SHARPER GENERALIZATION BOUNDS FOR LEARNING WITH GRADIENT-DOMINATED OBJECTIVE FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic optimization has become the workhorse behind many successful machine learning applications, which motivates a lot of theoretical analysis to understand its empirical behavior. As a comparison, there is far less work to study the generalization behavior especially in a non-convex learning setting. In this paper, we study the generalization behavior of stochastic optimization by leveraging the algorithmic stability for learning with β-gradient-dominated objective functions. We develop generalization bounds of the order $O(1/(n\beta))$ plus the convergence rate of the optimization algorithm, where n is the sample size. Our stability analysis significantly improves the existing non-convex analysis by removing the bounded gradient assumption and implying better generalization bounds. We apply our general results to various stochastic optimization algorithms, which show clearly how the variance-reduction techniques improve not only training but also generalization. Furthermore, our discussion explains how interpolation helps generalization for highly expressive models.

1 INTRODUCTION

Stochastic optimization has found tremendous applications in training highly expressive machine learning models including deep neural networks (DNNs) (Bottou et al., 2018), which are ubiquitous in modern learning architectures (LeCun et al., 2015). Oftentimes, the models trained in this way have not only very small training errors or even interpolate the training examples, but also surprisingly generalize well to testing examples (Zhang et al., 2017). While the low training error can be well explained by the over-parametrization of models and the efficiency of the optimization algorithm in identifying a local minimizer (Bassily et al., 2018; Vaswani et al., 2019; Ma et al., 2018), it is still unclear how the highly expressive models also achieve a low testing error (Ma et al., 2018). With the recent theoretical and empirical study, it is believed that a joint consideration of the interaction among the optimization algorithm, learning models and training examples is necessary to understand the generalization behavior (Neyshabur et al., 2017; Hardt et al., 2016; Lin et al., 2016).

The generalization error for stochastic optimization typically consists of an optimization error and an estimation error (see e.g. Bousquet & Bottou (2008)). Optimization errors arise from the suboptimality of the output of the chosen optimization algorithms, while estimation errors refer to the discrepancy between the testing error and training error at the output model. There is a large amount of literature on studying the optimization error (convergence) of stochastic optimization algorithms (Bottou et al., 2018; Karimi et al., 2016). In particular, the power of interpolation is clearly justified in boosting the convergence rate of stochastic gradient descent (SGD) (Bassily et al., 2018; Vaswani et al., 2019; Ma et al., 2018). In contrast, there is far less work on studying estimation errors of optimization algorithms. In a seminal paper (Hardt et al., 2016), the fundamental concept of algorithmic stability was used to study the generalization behavior of SGD, which was further improved and extended in Charles & Papailiopoulos (2018); Zhou et al. (2018b); Yuan et al. (2019); Kuzborskij & Lampert (2018). However, these results are still not quite satisfactory in the following three aspects. Firstly, the existing stability bounds in non-convex learning require very small step sizes (Hardt et al., 2016) and yield suboptimal generalization bounds (Yuan et al., 2019; Charles & Papailiopoulos, 2018; Zhou et al., 2018b). Secondly, majority of the existing work has focused on functions with a uniform Lipschitz constant which can be very large in practical models if not infinite (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Charles & Papailiopoulos, 2018; Kuzborskij & Lampert, 2018), e.g., DNNs. Thirdly, the existing stability analysis fails to explain how the highly

expressive models still generalize in an interpolation setting, which is observed for overparameterized DNNs (Arora et al., 2019; Brutzkus et al., 2017; Bassily et al., 2018; Belkin et al., 2019).

In this paper, we make attempts to address the above three issues using novel stability analysis approaches. Our main contributions are summarized as follows.

1. We develop general stability and generalization bounds for *any* learning algorithm to optimize (non-convex) β-gradient-dominated objectives. Specifically, we show that the excess generalization error is bounded by $O(1/(n\beta))$ plus the convergence rate of the algorithm, where n is the sample size. This general theorem implies that overfitting will never happen in this case, and generalization would always improve as we increase the training accuracy, which is due to an implicit regularization effect of gradient dominance condition. In particular, we show that interpolation actually improves generalization for highly expressive models. In contrast to the existing discussions based on either hypothesis stability or uniform stability which imply at best a bound of $O(1/\sqrt{n\beta})$, the main idea is to consider a weaker on-average stability measure which allows us to replace the uniform Lipschitz constant in Hardt et al. (2016); Kuzborskij & Lampert (2018); Charles & Papailiopoulos (2018) with the training error of the best model.

2. We apply our general results to various stochastic optimization algorithms, and highlight the advantage over existing generalization analysis. For example, we derive an exponential convergence of testing errors for SGD in an interpolation setting, which complements the exponential convergence of optimization errors (Bassily et al., 2018; Vaswani et al., 2019; Ma et al., 2018) and extends the existing results (Pillaud-Vivien et al., 2018; Nitanda & Suzuki, 2019) from a strongly-convex setting to a non-convex setting. In particular, we show that stochastic variance-reduced optimization outperforms SGD by achieving a significantly faster convergence of testing errors, while this advantage is only shown in terms of optimization errors in the literature (Reddi et al., 2016; Lei et al., 2017; Nguyen et al., 2017; Zhou et al., 2018a; Wang et al., 2019).

2 Related Work

Algorithmic Stability. We first review the related work on stability and generalization. Algorithmic stability is a fundamental concept in statistical learning theory (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005). The important uniform stability was introduced in Bousquet & Elisseeff (2002), where the authors showed that empirical risk minimization enjoys the uniform stability if the objective function is strongly convex. This concept was extended to study randomized algorithms such as bagging and bootstrap (Elisseeff et al., 2005). A much weaker on-average stability measure was explored thoroughly in Shalev-Shwartz et al. (2010), which was used to study unregularized learning algorithms (Gonen & Shalev-Shwartz, 2017). Very recently, almost optimal high-probability bounds were established for uniformly stable algorithms by developing elegant concentration inequalities for weakly-dependent random variables (Maurer, 2017; Feldman & Vondrak, 2019; Bousquet et al., 2020). Some other stability measures include the uniform argument stability (Liu et al., 2017), hypothesis stability (Bousquet & Elisseeff, 2002) and hypothesis set stability (Foster et al., 2019). The deep connection between algorithmic stability and learnability was identified (Shalev-Shwartz et al., 2010; Rakhlin et al., 2005). An interesting trade-off between stability and convergence was developed for iterative optimization algorithms, which was then used to study convergence lower bounds of different algorithms (Chen et al., 2018). Other than the standard classification and regression setting, algorithmic stability was very successfully to study transfer learning (Kuzborskij & Lampert, 2018), PAC-Bayesian bounds (London, 2017) and privacy learning (Bassily et al., 2019).

Generalization analysis. We now review related work on generalization analysis for stochastic optimization. In a seminal paper (Hardt et al., 2016), the authors use the nonexpansiveness of gradient mapping to develop uniform stability bounds for SGD to optimize convex, strongly convex and even non-convex objective functions. This inspired some interesting work on stochastic optimization. An interesting data-dependent stability bound was developed for SGD, a nice property of which is that it shows how the initialization would affect generalization (Kuzborskij & Lampert, 2018). These stability bounds were integrated into a PAC-Bayesian analysis of SGD, yielding generalization bounds for arbitrary posterior distributions (London, 2017). Almost optimal generalization bounds were developed for differentially private stochastic convex optimization (Bassily et al., 2019). The on-average variance of stochastic gradients was used to refine the stability analysis of SGD (Hardt et al., 2016) in non-convex optimization (Zhou et al., 2018b). The uniform stability was also studied for

SGD implemented in a stagewise manner (Yuan et al., 2019) and stochastic gradient Langevin dynamics in a non-convex setting (Li et al., 2020; Mou et al., 2018). Very recently, the discussions in Hardt et al. (2016) were extended to tackle non-smooth (Lei & Ying, 2020; Bassily et al., 2020) and non-Lipscthiz functions (Lei & Ying, 2020). The most related work is Charles & Papailiopoulos (2018), where some general hypothesis stability bounds were developed for learning algorithms that converge to optima. A very interesting point is that their bounds depend only on the convergence of the algorithm to a global minimum and the geometry of loss functions around the global minimum. Other than the stability approach, there is interesting generalization analysis of SGD based on either a uniform convergence approach (Lin et al., 2016), an integral operator approach (Lin & Rosasco, 2017; Ying & Pontil, 2008; Dieuleveut & Bach, 2016; Dieuleveut et al., 2017; Mücke et al., 2019) or an information-theoretic approach (Xu & Raginsky, 2017; Negrea et al., 2019; Bu et al., 2020).

3 Main Results

Let ρ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$, from which a training dataset $S = \big\{z_1, \ldots, z_n\big\}$ is drawn independently and identically. The aim is to find a good model $\mathbf{w}$ from a model parameter space $\mathcal{W}$ based on the training dataset S. The performance of a prescribed model $\mathbf{w}$ on a single example z can be measured by a nonnegative loss function $f(\mathbf{w}; z)$, where $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$. In machine learning we often apply an (randomized) algorithm $A : \cup_n \mathcal{Z}^n \mapsto \mathcal{W}$ to S to produce an output model $A(S) \in \mathcal{W}$. Oftentimes, the constructed model $\mathbf{w}$ would have a small *empirical risk* $F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}; z_i)$. However, we are mostly interested in the generalization performance of a model $\mathbf{w}$ on testing examples measured by the *population (true) risk* $F(\mathbf{w}) = \mathbb{E}_z\big[f(\mathbf{w}; z)\big]$, where $\mathbb{E}_z$ denotes the expectation with respect to (w.r.t.) z. The gap $\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big]$ between the population risk and empirical risk is called the *estimation error*, which is due to the approximation of ρ by sampling. A powerful tool to study the estimation error is the algorithmic stability (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Shalev-Shwartz et al., 2010; Hardt et al., 2016), which measures the sensitivity of the algorithm's output w.r.t. the perturbation of a training dataset. Below we give formal definitions of stability measures, whose connection to generalization is established in Theorem A.1. For any $m \in \mathbb{N}$, we denote $[m] = \{1, \ldots, m\}$. For any $i \in [n]$, denote $S \backslash z_i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$.

Definition 1 (Uniform Stability)**.** A randomized algorithm A has uniform stability ϵ if for all datasets $S, \widetilde{S} \in \mathcal{Z}^n$ that differ by at most one example, we have $\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(\widetilde{S}); z)\big] \leq \epsilon$.

Definition 2 (Pointwise Hypothesis Stability)**.** We say a randomized algorithm A has pointwise hypothesis stability ϵ if for all $i \in [n]$ there holds $\mathbb{E}_{S,A}\big[\big|f(A(S); z_i) - f(A(S\backslash z_i); z_i)\big|\big] \leq \epsilon$.

The following on-average stability is very similar to the average-RO stability introduced in Shalev-Shwartz et al. (2010). The difference is that we do not consider an absolute value here.

Definition 3 (On-average Stability)**.** Let $S = \{z_1, \ldots, z_n\}$ and $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be drawn independently from ρ. For each $i \in [n]$, denote $S^{(i)} = \{z_1, \ldots, z_{i-1}, \tilde{z}_i, z_{i+1}, \ldots, z_n\}$. We say an algorithm A has on-average stability ϵ if $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,\widetilde{S},A}\big[f(A(S^{(i)}); z_i) - f(A(S); z_i)\big] \leq \epsilon$.

In this paper, we are interested in the *excess generalization error* $F(A(S)) - F(\mathbf{w}^*)$, where $\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ is the best model with the least testing error (population risk). For this purpose, we introduce some basic assumptions. A basic assumption in non-convex learning is the smoothness of loss functions (Ghadimi & Lan, 2013; Karimi et al., 2016), meaning the gradients are Lipschitz continuous. Let $\|\cdot\|_2$ denote the Euclidean norm and ∇ denote the gradient operator.

Assumption 1 (Smoothness Assumption)**.** We assume for all $z \in \mathcal{Z}$, the differentiable function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is L-smooth, i.e., $\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}'; z)\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

Another assumption is the Polyak-Lojasiewicz (PL) condition on the objective function, which is common in non-convex optimization (Zhou et al., 2018b; Reddi et al., 2016; Karimi et al., 2016; Wang et al., 2019; Lei et al., 2017), and was shown to hold true for deep (linear) and shallow neural networks (Hardt & Ma, 2016; Charles & Papailiopoulos, 2018; Li & Yuan, 2017).

Assumption 2 (Polyak-Lojasiewicz Condition)**.** Denote $\hat{F}_S = \inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}')$. We assume F_S satisfies PL or gradient-dominated condition with parameter $\beta > 0$, i.e.,

$$\mathbb{E}_S\big[F_S(\mathbf{w}) - \hat{F}_S\big] \leq \frac{1}{2\beta}\mathbb{E}_S\big[\|\nabla F_S(\mathbf{w})\|_2^2\big], \quad \forall \mathbf{w} \in \mathcal{W}. \tag{3.1}$$

It is worthy of mentioning that our results in this section continue to hold if the global PL condition is relaxed to a local PL condition, i.e., (3.1) holds for $\mathbf{w}$ in a neighborhood of the minimizer of F_S.

The existing stability analysis often imposes a bounded gradient assumption below (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Charles & Papailiopoulos, 2018; Yuan et al., 2019; Kuzborskij & Lampert, 2018). Indeed, the resulting stability bounds depend on the uniform Lipschitz constant G (see eq. (3.4)), which can be prohibitively large in practical models, e.g., DNNs, or even infinite, e.g. least squares regression in an unbounded domain.

Assumption 3 (Bounded Gradient Assumption)**.** We assume $\|\nabla f(\mathbf{w}; z)\|_2 \leq G$ for all $\mathbf{w} \in \mathcal{W}$, $z \in \mathcal{Z}$ and a constant $G > 0$.

Our main result to be proved in Appendix B removes Assumption 3 and replaces the uniform Lipschitz constant G by the minimal empirical risk $\hat{F}_S$, which is significantly smaller than the Lipschitz constant. Note the assumption $L \leq n\beta/4$ is mild, and the previous generalization bounds become vacuous as $O(1)$ (Yuan et al., 2019; Charles & Papailiopoulos, 2018) if this assumption is violated.

Theorem 1 (Main Theorem)**.** *Let Assumptions 1, 2 hold and* $\mathbf{w}_S = A(S)$*. If* $L \leq n\beta/4$*, then*

$$\mathbb{E}\big[F(\mathbf{w}_S) - \hat{F}_S\big] \leq \frac{16L\mathbb{E}[\hat{F}_S]}{n\beta} + \frac{L\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]}{2\beta}. \tag{3.2}$$

An important implication is as follows. Since $\mathbb{E}\big[\hat{F}_S\big] \leq \mathbb{E}\big[F_S(\mathbf{w}^*)\big] = F(\mathbf{w}^*)$ and $\hat{F}_S \leq F_S(\mathbf{w}_S)$, Eq. (3.2) implies an upper bound on the excess generalization error $\mathbb{E}[F(\mathbf{w}_S)] - F(\mathbf{w}^*)$ and

$$\mathbb{E}\big[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)\big] = O\Big(\frac{1}{n\beta} + \frac{\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]}{\beta}\Big). \tag{3.3}$$

The above two terms can be explained as follows. The term $O(1/(n\beta))$ reflects the intrinsic complexity of the problem, while the term $\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]$ is called the *optimization error*. An interesting observation is that the overfitting phenomenon would never happen for learning under the PL condition (analogous to learning with strongly convex objectives where the global minimizer generalizes well (Bousquet & Elisseeff, 2002)). Indeed, if the optimization algorithm finds more and more accurate solutions, the algorithm achieves the limiting generalization bound $O(1/(n\beta))$. This shows an important message that optimization can be beneficial to generalization. This seemingly counter-intuitive phenomenon is due to the *implicit regularization* enforced by the PL condition (analogous to the strong convexity condition). Another notable property is that Theorem 1 applies to any optimization algorithm. We can plug any known optimization error bounds into it to immediately derive generalization bounds.

Note if A is empirical risk minimization, we get excess generalization error bounds of the order $O(1/(n\beta))$, which match the existing result in learning with strongly convex functions (Shalev-Shwartz et al., 2010).

Remark 1. We show that our result significantly improves the existing stability analysis. The work (Charles & Papailiopoulos, 2018) showed the pointwise hypothesis stability is controlled by $\frac{2G^2}{n\beta} + 2\sqrt{2}G\sqrt{\mathbb{E}[F_S(\mathbf{w}_S) - \hat{F}_S]/\beta}$, which together with the connection between stability and generalization (cf. (A.1)), implies with probability $1 - \delta$ that

$$F(\mathbf{w}_S) \leq F_S(\mathbf{w}_S) + \left(\frac{M^2}{n\delta} + \frac{24MG^2}{n\beta\delta} + \frac{24MG\sqrt{2\mathbb{E}[F_S(\mathbf{w}_S) - \hat{F}_S]}}{\sqrt{\beta}\delta}\right)^{\frac{1}{2}}. \tag{3.4}$$

The above bound requires the bounded gradient assumption $\|\nabla f(\mathbf{w}; z)\|_2 \leq G$ and the bounded loss assumption $0 \leq f(\mathbf{w}; z) \leq M$ for all $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$, which are successfully removed in our generalization analysis. Furthermore, our generalization bound significantly improves (3.4). Indeed, assume $\mathbb{E}[F_S(\mathbf{w}_S) - \hat{F}_S] \leq \epsilon^2\beta$ for some $\epsilon > 0$, then (3.3) implies

$$\mathbb{E}\big[F(\mathbf{w}_S)\big] = \mathbb{E}\big[F_S(\mathbf{w}_S)\big] + O\Big(\frac{1}{n\beta} + \epsilon^2\Big), \tag{3.5}$$

while (3.4) becomes $F(\mathbf{w}_S) = F_S(\mathbf{w}_S) + O\Big(\frac{1}{\sqrt{n\beta}} + \sqrt{\epsilon}\Big)$. To achieve the generalization guarantee $O(1/\sqrt{n\beta})$, the above bound requires the optimization accuracy $\epsilon = O(1/(n\beta))$, while our bound

(3.5) only requires the accuracy $\epsilon = O(1/\sqrt{n\beta})$ but gets the significantly better generalization bound $1/(n\beta)$. We actually develop a better stability bound. Specifically, the pointwise hypothesis stability is bounded by $O\big(\frac{1}{n\beta} + \epsilon\big)$ in Charles & Papailiopoulos (2018) while we show that the on-average stability is bounded by $O\big(\frac{1}{n\beta} + \epsilon^2\big)$, which is significantly tighter if $\frac{1}{n\beta} \leq \epsilon \leq 1$ (ignoring constant factors). It should be mentioned that the discussions in Charles & Papailiopoulos (2018) do not impose a smoothness assumption on loss functions. However, the smoothness assumption is widely used in non-convex optimization to derive meaningful rates (Ghadimi & Lan, 2013).

Remark 2 (Bounded gradient assumption)**.** Very recently, the bounded gradient assumption was also removed for the stability analysis (Lei & Ying, 2020). However, their analysis considered SGD applied to convex loss functions. As a comparison, we study stability and generalization in a non-convex learning setting, and our analysis applies to any stochastic optimization algorithms.

As a direct corollary, we can derive the following optimistic bound in the interpolation setting, which is the most intriguing case for over-parameterized or highly expressive DNN models.

Corollary 2. *Let Assumptions 1, 2 hold and* $\mathbf{w}_S = A(S)$*. If* $\mathbb{E}[\hat{F}_S] = 0$ *and* $L < n\beta/2$*, then*

$$\mathbb{E}\big[F(\mathbf{w}_S)\big] \leq \frac{L}{2\beta}\mathbb{E}\big[F_S(\mathbf{w}_S)\big].$$

Remark 3. Corollary 2 shows a clear benefit of interpolation in boosting the generalization by achieving a generalization bound $O(\epsilon)$ for any $\epsilon > 0$ if we minimize F_S sufficiently well. This benefit can not be explained by the existing discussions (Hardt et al., 2016; Charles & Papailiopoulos, 2018) as they imply the same generalization bound $O(1/\sqrt{n\beta})$ in the interpolation setting. Although it was observed that interpolation helps in training (Bassily et al., 2018; Vaswani et al., 2019; Ma et al., 2018; Oymak & Soltanolkotabi, 2020; Allen-Zhu et al., 2019; Zou et al., 2018), it is still largely unclear, as indicated in Ma et al. (2018), that how interpolation helps in generalization. Therefore, Corollary 2 illustrates new insights on how interpolation resultant from highly expressive models would help generalization.

We now move on to the discussion on the critical assumption in Corollary 2, i.e. $L < n\beta/2$. According to the proof, the two parameters L and β can be replaced by their local counterparts, i.e., the smoothness and PL condition related to a particular minimizer $\mathbf{w}'$ of $F_{S^{(i)}}$ (Eqs. (B.6), (B.7)). For example, β can be replaced by $\frac{1}{2}\|\nabla F_S(\mathbf{w}')\|_2^2/(F_S(\mathbf{w}') - \hat{F}_S)$, which can be larger than β. Below are some examples on explaining $L/\beta < n/2$. As we will see, the quantity L/β reflects the complexity of the problem (related to condition number as shown in Examples 1, 2). Therefore, the condition $L/\beta < n/2$ imposes implicitly a constraint on the complexity of the problems. This explains why the optimization algorithm would never overfit when applied to gradient-dominated objective functions if $L/\beta < n/2$, as shown in Theorem 1.

Example 1. Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}^m$ be a feature map, and $\ell : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}_+$ be a loss function which is L_ℓ-smooth and σ_ℓ-strongly convex w.r.t. the first argument. Consider $f(\mathbf{w}; z) = \ell(\langle \mathbf{w}, \phi(x_i)\rangle, y_i)$ with $\langle\cdot,\cdot\rangle$ being an inner product. Then, F_S satisfies the PL condition with the parameter $\sigma'_{\min}(\Sigma_S)\sigma_\ell$, where $\Sigma_S = \frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(x_i)^\top$ is the empirical covariance matrix, $A^\top$ denotes the transpose of a matrix A and $\sigma'_{\min}(A)$ denotes the minimal *non-zero* singular value of A. The term L/β is of the order of $\sigma_{\max}(\Sigma_S)/\sigma'_{\min}(\Sigma_S)$, where $\sigma_{\max}(A)$ denotes the maximal singular value of A (we give details in Appendix E.1). Then Corollary 2 shows that if $\sigma_{\max}(A)/\sigma'_{\min}(A) \leq cn$ for some appropriate $c > 0$ and there *exists* a model interpolating the training examples, the output model can generalize well to any accuracy if we improve the optimization accuracy.

Example 2. Consider neural networks with a single hidden layer with d inputs, m hidden neurons and a single output neuron, for which the prediction function takes the form $h_{\mathbf{v},\mathbf{w}} = \sum_{k=1}^m v_k\phi\big(\langle \mathbf{w}_k, x\rangle\big)$. Here $\mathbf{w}_k \in \mathbb{R}^d$ and $v_k \in \mathbb{R}$ denote the weight of the edges connecting the k-th hidden node to the input and output node, respectively, while $\phi : \mathbb{R} \mapsto \mathbb{R}$ is the activation function. Analogous to Arora et al. (2019); Oymak & Soltanolkotabi (2020), we fix $\mathbf{v} = (v_1, \ldots, v_m)^\top$ with $|v_k| = a$ for some $a > 0$ and train $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)^\top \in \mathbb{R}^{m\times d}$ from S. The loss function then takes the form $f(\mathbf{w}; z) = \big(\mathbf{v}^\top\phi(\mathbf{w}x) - y\big)^2$. If we consider the identity activation function, i.e., $\phi(t) = t$, then F_S satisfies the PL condition with the parameter $\sigma_{\min}(\Sigma_S)$, where $\sigma_{\min}(A)$ denotes the minimal singular value of A and $\Sigma_S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. The term L/β is of the order of $\sigma_{\max}(\Sigma_S)/\sigma_{\min}(\Sigma_S)$ (we give details in Appendix E.2 for a general activation function). Therefore, Corollary 2 applies if $\sigma_{\max}(\Sigma_S)/\sigma_{\min}(\Sigma_S) \leq cn$ for some appropriate c.

Finally, we present generalization bounds for any optimization algorithms applied to gradient-dominated and Lipschitz continuous functions. We do not require loss functions to be smooth here. It shows that the excess generalization bound can decay as fast as $O(1/(n\beta))$ if we solve the optimization problem to a sufficient accuracy, which is much better than the generalization bound $O(1/\sqrt{n\beta})$ in Charles & Papailiopoulos (2018). Recall the analysis in Charles & Papailiopoulos (2018) requires Assumptions 2, 3 and a further assumption on the boundedness of loss functions. The proof is given in Section C.

Theorem 3. *Let Assumptions 2, 3 hold and $\mathbf{w}_S = A(S)$. Then the following inequality holds*

$$\mathbb{E}\big[F(\mathbf{w}_S) - \hat{F}_S\big] \leq \frac{2G^2}{n\beta} + \frac{G\big(\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]\big)^{\frac{1}{2}}}{\sqrt{2\beta}}.$$

4 Applications

In this section, we apply Theorem 1 to different stochastic optimization algorithms such as stochastic gradient descent, randomized coordinate descent, and stochastic variance-reduced optimization. In particular, we study the number of stochastic gradient evaluations required to achieve a prescribed generalization bound, which is summarized in Table 1. We always assume $L \leq n\beta/4$ in this section.

4.1 Stochastic Gradient Descent

We need some notations to state results on SGD. Specifically, denote by $\mathbf{w}_1 \in \mathcal{W}$ an initial point of SGD. At the t-th iteration, we first randomly select an index $i_t \sim \text{unif}[n]$, and then update $\{\mathbf{w}_t\}_t$ by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t}), \tag{4.1}$$

where $\{\eta_t\}_t$ is a sequence of positive step sizes and unif$[n]$ denotes the uniform distribution over $[n]$. The proof of Theorem 4 is given in Appendix D.1.

Theorem 4. *Let Assumptions 1, 2 hold with $L \leq n\beta/4$. Let A be SGD with the step size sequence $\eta_t = \frac{2t+1}{2\beta(t+1)^2}$. Then $\mathbb{E}\big[F(\mathbf{w}_{T+1})\big] - F(\mathbf{w}^*) = O\Big(\frac{1}{n\beta} + \frac{1}{T\beta^3}\Big)$. We can take $O(\frac{n}{\beta^2})$ stochastic gradient evaluations to get excess generalization bounds $O(1/(n\beta))$.*

Remark 4. We compare Theorem 4 with the recent generalization analysis of SGD under the PL condition. Based on pointwise hypothesis stability analysis and the optimization error bound in Karimi et al. (2016), it was shown with probability at least $1-\delta$ (Charles & Papailiopoulos, 2018)

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = O\Big(\frac{1}{\sqrt{n\beta\delta}} + \frac{1}{T^{\frac{1}{4}}\beta^{\frac{3}{4}}\delta^{\frac{1}{2}}}\Big). \tag{4.2}$$

The above bound indicates that $O(n^2/\beta)$ stochastic gradient evaluations are needed to get the excess generalization bounds $O(1/\sqrt{n\beta})$. Based on the uniform stability bound in Hardt et al. (2016) and the optimization error bound in Karimi et al. (2016), it was shown in Yuan et al. (2019) that

$$\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) = O\Big(n^{-1}(\beta T)^{\frac{L/\beta}{1+L/\beta}}\Big) + O\Big(\frac{1}{T\beta^2}\Big). \tag{4.3}$$

By taking an optimal $T = n^{\frac{1+L/\beta}{1+2L/\beta}}\beta^{-\frac{2+3L/\beta}{1+2L/\beta}}$ (ignoring a constant factor) to balance the above two terms, we derive $\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) = O\Big(n^{-\frac{1+L/\beta}{1+2L/\beta}}\beta^{-\frac{L/\beta}{1+2L/\beta}}\Big)$. If L/β is moderately large, then this bound quickly becomes $\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) = O(1/\sqrt{n\beta})$. With high probability at least $1-\delta$, it was shown that SGD with the step size $\eta_t = \frac{c}{(t+2)\log(t+2)}$ gets the bound $F(\mathbf{w}_{T+1}) - F_S(\mathbf{w}_{T+1}) = O\big(\sqrt{c\log T}/\sqrt{n\delta}\big)$ (Zhou et al., 2018b). However, it is not clear how the optimization errors decay with such step sizes. Typically, c should be of the order $O(1/\beta)$ as shown in Karimi et al. (2016) and therefore the stability analysis in Zhou et al. (2018b) can at best achieve the generalization bounds $O\big(\sqrt{\log T}/\sqrt{n\beta}\big)$. To summarize, the existing stability analysis generally implies the generalization bound $O(1/\sqrt{n\beta})$ for SGD in learning with gradient-dominated objectives (Charles & Papailiopoulos, 2018; Zhou et al., 2018b; Yuan et al., 2019), which is significantly improved to $O(1/(n\beta))$ in our paper by the refined stability analysis. It is worth mentioning that, in this comparison, we have used the same optimization error bounds in Karimi et al. (2016), and the analysis in Charles & Papailiopoulos (2018); Zhou et al. (2018b); Yuan et al. (2019) requires a bounded gradient assumption and a bounded loss assumption, which are removed in our analysis.

Algorithm	Complexity for $1/(n\beta)$	Complexity for ϵ if $\mathbb{E}[\hat{F}_S] = 0$
SGD	$\frac{n}{\beta^2}$	$\frac{1}{\beta^2}\log\frac{1}{\beta\epsilon}$
RCD	$\frac{d\log n}{\beta}$	$\frac{d}{\beta}\log\frac{1}{\beta\epsilon}$
SVRG, SCSG	$\big(n + n^{\frac{2}{3}}/\beta\big)\log n$	$\big(n + n^{\frac{2}{3}}/\beta\big)\log\frac{1}{\beta\epsilon}$
SARAH, SpiderBoost	$\big(n + 1/\beta^2\big)\log n$	$\big(n + 1/\beta^2\big)\log\frac{1}{\beta\epsilon}$
SNVRG	$\big(n + \sqrt{n}/\beta\big)\log^4 n$	$\big(n + \sqrt{n}/\beta\big)\log^4\frac{1}{\beta\epsilon}$

Table 1: Iteration complexity for different optimization algorithms to achieve a stated generalization bound under Assumptions 1, 2. In the second column, we present the number of stochastic gradient evaluations to achieve excess generalization bounds $O(1/(n\beta))$. In the third column, we present the number of stochastic gradient evaluations to achieve generalization bounds $O(\epsilon)$ if $\mathbb{E}[\hat{F}_S] = 0$. We ignore constant factors. It is known that variance-reduction techniques improve the iteration complexity to achieve small training errors. Our stability analysis shows that such an improvement is also achieved for testing errors. Note that the stability analysis in Charles & Papailiopoulos (2018) can at most imply an excess generalization bound $O(1/\sqrt{n\beta})$ for these algorithms.

The above iteration complexity in Theorem 4 can be further improved if we impose a restricted secant inequality (Karimi et al., 2016) on F_S, which has been considered for non-convex optimization, e.g., optimizing neural networks (Li & Yuan, 2017). This is a slightly stronger assumption than the PL condition as shown in Karimi et al. (2016).

Definition 4. We say $F_S : \mathcal{W} \mapsto \mathbb{R}$ satisfies the restricted secant inequality with parameter β if for all $\mathbf{w} \in \mathcal{W}$ we have $\mathbb{E}\big[\langle \mathbf{w} - \mathbf{w}^{(S)}, \nabla F_S(\mathbf{w})\rangle\big] \geq \beta\mathbb{E}[\|\mathbf{w} - \mathbf{w}^{(S)}\|_2^2]$, where for any $\mathbf{w}$, we denote by $\mathbf{w}^{(S)}$ the Euclidean projection of $\mathbf{w}$ onto the set of global minimizers of F_S in $\mathcal{W}$.

Theorem 5. *Assume F_S satisfies the restricted secant inequality with parameter β. Let Assumption 1 hold with $L \leq n\beta/4$. Let A be SGD with $\eta_t = 1/(\beta(t+1))$. Then one can take $O(n/\beta)$ stochastic gradient evaluations to achieve the excess generalization bounds $O(1/(n\beta))$.*

Below we apply Theorem 1 to establish fast generalization bounds in an interpolation setting. Our analysis shows that interpolation actually boosts SGD by achieving an exponential convergence of testing errors, which can not be derived from the bound (3.4) in Charles & Papailiopoulos (2018).

Theorem 6. *Let Assumptions 1, 2 hold with $L \leq n\beta/4$, and $\mathbb{E}[\hat{F}_S] = 0$. Let A be SGD with $\eta_t = \beta/L^2$. Then $\mathbb{E}[F(\mathbf{w}_{T+1})] \leq \frac{L(1-\beta^2/L^2)^T}{2\beta}\mathbb{E}\big[F_S(\mathbf{w}_1)\big]$. We can take $O\big(\beta^{-2}\log(1/(\beta\epsilon))\big)$ stochastic gradient evaluations to achieve the generalization bound $O(\epsilon)$ for any $\epsilon > 0$.*

The above linear convergence does not contradict existing minimax lower bounds where the benefit of interpolation is not considered. The proofs for Theorems 5, 6 are given in Appendix D.1.

Remark 5. We discuss some recent work on error bounds in low-noise conditions. Optimization errors of SGD were studied for general non-convex objectives (Vaswani et al., 2019; Ma et al., 2018) and gradient-dominated objectives (Bassily et al., 2018). For binary classification problems with the specific squared loss, it was shown SGD achieves an exponential convergence of testing classification errors under a margin condition, i.e., positive and negative classes are separated by a margin that is strictly positive (Pillaud-Vivien et al., 2018). This was extended to general convex loss functions under the same margin condition (Nitanda & Suzuki, 2019). These discussions consider regularized objective functions (Pillaud-Vivien et al., 2018; Nitanda & Suzuki, 2019), which are strongly convex. The exponential convergence in Pillaud-Vivien et al. (2018); Nitanda & Suzuki (2019) was established for the testing classification errors, i.e., 0-1 loss. As a comparison, we establish an exponential convergence for the testing errors measured by loss functions used in training. In addition, the exponential convergence in Pillaud-Vivien et al. (2018); Nitanda & Suzuki (2019) comes into effect only after a sufficiently large number of iterations, which is not required in Theorem 6.

4.2 Randomized Coordinate Descent

Randomized coordinate descent (RCD) is an efficient optimization algorithm particularly useful for high-dimensional learning problems (Nesterov, 2012). At each iteration it firstly randomly selects a single coordinate $i_t \in \{1, \ldots, d\}$, and then performs the update along the i_t-th coordinate as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{i_t} F_S(\mathbf{w}_t)\mathbf{e}_{i_t},$$

where $\nabla_i F_S$ denotes the derivative of F_S w.r.t. the i-th coordinate and $\mathbf{e}_i$ is a vector in $\mathbb{R}^d$ with the i-th coordinate being 1 and other coordinates being 0. Theorem 7 develops generalization bound.

Theorem 7. *Let Assumptions 1 and 2 hold with $L \leq n\beta/4$. Let A be RCD with $\eta_t = 1/L$. Then $\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) = O\Big(\frac{1}{n\beta} + \frac{1}{\beta}\big(1 - \frac{\beta}{dL}\big)^T\Big)$. We take $O((d\log n)/\beta)$ stochastic gradient evaluations to get excess generalization bounds $O(1/(n\beta))$. If $\mathbb{E}[\hat{F}_S] = 0$, we take $O\big(\beta^{-1} d \log 1/(\beta\epsilon)\big)$ stochastic gradient evaluations to get generalization bounds $O(\epsilon)$ for any $\epsilon > 0$.*

The detailed proof for the above theorem is given in Appendix D.2. As indicated in Remark 1, the discussion in Charles & Papailiopoulos (2018) can only imply the generalization bound $O(1/\sqrt{n\beta})$.

4.3 Stochastic Variance-Reduced Optimization

SGD needs a diminishing step size due to the inherent variance of stochastic gradients, which generally yields a sublinear convergence rate (Bottou et al., 2018). Recently, there is a large amount of work to accelerate SGD by using some different gradient estimates with a reduced variance (Johnson & Zhang, 2013; Xiao & Zhang, 2014; Zhang et al., 2013; Allen-Zhu & Hazan, 2016; Fang et al., 2018; Wang et al., 2019; Nguyen et al., 2017; Zhou et al., 2018a; Schmidt et al., 2017; Defazio et al., 2014; Reddi et al., 2016). This class of algorithms proceeds in epochs. Let $\tilde{\mathbf{w}}_0$ be an initialization point. At the beginning of s-th epoch, we set a reference point $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$, draw a batch $\tilde{I}_s \subseteq [n]$ and compute $\mathbf{v}_0 = \nabla f_{\tilde{I}_s}(\mathbf{w}_0)$, where we denote $f_I(\mathbf{w}) = \frac{1}{|I|}\sum_{i\in I} f(\mathbf{w}; z_i)$ for $I \subseteq [n]$ and $|I|$ is the cardinality of I. The batch $\tilde{I}_s$ can be equal to $[n]$ (Johnson & Zhang, 2013; Xiao & Zhang, 2014; Wang et al., 2019; Reddi et al., 2016) or drawn with replacement according to the uniform distribution over $[n]$ (Lei et al., 2017; Fang et al., 2018). Then we proceed with m_s inner iterations by using some gradient estimators with reduced variances. At the t-th inner iteration, we first draw a batch $I_t \subseteq [n]$ from the uniform distribution over $[n]$. The original SVRG (Johnson & Zhang, 2013; Reddi et al., 2016; Xiao & Zhang, 2014) uses the gradient estimator (we omit the dependency on s)

$$\mathbf{v}_t = \nabla f_{I_t}(\mathbf{w}_t) - \nabla f_{I_t}(\mathbf{w}_0) + \mathbf{v}_0. \tag{4.4}$$

Recently a different update of gradient estimator is proposed (Nguyen et al., 2017; Fang et al., 2018)

$$\mathbf{v}_t = \nabla f_{I_t}(\mathbf{w}_t) - \nabla f_{I_t}(\mathbf{w}_{t-1}) + \mathbf{v}_{t-1}. \tag{4.5}$$

An important observation is that the variance of $\mathbf{v}_t$ diminishes to zero as we are approaching the minimum, which allows us to update the iterate with a constant step size $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{v}_t$ (Johnson & Zhang, 2013). The framework of stochastic variance-reduced optimization is described in Algorithm 1 in Appendix D.3. The following theorem gives generalization bounds $O(1/(n\beta))$ for stochastic variance-reduced optimization, which significantly improves the bound $O(1/\sqrt{n\beta})$ based on (3.4). The proof is given in Appendix D.3.

Theorem 8. *Let Assumptions 1 and 2 hold with $L \leq n\beta/4$. Let A be either the SARAH in Nguyen et al. (2017) or the SpiderBoost in Wang et al. (2019). We can take $O\big((n + 1/\beta^2)\log n\big)$ stochastic gradient evaluations to get excess generalization bounds $O(1/(n\beta))$. If $\mathbb{E}[\hat{F}_S] = 0$, we take $O\big((n + 1/\beta^2)\log 1/(\beta\epsilon)\big)$ stochastic gradient evaluations to get generalization bounds $O(\epsilon)$ for any $\epsilon > 0$.*

As compared to SGD (Section 4.1), Theorem 8 shows SARAH/SpiderBoost requires significantly fewer iterations to achieve the same testing errors. This shows a clear advantage of stochastic variance-reduced optimization over SGD in generalization other than training. Other than SARAH and SpiderBoost, we also develop generalization bounds for SVRG in Reddi et al. (2016), SCSG in Lei et al. (2017) (Theorem D.3) and SNVRG-PL in Zhou et al. (2018a) (Theorem D.4).

5 Conclusion

We study the generalization analysis of stochastic optimization under the PL condition. We show that the generalization errors can be bounded by $O(1/(n\beta))$ plus the convergence rate of algorithms. An observation is that the optimization always helps in generalization under the PL condition. Our analysis based on a weak on-average stability measure removes the bounded gradient assumption in the literature, and can imply significantly better bounds. In particular, we show how the interpolation accelerates the generalization. As an application to various stochastic optimization algorithms, we show that stochastic variance-reduced optimization algorithms outperform SGD by achieving a faster-decaying generalization errors, while this property was only shown for training errors in the literature. It is interesting to extend our discussions to the general non-convex case.

REFERENCES

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

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6155–6166, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pp. 11279–11288, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626, 2020.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 744–753, 2018.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

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

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279, 2019.

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, pp. 6726–6736. 2019.

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

Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):8245–8257, 2017.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

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

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

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2820–2829, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems 30*, pp. 2348–2358. 2017.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

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

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pp. 2340–2348, 2016.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167, 2017.

Ben London. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2931–2940, 2017.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334, 2018.

Andreas Maurer. A second-order look at stability and generalization. In *Conference on Learning Theory*, pp. 1461–1475, 2017.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pp. 605–638, 2018.

Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pp. 12568–12577, 2019.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pp. 1–39, 2018.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pp. 11015–11025, 2019.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

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

Atsushi Nitanda and Taiji Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *International Conference on Artificial Intelligence and Statistics*, pp. 1417–1426, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference On Learning Theory*, pp. 250–296, 2018.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Sashank Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pp. 314–323, 2016.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pp. 2199–2207, 2010.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204, 2019.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pp. 2403–2413. 2019.

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

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pp. 2604–2614, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, pp. 980–988, 2013.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3921–3932. 2018a.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization. *arXiv preprint arXiv:1802.06903*, 2018b.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

A Stability and Generalization

Theorem A.1 establishes the key connection between the generalization and various stability measures. Part (a) and part (b) show that the algorithm with either uniform stability or pointwise hypothesis stability generalizes well to testing examples (Bousquet & Elisseeff, 2002). Initially, they were developed for deterministic algorithms (Bousquet & Elisseeff, 2002), which were then extended to the setting of randomized algorithms (Elisseeff et al., 2005). Part (c) shows the connection between the generalization and the on-average stability (Shalev-Shwartz et al., 2010). Note part (b) involves a square root of $1/\delta$ instead of a $\log(1/\delta)$.

Theorem A.1 (Generalization by Stability)**.** *Let A be a randomized algorithm.*

(a) If A has uniform stability ϵ, then $\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq \epsilon$.

(b) Let $M > 0$. If A has pointwise hypothesis stability ϵ and $0 \leq f(\mathbf{w}; z) \leq M$ for all $\mathbf{w} \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then for all $\delta \in (0, 1)$ with probability at least $1 - \delta$

$$F(A(S)) \leq F_S(A(S)) + \Big(\frac{M^2 + 12Mn\epsilon}{n\delta}\Big)^{\frac{1}{2}}. \tag{A.1}$$

(c) If A has on-average stability ϵ, then $\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \epsilon$.

Proof. We prove Part (c) here due to its simplicity. Since z_i and $\tilde{z}_i$ are drawn from the same distribution, we know

$$\begin{aligned}\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[F(A(S^{(i)})) - F_S(A(S))\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[f(A(S^{(i)}); z_i) - f(A(S); z_i)\big],\end{aligned}$$

where the last identity holds since z_i is independent of $A(S^{(i)})$. The proof is complete by noting the definition of on-average stability. □

B Proof of Theorem 1

In this section, we prove Theorem 1. We begin our analysis with some useful properties of smooth functions. If g is L-smooth, we have the following self-bounding property (Srebro et al., 2010)

$$\|\nabla g(\mathbf{w})\|_2^2 \leq 2L\big(g(\mathbf{w}) - \inf_{\mathbf{w}'} g(\mathbf{w}')\big), \quad \forall \mathbf{w} \in \mathcal{W} \tag{B.1}$$

and the following elementary inequality for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{W}$ (Nesterov, 2012)

$$g(\mathbf{w}) \leq g(\tilde{\mathbf{w}}) + \langle \nabla g(\tilde{\mathbf{w}}), \mathbf{w} - \tilde{\mathbf{w}}\rangle + \frac{L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2}{2}. \tag{B.2}$$

In particular, if g is further nonnegative, then

$$\|\nabla g(\mathbf{w})\|_2^2 \leq 2Lg(\mathbf{w}), \quad \forall \mathbf{w} \in \mathcal{W}. \tag{B.3}$$

The following lemma follows directly from the self-bounding property of smooth loss functions.

Lemma B.1. *Assume F is L-smooth. Then ($\mathbf{w}$ can depend on S)*

$$\mathbb{E}[\|\nabla F(\mathbf{w})\|_2^2] \leq 2L\mathbb{E}\big[F(\mathbf{w}) - \hat{F}_S\big].$$

Proof. Recall $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. According to the self-bounding property (B.1) and the definition of $\mathbf{w}^*$ we know

$$\begin{aligned}\mathbb{E}[\|\nabla F(\mathbf{w})\|_2^2] &\leq 2L\mathbb{E}\big[F(\mathbf{w}) - F(\mathbf{w}^*)\big] = 2L\mathbb{E}\big[F(\mathbf{w}) - F_S(\mathbf{w}^*)\big] \\ &\leq 2L\mathbb{E}\big[F(\mathbf{w}) - \hat{F}_S\big],\end{aligned}$$

where we have used $\mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$ since $\mathbf{w}^*$ is independent of S, and $\hat{F}_S \leq F_S(\mathbf{w}^*)$ due to the definition of $\hat{F}_S$. The proof is complete. □

In the following lemma, we derive the on-average stability bounds under the PL condition. Recall for any $\mathbf{w}$, we denote by $\mathbf{w}^{(S)}$ the Euclidean projection of $\mathbf{w}$ onto the set of global minimizers of F_S in $\mathcal{W}$.

Lemma B.2. *If Assumptions 1, 2 hold, then A has on-average stability ϵ satisfying*

$$\epsilon \leq \frac{2L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + \mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big].$$

Proof. Let $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be drawn independently from ρ. For each $i \in [n]$, let $S^{(i)}$ be defined in Definition 3. For each $i \in [n]$, we denote $\mathbf{w}_{S^{(i)}} = A(S^{(i)})$ and $\mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ the projection of $\mathbf{w}_{S^{(i)}}$ onto the set of global minimizer of $F_{S^{(i)}}$. We decompose $f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_S; z_i)$ as follows

$$\begin{aligned}f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_S; z_i) &= \big(f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)\big) \\ &\quad + \big(f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big) + \big(f(\mathbf{w}_S^{(S)}; z_i) - f(\mathbf{w}_S; z_i)\big).\end{aligned} \tag{B.4}$$

We now address the above three terms separately. We **first** address $f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)$. According to the definition of $F_S, S, S^{(i)}$, we know

$$f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) = nF_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) - nF_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) + f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i).$$

Since z_i and $\tilde{z}_i$ follow from the same distribution, we know $\mathbb{E}[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i)] = \mathbb{E}[f(\mathbf{w}_S^{(S)}; z_i)]$ and further get

$$\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)\big] = n\mathbb{E}\big[F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\big] - n\mathbb{E}\big[F_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\big] + \mathbb{E}\big[f(\mathbf{w}_S^{(S)}; z_i)\big].$$

It then follows that

$$\begin{aligned}\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] &= n\mathbb{E}\big[F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) - F_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\big] \\ &= n\mathbb{E}\Big[F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) - \inf_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})\Big],\end{aligned} \tag{B.5}$$

where we have used the following identity due to the symmetry between z_i and $\tilde{z}_i$

$$\mathbb{E}[F_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})] = \mathbb{E}[\hat{F}_S] = \inf_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}).$$

By the PL condition of F_S, it then follows from (B.5) that

$$\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] \leq \frac{n}{2\beta}\mathbb{E}\big[\|\nabla F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\|_2^2\big]. \tag{B.6}$$

According to the definition of $\mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ we know $\nabla F_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) = 0$ and therefore ($(a+b)^2 \leq 2a^2 + 2b^2$)

$$\begin{aligned}\|\nabla F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\|_2^2 &= \Big\|\nabla F_{S^{(i)}}(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}) - \frac{1}{n}\nabla f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i) + \frac{1}{n}\nabla f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)\Big\|_2^2 \\ &\leq \frac{2}{n^2}\|\nabla f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i)\|_2^2 + \frac{2}{n^2}\|\nabla f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)\|_2^2 \\ &\leq \frac{4L}{n^2} f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i) + \frac{4L}{n^2} f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i), \end{aligned} \tag{B.7}$$

where we have used the self-bounding property of smooth loss functions (B.3). Since z_i and $\tilde{z}_i$ follow from the same distribution, we know

$$\mathbb{E}[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; \tilde{z}_i)] = \mathbb{E}[f(\mathbf{w}_S^{(S)}; z_i)], \qquad \mathbb{E}[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)] = \mathbb{E}[f(\mathbf{w}_S^{(S)}; \tilde{z}_i)].$$

It then follows that

$$\mathbb{E}\big[\|\nabla F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\|_2^2\big] \leq \frac{4L}{n^2}\mathbb{E}[f(\mathbf{w}_S^{(S)}; z_i)] + \frac{4L}{n^2}\mathbb{E}[f(\mathbf{w}_S^{(S)}; \tilde{z}_i)],$$

which, combined with (B.6), gives

$$\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] \leq \frac{2L}{n\beta}\Big(\mathbb{E}[f(\mathbf{w}_S^{(S)}; z_i)] + \mathbb{E}[f(\mathbf{w}_S^{(S)}; \tilde{z}_i)]\Big).$$

Taking a summation of the above inequality for $i = 1, \ldots, n$, we get

$$\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] \leq \frac{2L}{\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F_{\widetilde{S}}(\mathbf{w}_S^{(S)})]\Big). \tag{B.8}$$

We **then** address $f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)$. Since $\mathbf{w}_{S^{(i)}}$ and $\mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ are independent of z_i, we know

$$\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i)\big] = \mathbb{E}\big[F(\mathbf{w}_{S^{(i)}}) - F(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\big] = \mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big], \tag{B.9}$$

where we have used the symmetry between z_i and $\tilde{z}_i$.

Finally, we address $f(\mathbf{w}_S^{(S)}; z_i) - f(\mathbf{w}_S; z_i)$. By the definition of $\mathbf{w}_S^{(S)}$ we know

$$\sum_{i=1}^n \big(f(\mathbf{w}_S^{(S)}; z_i) - f(\mathbf{w}_S; z_i)\big) = n\big(\hat{F}_S - F_S(\mathbf{w}_S)\big). \tag{B.10}$$

Plugging (B.8), (B.9) and the above inequality back into (B.4), we derive

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_S; z_i)\big] \leq \frac{2L}{\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F_{\widetilde{S}}(\mathbf{w}_S^{(S)})]\Big) + \\ n\mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] + n\mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big].\end{aligned}$$

The proof is complete by recalling the definition of on-average stability and $\mathbb{E}[F_{\widetilde{S}}(\mathbf{w}_S^{(S)})] = \mathbb{E}[F(\mathbf{w}_S^{(S)})]$. □

We further require a lemma relating the convergence in terms of function values to the convergence in terms of models. This shows that the PL condition is stronger than a quadratic growth condition (Karimi et al., 2016).

Lemma B.3 (Karimi et al. 2016)**.** *If F_S satisfies the PL condition with parameter $\beta > 0$. Then for all $\mathbf{w} \in \mathcal{W}$ we have*

$$\mathbb{E}\big[F_S(\mathbf{w}) - F_S(\mathbf{w}^{(S)})\big] \geq 2\beta\mathbb{E}[\|\mathbf{w} - \mathbf{w}^{(S)}\|_2^2]. \tag{B.11}$$

We are now in a position to prove Theorem 1.

Proof of Theorem 1. Plugging the on-average stability established in Lemma B.2 back into Part (c) of Theorem A.1, we derive

$$\mathbb{E}\big[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)\big] \leq \frac{2L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + \mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big], \quad \text{(B.12)}$$

from which we derive

$$\mathbb{E}\big[F(\mathbf{w}_S^{(S)}) - \hat{F}_S\big] \leq \frac{2L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big). \quad \text{(B.13)}$$

By (B.2), we know the following inequality for all $\gamma > 0$

$$\begin{aligned} F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)}) &\leq \langle \nabla F(\mathbf{w}_S^{(S)}), \mathbf{w}_S - \mathbf{w}_S^{(S)} \rangle + \frac{L}{2}\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2 \\ &\leq \|\nabla F(\mathbf{w}_S^{(S)})\|_2 \|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2 + \frac{L}{2}\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2 \\ &\leq \frac{1}{4\gamma}\|\nabla F(\mathbf{w}_S^{(S)})\|_2^2 + \Big(\gamma + \frac{L}{2}\Big)\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2, \end{aligned}$$

where we have used the Cauchy-Schwartz inequality. This together with Lemma B.1 with $\mathbf{w} = \mathbf{w}_S^{(S)}$ implies that

$$\mathbb{E}[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})] \leq \frac{L}{2\gamma}\mathbb{E}\big[F(\mathbf{w}_S^{(S)}) - \hat{F}_S\big] + \Big(\gamma + \frac{L}{2}\Big)\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2].$$

Plugging (B.13) into the above inequality, we get

$$\mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] \leq \frac{L}{2\gamma}\frac{2L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + \Big(\gamma + \frac{L}{2}\Big)\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2].$$

Taking $\gamma = L/2$, we then get

$$\mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] \leq \frac{2L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + L\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2].$$

Plugging the above inequality back into (B.12), we derive the following inequality

$$\mathbb{E}\big[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)\big] \leq \frac{4L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + L\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2] + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big].$$

It then follows that

$$\mathbb{E}\big[F(\mathbf{w}_S) - \hat{F}_S\big] \leq \frac{4L}{n\beta}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big) + L\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2]. \quad \text{(B.14)}$$

Since $L \leq n\beta/4$, it follows from (B.13) that

$$\mathbb{E}\big[F(\mathbf{w}_S^{(S)}) - \hat{F}_S\big] \leq \frac{1}{2}\Big(\mathbb{E}[\hat{F}_S] + \mathbb{E}[F(\mathbf{w}_S^{(S)})]\Big)$$

and therefore

$$\mathbb{E}\big[F(\mathbf{w}_S^{(S)})\big] \leq 3\mathbb{E}[\hat{F}_S].$$

We can plug the above inequality back into (B.14) and derive

$$\mathbb{E}\big[F(\mathbf{w}_S) - \hat{F}_S\big] \leq \frac{16L\mathbb{E}[\hat{F}_S]}{n\beta} + L\mathbb{E}[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2^2]. \quad \text{(B.15)}$$

The stated bound then follows from (B.11). The proof is complete. □

C PROOF OF THEOREM 3

In this section, we present the proof of Theorem 3.

Proof of Theorem 3. Similar to (B.7) but using the boundedness of gradients, we know

$$\|\nabla F_S(\mathbf{w}_{S^{(i)}}^{(S^{(i)})})\|_2^2 \leq \frac{4G^2}{n^2}.$$

We can plug this inequality into (B.6) and derive

$$\mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] \leq \frac{n}{2\beta}\frac{4G^2}{n^2} = \frac{2G^2}{n\beta}.$$

Taking a summation of the above inequality gives

$$\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}^{(S^{(i)})}; z_i) - f(\mathbf{w}_S^{(S)}; z_i)\big] \leq 2G^2/\beta. \tag{C.1}$$

Plugging (C.1), (B.9) and (B.10) back into (B.4), we derive the following inequality

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}_{S^{(i)}}; z_i) - f(\mathbf{w}_S; z_i)\big] &\leq \frac{2G^2}{\beta} + n\mathbb{E}\big[F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)})\big] + n\mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big] \\ &\leq \frac{2G^2}{\beta} + nG\mathbb{E}\big[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2\big] + n\mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big],\end{aligned}$$

where in the last step we have used the inequality $F(\mathbf{w}_S) - F(\mathbf{w}_S^{(S)}) \leq G\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2$ due to the boundedness of gradients. According to the definition of on-average stability, we know that the on-average stability ϵ of A satisfies

$$\begin{aligned}\epsilon &\leq \frac{2G^2}{n\beta} + G\mathbb{E}\big[\|\mathbf{w}_S - \mathbf{w}_S^{(S)}\|_2\big] + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big] \\ &\leq \frac{2G^2}{n\beta} + \frac{G\big(\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]\big)^{\frac{1}{2}}}{\sqrt{2\beta}} + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big],\end{aligned}$$

where we have used Lemma B.3. According to Part (c) of Theorem A.1, it follows that

$$\mathbb{E}\big[F(\mathbf{w}_S) - F_S(\mathbf{w}_S)\big] \leq \frac{2G^2}{n\beta} + \frac{G\big(\mathbb{E}\big[F_S(\mathbf{w}_S) - \hat{F}_S\big]\big)^{\frac{1}{2}}}{\sqrt{2\beta}} + \mathbb{E}\big[\hat{F}_S - F_S(\mathbf{w}_S)\big].$$

The stated bound then follows directly. The proof is complete. □

D Proofs on Applications

In this section, we prove generalization bounds for various stochastic optimization algorithms.

D.1 Stochastic Gradient Descent

We consider here SGD. In the following proposition, we establish the variance of stochastic gradients for SGD under the PL condition.

Proposition D.1. *Let Assumptions 1, 2 hold. Let $\{\mathbf{w}_t\}_t$ be the sequence produced by SGD with step size sequence $\{\eta_t\}_{t\in\mathbb{N}}$. If there exists $t_0 \in \mathbb{N}$ such that $\eta_t \leq \beta/L^2$ for all $t \geq t_0$, then*

$$\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq 2L\max\{\mathbb{E}[F_S(\mathbf{w}_{t_0})], 2\mathbb{E}[\hat{F}_S]\} \quad \forall t \geq t_0.$$

Proof. By (B.2) and the update (4.1), we know

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

where we have used (B.3). Taking expectations on both sides we get the following inequality for all $t \geq t_0$

$$\begin{aligned}\mathbb{E}[F_S(\mathbf{w}_{t+1})] &\leq \mathbb{E}[F_S(\mathbf{w}_t)] - \eta_t \mathbb{E}[\|\nabla F_S(\mathbf{w}_t)\|_2^2] + L^2\eta_t^2 \mathbb{E}[f(\mathbf{w}_t; z_{i_t})] \\ &\leq \mathbb{E}[F_S(\mathbf{w}_t)] - 2\eta_t\beta\mathbb{E}[F_S(\mathbf{w}_t) - \hat{F}_S] + \eta_t\beta\mathbb{E}[F_S(\mathbf{w}_t)], \end{aligned} \tag{D.1}$$

where we have used the PL condition and $\eta_t \leq \beta/L^2$ in the last step. It then follows the following inequality for all $t \geq t_0$

$$\mathbb{E}[F_S(\mathbf{w}_{t+1})] \leq (1-\eta_t\beta)\mathbb{E}[F_S(\mathbf{w}_t)] + \eta_t\beta \cdot 2\mathbb{E}[\hat{F}_S] \leq \max\left\{\mathbb{E}[F_S(\mathbf{w}_t)], 2\mathbb{E}[\hat{F}_S]\right\}.$$

Applying this inequality recursively, we derive

$$\mathbb{E}[F_S(\mathbf{w}_{t+1})] \leq \max\{\mathbb{E}[F_S(\mathbf{w}_{t_0})], 2\mathbb{E}[\hat{F}_S]\} \quad \forall t \geq t_0.$$

This together with (B.3) implies the following inequality for all $t \geq t_0$

$$\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq 2L\mathbb{E}[f(\mathbf{w}_t; z_{i_t})] \leq 2L\max\{\mathbb{E}[F_S(\mathbf{w}_{t_0})], 2\mathbb{E}[\hat{F}_S]\}.$$

The proof is complete. □

We now prove generalization bounds in Theorem 4. We denote $B \asymp \widetilde{B}$ if there exist some constants c_1 and $c_2 > 0$ such that $c_1 B \leq \widetilde{B} \leq c_2 B$.

Proof of Theorem 4. Let $t_0 = \lfloor L^2/\beta^2 \rfloor$. It is clear that $\eta_t \leq \beta/L^2$ for all $t \geq t_0$. Let $\sigma = 2L\max\{\mathbb{E}[F_S(\mathbf{w}_1)], \ldots, \mathbb{E}[F_S(\mathbf{w}_{t_0})], 2\mathbb{E}[\hat{F}_S]\}$. According to the self-bounding property (B.3) and Proposition D.1, we know that $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq \sigma^2$ for all $t \in \mathbb{N}$. The following optimization error bound was established in Karimi et al. (2016)

$$\mathbb{E}\big[F_S(\mathbf{w}_{t+1}) - \hat{F}_S\big] \leq \frac{L\sigma^2}{2t\beta^2}. \tag{D.2}$$

We can plug the above inequality into (3.2) with $A(S) = \mathbf{w}_{T+1}$, and get

$$\mathbb{E}\big[F(\mathbf{w}_{T+1}) - \hat{F}_S\big] \leq \frac{16L\mathbb{E}\big[\hat{F}_S\big]}{n\beta} + \frac{L^2\sigma^2}{4T\beta^3}.$$

Since

$$\mathbb{E}\big[\hat{F}_S\big] \leq \mathbb{E}\big[F_S(\mathbf{w}^*)\big] = F(\mathbf{w}^*), \tag{D.3}$$

we further get

$$\mathbb{E}\big[F(\mathbf{w}_{T+1})\big] - F(\mathbf{w}^*) \leq \frac{16LF(\mathbf{w}^*)}{n\beta} + \frac{L^2\sigma^2}{4T\beta^3}.$$

By taking $T \asymp n/\beta^2$, we get $\mathbb{E}\big[F(\mathbf{w}_{T+1}) - \hat{F}_S\big] = O(1/(n\beta))$. This corresponds to $O(n/\beta^2)$ stochastic gradient evaluations. The proof is complete. □

Lemma D.2. *Assume F_S satisfies the restricted secant inequality with parameter β. Let A be SGD with the step size sequence $\eta_t = 1/(\beta(t+1))$. Then there exists some $\sigma \in \mathbb{R}$ such that*

$$\mathbb{E}\big[\|\mathbf{w}_T - \mathbf{w}_T^{(S)}\|_2^2\big] \leq \sigma^2/(\beta^2 T).$$

Proof of Lemma D.2. Analogous to the proof of Theorem 4, we can find $\sigma \in \mathbb{R}_+$ such that $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq \sigma^2$ for all $t \in \mathbb{N}$. Since $\mathbf{w}_{t+1}^{(S)}$ is a projection of $\mathbf{w}_{t+1}$ onto the set of global minimizer of F_S, we know

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

Taking an expectation and using $\mathbb{E}[\|\nabla f(\mathbf{w}_t; z_{i_t})\|_2^2] \leq \sigma^2$, we derive

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(S)}\|_2^2\big] &\leq \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] + \eta_t^2\sigma^2 + 2\eta_t\mathbb{E}\big[\langle \mathbf{w}_t^{(S)} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle\big] \\ &\leq \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] + \eta_t^2\sigma^2 - 2\eta_t\beta\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] \\ &= (1-2\eta_t\beta)\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] + \eta_t^2\sigma^2,\end{aligned}$$

where we have used the restricted secant inequality. For the step size $\eta_t = 1/(\beta(t+1))$, we have

$$\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(S)}\|_2^2\big] \leq \frac{t-1}{t+1}\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] + \frac{\sigma^2}{\beta^2(t+1)^2}.$$

Multiplying both sides by $t(t+1)$, we derive

$$t(t+1)\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(S)}\|_2^2\big] \leq (t-1)t\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(S)}\|_2^2\big] + \frac{\sigma^2}{\beta^2}.$$

Taking a summation of the above inequality from $t = 1$ to $T - 1$ gives

$$(T-1)T\mathbb{E}\big[\|\mathbf{w}_T - \mathbf{w}_T^{(S)}\|_2^2\big] \leq \sigma^2(T-1)/\beta^2.$$

The proof is complete. □

Proof of Theorem 5. It was shown that functions satisfying restricted secant inequality with parameter β also satisfies the PL condition with parameter β/L (Karimi et al., 2016). Therefore (B.15) holds with β there replaced by β/L. According to Lemma D.2, we know $\mathbb{E}\big[\|\mathbf{w}_T - \mathbf{w}_T^{(S)}\|_2^2\big] \leq \sigma^2/(\beta^2 T)$. We can plug this inequality back into (B.15) with $A(S) = \mathbf{w}_{T+1}$, and get

$$\mathbb{E}\big[F(\mathbf{w}_{T+1})\big] - F(\mathbf{w}^*) = O\Big(\frac{F(\mathbf{w}^*)}{n\beta} + \frac{1}{\beta^2 T}\Big),$$

where we have used (D.3). By taking $T \asymp n/\beta$, we get $\mathbb{E}\big[F(\mathbf{w}_{T+1})\big] - F(\mathbf{w}^*) = O(1/(n\beta))$. This corresponds to $O(n/\beta)$ stochastic gradient evaluations. The proof is complete. □

Proof of Theorem 6. Let $\eta = \beta/L^2$. According to the assumption $\mathbb{E}[\hat{F}_S] = 0$ and (D.1), we know

$$\mathbb{E}[F_S(\mathbf{w}_{t+1})] \leq \mathbb{E}[F_S(\mathbf{w}_t)] - 2\eta\beta\mathbb{E}[F_S(\mathbf{w}_t)] + \eta\beta\mathbb{E}[F_S(\mathbf{w}_t)] = \big(1 - \eta\beta\big)\mathbb{E}[F_S(\mathbf{w}_t)]. \tag{D.4}$$

Applying this inequality recursively, we get $\mathbb{E}[F_S(\mathbf{w}_{T+1})] \leq (1-\eta\beta)^T\mathbb{E}[F_S(\mathbf{w}_1)]$. We can plug the above inequality back into (3.2) with $A(S) = \mathbf{w}_{T+1}$ and get

$$\mathbb{E}[F(\mathbf{w}_{T+1})] \leq \frac{L\mathbb{E}[F_S(\mathbf{w}_{T+1})]}{2\beta} \leq \frac{L(1-\beta^2/L^2)^T}{2\beta}\mathbb{E}\big[F_S(\mathbf{w}_1)\big] \leq \frac{L\exp(-\beta^2 T/L^2)}{2\beta}\mathbb{E}\big[F_S(\mathbf{w}_1)\big],$$

where we have used the elementary inequality

$$1 - a \leq \exp(-a). \tag{D.5}$$

To achieve $\mathbb{E}[F(\mathbf{w}_{T+1})] \leq \epsilon$, we can take T such that

$$\exp\big(-\beta^2 T/L^2\big) \asymp \epsilon\beta \quad \Longleftrightarrow \quad T \asymp \beta^{-2}\log(1/(\beta\epsilon)).$$

The proof is complete. □

D.2 Randomized Coordinate Descent

We prove here the generalization bounds for randomized coordinate descent. We further assume that the gradient is coordinate-wise Lipschitz continuous in the sense that

$$F_S(\mathbf{w} + \alpha\mathbf{e}_i) \leq F_S(\mathbf{w}) + \alpha\nabla_i F_S(\mathbf{w}) + L\alpha^2/2, \quad \forall \alpha \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d, i \in [d].$$

Proof of Theorem 7. According to Theorem 3 in Karimi et al. (2016), we know

$$\mathbb{E}\big[F_S(\mathbf{w}_{T+1}) - \hat{F}_S\big] \leq \Big(1 - \frac{\beta}{dL}\Big)^T\mathbb{E}\big[F_S(\mathbf{w}_1) - \hat{F}_S\big]. \tag{D.6}$$

Plugging the above inequality back into (3.2) and using (D.3), we get

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{T+1})] - F(\mathbf{w}^*) &\leq \frac{16L\mathbb{E}\big[\hat{F}_S\big]}{n\beta} + \frac{L}{2\beta}\Big(1 - \frac{\beta}{dL}\Big)^T\mathbb{E}[F_S(\mathbf{w}_1)] \\ &= O\Big(\frac{F(\mathbf{w}^*)}{n\beta}\Big) + O\Big(\frac{1}{\beta}\exp\Big(-\frac{\beta T}{dL}\Big)\Big),\end{aligned}$$

where we have used (D.5). To achieve the excess generalization bounds $O(1/(n\beta))$, we require T satisfying

$$\exp\Big(-\frac{\beta T}{dL}\Big) \asymp n^{-1} \quad \Longleftrightarrow \quad T \asymp \frac{d\log n}{\beta}.$$

If $\mathbb{E}[\hat{F}_S] = 0$, then it follows from (3.2), (D.6) and (D.5) that

$$\mathbb{E}[F(\mathbf{w}_{T+1})] \leq \frac{L}{2\beta}\exp\Big(-\frac{T\beta}{dL}\Big)\mathbb{E}[F_S(\mathbf{w}_1)].$$

To achieve the generalization bound ϵ, we require T satisfying

$$\exp\Big(-\frac{T\beta}{dL}\Big) \asymp \beta\epsilon \quad \Longleftrightarrow \quad T \asymp \beta^{-1} d\log 1/(\beta\epsilon).$$

The proof is complete. □

D.3 Stochastic Variance-Reduced Optimization

We prove here generalization bounds for various stochastic variance-reduced optimization algorithms. We formulate the framework in Algorithm 1.

Algorithm 1: Stochastic Variance Reduced Optimization

Input: step size η, 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 (4.4) or (4.5)
    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

We now consider the stochastic variance-reduced gradient descent (SVRG) (Reddi et al., 2016) and stochastically controlled stochastic gradient (SCSG) (Lei et al., 2017).

Theorem D.3. *Let Assumptions 1 and 2 hold with $L \leq n\beta/4$. Let A be either the SVRG in Reddi et al. (2016) or the SCSG in Lei et al. (2017). Then we can take $O\big(\big(n + n^{\frac{2}{3}}/\beta\big)\log n\big)$ stochastic gradient evaluations to get excess generalization bounds $O(1/(n\beta))$. Furthermore, if $\mathbb{E}[\hat{F}_S] = 0$, then we can take $O\big(\big(n + n^{\frac{2}{3}}/\beta\big)\log 1/(\beta\epsilon)\big)$ stochastic gradient evaluations to achieve the generalization bound $O(\epsilon)$ for any $\epsilon > 0$.*

Proof. To achieve $\mathbb{E}[F_S(A(S)) - \hat{F}_S] \leq 2/n$, it was shown that SVRG and SCSG requires $O\big(\big(n + n^{\frac{2}{3}}/\beta\big)\log n\big)$ stochastic gradient evaluations (Reddi et al., 2016; Lei et al., 2017). We plug this optimization error bound into Theorem 1 and get $\mathbb{E}[F(A(S))] - F(\mathbf{w}^*) = O(1/(n\beta))$.

We now consider the case $\mathbb{E}[\hat{F}_S] = 0$. According to (3.2), to achieve generalization bound $O(\epsilon)$, it suffices that $\mathbb{E}[F(A(S)) - \hat{F}_S] = O(\beta\epsilon)$. This can be achieved by taking $O\big(\big(n + n^{\frac{2}{3}}/\beta\big)\log 1/(\beta\epsilon)\big)$ stochastic gradient evaluations (Reddi et al., 2016; Lei et al., 2017). The proof is complete. □

We now present the proof of Theorem 8 on the behavior of the stochastic recursive gradient algorithm (SARAH) (Nguyen et al., 2017) and SpiderBoost (Wang et al., 2019).

Proof of Theorem 8. To achieve $\mathbb{E}[F_S(A(S)) - \hat{F}_S] \leq 2/n$, it was shown that SARAH and SpiderBoost requires $O\big(\big(n + 1/\beta^2\big)\log n\big)$ stochastic gradient evaluations (Nguyen et al., 2017; Wang

et al., 2019). We plug this optimization error bound into Theorem 1 and get $\mathbb{E}[F(A(S))] - F(\mathbf{w}^*) = O(1/(n\beta))$.

We now consider the case $\mathbb{E}[\hat{F}_S] = 0$. According to (3.2), to achieve generalization bound $O(\epsilon)$, it suffices that $\mathbb{E}[F(A(S)) - \hat{F}_S] = O(\beta\epsilon)$. This can be achieved by taking $O\big(\big(n + 1/\beta^2\big) \log 1/(\beta\epsilon)\big)$ stochastic gradient evaluations (Nguyen et al., 2017; Wang et al., 2019). The proof is complete. □

Finally, we consider SNVRG-PL (Zhou et al., 2018a).

Theorem D.4. *Let Assumptions 1 and 2 hold with $L \leq n\beta/4$. Let A be the SNVRG-PL in Zhou et al. (2018a). Then we can take $O\big(\big(n + \sqrt{n}/\beta\big) \log^4 n\big)$ stochastic gradient evaluations to get excess generalization bounds $O(1/(n\beta))$. Furthermore, if $\mathbb{E}[\hat{F}_S] = 0$, then we can take $O\big(\big(n + \sqrt{n}/\beta\big) \log^4 \frac{1}{\beta\epsilon}\big)$ stochastic gradient evaluations to achieve the generalization bound $O(\epsilon)$ for any $\epsilon > 0$.*

Proof. To achieve $\mathbb{E}[F_S(A(S)) - \hat{F}_S] \leq 2/n$, it was shown that SNVRG-PL requires $O\big(\big(n + \sqrt{n}/\beta\big) \log^4 n\big)$ stochastic gradient evaluations (Zhou et al., 2018a). We plug this optimization error bound into Theorem 1 and get $\mathbb{E}[F(A(S))] - F(\mathbf{w}^*) = O(1/(n\beta))$.

We now consider the case $\mathbb{E}[\hat{F}_S] = 0$. According to (3.2), to achieve generalization bound $O(\epsilon)$, it suffices that $\mathbb{E}[F(A(S)) - \hat{F}_S] = O(\beta\epsilon)$. This can be achieved by taking $O\big(\big(n + \sqrt{n}/\beta\big) \log^4 1/(\beta\epsilon)\big)$ stochastic gradient evaluations (Zhou et al., 2018a). The proof is complete. □

E Discussions of Examples

In this section, we present some discussions on understanding the assumption $L/\beta < n/2$ in Theorem 2.

E.1 Discussion of Example 1

We first give the definition of strong convexity. For any differentiable function $g : \mathcal{W} \mapsto \mathbb{R}$, we say g is σ-strongly convex if for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ there holds

$$g(\mathbf{w}') \geq g(\mathbf{w}) + \langle \mathbf{w}' - \mathbf{w}, \nabla g(\mathbf{w}) \rangle + \frac{\sigma}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2.$$

Introduce $g : \mathbb{R}^n \mapsto \mathbb{R}_+$ by $g(\mathbf{v}) = \frac{1}{n} \sum_{i=1}^n \ell(v_i, y_i)$. Then the function F_S can be written as

$$F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\langle \mathbf{w}, \phi(x_i) \rangle, y_i) = g(A\mathbf{w}),$$

where $A = \big(\phi(x_1), \ldots, \phi(x_n)\big)^\top \in \mathbb{R}^{n \times m}$ is the matrix formed from the data. It is known that if g is σ_g-strongly convex, then F_S satisfies the PL condition (Karimi et al., 2016; Necoara et al., 2018)

$$F_S(\mathbf{w}) - \hat{F}_S \leq \frac{1}{2\sigma_g \big(\sigma'_{\min}(A)\big)^2} \|\nabla F_S(\mathbf{w})\|_2^2. \tag{E.1}$$

Since ℓ is σ_ℓ-strongly convex we know for any $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^n$

$$\begin{aligned} g(\mathbf{v}') = \frac{1}{n} \sum_{i=1}^n \ell(v_i', y_i) &\geq \frac{1}{n} \sum_{i=1}^n \ell(v_i, y_i) + \frac{1}{n} \sum_{i=1}^n \ell'(v_i, y_i)(v_i' - v_i) + \frac{\sigma_\ell}{2n} \sum_{i=1}^n (v_i - v_i')^2 \\ &= g(\mathbf{v}) + \langle \nabla g(\mathbf{v}), \mathbf{v}' - \mathbf{v} \rangle + \frac{\sigma_\ell}{2n} \|\mathbf{v}' - \mathbf{v}\|_2^2. \end{aligned} \tag{E.2}$$

That is, g is $\frac{\sigma_\ell}{n}$-strongly convex. This together with (E.1) shows that

$$F_S(\mathbf{w}) - \hat{F}_S \leq \frac{n}{2\sigma_\ell \big(\sigma'_{\min}(A)\big)^2} \|\nabla F_S(\mathbf{w})\|_2^2 = \frac{1}{2\sigma_\ell \sigma'_{\min}(\Sigma_S)} \|\nabla F_S(\mathbf{w})\|_2^2, \tag{E.3}$$

where we have used

$$\frac{1}{n}\big(\sigma'_{\min}(A)\big)^2 = \frac{1}{n}\sigma'_{\min}(A^\top A) = \sigma'_{\min}(\Sigma_S). \tag{E.4}$$

For any $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^n$, it follows from the L_ℓ-strong smoothness of ℓ that

$$\|\nabla g(\mathbf{v}) - \nabla g(\mathbf{v}')\|_2^2 = \frac{1}{n^2}\sum_{i=1}^n \big|\ell'(v_i, y_i) - \ell'(v_i', y_i)\big|^2 \le \frac{L_\ell^2}{n^2}\sum_{i=1}^n |v_i - v_i'|^2 = \frac{L_\ell^2}{n^2}\|\mathbf{v} - \mathbf{v}'\|_2^2.$$

That is, g is $\frac{L_\ell}{n}$-smooth. It then follows

$$\begin{aligned}\|\nabla F_S(\mathbf{w}) - \nabla F_S(\mathbf{w}')\|_2 &= \big\|A^\top \nabla g(A\mathbf{w}) - A^\top \nabla g(A\mathbf{w}')\big\|_2 \le \sigma_{\max}(A)\|\nabla g(A\mathbf{w}) - \nabla g(A\mathbf{w}')\|_2 \\ &\le \frac{L_\ell \sigma_{\max}(A)}{n}\|A(\mathbf{w} - \mathbf{w}')\|_2 \le \frac{L_\ell \sigma_{\max}^2(A)}{n}\|\mathbf{w} - \mathbf{w}'\|_2.\end{aligned}$$

This together with (E.4) shows ($\sigma'_{\min}$ replaced by $\sigma_{\max}$)

$$\|\nabla F_S(\mathbf{w}) - \nabla F_S(\mathbf{w}')\|_2 \le L_\ell \sigma_{\max}(\Sigma_S)\|\mathbf{w} - \mathbf{w}'\|_2. \tag{E.5}$$

It is reasonable to assume that L is of the order of the smoothness of F_S. In this case, it follows from (E.3) and (E.5) that the term L/β is of the order of $\sigma_{\max}(\Sigma_S)/\sigma'_{\min}(\Sigma_S)$.

E.2 Discussion of Example 2

We recall some notations in Example 2. Consider single-hidden-layer neural networks with d inputs, m hidden neurons and a single output, for which the prediction function takes the form $h_{\mathbf{v},\mathbf{w}} = \sum_{k=1}^m v_k \phi\big(\langle \mathbf{w}_k, x\rangle\big)$. Here $\mathbf{w}_k \in \mathbb{R}^d$ and $v_k \in \mathbb{R}$ denote the weight of the edges connecting the k-th hidden node to the input and output node, respectively, while $\phi : \mathbb{R} \mapsto \mathbb{R}$ is the activation function. We fix $\mathbf{v}$ with $|v_k| = a$ for some $a > 0$ and train $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)^\top \in \mathbb{R}^{m\times d}$ from S. Note we only use the PL condition $F_S(\mathbf{w}) - \hat{F}_S \le \frac{1}{2\beta}\|\nabla F_S(\mathbf{w})\|_2^2$ for $\mathbf{w} = \mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ in the proof of Theorem 1 (only in (B.6)). We fix $\mathbf{w} = \mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ here. Analogous to Soltanolkotabi et al. (2019), we define the Jacobian matrix $J = \big(J_1, J_2, \ldots, J_n\big) \in \mathbb{R}^{md\times n}$ at $\mathbf{w} = \mathbf{w}_{S^{(i)}}^{(S^{(i)})}$ with

$$J_j = \begin{pmatrix} v_1\phi'(\mathbf{w}_1^\top x_j)x_j \\ \vdots \\ v_m\phi'(\mathbf{w}_m^\top x_j)x_j \end{pmatrix}$$

and $r_j = \mathbf{v}^\top \phi(\mathbf{w}x_j) - y_j$ for $j \in [n]$. It was shown that (Soltanolkotabi et al., 2019)

$$\nabla F_S(\mathbf{w}) = \frac{1}{n}J\mathbf{r}, \quad \text{for} \quad \mathbf{r} = (r_1, \ldots, r_n)^\top, \tag{E.6}$$

and therefore

$$\|\nabla F_S(\mathbf{w})\|_2^2 = \frac{1}{n^2}\mathbf{r}^\top J^\top J\mathbf{r} = \frac{1}{n^2}\mathbf{r}^\top\big(J_j^\top J_{j'}\big)_{j,j'\in[n]}\mathbf{r}.$$

According to the definition J_j, we know

$$\begin{aligned}J_j^\top J_{j'} &= \Big(v_1\phi'(\mathbf{w}_1^\top x_j)x_j^\top, \ldots, v_m\phi'(\mathbf{w}_m^\top x_j)x_j^\top\Big)\begin{pmatrix} v_1\phi'(\mathbf{w}_1^\top x_{j'})x_{j'} \\ \vdots \\ v_m\phi'(\mathbf{w}_m^\top x_{j'})x_{j'} \end{pmatrix} \\ &= a^2\sum_{k=1}^m \phi'(\mathbf{w}_k^\top x_j)\phi'(\mathbf{w}_k^\top x_{j'})x_j^\top x_{j'}.\end{aligned}$$

It then follows that

$$J^\top J = a^2\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top), \tag{E.7}$$

where $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}$ is the data matrix and $\odot$ denotes the Hadamard (entry-wise) product of matrices. According to the definition of $\mathbf{r}$, we know $F_S(\mathbf{w}) = \frac{1}{n}\|\mathbf{r}\|_2^2$. Then, it follows from (E.6) and (E.7) that

$$\begin{aligned}
\|\nabla F_S(\mathbf{w})\|_2^2 &= \frac{a^2}{n^2}\mathbf{r}^\top\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)\mathbf{r} \\
&\geq \frac{a^2}{n^2}\sigma_{\min}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)\|\mathbf{r}\|_2^2 \\
&= \frac{a^2}{n}\sigma_{\min}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)F_S(\mathbf{w}).
\end{aligned}$$

That is, we can take the parameter of the PL condition as

$$\beta = \frac{a^2}{2n}\sigma_{\min}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big).$$

It is reasonable to assume that L is of the order of $\frac{a^2}{n}\sigma_{\max}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)$ (Soltanolkotabi et al., 2019). In this case, we have

$$L/\beta \asymp \frac{\sigma_{\max}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)}{\sigma_{\min}\Big(\sum_{k=1}^m \Big(\phi'(X\mathbf{w}_k)\big(\phi'(X\mathbf{w}_k)\big)^\top\Big) \odot (XX^\top)\Big)}.$$

If we consider the identify activation function, i.e., $\phi(t) = t$, then it follows from the definition of Σ_S that

$$L/\beta \asymp \frac{\sigma_{\max}\big(XX^\top\big)}{\sigma_{\min}\big(XX^\top\big)} = \frac{\sigma_{\max}\big(\Sigma_S\big)}{\sigma_{\min}\big(\Sigma_S\big)}.$$