\leq\widetilde{U}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\Big] + \sqrt{4b^2x'/n^2} + \frac{16bx'}{3n} \leq \frac{2b}{n} + 6\phi_n(\widetilde{U}) + \frac{8bx'}{n}.$$

If $\widetilde{U} \leq r_n^*$, then we have $P_n f \leq r_n^*$ by the definition of $\widetilde{U}$. Otherwise, we have $\phi_n(\widetilde{U}) \leq \sqrt{r_n^*\widetilde{U}}$ (by sub-root property) and therefore

$$\widetilde{U} \leq 6\sqrt{r_n^*\widetilde{U}} + \frac{8bx' + 2b}{n}.$$

We solve the above quadratic inequality of $\sqrt{\widetilde{U}}$ and derive $\widetilde{U} \leq 36r_n^* + \frac{16bx'+4b}{n}$ (Lemma B.7 with $x = \sqrt{\widetilde{U}}$). That is,

$$P_n f \leq 36r_n^* + \frac{16bx' + 4b}{n}, \quad \forall f \in \widetilde{U}.$$

We then combine the above two cases, and choose $x' = x + \log 3$ and derive the stated bound with probability at least $1 - \exp(-x)$. The proof is completed. □

## G. Problem Cases with Bounded Gradients

In this section, we provide some special problem cases where Assumption 3 holds.

### A. Shallow Neural Networks

Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be an activation function. Suppose that there exists a $G_\phi > 0$ such that $|\phi'(a)| \leq G_\phi$ for all $a \in \mathbb{R}$. Examples of such activation functions include the ReLU function and the sigmoid function. Consider shallow neural networks with the form

$$\Phi(\mathbf{W}; x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \phi(\langle \mathbf{w}^{(j)}, x \rangle),$$

where $m$ is the number of nodes in the hidden layer, $a_j \in \{-1, 1\}$ indicates the connection weight between the $j$-th node in the hidden layer to the node in the output layer, and $\mathbf{w}^{(j)} \in \mathbb{R}^d$ denotes the connection weight between the $j$th hidden node and the nodes in the input layer. Let $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ with $|\ell'(a, y)| \leq G_\ell$ for any $a \in \mathbb{R}$ and $y$. Examples of such $\ell$ include the logistic loss and the Huber's loss. Consider the following loss function

$$f(\mathbf{W}; z) = \ell(\Phi(\mathbf{W}; x), y).$$

Assume that $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$. Then, we know

$$\nabla \Phi(\mathbf{W}; x) = \frac{1}{\sqrt{m}} \begin{pmatrix} a_1 \phi'(\langle \mathbf{w}^{(1)}, x \rangle) x \\ \vdots \\ a_m \phi'(\langle \mathbf{w}^{(m)}, x \rangle) x \end{pmatrix},$$

from which we know

$$\|\nabla \Phi(\mathbf{W}; x)\|_F^2 = \frac{1}{m} \sum_{j=1}^{m} \big(\phi'(\langle \mathbf{w}^{(j)}, x \rangle)\big)^2 \|x\|_2^2 \leq G_\phi^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. By chain rule, we know $\nabla f(\mathbf{W}; \mathbf{z}) = \ell'(\Phi(\mathbf{W}; x), y) \nabla \Phi(\mathbf{W}; x)$ and therefore

$$\|\nabla f(\mathbf{W}; \mathbf{z})\|_2 \leq |\ell'(\Phi(\mathbf{W}; x), y)| \|\nabla \Phi(\mathbf{W}; x)\|_F \leq G_\phi G_\ell.$$

Therefore, Assumption 3 holds with $G = G_\phi G_\ell$.

### B. Robust Regression

For robust regression, we often consider loss functions of the form $f(\mathbf{w}; z) = \ell(y - \langle \mathbf{w}, x \rangle)$, where $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ is a potentially nonconvex function to improve robustness. A typical example is the Tukey's biweight loss, which is defined as

$$\ell(a) = \begin{cases} 1 - (1 - a^2/a_0^2)^3, & \text{if } |a| \leq a_0 \\ 1, & \text{otherwise,} \end{cases}$$

where $a_0 > 0$ is a hyperparameter. It is clear that for any $|a| \leq a_0$ we have

$$\ell'(a) = \frac{6a}{a_0^2} \Big(1 - \frac{a^2}{a_0^2}\Big)^2 \Longrightarrow |\ell'(a)| \leq \frac{6a}{a_0^2} \leq \frac{6}{a_0}.$$

If we assume $\|x\|_2 \leq 1$, then for any $\mathbf{w}$ and $z$ we have

$$\|\nabla f(\mathbf{w}; z)\|_2 = |\ell'(y - \langle \mathbf{w}, x \rangle)| \|x\|_2 \leq 6/a_0.$$

That is, Assumption 3 holds with $G = 6/a_0$.

### C. Generalized Linear Models

For generalized linear models, we consider loss functions of the form $f(\mathbf{w}; z) = \big(\ell(\mathbf{w}^\top x) - y\big)^2$, where $\ell : \mathbb{R} \mapsto \mathbb{R}$ is a link function. Generalized linear models have shown superior performance as compared to convex formulations in some applications. Standard choices of $\ell$ include the sigmoid link $\ell(a) = (1 + \exp(-a))^{-1}$ and the probit link $\ell(a) = \Phi(a)$, where $\Phi$ is the Gaussian cumulative distribution function. Now we show that under the assumption $\|x\|_2 \leq 1, |y| \leq 1$ and the choice $\ell(a) = (1 + \exp(-a))^{-1}$, Assumption 3 holds. Indeed, we have

$$|\ell(a)| = (1 + \exp(-a))^{-1} \leq 1$$

and

$$\ell'(a) = -\frac{-\exp(-a)}{(1 + \exp(-a))^2} \Longrightarrow |\ell'(a)| = \frac{\exp(-a)}{(1 + \exp(-a))^2} \leq \frac{\exp(-a)}{(2\exp(-a/2))^2} = \frac{1}{4}.$$

It then follows that

$$\|\nabla f(\mathbf{w}; z)\|_2 = 2\big|\ell(\mathbf{w}^\top x) - y\big|\big|\ell'(\mathbf{w}^\top x)\big| \|x\|_2 \leq 1.$$

That is, Assumption 3 holds with $G = 1$.

## REFERENCES


N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in *Advances in Neural Information Processing Systems*, 2010, pp. 2199–2207.

G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1999, vol. 94.

S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive approximation*, vol. 26, no. 2, pp. 153–172, 2007.

O. Bousquet, "Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms," Ph.D. dissertation, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

P. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

L. Zhang, T. Yang, and R. Jin, "Empirical risk minimization for stochastic convex optimization: $\mathrm{O}(1/n)$-and $\mathrm{O}(1/n^2)$-type of risk bounds," in *Conference on Learning Theory*, 2017, pp. 1954–1979.

P. Bartlett, M. Jordan, and J. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Berlin: Springer, 1991, vol. 23.

T. Zhang, "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, vol. 2, pp. 527–550, 2002.

P. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

Y. Lei and K. Tang, "Learning rates for stochastic gradient descent with nonconvex objectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4505–4511, 2021.

Y. Xu and A. Zeevi, "Towards optimal problem dependent generalization error bounds in statistical learning theory," *Mathematics of Operations Research*, 2024.

C. McDiarmid, "On the method of bounded differences," in *Surveys in combinatorics*, J. Siemous, Ed. Cambridge: Cambridge Univ. Press, 1989, pp. 148–188.

U. Haagerup, "The best constants in the khintchine inequality," *Studia Mathematica*, vol. 70, no. 3, pp. 231–283, 1981.

J. A. Tropp, "The expected norm of a sum of independent random matrices: An elementary approach," in *High Dimensional Probability VII: The Cargese Volume*. Springer, 2016, pp. 173–202.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

T. Zhang, "Data dependent concentration bounds for sequential prediction algorithms." in *Conference on Learning Theory*, 2005, pp. 173–187.

P. Tarres and Y. Yao, "Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5716–5735, 2014.

S. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International Conference on Machine Learning*, 2016, pp. 314–323.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *International Conference on Machine Learning*. JMLR. org, 2017, pp. 2613–2621.

C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Advances in Neural Information Processing Systems*, 2018, pp. 689–699.