Machine Learning (2022) 111:345–375
https://doi.org/10.1007/s10994-021-06056-w

Understanding generalization error of SGD in nonconvex optimization

Yi Zhou[1] · **Yingbin Liang**[2] · **Huishuai Zhang**[3]

Received: 20 April 2019 / Revised: 25 July 2021 / Accepted: 7 August 2021 /
Published online: 24 September 2021

Abstract
The success of deep learning has led to a rising interest in the generalization property of the stochastic gradient descent (SGD) method, and stability is one popular approach to study it. Existing generalization bounds based on stability do not incorporate the interplay between the optimization of SGD and the underlying data distribution, and hence cannot even capture the effect of randomized labels on the generalization performance. In this paper, we establish generalization error bounds for SGD by characterizing the corresponding stability in terms of the on-average variance of the stochastic gradients. Such characterizations lead to improved bounds on the generalization error of SGD and experimentally explain the effect of the random labels on the generalization performance. We also study the regularized risk minimization problem with strongly convex regularizers, and obtain improved generalization error bounds for the proximal SGD.

Keywords Generalization error · Stochastic gradient descent · Nonconvex machine learning

Editor: Tong Zhang.

✉ Yi Zhou
yi.zhou@utah.edu

Yingbin Liang
liang.889@osu.edu

Huishuai Zhang
huzhang@microsoft.com

[1] Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, USA

[2] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA

[3] Microsoft Research Asia, Beijing, China

1 Introduction

Many machine learning applications can be formulated as risk minimization problems, in which each data sample $\mathbf{z} \in \mathbb{R}^p$ is assumed to be generated by an underlying multivariate distribution $\mathcal{D}$. The loss function $\ell(\cdot;\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}$ measures the performance on the sample $\mathbf{z}$ and its form depends on specific applications, e.g., square loss for linear regression problems, logistic loss for classification problems and cross entropy loss for training deep neural networks, etc. The goal is to solve the following population risk minimization (PRM) problem over a certain parameter space $\Omega \subset \mathbb{R}^d$.

$$\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\, \ell(\mathbf{w};\mathbf{z}). \tag{PRM}$$

Directly solving the PRM can be difficult in practice, as either the distribution $\mathcal{D}$ is unknown or evaluation of the expectation of the loss function induces high computational cost. To avoid such difficulties, one usually samples a set of n data samples $S := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ from the distribution $\mathcal{D}$, and instead solves the following empirical risk minimization (ERM) problem.

$$\min_{\mathbf{w}\in\Omega} f_S(\mathbf{w}) := \frac{1}{n}\sum_{k=1}^{n} \ell(\mathbf{w};\mathbf{z}_k). \tag{ERM}$$

The ERM serves as an approximation of the PRM with finite samples. In particular, when the number n of data samples is large, one wishes that the solution $\mathbf{w}_S$ found by optimizing the ERM with the data set S has a good generalization performance, i.e., it also induces a small loss on the population risk. The gap between these two risk functions is referred to as the *generalization error* at $\mathbf{w}_S$, and is formally written as

$$\text{(generalization error)} := |f_S(\mathbf{w}_S) - f(\mathbf{w}_S)|. \tag{1}$$

Various theoretical frameworks have been established to study the generalization error from different aspects (see related work for references). This paper adopts the stability framework (Bousquet and Elisseeff 2002; Elisseeff et al. 2005), which has been applied to study the generalization property of the output produced by learning algorithms. More specifically, for a particular learning algorithm $\mathcal{A}$, its stability corresponds to how stable the output of the algorithm is with regard to the variations in the data set. As an example, consider two data sets S and $\overline{S}$ that differ at one data sample, and denote $\mathbf{w}_S$ and $\mathbf{w}_{\overline{S}}$ as the outputs of algorithm $\mathcal{A}$ when applied to solve the ERM with the data sets S and $\overline{S}$, respectively. Then, the stability of the algorithm measures the gap between the output function values of the algorithm on the perturbed data sets.

Recently, the stability framework has been further developed to study the generalization performance of the output produced by the stochastic gradient descent (SGD) method from various theoretical aspects (Hardt et al. 2016; Charles and Papailiopoulos 2017; Mou et al. 2017; Yin et al. 2017; Kuzborskij and Lampert 2017). These studies showed that the output of SGD can achieve a vanishing generalization error after multiple passes over the data set as the sample size $n \to \infty$. These results provide theoretical justifications in part to the success of SGD on training complex objectives such as deep neural networks.

However, as pointed out in Zhang et al. (2017), these bounds do not explain some experimental observations, e.g., they do not capture the change of the generalization performance as the fraction of random labels in training data changes. Thus, the aim of this paper is to develop better generalization bounds that incorporate both the optimization

information of SGD and the underlying data distribution, so that they can explain experimental observations. We summarize our contributions as follows.

1.1 Our contributions

For smooth nonconvex optimization problems, we propose a new analysis of the on-average stability of SGD that exploits the optimization properties as well as the underlying data distribution. Specifically, via upper-bounding the on-average stability of SGD, we provide a novel generalization error bound, which improves upon the existing bounds by incorporating the on-average *variance* of the stochastic gradient. We further corroborate the connection of our bound to the generalization performance of the recent experiments in Zhang et al. (2017), which were not explained by the existing bounds of the same type. In specific, our experiments demonstrate that the obtained generalization bound captures how the generalization error changes with the fraction of random labels via the on-average *variance* of SGD. Furthermore, our bound holds under probabilistic guarantee, which is statistically stronger than the bounds in expectation provided in, e.g., Hardt et al. (2016), Kuzborskij and Lampert (2017). Then, we study nonconvex optimization under gradient dominance condition, and show that the corresponding generalization bound for SGD can be improved by its fast convergence rate.

We further consider nonconvex problems with strongly convex regularizers, and study the role that the regularization plays in characterizing the generalization error bound of the proximal SGD. In specific, our generalization bound shows that strongly convex regularizers substantially improve the generalization bound of SGD for *nonconvex* loss functions to be as good as the strongly convex loss function. Furthermore, the uniform stability of SGD under a strongly convex regularizer yields a generalization bound for *nonconvex* problems with exponential concentration in probability. We also provide some experimental observations to support our result.

1.2 Related works

The stability approach was initially proposed by Bousquet and Elisseeff (2002) to study the generalization error, where various notions of stability were introduced to provide bounds on the generalization error with probabilistic guarantee. Elisseeff et al. (2005) further extended the stability framework to characterize the generalization error of randomized learning algorithms. Shalev-Shwartz et al. (2010) developed various properties of stability on learning problems. In Hardt et al. (2016), the authors first applied the stability framework to study the expected generalization error for SGD, and Kuzborskij and Lampert (2017) further provided a data dependent generalization error bound. In Mou et al. (2017), the authors studied the generalization error of SGD with additive Gaussian noise. Yin et al. (2017) studied the role that gradient diversity plays in characterizing the expected generalization error of SGD. All these works studied the expected generalization error of SGD. In Charles and Papailiopoulos (2017), the authors studied the generalization error of several first-order algorithms for loss functions satisfying the gradient dominance and the quadratic growth conditions. Poggio et al. (2011) studied the stability of online learning algorithms. This paper improves the existing bounds by incorporating the on-average variance of SGD into the generalization error bound and further corroborates its connection to the generalization performance via experiments. More detailed comparison with the existing bounds are given after the presentation of main results.

The PAC Bayesian theory (Valiant 1984; McAllester 1999) is another popular framework for studying the generalization error in machine learning. It was recently used to develop bounds on the generalization error of SGD (London 2017; Mou et al. 2017). Specifically, Mou et al. (2017) applied the PAC Bayesian theory to study the generalization error of SGD with additive Gaussian noise. London (2017) combined the stability framework with the PAC Bayesian theory and provided bound on the generalization error with probabilistic guarantee of SGD. The bound incorporates the divergence between the prior distribution and the posterior distribution of the parameters.

Recently, Russo and Zou (2016), Xu and Raginsky (2017) applied information-theoretic tools to characterize the generalization capability of learning algorithms, and Pensia et al. (2018) further extended the framework to study the generalization error of various first-order algorithms with noisy updates. Other approaches were also developed for characterizing the generalization error as well as the estimation error, which include, for example, the algorithm robustness framework (Xu and Mannor 2012; Zahavy et al. 2017), large margin theory (Bartlett et al. 2017; Neyshabur et al. 2018; Sokolić et al. 2017) and the classical VC theory (Vapnik 1995; Vapnik 1998). Also, some methods have been developed to study excessive risk of the output for a learning algorithm, which include the robust stochastic approach (Nemirovski et al. 2009), the sample average approximation approach (Shapiro and Nemirovski 2005; Lin and Rosasco 2017), etc.

2 Preliminary and on-average stability

Consider applying SGD to solve the empirical risk minimization (ERM) with a particular data set S. In particular, at each iteration t, the algorithm samples one data sample from the data set S uniformly at random. Denote the index of the sampled data sample at the t-th iteration as ξ_t. Then, with a stepsize sequence $\{\alpha_t\}_t$ and a fixed initialization $\mathbf{w}_0 \in \mathbb{R}^d$, the update rule of SGD can be written as, for $t = 0, \ldots, T-1$,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\xi_t}). \tag{SGD}$$

Throughout the paper, we denote the iterate sequence along the optimization path as $\{\mathbf{w}_{t,S}\}_t$, where S in the subscript indicates that the sequence is generated by the algorithm using the data set S. The stepsize sequence $\{\alpha_t\}_t$ is a decreasing and positive sequence, and typical choices for SGD are $\frac{1}{t}$, $\frac{1}{t \log t}$ Bottou (2010), which we adopt in our study.

Clearly, the output $\mathbf{w}_{T,S}$ is determined by the data set S and the sample path $\boldsymbol{\xi} := \{\xi_1, \ldots, \xi_{T-1}\}$ of SGD. We are interested in the generalization error of the T-th output generated by SGD, i.e., $|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})|$, and we adopt the following standard assumptions (Hardt et al. 2016; Kuzborskij and Lampert 2017) on the loss function ℓ in our study throughout the paper.

Assumption 1 For all $\mathbf{z} \sim \mathcal{D}$, the loss function satisfies:

1. Function $\ell(\cdot\,; \mathbf{z})$ is continuously differentiable;
2. Function $\ell(\cdot\,; \mathbf{z})$ is nonnegative and σ-Lipschitz, and $|\ell(\cdot\,; \mathbf{z})|$ is uniformly bounded by M;
3. The gradient $\nabla \ell(\cdot\,; \mathbf{z})$ is L-Lipschitz, and $\|\nabla \ell(\cdot\,; \mathbf{z})\|$ is uniformly bounded by σ, where $\|\cdot\|$ denotes the ℓ_2 norm.

The generalization error of SGD can be viewed as a nonnegative random variable whose randomnesses are due to the draw of the data set S and the sample path $\boldsymbol{\xi}$ of the algorithm. In particular, the mean square generalization error has been studied in Elisseeff et al. (2005) for general randomized learning algorithms. Specifically, an application of Lemma 11 (Elisseeff et al. 2005) to SGD under Assumption 1 yields the following result. Throughout the paper, we denote $\overline{S}$ as the data set that replaces one data sample of S with an i.i.d copy generated from the distribution $\mathcal{D}$ and denote $\mathbf{w}_{T,\overline{S}}$ as the output of SGD for solving the ERM with the data set $\overline{S}$.

Proposition 1 *Let Assumption 1 hold. Apply the SGD with the same sample path $\boldsymbol{\xi}$ to solve the ERM with the data sets S and $\overline{S}$, respectively. Then, the mean square generalization error of SGD satisfies*

$$\mathbb{E}[|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})|^2] \leq \frac{2M^2}{n} + 12M\sigma\mathbb{E}[\delta_{T,S,\overline{S}}],$$

where $\delta_{T,S,\overline{S}} := \|\mathbf{w}_{T,S} - \mathbf{w}_{T,\overline{S}}\|$ and the expectation is taken over the random variables $\overline{S}, S$ and $\boldsymbol{\xi}$.

Proposition 1 links the mean square generalization error of SGD to the quantity $\mathbb{E}_{\boldsymbol{\xi},S,\overline{S}}[\delta_{T,S,\overline{S}}]$. Intuitively, $\delta_{T,S,\overline{S}}$ captures the variation of the algorithm output with regard to the variation of the dataset. Hence, its expectation can be understood as the *on-average stability* of the iterates generated by SGD. We note that similar notions of stabilities were proposed in Kuzborskij and Lampert (2017), Shalev-Shwartz et al. (2010), Elisseeff et al. (2005), which are based on the variation of the function values at the output instead.

3 Generalization bound for SGD in nonconvex optimization

In this section, we develop the generalization error of SGD by characterizing the corresponding on-average stability of the algorithm.

An intrinsic quantity that affects the optimization path of SGD is the variance of the stochastic gradients. To capture the impact of the variance of the stochastic gradients, we adopt the following standard assumption from the stochastic optimization theory (Bottou 2010; Nemirovski et al. 2009; Ghadimi et al. 2016).

Assumption 2 For any fixed training set S and any ξ that is generated uniformly from $\{1, \ldots, n\}$ at random, there exists a constant $\nu_S > 0$ such that for all $\mathbf{w} \in \Omega$ one has

$$\mathbb{E}_{\xi}\left\|\nabla \ell(\mathbf{w};\mathbf{z}_{\xi}) - \frac{1}{n}\sum_{k=1}^{n}\nabla \ell(\mathbf{w};\mathbf{z}_k)\right\|^2 \leq \nu_S^2. \tag{2}$$

Assumption 2 essentially bounds the variance of the stochastic gradients for the particular data set S. The variance ν_S^2 of the stochastic gradient is typically much smaller than the uniform upper bound σ in Assumption 1 for the norm of the stochastic gradient, e.g., normal random variable has unit variance and is unbounded, and hence may provide a tighter bound on the generalization error.

Based on Assumption 2 and Assumption 1, we obtain the following generalization bound of SGD by exploring its optimization path to study the corresponding stability.

Theorem 1 (Bound with Probabilistic Guarantee) *Suppose ℓ is nonconvex. Let Assumptions 1 and 2 hold. Apply the SGD to solve the ERM with the data set S, and choose the step size $\alpha_t = \frac{c}{(t+2)\log(t+2)}$ with $0 < c < \frac{1}{L}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})| \leq \sqrt{\frac{1}{n\delta}\left(2M^2 + 24M\sigma c\sqrt{2Lf(\mathbf{w}_0) + \frac{1}{2}\mathbb{E}_S[v_S^2]}\log T\right)}.$$

An important variable in the above generalization bound is the on-average stochastic variance $\mathbb{E}_S[v_S^2]$. We can compare the above bound with the generalization bound developed in the recent literature. Specifically, Hardt et al. (2016), Kuzborskij and Lampert (2017), Yin et al. (2017) all developed bounds for the expected generalization error of SGD and choose the step size $\alpha_t = \frac{c}{t}$, while our generalization bound in the above theorem is probabilistic and hence provides stronger guarantee, and we use a slightly smaller step size $\alpha_t = \frac{c}{(t+2)\log(t+2)}$. The generalization bound in Hardt et al. (2016) is based on the uniform stability $\sup_{S,\bar{S}} \mathbb{E}_{\boldsymbol{\xi}}[\delta_{T,S,\bar{S}}]$ and assumes an upper bound σ of the norm of all gradients. Kuzborskij and Lampert (2017) develops a data-dependent bound on expected generalization error by leveraging the notion of on-average stability, and they adopt an additional assumption on the Lipschitz continuity of the Hessian matrix. Yin et al. (2017) characterizes the expected generalization error of SGD using the notion of uniform stability and gradient diversity, but their analysis requires the function to be (strongly)-convex. In comparison, our generalization bound is based on the more relaxed on-average stability $\mathbb{E}_{S,\bar{S}}\mathbb{E}_{\boldsymbol{\xi}}[\delta_{T,S,\bar{S}}]$ that allows us to introduce the on-average variance, which is generally smaller and tighter than the uniform gradient bound σ used in Hardt et al. (2016). Moreover, the generalization error bounds in all these works have a polynomial dependence on T, whereas our generalization error bound only scales with $\log T$. Next, we outline the proof of Theorem 1 below and discuss other implications.

Outline of the Proof of Theorem 1 We provide an outline of the proof here, and relegate the detailed proof in the supplementary materials.

The central idea is to bound the on-average stability $\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{T,S,\bar{S}}]$ of the iterates in Proposition 1. Hence, suppose we apply SGD with the same sample path $\boldsymbol{\xi}$ to solve the ERM with the data sets S and $\bar{S}$, respectively. We first obtain the following recursive property of the on-average iterate stability (Lemma 2 in the appendix):

$$\begin{aligned}\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{t+1,S,\bar{S}}] \leq (1+\alpha_t L)\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{t,S,\bar{S}}] \\ + \frac{2\alpha_t}{n}\mathbb{E}_{S,\boldsymbol{\xi}}\left[\left\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right].\end{aligned} \tag{3}$$

We then further derive the following bound on $\mathbb{E}_{S,\boldsymbol{\xi}}\left[\left\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right]$ by exploiting the optimization path of SGD (Lemma 3 in the appendix):

$$\mathbb{E}_{\boldsymbol{\xi},S}\left[\left\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right] \leq \sqrt{2Lf(\mathbf{w}_0) + \frac{1}{2}\mathbb{E}_S[v_S^2]}. \tag{4}$$

Substituting (4) into (3) and telescoping, we obtain an upper bound on $\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{T,S,\bar{S}}]$. Further substituting such a bound into Proposition 1, we obtain an upper bound on the second

moment of the generalization error. Then, the result in Theorem 1 follows from the Chebyshev's inequality. □

The proof of Theorem 1 is to characterize the on-average stability of SGD, and it explores the optimization path by applying the technical tools developed in stochastic optimization theory. Comparing to the generalization bound developed in Hardt et al. (2016) that characterizes the expected generalization error based on the uniform stability $\sup_{S,\bar{S}} \mathbb{E}_{\xi}[\delta_{T,S,\bar{S}}]$, our generalization bound in Theorem 1 provides a probabilistic guarantee, and is based on the more relaxed on-average stability $\mathbb{E}_{S,\bar{S}}\mathbb{E}_{\xi}[\delta_{T,S,\bar{S}}]$ which yields a tighter bound. Intuitively, the on-average variance term $\mathbb{E}_S[v_S^2]$ in our bound measures the 'stability' of the stochastic gradients over all realizations of the dataset S. If such on-average variance of SGD is large, then the optimization paths of SGD on two slightly different datasets are diverse from each other, leading to the bad stability of SGD and in turn yielding a high generalization error.

Remark on optimization convergence rate: We note that the generalization error bound in Theorem 1 is derived based on the step size $\alpha_t = c/[(t+2)\log(t+2)]$. With this step size, the standard nonconvex optimization convergence rate of SGD (Bottou 2010) is in the order of

$$\frac{\sum_{t=0}^{T-1} \alpha_t^2}{\sum_{t=0}^{T-1} \alpha_t} = O\left(\frac{1}{\log\log T}\right),$$

which is very slow. However, it is possible to choose a proper step size to achieve a similar generalization error bound and a faster optimization convergence rate. Specifically, one can choose $\alpha_t = \frac{c}{t+2}$ with constant $c = \frac{\log\log(T+2)}{2L\log(T+2)}$, where T is the total number of iterations. Then, following the same proof of Theorem 1, one can instead prove the following stability bound

$$\mathbb{E}[\delta_T] \leq \frac{2\sqrt{2Lf(\mathbf{w}_0) + \frac{\mathbb{E}[v_S^2]}{2}}}{nL}(T+2)^{2cL} = O\left(\frac{\sqrt{Lf(\mathbf{w}_0) + \mathbb{E}[v_S^2]}}{nL}\log(T+2)\right).$$

Therefore, the generalization error bound still scales with $\log T$. On the other hand, the optimization convergence rate is now in the order of

$$\frac{\sum_{t=0}^{T-1} \alpha_t^2}{\sum_{t=0}^{T-1} \alpha_t} = O\left(\frac{\log\log(T+2)}{\log^2(T+2)}\right).$$

Remark on choice of step size: One can also adopt a constant step size in Theorem 1, which will lead to a very different line of proof and a different final bound. In this case, one can choose a sufficiently small constant step size (with polynomial dependence on the total number of iterations T) and obtain a comparable generalization bound.

Discussion: We next elaborate on how our generalization bound can help explain the observations in classification experiments with randomized labels (Zhang et al. 2017). Specifically, consider the following binary classification problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \exp(-y_i \mathbf{w}^T \mathbf{x}_i),$$

where $\mathbf{w}$ corresponds to the linear classifier and $(\mathbf{x}_i, y_i)$ denotes the i-th data sample (y_i is a binary label). Consider a simplified case where the feature dimension $d = 1$ and total sample size $N = 2n$. Assume the features $x_1 = x_2 = \cdots = x_n = 1$ and $x_{n+1} = x_{n+2} = \cdots = x_{2n} = -1$. In particular, assume that $\alpha \in (0, 0.5)$ portion of the $2n$ samples have incorrect labels, i.e., α portion of the samples in $\{x_1, x_2, \ldots, x_n\}$ are incorrectly labeled as '-1' (true label is '$+1$') and α portion of the samples in $\{x_{n+1}, x_{n+2}, \ldots, x_{2n}\}$ are incorrectly labeled as '$+1$' (true label is '-1'). In this setting, it can be calculated that the full gradient of the empirical loss is $\nabla f_S(\mathbf{w}) = \alpha \exp(\mathbf{w}) - (1-\alpha)\exp(-\mathbf{w})$. Then, the empirical gradient variance of any classifier $\mathbf{w}$ takes the value

$$\begin{aligned}
\mathbb{E}[v_S^2] &= \frac{1}{N}\sum_{i=1}^{N} \left\|\nabla \ell_i(\mathbf{w}) - \nabla f_S(\mathbf{w})\right\|^2 \\
&= \frac{1}{N}\left(\sum_{\substack{i\in\{1,\ldots,n\} \\ y_i = 1}} + \sum_{\substack{i\in\{1,\ldots,n\} \\ y_i = -1}} + \sum_{\substack{i\in\{n+1,\ldots,2n\} \\ y_i = 1}} + \sum_{\substack{i\in\{n+1,\ldots,2n\} \\ y_i = -1}} \right) \left\|\nabla \ell_i(w) - \nabla f_S(\mathbf{w})\right\|^2 \\
&= \alpha(1-\alpha)(\exp(\mathbf{w}) + \exp(-\mathbf{w}))^2.
\end{aligned}$$

Hence, as the random label probability α increases (from 0 to 0.5), the above empirical gradient variance keeps increasing and the generalization error also increases. In particular, the maximum variance is achieved when half of the data are incorrectly labeled, and this gives the maximum classification uncertainty. This example shows that the optimization gradient variance term in our Theorem 1 properly captures the impact of data distribution on the generalization performance. We note that one can generalize this example to high dimensional space $d > 1$ where the features follow two distinct normal distributions, and the conclusion will be the same but requires dedicated calculations.

4 Generalization bound for SGD under gradient dominant condition

In this section, we consider nonconvex loss functions with the empirical risk function f_S further satisfying the following gradient dominance condition.

Definition 1 Denote $f^* := \inf_{\mathbf{w}\in\Omega} f(\mathbf{w})$. Then, the function f is said to be γ-gradient dominant for $\gamma > 0$ if

$$f(\mathbf{w}) - f^* \le \frac{1}{2\gamma}\|\nabla f(\mathbf{w})\|^2, \ \forall \mathbf{w} \in \Omega. \tag{5}$$

The gradient dominance condition (also referred to as Polyak-Łojasiewicz condition (Polyak 1963; Łojasiewicz 1963) guarantees a linear convergence of the function value sequence generated by gradient-based first-order methods (Karimi et al. 2016). It is a condition that is much weaker than the strong convexity, and many nonconvex machine learning problems satisfy this condition around the global minimizers (Li et al. 2016; Zhou et al. 2016).

The gradient dominance condition helps to improve the bound on the on-average stochastic gradient norm $\mathbb{E}_{\xi,S}\big[\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\|\big]$ (see Lemma 4 in the appendix), which is given by

$$\mathbb{E}_{\xi,S}\big[\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\|\big] \leq \sqrt{2L\mathbb{E}_S[f_S^*] + \frac{1}{t}\big(2Lf(\mathbf{w}_0) + \mathbb{E}_S[v_S^2]\big)}. \tag{6}$$

Compared to (4) for general nonconvex functions, the above bound is further improved by a factor of $\frac{1}{t}$. This is because SGD converges sub-linearly to the optimum function value f_S^* under the gradient dominance condition, and $\frac{1}{t}$ is essentially the convergence rate of SGD. In particular, for sufficiently large t, the on-average stochastic gradient norm is essentially bounded by $\sqrt{2L\mathbb{E}_S[f_S^*]}$, which is much smaller then the bound in (4). With the bound in (6), we obtain the following theorem.

Theorem 2 (Mean Square Bound) *Suppose ℓ is nonconvex, and f_S is γ-gradient dominant ($\gamma < L$). Let Assumptions 1 and 2 hold. Apply the SGD to solve the ERM with the data set S and choose $\alpha_t = \frac{c}{(t+2)\log(t+2)}$ with $0 < c < \min\{\frac{1}{L}, \frac{1}{2\gamma}\}$. Then, the following bound holds.*

$$\mathbb{E}_{\xi,S}[|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})|^2] \leq \frac{2M^2}{n} + \frac{24M\sigma c}{n}\left(\sqrt{2L\mathbb{E}_S[f_S^*]}\log T + \sqrt{2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]}\right).$$

The above bound for the mean square generalization error under gradient dominance condition improves that for general nonconvex functions in Theorem 1, as the dominant term (i.e., $\log T$-dependent term) has coefficient $\sqrt{2L\mathbb{E}_S[f_S^*]}$, which is much smaller than the term $\sqrt{2Lf(\mathbf{w}_0) + \frac{1}{2}\mathbb{E}_S[v_S^2]}$ in the bound of Theorem 1. As an intuitively understanding, the on-average variance of the SGD is further reduced by its fast convergence rate $\frac{1}{t}$ under the gradient dominance condition. This results in a more stable on-average iterate stability which in turn improves the mean square generalization error. We note that Charles and Papailiopoulos (2017) also studied the generalization error of SGD for loss functions satisfying both the gradient dominance condition and an additional quadratic growth condition. They also assumed that the algorithm converges to a global minimizer point, which may not always hold for noisy algorithms like SGD.

Remark on optimization convergence rate: The optimization convergence rate of SGD under the gradient dominant condition has been characterized by the Theorem 4 of Karimi et al. (2016). In particular, with the step size $\alpha_t = O(\frac{1}{t})$, Karimi et al. (2016) proved that the convergence rate of SGD is in the order of $\mathbb{E}[f_S(\mathbf{w}_{t,S}) - f_S^*] \leq O(\frac{1}{t})$. Note that the generalization error bound in Theorem 2 is derived based on a slightly smaller step size $\alpha_t = O(\frac{1}{t\log t})$, which leads to the same order of convergence rate $\widetilde{O}(\frac{1}{t})$ up to certain logarithmic factors. Hence, under the gradient dominant condition, SGD can achieve a small generalization error as well as a fast convergence.

Theorem 2 directly implies the following *probabilistic* guarantee for the generalization error of SGD.

Theorem 3 *(Bound with Probabilistic Guarantee) Suppose ℓ is nonconvex, and f_S is γ-gradient dominant ($\gamma < L$). Let Assumptions 1 and 2 hold. Apply the SGD to solve the ERM with the data set S, and choose $\alpha_t = \frac{c}{(t+2)\log(t+2)}$ with $0 < c < \min\{\frac{1}{L}, \frac{1}{2\gamma}\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})| \leq \sqrt{\frac{2M^2}{n\delta} + \frac{24M\sigma c}{n\delta}\left(\sqrt{2L\mathbb{E}_S[f_S^*]}\log T + \sqrt{2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]}\right)}.$$

5 Regularized nonconvex optimization

In practical applications, regularization is usually applied to the risk minimization problem in order to either promote certain structures on the desired solution or to restrict the parameter space. In this section, we explore how regularization can improve the generation error, and hence help to avoid overfitting for SGD.

Here, for any weight $\lambda > 0$, we consider the regularized population risk minimization (R-PRM) and the regularized empirical risk minimization (R-ERM):

$$\min_{\mathbf{w}\in\Omega} \boldsymbol{\Phi}(\mathbf{w}) := f(\mathbf{w}) + \lambda h(\mathbf{w}), \tag{R-PRM}$$

$$\min_{\mathbf{w}\in\Omega} \boldsymbol{\Phi}_S(\mathbf{w}) := f_S(\mathbf{w}) + \lambda h(\mathbf{w}), \tag{R-ERM}$$

where h corresponds to the regularizer and f, f_S are the population and empirical risks, respectively. In particular, we are interested in the following class of regularizers.

Assumption 3 The regularizer function h is 1-strongly convex and nonnegative.

Without loss of generality, we assume that the strongly convex parameter of h is 1, and this can be adjusted by scaling the weight parameter λ. Strongly convex regularizers are commonly used in machine learning applications, and typical examples include $\frac{\lambda}{2}\|\mathbf{w}\|^2$ for ridge regression, Tikhonov regularization $\frac{\lambda}{2}\|\Gamma\mathbf{w}\|^2$ and elastic net $\lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|^2$, etc. Here, we allow the regularizer h to be non-differentiable (e.g., the elastic net), and introduce the following proximal mapping with parameter $\alpha > 0$ to deal with the non-smoothness.

$$\mathrm{prox}_{\alpha h}(\mathbf{w}) := \arg\min_{\mathbf{u}\in\Omega} h(\mathbf{u}) + \frac{1}{2\alpha}\|\mathbf{u} - \mathbf{w}\|^2. \tag{7}$$

The proximal mapping is the core of the proximal method for solving convex problems (Parikh and Boyd 2014; Beck and Teboulle 2009) and nonconvex ones (Li et al. 2017; Attouch et al. 2013). In particular, we apply the proximal SGD to solve the R-ERM. With the same notations as those defined in the previous section, the update rule of the proximal SGD can be written as, for $t = 0, \ldots, T-1$

$$\mathbf{w}_{t+1} = \mathrm{prox}_{\alpha_t h}\left(\mathbf{w}_t - \alpha_t \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\xi_t})\right). \tag{proximal-SGD}$$

Similarly, we denote $\{\mathbf{w}_{t,S}\}_t$ as the iterate sequence generated by the proximal SGD with the data set S.

It is clear that the generalization error of the function value for the regularized risk minimization, i.e., $|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})|$, is the same as that for the un-regularized risk minimization. Hence, Theorem 1 is also applicable to the mean square generalization error of the regularized risk minimization. However, the development of the generalization error bound

is different from the analysis in the previous section from two aspects. First, the analysis of the on-average iterate stability of the proximal SGD is technically more involved than that of SGD due to the possibly non-smooth regularizer. Secondly, the proximal mappings of strongly convex functions are strictly contractive (see item 2 of Proposition 5 in the appendix). Thus, the proximal step in the proximal SGD enhances the stability between the iterates $\mathbf{w}_{t,S}$ and $\mathbf{w}_{t,\bar{S}}$ that are generated by the algorithm using perturbed datasets, and this further improves the generalization error. The next result provides a quantitative statement.

Theorem 4 *Consider the regularized risk minimization. Suppose ℓ is nonconvex. Let Assumptions 1, 2 and3 hold, and apply the proximal SGD to solve the R-ERM with the dataset S. Let $\lambda > L$ and $\alpha_t = \frac{c}{t+2}$ with $0 < c < \frac{1}{L}$. Then, the following bound holds with probability at least $1 - \delta$.*

$$|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})| \leq \sqrt{\frac{1}{n\delta}\left(2M^2 + \frac{24M\sigma}{(\lambda - L)}\sqrt{L\boldsymbol{\Phi}(\mathbf{w}_0) + \mathbb{E}_S[v_S^2]}\right)}.$$

Theorem 4 provides *probabilistic* guarantee for the generalization error of the proximal SGD in terms of the on-average variance of the stochastic gradients. Comparison of Theorem 4 with Theorem 1 indicates that a strongly convex regularizer substantially improves the generalization error bound of SGD for nonconvex loss functions by removing the logarithm dependence on T. It is also interesting to compare Theorem 4 with [Proposition 4 and Theorem 1, London 2017], which characterize the generalization error of SGD for strongly convex functions with probabilistic guarantee. The two bounds have the same order in terms of n and T, indicating that a strongly convex regularizer even improves the generalization error for a nonconvex function to be the same as that for a strongly convex function. In practice, the regularization weight λ should be properly chosen to balance between the generalization error and the training loss, as otherwise the parameter space can be too restrictive to yield a good solution for the risk function.

5.1 Generalization bound with high-probability guarantee

The studies of the previous sections explore the *probabilistic* guarantee for the generalization errors of nonconvex loss functions and nonconvex loss functions with strongly convex regularizers. For example, apply SGD to solve a generic nonconvex loss function, then Theorem 1 suggests that for any $\epsilon > 0$,

$$\mathbb{P}(|f(\mathbf{w}_{T,S}) - f_S(\mathbf{w}_{T,S})| > \epsilon) < O\left(\frac{\log T}{n\epsilon^2}\right),$$

which decays sublinearly as $\frac{n}{\log T} \to \infty$. In this subsection, we study a stronger probabilistic guarantee for the generalization error, i.e., the probability for it to be less than ϵ decays *exponentially*. We refer to such a notion as high-probability guarantee. In particular, we explore for which cases of nonconvex loss functions we can establish such a stronger performance guarantee.

Towards this end, we adopt the uniform stability framework proposed in Elisseeff et al. (2005). Note that Hardt et al. (2016) also studied the uniform stability of SGD, but only characterized the generalization error in expectation, which is weaker than the *exponential* probabilistic concentrtion bound that we study here.

Suppose we apply SGD with the same sample path $\boldsymbol{\xi}$ to solve the ERM with the datasets S and $\overline{S}$, respectively, and denote $\mathbf{w}_{T,S,\boldsymbol{\xi}}$ and $\mathbf{w}_{T,\overline{S},\boldsymbol{\xi}}$ as the corresponding outputs. Also, suppose we apply the SGD with different sample paths $\boldsymbol{\xi}$ and $\overline{\boldsymbol{\xi}}$ to solve the same problem with the dataset S, respectively, and denote $\mathbf{w}_{T,S,\boldsymbol{\xi}}$ and $\mathbf{w}_{T,S,\overline{\boldsymbol{\xi}}}$ as the corresponding outputs. Here, $\overline{\boldsymbol{\xi}}$ denotes the sample path that replaces one of the sampled indices, say ξ_{t_0}, with an i.i.d copy ξ'_{t_0}. The following result is a variant of Theorem 15 (Elisseeff et al. 2005).

Lemma 1 *Let Assumption 1 hold. If SGD satisfies the following conditions*[1]

$$\sup_{S,\overline{S},\mathbf{z}} \mathbb{E}_{\boldsymbol{\xi}} |\ell(\mathbf{w}_{T,S,\boldsymbol{\xi}};\mathbf{z}) - \ell(\mathbf{w}_{T,\overline{S},\boldsymbol{\xi}};\mathbf{z})| \leq \beta,$$

$$\sup_{\boldsymbol{\xi},\overline{\boldsymbol{\xi}},S,\mathbf{z}} |\ell(\mathbf{w}_{T,S,\boldsymbol{\xi}};\mathbf{z}) - \ell(\mathbf{w}_{T,S,\overline{\boldsymbol{\xi}}};\mathbf{z})| \leq \rho.$$

Then, the following bound holds with probability at least $1-\delta$.

$$|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})| \leq 2\beta + \left(\frac{M+4n\beta}{\sqrt{2n}} + \sqrt{2T}\rho\right)\sqrt{\log\tfrac{2}{\delta}}.$$

Note that Theorem 1 implies that

$$\mathbb{P}(|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})| > \epsilon) \leq O\left(\exp\left(\frac{-\epsilon^2}{\sqrt{n}\beta+\sqrt{T}\rho}\right)\right).$$

Hence, if $\beta = o(n^{-\frac{1}{2}})$ and $\rho = o(T^{-\frac{1}{2}})$, then we have exponential decay in probability as $n \to \infty$ and $T \to \infty$. It turns out that our analysis of the uniform stability of SGD for general nonconvex functions yields that $\beta = O(n^{-1})$, $\rho = O(\log T)$, which does not lead to the desired high-probability guarantee for the generalization error. On the other hand, the analysis of the uniform stability of the proximal SGD for nonconvex loss functions with strongly convex regularizers yields that $\beta = O(n^{-1})$, $\rho = O(T^{-c(\lambda-L)})$, which leads to the high-probability guarantee if we choose $\lambda > L$ and $c > \frac{1}{2(\lambda-L)}$. This further demonstrates that a strongly convex regularizer can significantly improve the quality of the probabilistic bound for the generalization error. The following result is a formal statement of the above discussion.

Theorem 5 *Consider the regularized risk minimization with the nonconvex loss function* ℓ. *Let Assumptions 1 and 3 hold, and apply the proximal SGD to solve the R-ERM with the data set S. Choose* $\lambda > L$ *and* $\alpha_t = \frac{c}{t+2}$ *with* $\frac{1}{2(\lambda-L)} < c < \frac{1}{\lambda-L}$. *Then, the following bound holds with probability at least* $1-\delta$

$$|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})| \leq \left(\frac{M}{\sqrt{n}} + \frac{4\sigma^2}{\sqrt{n}(\lambda-L)} + \frac{4\sigma^2 c}{T^{c(\lambda-L)-\frac{1}{2}}}\right)\sqrt{\log\frac{2}{\delta}}.$$

Theorem 5 implies that

[1] Theorem 1 is slightly different from that in Theorem 15 (Elisseeff et al. 2005), in which $\overline{S}$ excludes a particular sample instead of replacing it. The proof follows the same idea and we omit it for simplicity.

$$\mathbb{P}(|\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S})| > \epsilon) \leq O\left(\exp\left(\frac{-\epsilon^2}{n^{-\frac{1}{2}} + T^{\frac{1}{2}-c(\lambda-L)}}\right)\right).$$

Hence, if we choose $c = \frac{1}{\lambda - L}$ and run the proximal SGD for $T = O(n)$ iterations (i.e., constant passes over the data), then the probability of the event decays exponentially as $O(\exp(-\sqrt{n}\epsilon^2))$.

The proof of Theorem 5 characterizes the uniform iterate stability of the proximal SGD with regard to the perturbations of both the dataset and the sample path. Unlike the on-average stability in Theorem 1 where the stochastic gradient norm is bounded by the on-average variance of the stochastic gradients, the uniform stability captures the worst case among all datasets, and hence uses the uniform upper bound σ for the stochastic gradient norm. We note that Theorem 3 (London 2017) also established a probabilistic bound under the PAC Bayesian framework. However, their result yields exponential concentration guarantee only for strongly convex loss functions. As a comparison, Theorem 5 relaxes the requirement of strong convexity for loss functions to *nonconvex* loss functions with strongly convex regularizers, and hence serves as a complementary result to theirs. Also, Mou et al. (2017) establishes the high-probability bound for the generalization error of SGD with regularization. However, their result holds only for the particular regularizer $\frac{1}{2}\|\mathbf{w}\|^2$, and high-probability bound holds only with regard to the random draw of the data. As a comparison, our result holds for all strongly convex regularizers, and the high-probability bound hold with regard to both the draw of data and randomness of algorithm.

6 Experiments

In this section, we conduct deep learning experiments to demonstrate that the on-average variance of SGD does correlate with the generalization performance in practice. Specifically, it has been observed that a classification dataset with randomized labels can substantially degrade the generalization performance of the trained deep model (Zhang et al. 2017). Following this observation, we consider training a three-layer MLP neural network and a ResNet-18 network (He et al. 2016) using the MNIST dataset (Lecun et al. 1998) and the CIFAR10 dataset (Krizhevsky 2009), respectively. For all the data labels in each dataset, we replace their underlying true labels with random labels with probability $p \in [0.0, 0.4]$. During the SGD training, we evaluate the on-average variance of SGD for the last multiple iterations of the training process. In all the experiments, we train the networks for a sufficient number of epochs until the training error is saturated. Also, as the on-average variance involves an expectation over the data distribution, we use the corresponding sample mean over the random draw of the training data as an approximation.

6.1 Generalization error and stability under random labels

In Fig. 1, we present the results of training MLP and ResNet-18 under the random label probability p ranging from 0.1 to 0.4. We use the learning rate 0.01 and batch size 256 for both experiments. It can be seen from these results that the on-average variance (blue) consistently increases as the fraction of random labels increases. At the same time, the generalization error (red) also increases. Thus, our empirical study confirms that the on-average variance captured in our generalization bound is correlated with the generalization performance in the experiments.

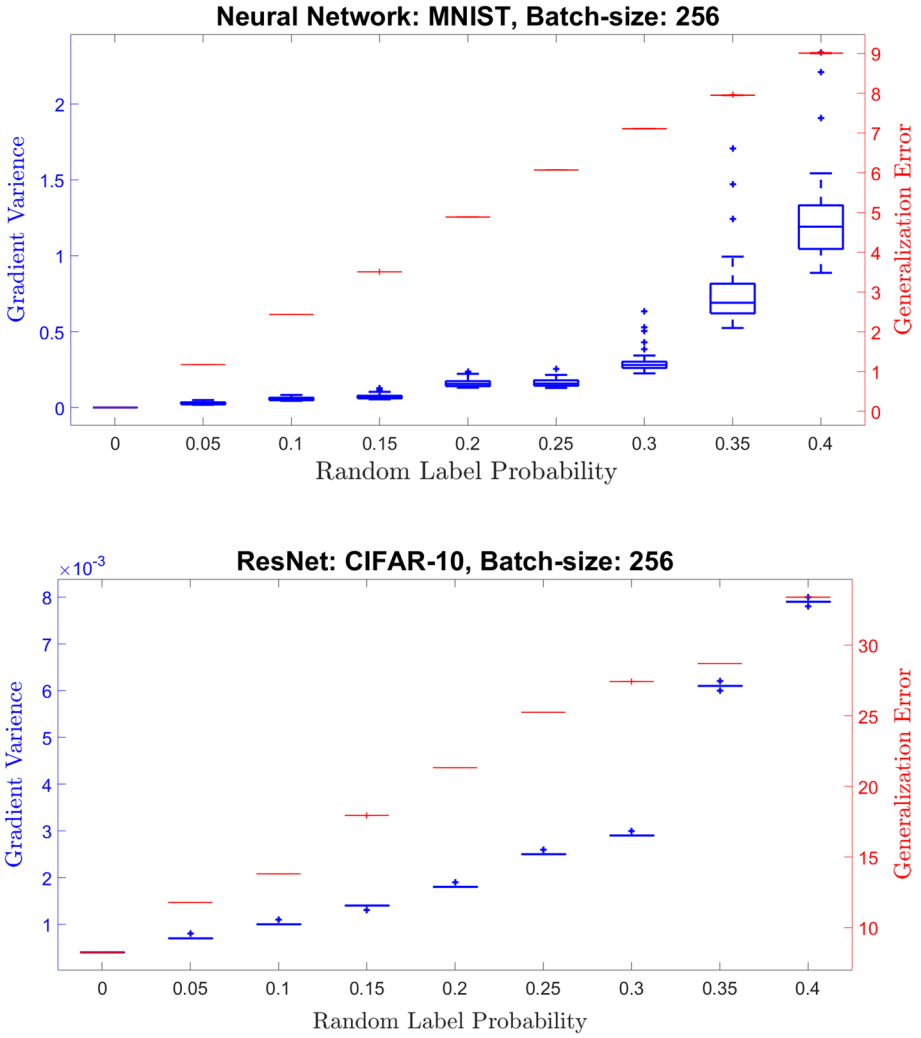


Fig. 1 Relation of on-average variance (left y axis), generalization error (right y axis) and random label probability (x axis) in training MLP and ResNet using SGD

We note that from the numerical result shown in Fig. 1, it seems that the generalization error does not exactly scale with the on-average variance in a way as predicted by Theorem 1. This is the nature and limit of the proposed statistical generalization theory, which only establishes bounds for a general class of functions. Characterizing the precise numerical dependence between generalization error and on-average variance is out of the scope of this work.

6.2 Impact of batch size and data augmentation

We further explore how the batch size and data augmentation affect the generalization error and on-average variance of SGD under random labels.

First, we explore the impact of batch size by considering three different batch sizes, i.e., 128, 192 and 256. We use the same learning rate 0.01 and vary the random label probability from 0.1 to 0.4. Figure 2 shows the training results of MLP and ResNet-18 with different batch sizes. It can be seen that the generalization error consistently correlates with the on-average variance under all random label probabilities. These observations support our theoretical findings. Also, by comparing these figures, it seems that the generalization error roughly stays at the same level as the batch size increases, while the on-average variance increases as the batch size increases. We think that this is because training with larger batch size with noisy labels makes it more challenging to reach the global minimum, and therefore the gradient variance remains large.

Next, we explore the impact of data augmentation on the generalization error and on-average variance. We train MLP and ResNet-18 using learning rate 0.01 and batch size 128 with the original datasets and their augmented versions. For the data augmentation, we apply the random rotation augmentation method to modify the images. Specifically for each image, we randomly rotate the image with a degree uniformly generated between -20 and 20 degrees. Figure 3 shows the training results with the original and augmented data. It can be seen that the generalization error consistently correlates with the on-average variance under all random label probabilities and data augmentation. In the MLP training with the MNIST dataset, data augmentation does not yield a substantial decrease of the generalization error, and the on-average variance is larger with augmented data than that with the original data. In the ResNet-18 training with the CIFAR10 dataset, data augmentation does lead to a consistent decrease of the generalization error under all random label probabilities, but the on-average variance is larger with augmented data. We think that this is because a subset of the augmented data samples that are assigned random labels increase the gradient uncertainty in optimization, and is not captured by the current theoretical framework. This suggests a research direction for future study.

6.3 Effect of regularization

We further conduct experiments to explore the effect of regularization on the generalization error by adding the regularizer $\frac{\lambda}{2}\|\mathbf{w}\|^2$ to the objective functions. In particular, we apply the proximal SGD to solve the logistic regression (with dataset a9a Chang and Lin 2011) and train the MLP network (with dataset MNIST). Figure 4 shows the results where the left axis denotes the scale of the training error and the right axis denotes the scale of the generalization error. It can be seen that the corresponding generalization errors improve as the regularization weight gets large. This matches our theoretical finding on the impact of regularization. On the other hand, the training performances for both problems degrade as the regularization weight increases, which is reasonable because in such a case the optimization focuses too much on the regularizer and the obtained solution does not minimize the loss function well. Hence, there is a trade-off between the training and generalization performance in tuning the regularization parameter.

7 Conclusion

In this paper, we develop the generalization error bound of SGD with probabilistic guarantee for nonconvex optimization. We obtain the improved bounds based on the variance of the stochastic gradients by exploiting the optimization path of SGD. Our generalization

Fig. 2 Comparison of on-average variance, generalization error and random label probability under different choices of batch size of SGD ▶

bound is consistent with the effect of random labels on the generalization error that observed in practical experiments. We further show that strongly convex regularizers can significantly improve the probabilistic concentration bounds for the generalization error from the sub-linear rate to the exponential rate. Our study demonstrates that the geometric structure of the problem can be an important factor in improving the generalization performance of algorithms. Thus, it is of interest to explore the generalization error under various geometric conditions of the objective function in the future work.

Appendix: Proof of main results

Proof of Proposition 1

The proof is based on Lemma 11 (Elisseeff et al. 2005) and Assumption 1. Denote S^i as the data set that replaces the i-th sample of S with an i.i.d. copy generated from the distribution $\mathcal{D}$. Following from Lemma 11 of Elisseeff et al. (2005), we obtain

$$\begin{aligned}\mathbb{E}_{S,\xi}|f_S(\mathbf{w}_{T,S}) - f(\mathbf{w}_{T,S})|^2 &\le \frac{2M^2}{n} + \frac{12M}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi,S,S^i}\left[|\ell(\mathbf{w}_{T,S};\mathbf{z}_i) - \ell(\mathbf{w}_{T,S^i};\mathbf{z}_i)|\right]\\ &\le \frac{2M^2}{n} + \frac{12M\sigma}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi,S,S^i}\|\mathbf{w}_{T,S} - \mathbf{w}_{T,S^i}\|\\ &= \frac{2M^2}{n} + 12M\sigma\mathbb{E}_{\xi,S,\bar{S}}\|\mathbf{w}_{T,S} - \mathbf{w}_{T,\bar{S}}\|,\end{aligned}$$

where the second inequality uses the Lipschitz property of the loss function in Assumption 1, and the last equality is due to the fact that the perturbed samples in S^i and $\bar{S}$ are generated i.i.d from the underlying distribution.

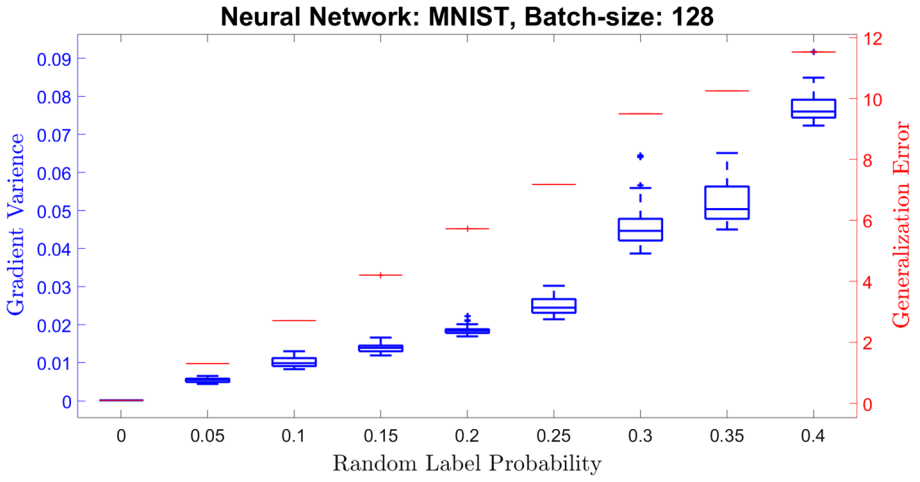
Proof of Theorem 1

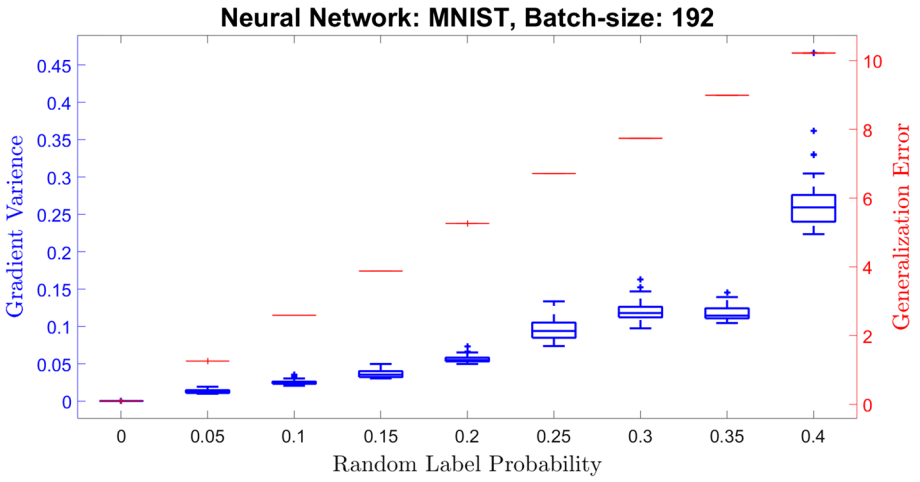
The proof is based on the following two important lemmas, which we prove first.

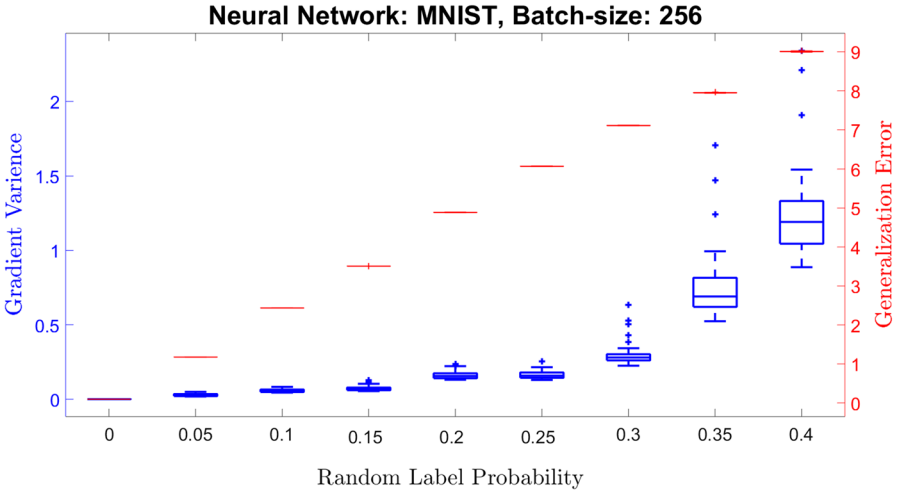
Lemma 2 *Let Assumption 1 hold. Apply SGD with the same sample path $\boldsymbol{\xi}$ to solve the ERM with data sets S and $\bar{S}$, respectively. Choose $\alpha_t = \frac{c}{(t+2)\log(t+2)}$ with $0 < c < \frac{1}{L}$, then the following bound holds.*

$$\mathbb{E}_{S,\bar{S},\xi}[\delta_{t+1,S,\bar{S}}] \le (1+\alpha_t L)\mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t}{n}\mathbb{E}_{S,\xi}\left[\left\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right].$$

Proof of Lemma 2 Consider the two fixed data sets S and $\bar{S}$ that differ at, say, the first data sample. At the t-th iteration, we consider two cases of the sampled index ξ_t. In the first case, $1 \notin \xi_t$ (w.p. $\frac{n-1}{n}$), i.e., the sampled data from S and $\bar{S}$ are the same, and we obtain that


Neural Network: MNIST, Batch-size: 128
Gradient Varience
Generalization Error
Random Label Probability
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
0 2 4 6 8 10 12


Neural Network: MNIST, Batch-size: 192
Gradient Varience
Generalization Error
Random Label Probability
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45
0 2 4 6 8 10


Neural Network: MNIST, Batch-size: 256
Gradient Varience
Generalization Error
Random Label Probability
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4
0 0.5 1 1.5 2
0 1 2 3 4 5 6 7 8 9

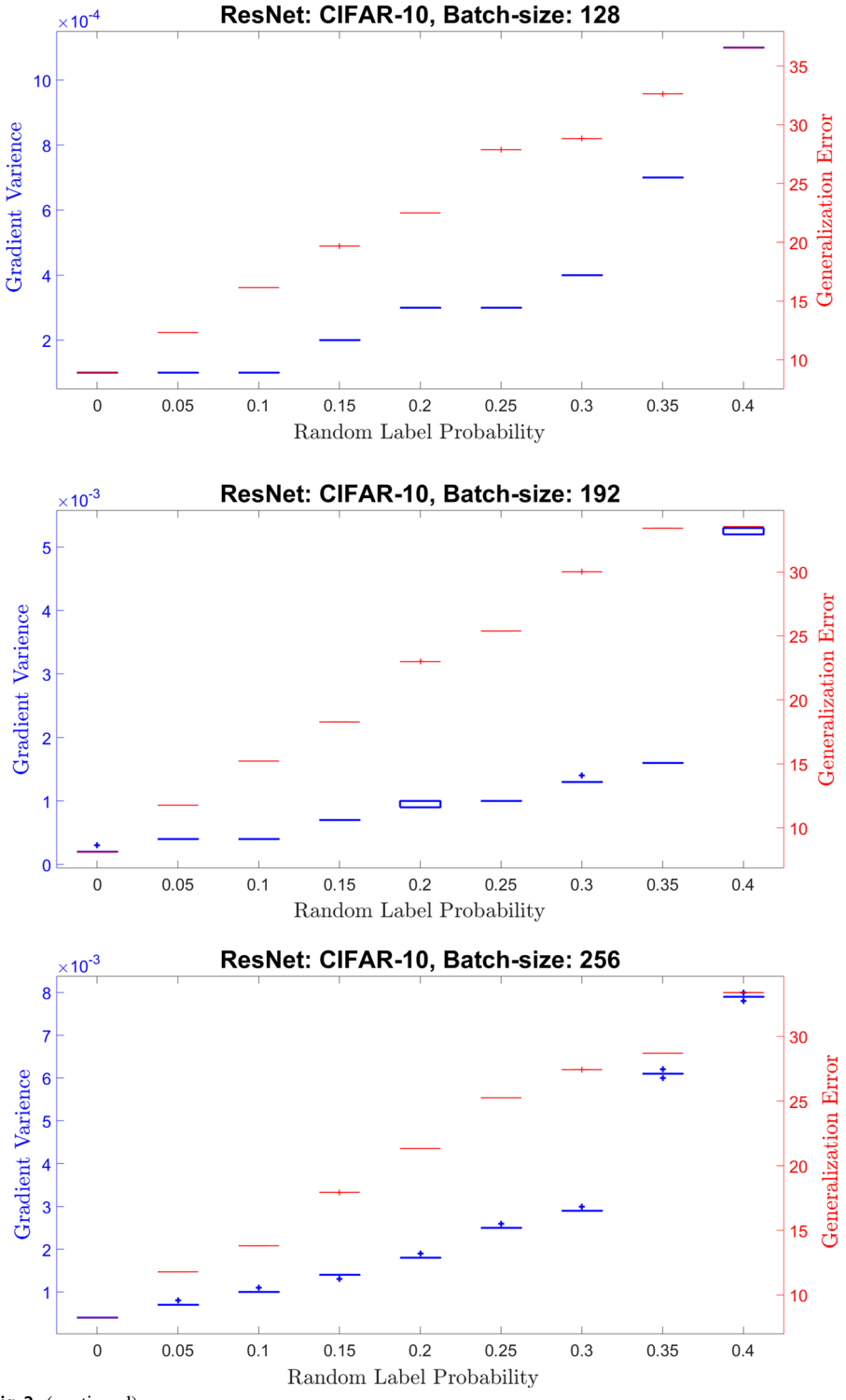


Fig. 2 (continued)

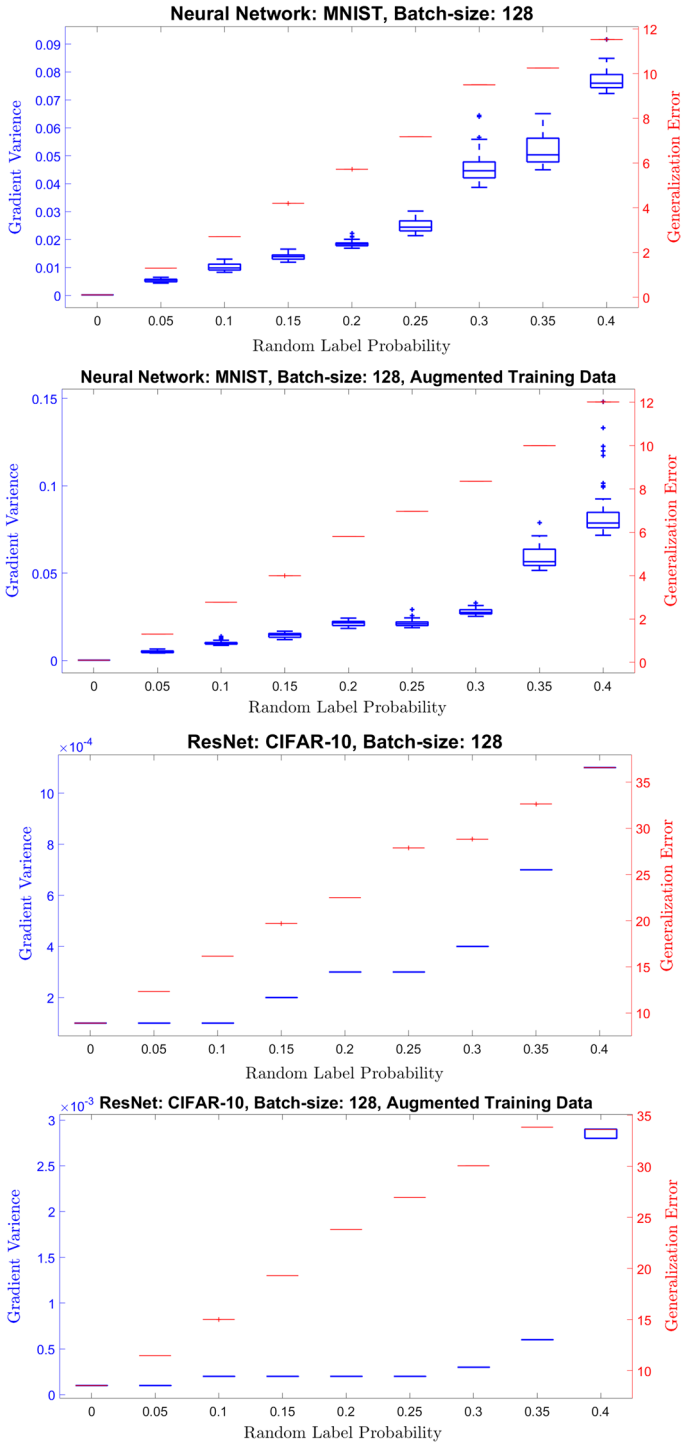


Fig. 3 Comparison of on-average variance, generalization error and random label probability with/without data augmentation

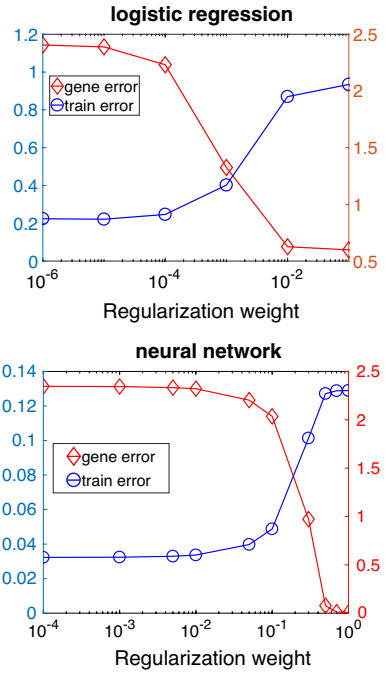


Fig. 4 Generalization error vs. regularization parameter

$$\begin{aligned}\delta_{t+1,S,\bar{S}} &= \left\|\mathbf{w}_{t,S} - \alpha_t \nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_{\xi_t}) - \mathbf{w}_{t,\bar{S}} + \alpha_t \nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_{\xi_t})\right\| \\ &\leq \delta_{t,S,\bar{S}} + \alpha_t \left\|\nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_{\xi_t}) - \nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_{\xi_t})\right\| \\ &\leq (1 + \alpha_t L)\delta_{t,S,\bar{S}},\end{aligned} \tag{8}$$

where the last inequality uses the L-Lipschitz property of $\nabla \ell$. In the other case, $1 \in \xi_t$ (w.p. $\frac{1}{n}$), we obtain that

$$\begin{aligned}\delta_{t+1,S,\bar{S}} &= \left\|\mathbf{w}_{t,S} - \alpha_t \nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_1) - \mathbf{w}_{t,\bar{S}} + \alpha_t \nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_1')\right\| \\ &\leq \delta_{t,S,\bar{S}} + \alpha_t \left\|\nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_1) - \nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_1')\right\| \\ &\leq \delta_{t,S,\bar{S}} + \alpha_t \Big(\left\|\nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_1)\right\| + \left\|\nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_1')\right\| \Big).\end{aligned} \tag{9}$$

Combining the above two cases and taking expectation with respect to all randomness, we obtain that

$$\begin{aligned}\mathbb{E}_{S,\bar{S},\xi}[\delta_{t+1,S,\bar{S}}] &\leq \left[\frac{n-1}{n}(1 + \alpha_t L) + \frac{1}{n}\right] \mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{1}{n}\alpha_t \mathbb{E}_{S,\bar{S},\xi}\Big(\left\|\nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_1)\right\| + \left\|\nabla \ell(\mathbf{w}_{t,\bar{S}}; \mathbf{z}_1')\right\| \Big) \\ &\overset{(i)}{\leq} (1 + \alpha_t L)\mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t}{n} \mathbb{E}_{S,\xi}\left[\left\|\nabla \ell(\mathbf{w}_{t,S}; \mathbf{z}_1)\right\|\right],\end{aligned} \tag{10}$$

where (i) uses the fact that $\mathbf{z}_1'$ is an i.i.d. copy of $\mathbf{z}_1$. □

Lemma 3 *Let Assumptions 1 and 2 hold. Apply SGD to solve the ERM with data set S and choosing $\alpha_t \le \frac{c}{t+2}$ for some $0 < c < \frac{1}{L}$. Then, the following bound holds.*

$$\mathbb{E}_{\boldsymbol{\xi},S}\left[\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\|\right] \le \sqrt{2Lf(\mathbf{w}_0) + \frac{1}{2}\mathbb{E}_S[v_S^2]}.$$

Proof of Lemma 3 By Assumption 1, ℓ is nonnegative and $\nabla\ell$ is L-Lipschitz. Then, eq. (12.6) of Shalev-Shwartz and Ben-David (2014) shows that

$$\forall \mathbf{w}, \quad \|\nabla \ell(\mathbf{w};\mathbf{z})\| \le \sqrt{2L\ell(\mathbf{w};\mathbf{z})}. \tag{11}$$

Based on (11), we further obtain that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi},S}\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\| &\le \sqrt{2L}\mathbb{E}_{\boldsymbol{\xi},S}\sqrt{\ell(\mathbf{w}_{t,S};\mathbf{z}_1)} \overset{(i)}{\le} \sqrt{2L}\sqrt{\mathbb{E}_{\boldsymbol{\xi},S}\ell(\mathbf{w}_{t,S};\mathbf{z}_1)} \\ &\overset{(ii)}{\le} \sqrt{2L}\sqrt{\mathbb{E}_{\boldsymbol{\xi},S}\frac{1}{n}\sum_{j=1}^{n}\ell(\mathbf{w}_{t,S};\mathbf{z}_j)} = \sqrt{2L}\sqrt{\mathbb{E}_{\boldsymbol{\xi},S}f_S(\mathbf{w}_{t,S})},\end{aligned} \tag{12}$$

where (i) uses the Jesen's inequality and (ii) uses the fact that all samples in S are generated i.i.d. from $\mathcal{D}$.

Next, consider a fixed data set S and denote $\mathbf{g}_{t,S} = \nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_{\xi_t})$ as the sampled stochastic gradient at iteration t. Then, by smoothness of ℓ and the update rule of the SGD, we obtain that

$$\begin{aligned}f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S}) &\le \langle \mathbf{w}_{t+1,S} - \mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})\rangle + \frac{L}{2}\|\mathbf{w}_{t+1,S} - \mathbf{w}_{t,S}\|^2 \\ &= \langle -\alpha_t \mathbf{g}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})\rangle + \frac{L\alpha_t^2}{2}\left\|\mathbf{g}_{t,S}\right\|^2.\end{aligned}$$

Conditioning on $\mathbf{w}_{t,S}$ and taking expectation with respect to $\boldsymbol{\xi}$, we further obtain from the above inequality that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}\left[f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S})|\mathbf{w}_{t,S}\right] \\ &\le \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\left\|\nabla f_S(\mathbf{w}_{t,S})\right\|^2 + \frac{L\alpha_t^2}{2}\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\mathbf{g}_{t,S}\right\|^2 - \left\|\nabla f_S(\mathbf{w}_{t,S})\right\|^2|\mathbf{w}_{t,S}\right].\end{aligned} \tag{13}$$

Note that $\frac{L\alpha_t^2}{2} - \alpha_t < 0$ by our choice of stepsize. Further taking expectation with respect to the randomness of $\mathbf{w}_{t,S}$ and S, and telescoping the above inequality over $0, \ldots, t-1$, we obtain that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi},S}\left[f_S(\mathbf{w}_{t,S})\right] &\overset{(i)}{\le} \mathbb{E}_S f_S(\mathbf{w}_0) + \sum_{t'=0}^{t-1}\frac{L\alpha_{t'}^2}{2}\mathbb{E}_S[v_S^2] \\ &= f(\mathbf{w}_0) + \sum_{t'=0}^{t-1}\frac{Lc^2\mathbb{E}_S[v_S^2]}{2(t'+2)^2} \overset{(ii)}{\le} f(\mathbf{w}_0) + \frac{Lc^2\mathbb{E}_S[v_S^2]}{4},\end{aligned}$$

where (i) uses the fact that the variance of the stochastic gradients is bounded by $\mathbb{E}_S[v_S^2]$, and (ii) upper bounds the summation by the integral, i.e., $\sum_{t'=0}^{t-1}\frac{1}{(t'+2)^2} \lesssim \int_1^t \frac{1}{t'^2}dt'$. Substituting the above result into (12) and noting that $cL \le 1$, we obtain the desired result. □

Now by Lemma 2, we obtain that

$$\begin{aligned}\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{t+1,S,\bar{S}}] &\le (1+\alpha_t L)\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t}{n}\mathbb{E}_{S,\boldsymbol{\xi}}\left[\left\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right] \\ &\overset{(i)}{\le} (1+\alpha_t L)\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t\sqrt{2Lf(\mathbf{w}_0)+\frac{\mathbb{E}_S[v_S^2]}{2}}}{n},\end{aligned} \tag{14}$$

where (i) applies Lemma 3. Recursively applying (14) over $t = 0, \ldots, T-1$ and noting that $\delta_0 = 0$ and $\alpha_t = \frac{c}{(t+2)\log(t+2)}$, we obtain

$$\begin{aligned}\mathbb{E}_{S,\bar{S},\boldsymbol{\xi}}[\delta_T] &\le \sum_{t=0}^{T-1}\left[\prod_{k=t+1}^{T-1}(1+\alpha_k L)\right]\frac{2c\sqrt{2Lf(\mathbf{w}_0)+\frac{\mathbb{E}_S[v_S^2]}{2}}}{(t+2)\log(t+2)n} \\ &\overset{(i)}{\le} \sum_{t=0}^{T-1}\left[\exp\left(\sum_{k=t+1}^{T-1}\frac{cL}{(k+2)\log(k+2)}\right)\right]\frac{2c\sqrt{2Lf(\mathbf{w}_0)+\frac{\mathbb{E}_S[v_S^2]}{2}}}{(t+2)\log(t+2)n} \\ &\overset{(ii)}{\le} \sum_{t=0}^{T-1}\left(\frac{\log T}{\log(t+2)}\right)^{cL}\frac{2c\sqrt{2Lf(\mathbf{w}_0)+\frac{\mathbb{E}_S[v_S^2]}{2}}}{(t+2)\log(t+2)n} \\ &\overset{(iii)}{\le} \frac{2c\sqrt{2Lf(\mathbf{w}_0)+\frac{\mathbb{E}_S[v_S^2]}{2}}}{n}\log T,\end{aligned}$$

where (i) uses the fact that $1+x \le \exp(x)$. For (ii) and (iii), we apply the integral upper bounds to bound the summations, i.e., $\sum_{k=t+1}^{T-1}\frac{cL}{(k+2)\log(k+2)} \lesssim \int_t^T \frac{cL}{k\log k}dk$, $\sum_{t=0}^{T-1}(t+2)^{-1}\log^{-1-cL}(t+2) \lesssim \int_{t=1}^{T} t^{-1}\log^{-1-cL} t dt$, and use the fact that $cL < 1$. Substituting the above result into Proposition 1 and applying the Chebyshev's inequality yields the desired result.

Proof of Theorem 2

We first prove a useful lemma.

Lemma 4 *Let Assumptions 1 and 2 hold. Apply the SGD to solve the ERM with data set S, where f_S is γ-gradient dominant ($\gamma < L$) with the minimum function value f_S^*. Suppose we choose $\alpha_t \le \frac{c}{t+2}$ for some $0 < c < \min\{\frac{2}{\gamma}, \frac{1}{L}\}$. Then the following bound holds.*

$$\mathbb{E}_{\boldsymbol{\xi},S}\left[\left\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right] \le \sqrt{2L\mathbb{E}_S[f_S^*] + \frac{1}{t}\left(2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\right)}.$$

Proof of Lemma 4 We first note that (12) and (13) both hold here, which we rewritten below for convenience.

$$\mathbb{E}_{\boldsymbol{\xi},S}\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\| \le \sqrt{2L}\sqrt{\mathbb{E}_{\boldsymbol{\xi},S} f_S(\mathbf{w}_{t,S})}, \tag{15}$$

$$\mathbb{E}_{\xi}\left[f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S})|\mathbf{w}_{t,S}\right] \leq \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\left\|\nabla f_S(\mathbf{w}_{t,S})\right\|^2 + \frac{L\alpha_t^2}{2}\mathbb{E}_{\xi}\left[\left\|\mathbf{g}_{t,S}\right\|^2 - \left\|\nabla f_S(\mathbf{w}_{t,S})\right\|^2|\mathbf{w}_{t,S}\right]. \tag{16}$$

Following from (16) and the fact that f_S is γ-gradient dominant, we obtain

$$\mathbb{E}_{\xi}\left[f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S}) \mid \mathbf{w}_{t,S}\right] \leq \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)2\gamma(f_S(\mathbf{w}_{t,S}) - f_S^*) + \frac{L\alpha_t^2}{2}\mathbb{E}_{\xi}\left[\left\|g_{t,S}\right\|^2 - \left\|\nabla f_S(\mathbf{w}_{t,S})\right\|^2 \mid \mathbf{w}_{t,S}\right]. \tag{17}$$

Further taking expectation with respect to the randomness of $\mathbf{w}_{t,S}$ and S, we obtain from the above inequality that

$$\begin{aligned}\mathbb{E}_{\xi,S}\left[f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S})\right] &\leq \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)2\gamma\mathbb{E}_{\xi,S}(f_S(\mathbf{w}_{t,S}) - f_S^*) + \frac{L\alpha_t^2}{2}\mathbb{E}_S[\nu_S^2] \\ &\leq -\alpha_t\gamma\mathbb{E}_{\xi,S}(f_S(\mathbf{w}_{t,S}) - f_S^*) + \frac{L\alpha_t^2\mathbb{E}_S[\nu_S^2]}{2},\end{aligned}$$

where the last inequality uses the fact that $\frac{L\alpha_t^2}{2} \leq \alpha_t/2$ for $c < \frac{1}{L}$. Rearranging the above inequality, we further obtain that

$$\begin{aligned}\mathbb{E}_{\xi,S}\left[f_S(\mathbf{w}_{t+1,S}) - f_S^*\right] &\leq (1 - \alpha_t\gamma)\mathbb{E}_{\xi,S}(f_S(\mathbf{w}_{t,S}) - f_S^*) + \frac{L\alpha_t^2\nu^2}{2} \\ &\leq \prod_{t'=0}^{t}(1 - \alpha_{t'}\gamma)\mathbb{E}_S(f_S(\mathbf{w}_0) - f_S^*) + \sum_{t'=0}^{t}\prod_{k=t'+1}^{t-1}(1 - \alpha_k\gamma)\frac{L\alpha_{t'}^2\mathbb{E}_S[\nu_S^2]}{2} \\ &\overset{(i)}{\leq} t^{-c\gamma}\mathbb{E}_S(f_S(\mathbf{w}_0) - f_S^*) + \frac{Lc^2\mathbb{E}_S[\nu_S^2]}{t^{c\gamma}} \\ &\overset{(ii)}{\leq} \frac{1}{t^{c\gamma}}\left[f(\mathbf{w}_0) + Lc^2\mathbb{E}_S[\nu_S^2]\right],\end{aligned}$$

where (i) uses the fact that $1 - x \leq \exp(-x)$ and upper bounds the summations by the corresponding integrals, i.e., $\exp(-c\gamma\sum_{t'=0}^{t}\frac{1}{t'+2}) \lesssim \exp(-c\gamma\int_0^t\frac{1}{t'}dt')$ and (ii) uses the fact that $c\gamma < 1/2$. We then conclude that

$$\mathbb{E}_{\xi,S}f_S(\mathbf{w}_{t,S}) \leq \mathbb{E}_S[f_S^*] + \frac{1}{t^{c\gamma}}\left[f(\mathbf{w}_0) + L\nu^2c^2\right].$$

Substituting this bound into (15) and noting that $cL \leq 1$, we obtain the desired result.

□

To continue our proof, by Lemma 2, we obtain that

$$\begin{aligned}\mathbb{E}_{S,\bar{S},\xi}[\delta_{t+1,S,\bar{S}}] &\leq (1 + \alpha_t L)\mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t}{n}\mathbb{E}_{S,\xi}\left[\left\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right\|\right] \\ &\leq (1 + \alpha_t L)\mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{2\alpha_t}{n}\sqrt{2L\mathbb{E}_S[f_S^*] + \frac{1}{t^{c\gamma}}\left(2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[\nu_S^2]\right)},\end{aligned} \tag{18}$$

where the last line applies Lemma 4. Applying (18) recursively over $t = 0, \ldots, T-1$ and noting that $\delta_0 = 0, \alpha_t = \frac{c}{(t+2)\log(t+2)}$, we obtain that

$$\begin{aligned}
\mathbb{E}_{S,\bar{S},\xi}[\delta_T] &\leq \sum_{t=0}^{T-1}\left[\prod_{k=t+1}^{T-1}(1+\alpha_k L)\right]\frac{2c}{(t+2)\log(t+2)n}\sqrt{2L\mathbb{E}_S[f_S^*] + \frac{1}{t^{c\gamma}}\left(2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\right)} \\
&\leq \frac{2c}{n}\sum_{t=0}^{T-1}\left(\frac{\log T}{\log(t+2)}\right)^{cL}\frac{\sqrt{2L\mathbb{E}_S[f_S^*]} + \sqrt{\frac{1}{t^{c\gamma}}\left(2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\right)}}{(t+2)\log(t+2)} \\
&\leq \frac{2c}{n}\left(\sqrt{2L\mathbb{E}_S[f_S^*]}\log T + \sqrt{2Lf(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]}\right).
\end{aligned}$$

Substituting the above result into Proposition 1 yields the desired result.

Proof of Theorem 4

Consider the fixed data sets S and $\bar{S}$ that are differ at the first sample. At the t-th iteration, if $1 \notin \xi_t$ (w.p. $\frac{n-1}{n}$), we obtain that

$$\begin{aligned}
\delta_{t+1,S,\bar{S}} &= \left\|\mathrm{prox}_{\alpha_t h}\left(\mathbf{w}_{t,S} - \alpha_t\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_{\xi_t})\right) - \mathrm{prox}_{\alpha_t h}\left(\mathbf{w}_{t,\bar{S}} - \alpha_t\nabla\ell(\mathbf{w}_{t,\bar{S}};\mathbf{z}_{\xi_t})\right)\right\| \\
&\overset{(i)}{\leq} \frac{1}{1+\alpha_t\lambda}\left\|\mathbf{w}_{t,S} - \alpha_t\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_{\xi_t}) - \mathbf{w}_{t,\bar{S}} + \alpha_t\nabla\ell(\mathbf{w}_{t,\bar{S}};\mathbf{z}_{\xi_t})\right\| \\
&\leq \frac{1+\alpha_t L}{1+\alpha_t\lambda}\delta_{t,S,\bar{S}},
\end{aligned} \tag{19}$$

where (i) uses item 2 of Lemma 5. On the other hand, if $1 \in \xi_t$ (w.p. $\frac{1}{n}$), we obtain that

$$\begin{aligned}
\delta_{t+1,S,\bar{S}} &= \left\|\mathrm{prox}_{\alpha_t h}\left(\mathbf{w}_{t,S} - \alpha_t\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\right) - \mathrm{prox}_{\alpha_t h}\left(\mathbf{w}_{t,\bar{S}} - \alpha_t\nabla\ell(\mathbf{w}_{t,\bar{S}};\mathbf{z}_1')\right)\right\| \\
&\overset{(i)}{\leq} \frac{1}{1+\alpha_t\lambda}\left\|\mathbf{w}_{t,S} - \alpha_t\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1) - \mathbf{w}_{t,\bar{S}} + \alpha_t\nabla\ell(\mathbf{w}_{t,\bar{S}};\mathbf{z}_1')\right\| \\
&\leq \frac{1}{1+\alpha_t\lambda}\delta_{t,S,\bar{S}} + \frac{\alpha_t}{1+\alpha_t\lambda}\left(\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\| + \|\nabla\ell(\mathbf{w}_{t,\bar{S}};\mathbf{z}_1')\|\right),
\end{aligned} \tag{20}$$

where (i) uses item 2 of Lemma 5. Combining the above two cases and taking expectation with respect to the randomness of $\boldsymbol{\xi}$, S and $\bar{S}$, we obtain that

$$\begin{aligned}
\mathbb{E}_{S,\bar{S},\xi}[\delta_{t+1,S,\bar{S}}] &\leq \left[\frac{n-1}{n}\frac{1+\alpha_t L}{1+\alpha_t\lambda} + \frac{1}{n}\frac{1}{1+\alpha_t\lambda}\right]\mathbb{E}_{S,\bar{S},\xi}[\delta_{t,S,\bar{S}}] + \frac{1}{n}\frac{2\alpha_t}{1+\alpha_t\lambda}\mathbb{E}_{S,\xi}\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\| \\
&\overset{(i)}{\leq} \frac{1+\alpha_t L}{1+\alpha_t\lambda}\mathbb{E}_{S,\bar{S},\xi}[\delta_{S,\bar{S},\xi}] + \frac{2\alpha_t}{n}\frac{1}{1+\alpha_t\lambda}\sqrt{2L\boldsymbol{\Phi}(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\log t} \\
&\lesssim \exp(\alpha_t(L-\lambda))\mathbb{E}_{S,\bar{S},\xi}[\delta_{S,\bar{S},\xi}] + \frac{2\alpha_t}{n}\sqrt{2L\boldsymbol{\Phi}(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\log t},
\end{aligned}$$

where (i) uses Lemma 6. Recursively applying the above inequality over $t = 0, \ldots, T-1$ and noting that $\delta_0 = 0, \alpha_t = \frac{c}{t+2}$, we obtain that

$$\mathbb{E}_{S,\bar{S},\xi}[\delta_{T,S,\bar{S}}] \leq \sum_{t=0}^{T-1}\left[\prod_{k=t+1}^{T-1}\exp(\alpha_k(L-\lambda))\right]\frac{2c\sqrt{2L\Phi(\mathbf{w}_0)+2\mathbb{E}_S[v_S^2]\log t}}{(t+2)n}$$
$$\overset{(i)}{\leq} \sum_{t=0}^{T-1}\left(\frac{t+2}{T}\right)^{c(\lambda-L)}\frac{2c\sqrt{2L\Phi(\mathbf{w}_0)+2\mathbb{E}_S[v_S^2]}}{(t+2)n}\log t$$
$$\overset{(ii)}{\leq} \frac{2}{n(\lambda-L)}\sqrt{2L\Phi(\mathbf{w}_0)+2\mathbb{E}_S[v_S^2]},$$

where the $\log t$ term in (i) is ignored as it is order-wise smaller than other polynomial terms (In particular, for any $\delta > 0$ we have $\lim_{t\to\infty}\log t/t^{\delta} = 0$), and (ii) further upper bounds the summation with the integral, i.e., $\sum_{t=0}^{T-1}(t+2)^{c(\lambda-L)-1} \lesssim \int_1^T t^{c(\lambda-L)-1}dt$, and uses the fact that $c < \frac{1}{L}$. Then, applying Proposition 1 to the regularized risk minimization, we further obtain that

$$\mathbb{E}_{\xi,S}\left[|\Phi_S(\mathbf{w}_{T,S}) - \Phi(\mathbf{w}_{T,S})|^2\right] \leq \frac{1}{n}\left(2M^2 + \frac{24M\sigma}{(\lambda-L)}\sqrt{L\Phi(\mathbf{w}_0)+\mathbb{E}_S[v_S^2]}\right).$$

The desired result then follows by applying Chebyshev's inequality.

Proof of Theorem 5

The idea of the proof is to apply Lemma 1 by developing the uniform stability bounds β and γ. The proof also applies two useful lemmas on the proximal SGD.

We first evaluate β. Following the proof logic of Theorem 4 and replacing the bound for the on-average stochastic gradient norm $\mathbb{E}_{S,\xi}\|\nabla\ell(\mathbf{w}_{t,S};\mathbf{z}_1)\|$ with the uniform upper bound σ, we obtain that

$$\sup_{S,\bar{S},\mathbf{z}}\mathbb{E}_{\xi}|\ell(\mathbf{w}_{T,S};\mathbf{z}) - \ell(\mathbf{w}_{T,\bar{S}};\mathbf{z})| \leq \sigma\sup_{S,\bar{S},\mathbf{z}}\mathbb{E}_{\xi}[\delta_{T,S,\bar{S}}] \leq \frac{2\sigma^2}{n(\lambda-L)} := \beta.$$

Next, we evaluate ρ. Consider any two sample paths $\boldsymbol{\xi} := \{\xi_1,\ldots,\xi_{t_0},\ldots,\xi_{T-1}\}$ and $\bar{\boldsymbol{\xi}} := \{\xi_1,\ldots,\xi'_{t_0},\ldots,\xi_{T-1}\}$, which are different at the t_0-th mini-batch. Note that

$$\sup_{\xi,\bar{\xi},S,\mathbf{z}}|\ell(\mathbf{w}_{T,S,\xi};\mathbf{z}) - \ell(\mathbf{w}_{T,S,\bar{\xi}};\mathbf{z})| \leq \sup_{\xi,\bar{\xi},S,\mathbf{z}}\sigma\|\mathbf{w}_{T,S,\xi} - \mathbf{w}_{T,S,\bar{\xi}}\|. \tag{21}$$

Since the two sample paths only differ at the t_0-th iteration, we have that $\mathbf{w}_{t,S,\xi} - \mathbf{w}_{t,S,\bar{\xi}} = \mathbf{0}$ for $t = 0,\ldots,t_0$. In particular, for $t = t_0$ we obtain that

$$
\begin{aligned}
&\|\mathbf{w}_{t_0+1,S,\xi} - \mathbf{w}_{t_0+1,S,\bar{\xi}}\| \\
&= \left\| \mathrm{prox}_{\alpha_{t_0} h}\Big(\mathbf{w}_{t_0,S,\xi} - \alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\xi};\mathbf{z}_{\xi_{t_0}})\Big) - \mathrm{prox}_{\alpha_{t_0} h}\Big(\mathbf{w}_{t_0,S,\bar{\xi}} - \alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\bar{\xi}};\mathbf{z}_{\xi'_{t_0}})\Big)\right\| \\
&\overset{(i)}{\leq} \frac{1}{1+\alpha_{t_0}\lambda}\left\|\mathbf{w}_{t_0,S,\xi} - \alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\xi};\mathbf{z}_{\xi_{t_0}}) - \mathbf{w}_{t_0,S,\bar{\xi}} + \alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\bar{\xi}};\mathbf{z}_{\xi'_{t_0}})\right\| \\
&= \frac{1}{1+\alpha_{t_0}\lambda}\left\|\alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\xi};\mathbf{z}_{\xi_{t_0}}) - \alpha_{t_0}\nabla\ell(\mathbf{w}_{t_0,S,\bar{\xi}};\mathbf{z}_{\xi'_{t_0}})\right\| \\
&\overset{(ii)}{\leq} 2\alpha_{t_0}\sigma,
\end{aligned}
$$

where (i) uses Lemma 5 and (ii) uses the σ-bounded property of $\|\nabla\ell\|$. Now consider $t > t_0 + 1$. Note that in this case the sampled indices in $\boldsymbol{\xi}$ and $\bar{\boldsymbol{\xi}}$ are the same, and we further obtain that

$$
\begin{aligned}
&\|\mathbf{w}_{t+1,S,\xi} - \mathbf{w}_{t+1,S,\bar{\xi}}\| \\
&= \left\| \mathrm{prox}_{\alpha_t h}\Big(\mathbf{w}_{t,S,\xi} - \alpha_t\nabla\ell(\mathbf{w}_{t,S,\xi};\mathbf{z}_{\xi_t})\Big) - \mathrm{prox}_{\alpha_t h}\Big(\mathbf{w}_{t,S,\bar{\xi}} + \alpha_t\nabla\ell(\mathbf{w}_{t,S,\bar{\xi}};\mathbf{z}_{\xi_t})\Big)\right\| \\
&\leq \frac{1}{1+\alpha_t\lambda}\left\|\mathbf{w}_{t,S,\xi} - \alpha_t\nabla\ell(\mathbf{w}_{t,S,\xi};\mathbf{z}_{\xi_t}) - \mathbf{w}_{t,S,\bar{\xi}} + \alpha_t\nabla\ell(\mathbf{w}_{t,S,\bar{\xi}};\mathbf{z}_{\xi_t})\right\| \\
&\leq \frac{1+\alpha_t L}{1+\alpha_t\lambda}\|\mathbf{w}_{t,S,\xi} - \mathbf{w}_{t,S,\bar{\xi}}\| \lesssim \exp(-\alpha_t(\lambda - L))\|\mathbf{w}_{t,S,\xi} - \mathbf{w}_{t,S,\bar{\xi}}\|.
\end{aligned}
$$

Telescoping over $t = t_0, \ldots, T-1$, we further obtain that

$$
\begin{aligned}
\|\mathbf{w}_{T,S,\xi} - \mathbf{w}_{T,S,\bar{\xi}}\| &\leq 2\alpha_{t_0}\sigma\exp\left(-(\lambda - L)\sum_{t=t_0+1}^{T-1}\alpha_t\right) \\
&\lesssim \frac{2\sigma c}{(t_0+2)}\exp\left(-(\lambda - L)c\log\frac{T}{(t_0+2)}\right) \\
&= \frac{2\sigma c}{(t_0+2)^{1-c(\lambda-L)}T^{c(\lambda-L)}} \\
&\leq \frac{2\sigma c}{T^{c(\lambda-L)}}.
\end{aligned}
$$

Thus, from (21) we obtain that $\rho = \frac{2\sigma^2 c}{T^{c(\lambda-L)}}$. Substituting the expressions of β and ρ into Lemma 1, we conclude that with probability at least $1-\delta$

$$
\begin{aligned}
\boldsymbol{\Phi}(\mathbf{w}_{T,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{T,S}) &\leq \frac{4\sigma^2}{n(\lambda - L)} + \left(\frac{M}{\sqrt{n}}2\sqrt{n}\frac{2\sigma^2}{n(\lambda - L)} + \sqrt{2T}\frac{2\sigma^2 c}{T^{c(\lambda-L)}}\right)\sqrt{\log\frac{2}{\delta}} \\
&\leq \left(\frac{M}{\sqrt{n}} + \frac{4\sigma^2}{\sqrt{n}(\lambda - L)} + \frac{4\sigma^2 c}{T^{c(\lambda-L)-\frac{1}{2}}}\right)\sqrt{\log\frac{2}{\delta}}.
\end{aligned}
$$

Proof of technical Lemmas for proximal SGD

For any vector $\mathbf{g} \in \mathbb{R}^d$, we define the following quantity:

$$G^{\alpha}(\mathbf{w},\mathbf{g}) := \tfrac{1}{\alpha}\big(\mathbf{w} - \mathrm{prox}_{\alpha h}(\mathbf{w}-\alpha\mathbf{g})\big). \tag{22}$$

Lemma 5 *Let h be a convex and possibly non-smooth function. Then, the following statements hold.*

1. *For any $\mathbf{w},\mathbf{g}_1,\mathbf{g}_2 \in \Omega$, it holds that*

$$\big\|G^{\alpha}(\mathbf{w},\mathbf{g}_1) - G^{\alpha}(\mathbf{w},\mathbf{g}_2)\big\| \le \big\|\mathbf{g}_1 - \mathbf{g}_2\big\|.$$

2. *If h is λ strongly convex, then for all $\mathbf{w},\mathbf{v} \in \Omega$ and $\alpha > 0$, it holds that*

$$\|\mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\| \le \tfrac{1}{1+\alpha\lambda}\|\mathbf{w}-\mathbf{v}\|.$$

Proof of Lemma 5 Consider the first item. By definition, we have

$$\begin{aligned}\big\|G^{\alpha}(\mathbf{w},\mathbf{g}_1) - G^{\alpha}(\mathbf{w},\mathbf{g}_2)\big\| &= \frac{1}{\alpha}\big\|\mathrm{prox}_{\alpha h}(\mathbf{w}-\alpha\mathbf{g}_1) - \mathrm{prox}_{\alpha h}(\mathbf{w}-\alpha\mathbf{g}_2)\big\| \\ &\le \frac{1}{\alpha}\big\|(\mathbf{w}-\alpha\mathbf{g}_1) - (\mathbf{w}-\alpha\mathbf{g}_2)\big\| \\ &= \big\|\mathbf{g}_1 - \mathbf{g}_2\big\|,\end{aligned} \tag{23}$$

where the inequality uses the 1-Lipschitz property of the proximal mapping for convex functions.

Next, consider the second item. Recall the resolvent representation Bauschke and Combettes (2011) of the proximal mapping for convex functions, i.e.,

$$\mathrm{prox}_{\alpha h}(\mathbf{w}) = (I + \alpha\nabla h)^{-1}(\mathbf{w}),$$

where I denotes the identity operator. Applying the operator $(I+\alpha\nabla h)$ on both sides of the above equation, we obtain that $(I+\alpha\nabla h)(\mathrm{prox}_{\alpha h}(\mathbf{w})) = \mathbf{w}$. Thus, we conclude that

$$\mathbf{w} - \mathrm{prox}_{\alpha h}(\mathbf{w}) = \alpha\nabla h(\mathrm{prox}_{\alpha h}(\mathbf{w})),$$

which further implies that

$$\begin{aligned}&\langle[\mathbf{w} - \mathrm{prox}_{\alpha h}(\mathbf{w})] - [\mathbf{v} - \mathrm{prox}_{\alpha h}(\mathbf{v})], \mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\rangle \\ &\quad= \alpha\langle\nabla h(\mathrm{prox}_{\alpha h}(\mathbf{w})) - \nabla h(\mathrm{prox}_{\alpha h}(\mathbf{v})), \mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\rangle \\ &\quad\ge \alpha\lambda\|\mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\|^2,\end{aligned}$$

where the last inequality uses the fact that h is λ-strongly convex. Rearranging the above inequality, we obtain that

$$\begin{aligned}&\langle\mathbf{w}-\mathbf{v}, \mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\rangle \\ &\quad\ge (1+\alpha\lambda)\|\mathrm{prox}_{\alpha h}(\mathbf{w}) - \mathrm{prox}_{\alpha h}(\mathbf{v})\|^2.\end{aligned}$$

Applying Cauchy-Swartz inequality on the left hand side, we obtain the desired result. □

Lemma 6 *Let Assumptions 1, 2 and 3 hold. Applying the proximal SGD to solve the R-ERM with data set S and choosing* $\alpha_t \leq \frac{c}{t+2}$ *with* $0 < c < \frac{1}{L}$. *Then, it holds that*

$$\mathbb{E}_{S,\boldsymbol{\xi}}\left[\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\|\right] \leq \sqrt{2L\boldsymbol{\Phi}(\mathbf{w}_0) + 2\mathbb{E}_S[v_S^2]\log t}.$$

Proof of Lemma 6 The proof is based on the technical tools developed in Ghadimi et al. (2016) for analyzing the optimization path of the proximal SGD.

Under the assumptions of the lemma, we first recall the following result from [Lemma 1, Ghadimi et al. 2016]: For any $\mathbf{w} \in \Omega, \mathbf{g} \in \mathbb{R}^d$, it holds that

$$\langle \mathbf{g}, G^{\alpha}(\mathbf{w},\mathbf{g})\rangle \geq \|G^{\alpha}(\mathbf{w},\mathbf{g})\|^2 + \frac{1}{\alpha}\big(h(\mathrm{prox}_{\alpha h}(\mathbf{w}-\alpha\mathbf{g})) - h(\mathbf{w})\big).$$

Denoting $\mathbf{g}_{t,S} = \nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_{\xi_t})$ as the stochastic gradient sampled at iteration t and setting $\mathbf{w} = \mathbf{w}_{t,S}, \mathbf{g} = \mathbf{g}_{t,S}$ in the above inequality, we obtain that

$$\langle \mathbf{g}_{t,S}, G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S})\rangle \geq \|G^{\alpha}(\mathbf{w}_{t,S},\mathbf{g}_{t,S})\|^2 + \frac{1}{\alpha_t}\big(h(\mathbf{w}_{t+1,S}) - h(\mathbf{w}_{t,S})\big). \tag{24}$$

On the other hand, using (11) and non-negativity of h, we obtain

$$\mathbb{E}_{\xi,S}\|\nabla \ell(\mathbf{w}_{t,S};\mathbf{z}_1)\| \leq \sqrt{2L}\sqrt{\mathbb{E}_{\xi,S} f_S(\mathbf{w}_{t,S})} \leq \sqrt{2L}\sqrt{\mathbb{E}_{\xi,S}\boldsymbol{\Phi}_S(\mathbf{w}_{t,S})}. \tag{25}$$

Next, consider a fixed S, by the smoothness of ℓ we obtain

$$\begin{aligned}
&f_S(\mathbf{w}_{t+1,S}) - f_S(\mathbf{w}_{t,S}) \\
&\quad\leq \langle \mathbf{w}_{t+1,S} - \mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})\rangle + \frac{L}{2}\|\mathbf{w}_{t+1,S} - \mathbf{w}_{t,S}\|^2 \\
&\quad= \langle -\alpha_t G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S}), \nabla f_S(\mathbf{w}_{t,S})\rangle + \frac{L\alpha_t^2}{2}\left\|G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S})\right\|^2 \\
&\quad= -\alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S}), \mathbf{g}_{t,S}\rangle - \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S}), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle + \frac{L\alpha_t^2}{2}\left\|G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S})\right\|^2 \\
&\quad= -\alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S}), \mathbf{g}_{t,S}\rangle - \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S},\nabla f_S(\mathbf{w}_{t,S})), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle \\
&\qquad+ \frac{L\alpha_t^2}{2}\left\|G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S})\right\|^2 \\
&\qquad+ \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S},\nabla f_S(\mathbf{w}_{t,S})) - G^{\alpha_t}(\mathbf{w}_{t,S},\mathbf{g}_{t,S}), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle.
\end{aligned} \tag{26}$$

Now combining with (24) and rearranging, we obtain that

$$
\begin{aligned}
&\boldsymbol{\Phi}_S(\mathbf{w}_{t+1,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{t,S}) \\
&\quad \le \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\left\|G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S})\right\|^2 - \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle \\
&\qquad + \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})) - G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S}), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle \\
&\quad \le \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\left\|G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S})\right\|^2 - \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle \\
&\qquad + \alpha_t\|G^{\alpha_t}(\mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})) - G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S})\|\|\nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\| \\
&\quad \le \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\left\|G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S})\right\|^2 - \alpha_t\langle G^{\alpha_t}(\mathbf{w}_{t,S}, \nabla f_S(\mathbf{w}_{t,S})), \nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\rangle \\
&\qquad + \alpha_t\|\nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\|^2,
\end{aligned}
$$

where the last line uses item 1 of Lemma 5. Conditioning on $\mathbf{w}_{t,S}$, and taking expectation with respect to $\boldsymbol{\xi}$, we further obtain from the above inequality that

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\Phi}_S(\mathbf{w}_{t+1,S}) - \boldsymbol{\Phi}_S(\mathbf{w}_{t,S}) \mid \mathbf{w}_{t,S}] \\
&\quad \le \left(\frac{L\alpha_t^2}{2} - \alpha_t\right)\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|G^{\alpha_t}(\mathbf{w}_{t,S}, \mathbf{g}_{t,S})\right\|^2 \mid \mathbf{w}_{t,S}\right] + \alpha_t\mathbb{E}_{\boldsymbol{\xi}}\left[\|\nabla f_S(\mathbf{w}_{t,S}) - \mathbf{g}_{t,S}\|^2 \mid \mathbf{w}_{t,S}\right].
\end{aligned}
$$

Further taking expectation with respect to the randomness of $\mathbf{w}_{t,S}$ and S, telescoping the above inequality over $0, \ldots, t-1$ and noting that $\frac{L\alpha_t^2}{2} < \alpha_t$, we obtain that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi},S}\left[\boldsymbol{\Phi}_S(\mathbf{w}_{t,S})\right] &\le \mathbb{E}_S\boldsymbol{\Phi}_S(\mathbf{w}_0) + \sum_{t'=0}^{t-1}\frac{c\mathbb{E}_S[v_S^2]}{t'+2} \\
&\le \boldsymbol{\Phi}(\mathbf{w}_0) + c\mathbb{E}_S[v_S^2]\log t,
\end{aligned}
$$

where we have used the bound for the variance of the stochastic gradients. Substituting the above expression into (25) and note that $cL < 1$, we obtain the desired result. □

Author Contributions The corresponding author Yi Zhou is the main contributor of this work, he develops all the main technical results. All the authors contributed to the development of the main idea of this paper, i.e., introducing the on-average stability. The co-authors Yingbin Liang and Huishuai Zhang provided technical suggestions to improve the dependence of Theorem 1 on the step size and help checked the technical proof.

Availability of data and material All the data used in this work (i.e., MNIST and CIFAR10) are open-source data that are publicly available.

Declarations

Conflict of interest The author declare that they have no conflict of interest.

Code availability All the codes are made publicly available on GitHub at https://github.com/ccheng21/Generalization_of_Nonconvex_SGD.git.

References

Attouch, H., Bolte, J., & Svaiter, B. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming, 137*(1), 91–129.

Bartlett, P., Foster, D.J., & Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. In: *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 6240–6249.

Bauschke, H., & Combettes, P. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences, 2*(1), 183–202.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics*, pp. 177–186.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research, 2,* 499–526.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Charles, Z., & Papailiopoulos, D. (2017). Stability and generalization of learning algorithms that converge to global optima. ArXiv: 1710.08402.

Elisseeff, A., Evgeniou, T., & Pontil, M. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research, 6,* 55–79.

Ghadimi, S., Lan, G., & Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming, 155*(1), 267–305.

Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1225–1234.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Karimi, H., Nutini, J., & Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. *Machine Learning and Knowledge Discovery in Databases: European Conference*, pp. 795–811.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

Kuzborskij, I., & Lampert, C. H. (2017). Data-dependent stability of stochastic gradient descent. ArXiv: 1703.01678v3.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Li, X., Ling, S., Strohmer, T., & Wei, K. (2016). Rapid, robust, and reliable blind deconvolution via nonconvex optimization. arXiv: 1606.04933v1.

Li, Q., Zhou, Y., Liang, Y., & Varshney, P. K. (2017). Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Lin, J., & Rosasco, L. (2017). Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research, 18,* 1–47.

Łojasiewicz, S. (1963). A topological property of real analytic subsets. *Colloid du CNRS, Les equations aux derivees partielles*, pp. 87–89.

London, B. (2017). A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*.

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pp. 164–170.

Mou, W., Wang, L., Zhai, X., & Zheng, K. (2017). Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. ArXiv: 1707.05947.

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization, 19*(4), 1574–1609.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2018). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the International Conference on Learning Representations(ICLR)*.

Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization, 1*(3), 127–239.

Pensia, A., Jog, V., & Loh, P. (2018). Generalization error bounds for noisy, iterative algorithms. arXiv: 1801.04295v1.

Poggio, T., Voinea, S., & L., R. (2011). Online learning, stability, and stochastic gradient descent. ArXiv: 1105.4701v3.

Polyak, B. (1963). Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics, 3*(4), 864–878.

Russo, D., & Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 51, pp. 1232–1240.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research, 11,* 2635–2670.

Shapiro, A., & Nemirovski, A. (2005). *On Complexity of Stochastic Programming Problems*. New York: Springer.

Sokolić, J., Giryes, R., Sapiro, G., & Rodrigues, M. R. D. (2017). Robust large margin deep neural networks. *IEEE Transactions on Signal Processing, 65*(16), 4265–4280.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Hoboken: Wiley.

Xu, A., & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Proceedings of the 30th Advances in Neural Information Processing Systems (NIPS)*, pp. 2521–2530.

Xu, H., & Mannor, S. (2012). Robustness and generalization. *Machine Learning, 86*(3), 391–423.

Yin, D., Pananjady, A., Lam, M., Papailiopoulos, D., Ramchandran, K., & Bartlett, P. L. (2017). Gradient diversity: a key ingredient for scalable distributed learning. ArXiv: 1706.05699v3.

Zahavy, T., Kang, B., Sivak, A., Feng, J., Xu, H., & Mannor, S. (2017). Ensemble robustness and generalization of stochastic deep learning algorithms. ArXiv: 1602.02389v4.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhou, Y., Zhang, H., & Liang, Y. (2016). Geometrical properties of phase retrieval and convergence of accelerated reshaped wirtinger flow. In *Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.