Minibatch and Local SGD: Algorithmic Stability and Linear Speedup in Generalization

Yunwen Lei[a,*], Tao Sun[b], Mingrui Liu[c]

[a]*Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong, China*
[b]*College of Computer, National University of Defense Technology, Changsha, China*
[c]*Department of Computer Science, George Mason University, Fairfax, USA*

Abstract

The increasing scale of data propels the popularity of leveraging parallelism to speed up the optimization. Minibatch stochastic gradient descent (minibatch SGD) and local SGD are two popular methods for parallel optimization. The existing theoretical studies show a linear speedup of these methods with respect to the number of machines, which, however, is measured by optimization errors in a multi-pass setting. As a comparison, the stability and generalization of these methods are much less studied. In this paper, we study the stability and generalization analysis of minibatch and local SGD to understand their learnability by introducing an expectation-variance decomposition. We incorporate training errors into the stability analysis, which shows how small training errors help generalization for overparameterized models. We show minibatch and local SGD achieve a linear speedup to attain the optimal risk bounds.

Keywords:
Learning Theory, Algorithmic Stability, Stochastic Gradient Descent, Generalization Analysis

1. Introduction

Modern machine learning often comes along with models and datasets of massive scale (e.g., millions or billions of parameters over enormous training datasets) [52, 28, 38, 22], which renders the training with sequential algorithms impractical for large-scale data analysis. To speed up the computation, it is appealing to develop learning schemes that can leverage parallelism to reduce the amount of time in the training stage [44]. First-order stochastic optimization is especially attractive for parallelism since the gradient computation is easy to parallelize across multiple computation devices [38, 27, 45]. For distributed optimization, communication has been reported to be a major bottleneck for large-scale applications [41]. Therefore, increasing the computation to communication ratio is a major concern in developing parallelizable optimization algorithms.

A simple stochastic first-order method is the minibatch stochastic gradient descent (minibatch SGD) [38, 12, 11, 28, 47], where the update at each round is performed based on an average of

*Corresponding author
Email addresses: `leiyw@hku.hk` (Yunwen Lei), `suntao.saltfish@outlook.com` (Tao Sun), `mingruil@gmu.edu` (Mingrui Liu)

gradients over several training examples rather than a single example. Using a minibatch helps in reducing the variance, and therefore accelerates the optimization. The computation over a minibatch of size b can be distributed over M machines, where each machine computes a minibatch of size $K = b/M$ before communication. This increases the computation to communication ratio. Due to its simplicity, minibatch SGD has found successful applications in a variety of settings [45, 38].

An orthogonal approach to increase the computation to communication ratio is the local SGD [32, 41, 50]. For local SGD with M machines, we divide the implementation into R rounds. At each round, each machine conducts SGD independently in K iterations, after which an average over M machines is taken to get a consensus point. Unlike minibatch SGD, local SGD is constantly improving its behavior even when the machines are not communicating with each other. Due to this appealing property, local SGD has been widely deployed in many applications [32].

The promising applications of minibatch SGD and local SGD motivate a lot of theoretical work to understand the performance of these methods. A linear speedup with respect to (w.r.t.) the batch size was established for minibatch SGD in both online [12] and stochastic setting [38, 11], which is further extended to its accelerated variants [12, 45]. The analysis for local SGD is more challenging. A linear speedup w.r.t. the number of machines was developed for local SGD with strongly convex [41] and convex problems [44, 21]. These results on linear speedup build the theoretical foundation for using the parallelism to reduce the computation for large-scale problems.

The above results on linear speedup are obtained for optimization errors in a multi-pass setting, i.e., the performance of models on training examples. However, in machine learning we care more about the generalization behavior of these models on testing examples, which have been scarcely touched for both minibatch and local SGD with multi-passes over the data. To our knowledge, other than regression with the specific least squares loss [35, 6, 29, 17], there is no generalization analysis of minibatch and local SGD that shows a linear speedup measured by testing errors. In this paper, we conduct the generalization analysis of minibatch and local SGD based on the concept of algorithmic stability [4]. Our aim is to show the linear speedup observed in optimization errors also holds for testing errors. Our main contributions are summarized as follows.

1. We develop stability bounds of minibatch SGD for convex, strongly convex, and nonconvex problems. Our stability bounds incorporate the property of small training errors, which are often the case for overparamterized models. For strongly convex problems, we develop stability bounds independent of the iteration number, which is also novel for the vanilla SGD in the sense of removing the Lipschitz continuity assumption. Based on these stability bounds, we further develop optimistic bounds on excess population risks which imply fast rates under a low noise condition.

2. We develop stability bounds of local SGD for both convex and strongly convex problems, based on which we develop excess risk bounds. This gives the first stability and generalization bounds for local SGD.

3. Our risk bounds for both minibatch SGD and local SGD are optimal. For convex problems our

bounds are of the order $O(1/\sqrt{n})$, while for μ-strongly convex problems our bounds are of the order $O(1/(n\mu))$, where n is the sample size. These match the existing minimax lower bounds for the statistical guarantees [1]. Furthermore, we show that minibatch SGD achieves a linear speedup w.r.t. the batch size, and local SGD achieves a linear speedup w.r.t. the number of machines. To our knowledge, these are the first linear speedup for minibatch and local SGD in generalization for general problems in the multi-pass setting.

To achieve these results, we develop techniques by introducing the *expectation-variance decomposition* and self-bounding property [24, 25] into the stability analysis based on a reformulation of minibatch SGD with binomial variables [14]. Indeed, the existing stability analysis of the vanilla SGD [18, 24, 25] does not apply to minibatch SGD. Furthermore, even with our formulation, the techniques in [25] would imply suboptimal stability bounds.

The paper is organized as follows. We survey the related work in Section 2, and formulate the problem in Section 3. We study the stability and generalization for minibatch SGD in Section 4, and extend these discussions to local SGD in Section 5. We present the proof of minibatch SGD in Section 6 and the proof of local SGD in Section 7. We conclude the paper in Section 8.

2. Related Work

In this section, we survey the related work on algorithmic stability, minibatch and local SGD.

Algorithmic stability. As a fundamental concept in statistical learning theory (SLT), algorithmic stability measures the sensitivity of an algorithm w.r.t. the perturbation of a training dataset. Various concepts of stability have been introduced into the literature, including uniform stability [4], hypothesis stability [4], on-average stability [37, 24] and on-average model stability [25]. One of the most widely used stability concept is the uniform stability, which can imply almost optimal high-probability bounds [14, 5, 13]. Stability has found wide applications in stochastic optimization [18, 25, 24, 7, 34, 43, 10, 9]. An important property of the stability analysis is that it considers only the particular model produced by the algorithm, and therefore can use the property of the learning algorithm to imply capacity-independent generalization bounds. Lower bounds on the stability of gradient methods also draw increasing attention [3, 23].

Minibatch algorithm. Minibatch algorithms are efficient in speeding up optimization for smooth problems. Shamir and Srebro [38] showed that minibatch distributed optimization can attain a linear speedup w.r.t. the batch size, which was also observed for general algorithms in an online learning setting [12]. These results were improved in [11], where the convergence rates involve the training error of the best model and would decay fast in an interpolation setting. The above speedup was derived if the batch size is not large. Indeed, a large batch size may negatively affect the performance of the algorithm [20, 31]. Minibatch stochastic approximation methods were studied for stochastic composite optimization problems [15] and nonconvex problems [16]. Recently, minibatch algorithms have been shown to be immune to the heterogeneity of the problem [45]. For problems with nonsmooth loss functions, minibatch algorithms do not get any speedup [38].

Local SGD. Local SGD, also known as "parallel SGD" or "federated averaging", is widely used to solve large-scale convex and nonconvex optimization problems [32]. A linear speedup in the number (M) of machines was obtained for local SGD on strongly convex problems [41]. The key observation is that local SGD can roughly yield a reduction in the variance by a factor of M. Despite its promising performance in practice, the theoretical guarantees on convergence rates are still a bit weak and are often dominated by minibatch SGD. Indeed, initial analysis of local SGD failed to derive a convergence rate matching minibatch SGD's performance, due to an additional term proportional to the dispersion of the individual machine's iterates for local SGD [44]. For example, the work [44] also presented a lower bound on the performance of local SGD that is worse than the minibatch SGD guarantee in a certain regime, showing that local SGD does not dominate minibatch SGD. Until recently, the guarantees better than minibatch SGD were obtained under some cases (e.g., case with rare communication) [44, 21, 39]. These discussions impose different assumptions: Woodworth et al. [44] imposed a bounded variance assumption, while Khaled et al. [21] considered an almost sure smoothness assumption without the bounded variance assumption. These results were extended to a heterogeneous distributed learning setting [21, 45], for which heterogeneity was shown to be particularly problematic for local SGD. A linear speedup w.r.t. M was also observed for nonconvex loss functions under a more restrictive constraint on the synchronization delay than that in the convex case [49]. Lower bounds of local SGD were established [44]. Generalization bounds of federated learning were recently studied based on Rademacher complexity [33] and stability [42, 8].

The above results on the linear speedup for minibatch and local SGD were obtained for optimization errors, which is the focus of the paper. The benefit of minibatch in generalization was studied for SGD with the square loss function [35, 29, 6]. These discussions use the analytic representation of iterators in terms of integral operators, which do not apply to general problems considered here.

3. Problem Setup

Let ρ be a probability measure defined on a sample space $\mathcal{Z}$, from which we independently draw a dataset $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ of n examples. Based on S, we wish to learn a model $\mathbf{w}$ in a model space $\mathcal{W} = \mathbb{R}^d$ for prediction, where $d \in \mathbb{N}$ is the dimension. The performance of $\mathbf{w}$ on a single example $z \in \mathcal{Z}$ can be measured by a nonnegative loss function $f(\mathbf{w}; z)$. The empirical behavior of $\mathbf{w}$ can be quantified by the empirical risk $F_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i)$. Usually, we apply a randomized algorithm A to minimize F_S over $\mathcal{W}$ to get a model $A(S)$. Then an algorithm can be considered as a map from the set of samples to $\mathcal{W}$, i.e., $A : \cup_{n=1}^{\infty} \mathcal{Z}^n \mapsto \mathcal{W}$. A good behavior on training examples does not necessarily mean a good behavior on testing examples, which is the quantity of real interest in machine learning and can be quantified by the population risk $F(\mathbf{w}) := \mathbb{E}_Z[f(\mathbf{w}; Z)]$. Here $\mathbb{E}_Z[\cdot]$ denotes the expectation w.r.t. Z. In this paper, we study the excess population risk of a model $\mathbf{w}$ defined by $F(\mathbf{w}) - F(\mathbf{w}^*)$, which measures the suboptimality as compared to the best model $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$. Our basic strategy is to use the following error decomposition

$$\mathbb{E}_{S,A}\big[F(A(S)) - F(\mathbf{w}^*)\big] = \mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] + \mathbb{E}_{S,A}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big], \qquad (3.1)$$

where we have used the identity $\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$ and $\mathbb{E}_{S,A}[\cdot]$ denotes the expectation w.r.t. S and A. We refer to the first term $\mathbb{E}\big[F(A(S)) - F_S(A(S))\big]$ as the generalization error (generalization gap), which measures the discrepancy between training and testing at the output model $A(S)$. We call the second term $\mathbb{E}\big[F_S(A(S)) - F_S(\mathbf{w}^*)\big]$ the optimization error, which measures the suboptimality in terms of the empirical risk. One can control the optimization error by tools in optimization theory. As a comparison, there is little work on the generalization error of minibatch SGD and local SGD in the multi-pass setting, the key challenge of which is the dependency of $A(S)$ on S.

In this paper, we will use a specific algorithmic stability —on-average model stability— to address the generalization error. We use $\|\cdot\|_2$ to denote the Euclidean norm. We denote $S \sim S'$ if S and S' differ by at most a single example.

Definition 1 (Uniform Stability)**.** Let $\epsilon > 0$. We say a randomized algorithm A is ϵ-uniformly stable if $\sup_{S\sim S',z} \mathbb{E}_A[|f(A(S);z) - f(A(S');z)|] \le \epsilon$.

Definition 2 (On-average Model Stability [25])**.** Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from ρ. For any $i \in [n] := \{1, \ldots, n\}$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$ as the set formed from S by replacing the i-th element with z'_i. Let $\epsilon > 0$. We say a randomized algorithm A is ℓ_1 on-average model ϵ-stable if $\mathbb{E}_{S,S',A}\big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2\big] \le \epsilon$, and ℓ_2 on-average model ϵ-stable if $\mathbb{E}_{S,S',A}\big[\frac{1}{n}\sum_{i=1}^n \|A(S) - A(S^{(i)})\|_2^2\big] \le \epsilon^2$.

According to the above definition, on-average model stability considers the perturbation of each single example, and measures how these perturbations would affect the output models on average. Lemma 1 gives a quantitative connection between the generalization error and on-average model stability. We first introduce some necessary definitions. We use ∇g to denote the gradient of g.

Definition 3. Let $g : \mathcal{W} \mapsto \mathbb{R}$, $G, L > 0$ and $\mu \ge 0$.

1. We say g is G-Lipschitz continuous if $|g(\mathbf{w}) - g(\mathbf{w}')| \le G\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

2. We say g is L-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \le L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.

3. We say g is μ-strongly convex if $g(\mathbf{w}) \ge g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}')\rangle + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is convex if it is μ-strongly convex with $\mu = 0$.

A non-negative and L-smooth function g enjoys the self-bounding property, meaning $\|\nabla g(\mathbf{w})\|_2^2 \le 2Lg(\mathbf{w})$ [40]. Examples of smooth and convex loss functions include the logistic loss, least square loss and Huber loss. Examples of Lipschitz and convex loss functions include the hinge loss, logistic loss and Huber loss.

Lemma 1 ([25])**.** *Let S, S' and $S^{(i)}$ be constructed as in Definition 2, and $\gamma > 0$.*

(a) Suppose for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex. If A is ℓ_1 on-average model ϵ-stable and $\sup_z \|\nabla f(A(S); z)\|_2 \leq G$ for any S, then $\big|\mathbb{E}_{S,A}\big[F_S(A(S)) - F(A(S))\big]\big| \leq G\epsilon$.

(b) Suppose for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. If A is ℓ_2 on-average model ϵ-stable, then the following inequality holds

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{L}{\gamma}\mathbb{E}_{S,A}\big[F_S(A(S))\big] + \frac{L+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,S',A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big].$$

Part (a) gives the connection between generalization and ℓ_1 on-average model stability under a convexity condition, while Part (b) relates generalization to ℓ_2 on-average model stability under a smoothness condition (without a Lipschitzness condition). Note Part (a) differs slightly from that in [25] by replacing the Lipschitz condition with a convexity condition and $\sup_z \|\nabla f(A(S); z)\|_2 \leq G$. However, the analysis is almost identical and we omit the proof. An advantage of ℓ_2 on-average model stability is that the upper bound involves the training errors, and improves if $F_S(A(S))$ is small.

4. Generalization of Minibatch SGD

In this section, we consider the minibatch SGD for convex, strongly convex and nonconvex problems. Minibatch SGD is implemented in several rounds/iterations. Let $\mathbf{w}_1 \in \mathcal{W}$ be an initial point. At the t-th round, minibatch SGD randomly draws (with replacement) b numbers $i_{t,1}, \ldots, i_{t,b}$ independently from the uniform distribution over $[n]$, where $b \in [n]$ is the batch size. Then it updates $\{\mathbf{w}_t\}$ by ($t \in [R] = \{1, 2, \ldots, R\}$)

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{b}\sum_{j=1}^{b}\nabla f(\mathbf{w}_t; z_{i_{t,j}}), \tag{4.1}$$

where $\{\eta_t\}$ is a positive step size sequence. If $b = 1$, then Eq. (4.1) recovers the vanilla SGD. If $b = n$, the above scheme is still different from gradient descent since we consider selection with replacement. For simplicity, we always assume $b \geq 2$. We summarize the results of minibatch SGD in Table 1.

4.1. Convex Case

We first present stability bounds to be proved in Section 6.1. Eq. (4.2) considers the ℓ_1 on-average model stability, while Eq. (4.3) considers the ℓ_2 on-average model stability. An advantage of the analysis with ℓ_2 on-average model stability over ℓ_1 on-average model stability is that it can imply generalization bounds without a Lipschitzness condition. We denote $A \lesssim B$ if there exists a universal constant C such that $A \leq CB$. We denote $A \gtrsim B$ if there exists a universal constant C such that $A \geq CB$. We denote $A \asymp B$ if $A \lesssim B$ and $A \gtrsim B$.

Theorem 2 (Stability Bounds for Minibatch SGD: Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let S, S' and $S^{(m)}$ be given in Definition 2. Let*

Table 1: Excess population risks of Minibatch SGD for convex, strongly convex and gradient-dominated problems. We consider smooth problems and only show the dependency on n, b, μ and $F(\mathbf{w}^*)$. The column "Risk" denotes the excess population risk, the column "R" denotes the number of iterations, the column "Constraint" indicates the constraint on the batch size b and the column "Optimal R" is derived by putting the largest b in R. We achieve a linear speedup w.r.t. the batch size for convex, strongly convex and nonconvex problems (PL condition is defined in Eq. (4.7)). For convex problems, we derive optimistic bounds which improve to $O(n^{-1})$ in a low noise case, i.e., $F(\mathbf{w}^*) < n^{-1}$.

Assumption		Risk	R	Constraint	Optimal R
convex	$F(\mathbf{w}^*) \geq 1/n$	$\sqrt{F(\mathbf{w}^*)/n}$	n/b	$b \leq \frac{\sqrt{nF(\mathbf{w}^*)}}{2L}$	$\frac{\sqrt{n}}{\sqrt{F(\mathbf{w}^*)}}$
	$F(\mathbf{w}^*) < 1/n$	$\frac{1}{n}$	n	—	n
μ-strongly convex		$1/(n\mu)$	$\max\{n/b, \mu^{-1}\log n\}$	—	$\mu^{-1}\log n$
μ-PL condition		$1/(n\mu)$	$n/(b\mu^2)$	$b \leq \sqrt{n}/\mu$	$\mu^{-1}\log n$

$\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(m)}\}$ be produced by (4.1) with $\eta_t \leq 2/L$ based on S and $S^{(m)}$, respectively. Then

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2] \leq \sum_{k=1}^{t}\frac{2\eta_k\sqrt{2L\mathbb{E}[F_S(\mathbf{w}_k)]}}{n} \tag{4.2}$$

and

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2^2\big] \leq \frac{16L}{nb}\sum_{k=1}^{t}\eta_k^2\mathbb{E}\big[F_S(\mathbf{w}_k)\big] + \frac{8}{n^3}\sum_{m=1}^{n}\mathbb{E}\Big[\Big(\sum_{k=1}^{t}\eta_k\|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big]. \tag{4.3}$$

Remark 1 (Explanation and comparison)**.** A property of these stability bounds is that they involve the empirical risks of $\mathbf{w}_k$, which would be small since we are minimizing the empirical risk by stochastic optimization algorithms. Similar stability bounds involving $F_S(\mathbf{w}_k)$ were developed for the vanilla SGD [25]. Their argument needs to distinguish two cases according to whether the algorithm chooses a particular example at each iteration. This argument does not work for the minibatch SGD since we draw b examples per iteration, and we can draw the particular example several times. We bypass this difficulty by introducing the *expectation-variance decomposition* and self-bounding property into the stability analysis based on a reformulation of minibatch SGD [24, 25, 14]. We refer the readers to Remark 8 for the detailed discussions on the novelty of our analysis.

The stability of minibatch SGD with $\eta_t = \eta$ has also been studied recently [47, 2]. The discussions in Theorem 9 in [47] give a stability bound of the order $O(\eta t/n + \gamma\eta t)$, where $\gamma = \Pr\{\inf_{\mathbf{w},\mathbf{w}'}\bar{B}_S(\mathbf{w},\mathbf{w}') < (b-1)/(2/(L\eta) - n/(n-1))\}$ and $\bar{B}_S(\mathbf{w},\mathbf{w}')$ is a measure on the gradient diversity defined below

$$\bar{B}_S(\mathbf{w},\mathbf{w}') := \frac{n\sum_{i=1}^{n}\|\nabla f(\mathbf{w};z_i)-\nabla f(\mathbf{w}';z_i)\|_2^2}{\|\sum_{i=1}^{n}(\nabla f(\mathbf{w};z_i)-\nabla f(\mathbf{w}';z_i))\|_2^2}.$$

If γ is not very small, their stability bounds would be vacuous due to the term $\gamma\eta t$. The stability bound order $O(\eta t/n)$ was developed in [2]. These discussions require f to be convex, smooth and Lipschitz continuous. Furthermore, these discussions do not incorporate training errors into the stability bounds, and cannot imply optimistic bounds. We remove the Lipschitz condition in our analysis and obtain optimistic bounds.

We plug the stability bounds in Theorem 2 into Lemma 1 to control generalization errors, which together with the optimization error bounds in Lemma 15, implies the following excess risk bounds. It should be noted that we do not require the function f to be Lipschitz continuous. The proof is given in Section 6.2.

Theorem 3 (Risk Bounds for Minibatch SGD: Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_t\}$ be produced by (4.1) with $\eta_t = \eta \leq 1/(2L)$. Then the following inequality holds for $\bar{\mathbf{w}}_R := \frac{1}{R}\sum_{t=1}^R \mathbf{w}_t$ and any $\gamma > 0$*

$$\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{\eta L F(\mathbf{w}^*)}{b} + \frac{\|\mathbf{w}^*\|_2^2}{\eta R} + L\Big(F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{\eta R}\Big)\Big(\frac{1}{\gamma} + (L+\gamma)\eta^2\Big(\frac{R}{nb} + \frac{R^2}{n^2}\Big)\Big).$$

Note the above excess risk bounds involve $F(\mathbf{w}^*)$ and would improve if $F(\mathbf{w}^*)$ is small, which is true in many learning problems. The terms involving $F(\mathbf{w}^*)$ also correspond to gradient noise since the variance of gradients can be bounded by function values according to the self-bounding property of smooth functions. The risk bounds of this type are called optimistic bounds in the literature [40].

As a corollary, we develop explicit excess risk bounds by choosing suitable step sizes and number of rounds, using the idea of early-stopping [46]. Note the step size depends on $F(\mathbf{w}^*)$ which is unknown to us. However, this is not a big issue since we can choose step sizes independent of $F(\mathbf{w}^*)$ to derive bounds of the same order of n but worse order of $F(\mathbf{w}^*)$. It shows that minibatch SGD can achieve the excess risk bounds of the order $\sqrt{F(\mathbf{w}^*)/n}$ if $F(\mathbf{w}^*) \geq 1/n$, and can imply much better error bounds of the order $1/n$ if $F(\mathbf{w}^*) < 1/n$. The proof is given in Section 6.2.

Corollary 4. *Let assumptions in Theorem 3 hold and $\eta = \min\big\{\frac{\|\mathbf{w}^*\|_2 b}{\sqrt{LnF(\mathbf{w}^*)}}, \frac{1}{2L}\big\}$.*

1. If $F(\mathbf{w}^*) \geq 4Lb^2\|\mathbf{w}^*\|_2^2/n$, we can take $R \asymp \frac{n}{b}$ to derive $\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{(LF(\mathbf{w}^*))^{\frac{1}{2}}\|\mathbf{w}^*\|_2}{\sqrt{n}}$.

2. If $F(\mathbf{w}^*) \leq 4Lb^2\|\mathbf{w}^*\|_2^2/n$, we take $R \asymp n$ to get $\mathbb{E}[F(\bar{\mathbf{w}}_R)] \lesssim F(\mathbf{w}^*) + \frac{L\|\mathbf{w}^*\|_2^2}{n}$.

Remark 2 (Linear speedup)**.** We now give some explanations on linear speedup. For the case $F(\mathbf{w}^*) \gtrsim 1/n$, a larger batch size allows for a larger step size, which further decreases the number R of rounds. It shows that minibatch SGD achieves a linear speedup if the batch size is not large, i.e., it only requires $O(n/b)$ rounds to achieve the excess risk bound $O(n^{-\frac{1}{2}})$ if $b \lesssim \sqrt{nF(\mathbf{w}^*)}/(\sqrt{L}\|\mathbf{w}^*\|_2)$. Such a linear speedup was observed for optimization errors for multi-pass SGD [11]. Indeed, it was shown that minibatch SGD requires $O(n/b)$ rounds to achieve the optimization error bounds $\mathbb{E}[F_S(\bar{\mathbf{w}}_R)] - F_S(\mathbf{w}^*) \lesssim \sqrt{F_S(\mathbf{w}^*)/n}$ if $b \lesssim \sqrt{nF(\mathbf{w}^*)}/(\sqrt{L}\|\mathbf{w}^*\|_2)$. We extend the existing optimization error analysis to generalization, and develop the first linear speedup of the minibatch multi-pass SGD as measured by risks for general convex problems. In particular, our regime $b \lesssim \sqrt{nF(\mathbf{w}^*)}$ for linear speedup in generalization matches the regime $b \lesssim \sqrt{nF_S(\mathbf{w}^*)}$ for the linear speedup in optimization [11].

For the case $F(\mathbf{w}^*) \lesssim 1/n$, Corollary 4 shows that a larger batch size does not bring any gain in speeding up the risk bounds. The underlying reason is that the variance is already very small in this case, and a further reduction of variance by minibatch does not bring essential benefits in the learning process.

4.2. Strongly Convex Case

We now consider strongly convex problems. Theorem 5 gives stability bounds, while Theorem 6 gives excess population risk bounds. The proofs are given in Section 6.3.

Theorem 5 (Stability Bounds for Minibatch SGD: Strongly Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, μ-strongly convex and L-smooth. Let S, S' and $S^{(m)}$ be constructed as in Definition 2. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(m)}\}$ be produced by* (4.1) *based on S and $S^{(m)}$, respectively. Then*

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2] \le \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t}\eta_k\sqrt{\mathbb{E}[F_S(\mathbf{w}_k)]}\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2), \tag{4.4}$$

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2] \lesssim 1/(n\mu), \tag{4.5}$$

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2^2\big] \le \sum_{k=1}^{t}\Big(\frac{16L\eta_k^2}{nb}+\frac{32L\eta_k}{n^2\mu}\Big)\mathbb{E}[F_S(\mathbf{w}_k)]\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2). \tag{4.6}$$

Remark 3 (Explanation)**.** Eq. (4.4) and Eq. (4.5) consider the ℓ_1 on-average stability. The former involves the empirical risks in the upper bound and therefore can benefit from small empirical risks, while the latter shows minibatch SGD is always stable in the strongly convex case, no matter how many iterations it takes. Eq. (4.5) is also new in the vanilla SGD case with $b = 1$. Indeed, the work [18] also derived the iteration-independent stability bound $O(1/n\mu)$. However, their discussion requires the function f to be strongly-convex, smooth and Lipschitz. We show that the Lipschitz condition can be removed without affecting the stability bounds. Eq. (4.6) addresses the ℓ_2 on-average stability, which shows that increasing the batch size is beneficial to stability.

Theorem 6 (Risk Bounds for Minibatch SGD: Strongly Convex Case)**.** *Let assumptions in Theorem 5 hold and assume $\sup_z \|\nabla f(A(S); z)\|_2 \le G$. Let $\sigma_*^2 = \mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}^*; z_{i_t})\|_2^2]$. If $R \ge \frac{L}{\mu}\log\frac{nL}{G}$ and $b \ge n\sigma_*^2/(GR)$, then we can find appropriate step size sequences and an average $\hat{\mathbf{w}}_R$ of $\{\mathbf{w}_t\}_{t=1}^R$ such that $\mathbb{E}[F(\hat{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim G/(n\mu)$.*

Note that the assumption $\sup_z \|\nabla f(A(S); z)\|_2 \le G$ is much milder than the Lipschitz condition since it only requires a bound of the gradient on the output model, which can be achieved by a projection to the final output. To obtain the excess population risk bounds of the order $O(G/(n\mu))$, we require $R = \max\{\frac{n\sigma_*^2}{Gb}, \frac{L}{\mu}\log\frac{nL}{G}\}$. Then, if $b \lesssim n\mu\sigma_*^2/(GL\log(nL/G))$, we know $\frac{L}{\mu}\log\frac{nL}{G} \lesssim \frac{n\sigma_*^2}{Gb}$ and choose $R \asymp \frac{n\sigma_*^2}{Gb}$ to obtain a linear speedup w.r.t. the batch size.

4.3. Nonconvex Case

In this subsection, we consider minibatch SGD for nonconvex problems. The following theorem presents the stability bounds for smooth problems without the convexity and Lipschitzness assumption. The proof is given in Section 6.4.

Theorem 7. *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. Let S, S' and $S^{(m)}$ be given in Definition 2. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(m)}\}$ be produced by (4.1) with $\eta_t \leq 2/L$ based on S and $S^{(m)}$, respectively. Then*

$$\frac{1}{n}\sum_{m=1}^{n} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] \leq \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t} \eta_k \mathbb{E}\big[\sqrt{F_S(\mathbf{w}_k)}\big] \prod_{k'=k+1}^{t}(1+\eta_{k'}L).$$

Now, we consider a special nonconvex problem under a Polyak-Łojasiewicz (PL) condition. The PL condition was shown to hold for deep (linear) and shallow neural networks [7].

Assumption 1 (Polyak-Łojasiewicz Condition)**.** Let $\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. We assume F_S satisfies the PL condition with parameter $\mu > 0$, i.e., for all $\mathbf{w} \in \mathcal{W}$

$$\mathbb{E}_S\big[F_S(\mathbf{w}) - F_S(\mathbf{w}_S)\big] \leq \frac{1}{2\mu}\mathbb{E}_S\big[\|\nabla F_S(\mathbf{w})\|_2^2\big]. \tag{4.7}$$

Theorem 8 gives risk bounds for minibatch SGD under the PL condition, whose proof is given in Section 6.4.

Theorem 8 (Risk Bounds for Minibatch SGD: PL Condition)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. Let $\{\mathbf{w}_t\}$ be produced by Eq. (4.1) with $\eta_t = 2/(\mu(t+a))$ and $a \geq 4L/\mu$. Let Assumption 1 hold and $\mathbb{E}_z\big[\|\nabla f(\mathbf{w}_t; z_{i_k}) - \nabla F_S(\mathbf{w}_t)\|_2^2\big] \leq \sigma^2$, where i_k follows from the uniform distribution over $[n]$. If $R \geq \max\big\{L\sqrt{n}/\mu, L\sigma^2 n/(b\mu^2)\big\}$, then $\mathbb{E}[F(\mathbf{w}_R)] - F(\mathbf{w}^*) \lesssim L/(n\mu)$.*

According to Theorem 8, we require $R \geq \max\big\{L\sqrt{n}/\mu, L\sigma^2 n/(b\mu^2)\big\}$ to obtain the excess risk bounds $O(1/(n\mu))$. If $b \leq \sigma^2\sqrt{n}/\mu$, we have $L\sigma^2 n/(b\mu^2) \geq L\sqrt{n}/\mu$ and therefore we can choose $R \asymp L\sigma^2 n/(b\mu^2)$ to obtain a linear speedup w.r.t. the batch size. In particular, we can choose $b \asymp \sigma^2\sqrt{n}/\mu$ and $R \asymp L\sqrt{n}/\mu$ to get the bound $\mathbb{E}[F(\mathbf{w}_R)] - F(\mathbf{w}^*) \lesssim L/(n\mu)$.

5. Generalization of Local SGD

In this section, we consider local SGD with M machines and R rounds. At the r-th round, each machine starts with the same iterate $\mathbf{w}_r$ and independently applies SGD with K steps. After that, we take an average of the iterates in each machine to get a consensus point $\mathbf{w}_{r+1}$. Let $\mathbf{w}_{m,r,t+1}$ be the $(t+1)$-th iterate in the machine m at round r. Then, the formulation of local SGD is given below

$$\begin{aligned}
\mathbf{w}_{m,r,1} &= \mathbf{w}_r, \quad m \in [M],\\
\mathbf{w}_{m,r,t+1} &= \mathbf{w}_{m,r,t} - \eta_{r,t}\nabla f(\mathbf{w}_{m,r,t}; z_{i_{m,r,t}}), \quad t \in [K],\\
\mathbf{w}_{r+1} &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,K+1}, \quad r \in [R],
\end{aligned} \tag{5.1}$$

where $\eta_{r,t}$ is the step size for the t-th update at round r, and $i_{m,r,t}$ is drawn independently from the uniform distribution over $[n]$. The pseudo-code is given in Algorithm 1. If $R = 1$, then local SGD becomes the one-shot SGD, i.e., one only takes an average once in the end of the optimization [52, 30, 19]. If $K = 1$, then local SGD becomes the minibatch SGD. Note that the computation cost per machine is KR. We summarize the results on local SGD in Table 2, where we consider smooth problems and ignore constant factors.

Algorithm 1 Local SGD

Inputs: step sizes $\{\eta_{m,r,t}\}$ and S
Initialize: $\mathbf{w}_1 \in \mathcal{W}$
for $r = 1, 2, \ldots, R$ **do**
  for $m = 1, 2, \ldots, M$ **in parallel do**
    $\mathbf{w}_{m,r,1} = \mathbf{w}_r$
    for $t = 1, 2, \ldots, K$ **do**
      $\mathbf{w}_{m,r,t+1} = \mathbf{w}_{m,r,t} - \eta_{r,t} \nabla f(\mathbf{w}_{m,r,t}; z_{i_{m,r,t}})$
    end for
  end for
  $\mathbf{w}_{r+1} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{w}_{m,r,K+1}$
end for
Outputs: an average of $\mathbf{w}_{m,r,t}$

Table 2: Excess population risks of Local SGD for convex and strongly convex problems. The column "Risk" denotes the excess population risk, the column "KR" denotes the number of iterations per local machine, the column "R" denotes the communication cost, the column "Constraint" indicates the constraint on the number of machines M and the column "Optimal KR" is derived by putting the largest M in KR. We achieve a linear speedup w.r.t. the number of machines for both convex and strongly convex problems, under different regimes of M.

Assumption	Risk	KR	R	Constraint	Optimal KR
convex	$O(1/\sqrt{n})$	n/M	$n/(KM)$	$M \leq n^{\frac{1}{2}}$	$\sqrt{n}$
μ-strongly convex	$O((n\mu)^{-1}\log(KR))$	n/M	$n/(KM)$	$M \leq \sqrt{n\mu}$	$\sqrt{n/\mu}$

In the following theorem, we develop the stability bounds for local SGD to be proved in Section 7.1. We consider both ℓ_1 and ℓ_2 on-average model stabilities.

Theorem 9 (Stability Bound for Local SGD)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let S, S' and $S^{(k)}$ be constructed as in Definition 2. Let $\{\mathbf{w}_r\}$ and $\{\mathbf{w}_r^{(k)}\}$ be produced by (5.1) with $\eta_{r,t} \leq 2/L$ based on S and $S^{(k)}$, respectively. Then*

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\big[\|\mathbf{w}_{R+1} - \mathbf{w}_{R+1}^{(k)}\|_2\big] \leq \frac{2\sqrt{2L}}{nM}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K} \eta_{r,t}\mathbb{E}\Big[\sqrt{F_S(\mathbf{w}_{m,r,t})}\Big], \tag{5.2}$$

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\big\|\mathbf{w}_{R+1}-\mathbf{w}_{R+1}^{(k)}\big\|_2^2\big]\le\frac{16L}{nM^2}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}^2\mathbb{E}\big[F_S(\mathbf{w}_{m,r,t})\big]$$
$$+\frac{2}{n^3M^2}\sum_{k=1}^{n}\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\|\nabla f(\mathbf{w}_{m,r,t};z_k)-\nabla f(\mathbf{w}_{m,r,t}^{(k)};z_k')\|_2\Big)^2\Big]. \quad (5.3)$$

Remark 4 (Simplification)**.** Note that the above stability bounds involve empirical risks, and can benefit from small empirical risks. Assume $\eta_{r,t}=\eta$ and $\mathbb{E}\big[\sqrt{F_S(\mathbf{w}_{m,r,t})}\big]\lesssim 1$ (this is a reasonable assumption since we are minimizing F_S). Then Eq. (5.2) implies $\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{R+1}-\mathbf{w}_{R+1}^{(k)}\|_2\big]\lesssim KR\eta/n$. Eq. (5.3) implies $\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\big\|\mathbf{w}_{R+1}-\mathbf{w}_{R+1}^{(k)}\big\|_2^2\big]\lesssim KR\eta^2/(nM)+R^2K^2\eta^2/n^2$, which shows that increasing the number of machines improves the stability and generalization. It was shown that increasing M can improve the optimization [44]. For example, the optimization error bound of the order $O\big(\frac{1}{K^{\frac{1}{3}}R^{\frac{2}{3}}}+\frac{1}{\sqrt{MKR}}\big)$ was developed in [44]. Therefore, we expect that increasing M would accelerate the learning process.

Remark 5 (Effect of M)**.** We give some explanation on the effect of M on stability analysis. Note the above ℓ_1 on-average stability bounds are independent of M, while the ℓ_2 on-average stability bounds improve as M increases. These phenomena can be explained by how the average operator affects the expectation and variance. Indeed, both the ℓ_1 and ℓ_2 stability analysis are based on the following inequality in Eq. (7.3)

$$\big\|\mathbf{w}_{R+1}-\mathbf{w}_{R+1}^{(k)}\big\|_2\le\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}, \quad (5.4)$$

where $\mathfrak{C}_{m,r,t,k}=\|\nabla f(\mathbf{w}_{m,r,t};z_k)-\nabla f(\mathbf{w}_{m,r,t}^{(k)};z_k')\|_2$, and $\mathbb{I}_{[i_{m,r,t}=k]}=1$ if $i_{m,r,t}=k$, and 0 otherwise. Note the above upper bound is an average of $\xi_m:=\sum_{r=1}^{R}\sum_{t=1}^{K}\eta_{r,t}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}$ over $m\in[M]$, which comes from the average scheme in local SGD. We take an expectation over both sides of Eq. (5.4) to get ℓ_1 on-average stability bounds. An average operator does not affect the expectation, which explains why the ℓ_1 on-average stability bounds are independent of M. We take an expectation-variance decomposition to conduct the ℓ_2 stability analysis, and the resulting bound involves a term related to variance and a term related to expectation. The variance of an average of M random variables decreases by a factor of M, which explains why the first term on the right-hand side of Eq. (5.3) involves a factor of $1/M$. The second term in Eq. (5.3) is independent of M since the average does not affect expectation. This phenomenon also happens for minibatch SGD, where the average over a batch of size b decreases the variance by a factor of b, and does not affect the expectation.

In the following table, we summarize the comparison on the stability bounds of minibatch and local SGD for convex and smooth problems. Here T is the number of iterations per machine, which is R for minibatch SGD and RK for local SGD. For simplicity, we ignore the discussion with optimistic bounds, and simply assume the empirical risks are bounded in expectation.

Problems	ℓ_1 on-average model stability	ℓ_2 on-average model stability
minibatch SGD	$\frac{T\eta}{n}$	$\frac{\sqrt{T}\eta}{\sqrt{nb}}+\frac{T\eta}{n}$
local SGD	$\frac{T\eta}{n}$	$\frac{\sqrt{T}\eta}{\sqrt{nM}}+\frac{T\eta}{n}$

Note that all the above bounds involve $\frac{T\eta}{n}$, which corresponds to an *expectation* term in controlling the distance between two sequences of SGD iterates. We have either the term $\frac{\sqrt{T}\eta}{\sqrt{nb}}$ or $\frac{\sqrt{T}\eta}{\sqrt{nM}}$ for ℓ_2 stability analysis, which corresponds to a *variance* and decreases as the batch size (number of machines) increases.

We now use the above stability bounds to develop excess population risk bounds for local SGD. We first consider a convex case. The proof is given in Section 7.2. Note our stability analysis for local SGD is data-dependent in the sense of involving training errors. Our excess risk bounds are not data-dependent since the existing optimization error bounds are not data-dependent [44]. It is interesting to develop data-dependent bounds for local SGD.

Theorem 10 (Risk Bound for Local SGD: Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_{m,r,t}\}$ be produced by the algorithm A defined in* (5.1) *with $\eta_{r,t} = \eta \leq 2/L$. Assume for all $r \in [R], t \in [K]$, $\mathbb{E}_{i_{m,r,t}}[\|\nabla f(\mathbf{w}_{m,r,t}; z_{i_{m,r,t}}) - \nabla F_S(\mathbf{w}_{m,r,t})\|_2^2] \leq \sigma^2$. Suppose we choose $\eta \asymp \|\mathbf{w}^*\|_2\sqrt{n}/(KR\sqrt{L})$. If $KRM \asymp n$, $\eta \lesssim (K-1)^{-\frac{1}{2}}\|\mathbf{w}^*\|_2^{\frac{1}{2}}/(nL)^{\frac{1}{4}}$ and $\eta \leq 1/(2L)$, then $\mathbb{E}[F(\bar{\mathbf{w}}_{R,1})] - F(\mathbf{w}^*) \lesssim \frac{\sqrt{L}\|\mathbf{w}^*\|_2}{\sqrt{n}}$, where $\bar{\mathbf{w}}_{R,1} = \frac{1}{MKR}\sum_{m=1}^M\sum_{r=1}^R\sum_{t=1}^K \mathbf{w}_{m,r,t}$.*

Remark 6 (Linear speedup)**.** Theorem 10 shows that local SGD can achieve the minimax optimal excess population risk bounds $1/\sqrt{n}$ in the sense of matching the existing lower bounds [1]. We now discuss the speedup in the computation and we have $\eta \asymp \|\mathbf{w}^*\|_2 M/\sqrt{nL}$. Note $\eta \leq 2/L$ requires $M \lesssim \sqrt{nL}/\|\mathbf{w}^*\|_2$. Furthermore, the condition $\eta \lesssim (K-1)^{-\frac{1}{2}}\|\mathbf{w}^*\|_2^{\frac{1}{2}}/(nL)^{\frac{1}{4}}$ requires $M \lesssim (nL)^{\frac{1}{4}}/\sqrt{(K-1)\|\mathbf{w}^*\|_2}$. Under these conditions, local SGD achieves a linear speedup in the sense that the computation per machine is of the order of $KR \asymp n/M$.

Finally, we give risk bounds of local SGD for strongly convex problems to be proved in Section 7.3.

Theorem 11 (Risk Bounds for Local SGD: Strongly Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, μ-strongly convex and L-smooth. Let $\{\mathbf{w}_{m,r,t}\}$ be produced by the algorithm A defined in* (5.1) *with $\eta_{r,t} = \frac{4}{\mu(a+(r-1)K+t)} \leq 2/L$ and $a \geq 2L/\mu$. Assume for all $r \in [R], t \in [K]$, $\mathbb{E}_{i_{m,r,t}}[\|\nabla f(\mathbf{w}_{m,r,t}; z_{i_{m,r,t}}) - \nabla F_S(\mathbf{w}_{m,r,t})\|_2^2] \leq \sigma^2$. Assume $\sup_z \|\nabla f(A(S); z)\|_2 \leq G$. If $KR \gtrsim \frac{n\sigma^2}{MG\sqrt{L}}$ and $\mu KR^2 \gtrsim \frac{n\sqrt{L}}{G}$, then*

$$\mathbb{E}[F(\bar{\mathbf{w}}_{R,2})] - F(\mathbf{w}^*) \lesssim G\sqrt{L}\log(KR)/(n\mu),$$

where

$$S_R = \sum_{r=1}^R\sum_{t=1}^K(a+(r-1)K+t) \text{ and } \bar{\mathbf{w}}_{R,2} = \frac{1}{MS_R}\sum_{m=1}^M\sum_{r=1}^R\sum_{t=1}^K(a+(r-1)K+t)\mathbf{w}_{m,r,t}.$$

If $M \lesssim \frac{\sqrt{n\mu}\sigma^2}{\sqrt{G}L^{\frac{3}{4}}\sqrt{K}}$, we can choose $R \asymp \frac{n\sigma^2}{G\sqrt{L}KM}$ to show that $\mu KR^2 \asymp \frac{\mu n^2\sigma^4}{G^2LKM^2} \gtrsim \frac{n\sqrt{L}}{G}$. Therefore, all the conditions of Theorem 11 hold, and we get the rate $G\sqrt{L}\log(KR)/(n\mu)$.

Remark 7 (Comparison)**.** Generalization bounds for agnostic federated learning were developed from a uniform convergence approach [33], which involve Rademacher complexities of function spaces and are algorithm-independent. As a comparison, we study generalization from an algorithmic stability approach and get complexity-independent bounds.

A federated stability was introduced to study the generalization of federated learning algorithms [8] in a strongly convex setting. As a comparison, our analysis also applies to general convex problems. Furthermore, their stability analysis was conducted for abstract approximate minimizers, while our stability analysis is developed for local SGD. Finally, their bound involves an upper bound of the loss function over a compact domain, and therefore cannot imply optimistic bounds.

There is a recent work on the generalization of federated learning algorithms on a heterogeneous setup where the i-th local machine has its own dataset S_i [42]. For local SGD with a constant step size η, their generalization bounds are of the order of $O(n^{-1}RK\sigma\eta(1+K\eta))$ under a Lipschitz continuity assumption and a bounded variance assumption $\mathbb{E}[\|\nabla f(\mathbf{w};z_i)-\nabla F_{S_i}(\mathbf{w})\|_2^2]\leq\sigma^2$, where z_i is drawn uniformly from S_i. While the bounds in [42] also involve $\|\nabla F(\mathbf{w}_t)\|$, it is dominated by σ and therefore cannot imply fast rates in an interpolation setting. As a comparison, our bounds in Eq. (5.3) are optimistic and decay fast if $F_S(\mathbf{w}_{m,r,t})$ decays to 0. Furthermore, the analysis in [42] requires a Lipschitz condition on the loss function, which is removed in our analysis. Finally, we also develop ℓ_2 on-average stability bounds, which are more challenging and illustrate the second-order information on the stability.

6. Proofs on Minibatch SGD

6.1. Proof of Theorem 2

To prove Theorem 2, we first introduce several lemmas. The following lemma shows the self-bounding property for nonnegative and smooth functions, meaning the magnitude of gradients can be bounded by function values [40, 48].

Lemma 12 ([40])**.** *Assume for all z, the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is nonnegative and L-smooth. Then $\|\nabla f(\mathbf{w};z)\|_2^2\leq 2Lf(\mathbf{w};z)$.*

In our analysis, we will use the concept of binomial distribution. Let $\mathrm{Var}(X)$ denote the variance of a random variable X.

Definition 4 (Binomial distribution)**.** The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent trials, with the probability of success on a single trial denoted by p. We use $B(n,p)$ to denote the binomial distribution with parameters n and p.

Lemma 13. *If $X\sim B(n,p)$, then*

$$\mathbb{E}[X]=np\quad\textit{and}\quad\mathit{Var}(X)=np(1-p).$$

A key property on establishing the stability of SGD is the non-expansiveness of the gradient-update operator established in the following lemma.

Lemma 14 ([18])**.** *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex and L-smooth. Then for $\eta \leq 2/L$ we know*

$$\|\mathbf{w} - \eta \nabla f(\mathbf{w}; z) - \mathbf{w}' + \eta \nabla f(\mathbf{w}'; z)\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2. \tag{6.1}$$

Furthermore, if $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is μ-strongly convex and $\eta \leq 1/L$ then

$$\|\mathbf{w} - \eta \nabla f(\mathbf{w}; z) - \mathbf{w}' + \eta \nabla f(\mathbf{w}'; z)\|_2 \leq (1 - \eta\mu/2)\|\mathbf{w} - \mathbf{w}'\|_2, \tag{6.2}$$

$$\|\mathbf{w} - \eta \nabla f(\mathbf{w}; z) - \mathbf{w}' + \eta \nabla f(\mathbf{w}'; z)\|_2^2 \leq (1 - \eta\mu)\|\mathbf{w} - \mathbf{w}'\|_2^2. \tag{6.3}$$

We are now ready to prove Theorem 2. The analysis for ℓ_1-stability bounds is standard [25]. As a comparison, the analysis with the ℓ_2-stability bounds requires new techniques such as the expectation-variance decomposition based on a representation of SGD with Binomial random variables. For simplicity, we define $J_t = \{i_{t,1}, \ldots, i_{t,b}\}, t \in \mathbb{N}$.

Proof of Theorem 2. Define

$$\alpha_{t,m} = \big|\{j : i_{t,j} = m\}\big|, \quad \forall t \in \mathbb{N}, m \in [n], \tag{6.4}$$

where we use $|S'|$ to denote the cardinality of a set S'. That is, $\alpha_{t,m}$ is the number of indices equal to m in the t-th iteration. Then the SGD update in Eq. (4.1) can be reformulated as

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \frac{\eta_t}{b} \sum_{k=1}^n \alpha_{t,k} \nabla f(\mathbf{w}_t; z_k), \\ \mathbf{w}_{t+1}^{(m)} &= \mathbf{w}_t^{(m)} - \frac{\eta_t}{b} \sum_{k: k \neq m} \alpha_{t,k} \nabla f(\mathbf{w}_t^{(m)}; z_k) - \frac{\eta_t \alpha_{t,m}}{b} \nabla f(\mathbf{w}_t^{(m)}; z_m'), \end{aligned} \tag{6.5}$$

from which we know

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2 = \big\|\mathbf{w}_t &- \frac{\eta_t}{b} \sum_{k: k \neq m} \alpha_{t,k} \nabla f(\mathbf{w}_t; z_k) - \frac{\eta_t \alpha_{t,m}}{b} \nabla f(\mathbf{w}_t; z_m) \\ &- \mathbf{w}_t^{(m)} + \frac{\eta_t}{b} \sum_{k: k \neq m} \alpha_{t,k} \nabla f(\mathbf{w}_t^{(m)}; z_k) + \frac{\eta_t \alpha_{t,m}}{b} \nabla f(\mathbf{w}_t^{(m)}; z_m')\big\|_2. \end{aligned} \tag{6.6}$$

For simplicity, introduce the notations for any $t \in [T]$ and $m \in [n]$

$$\Delta_{t,m} = \|\mathbf{w}_t - \mathbf{w}_t^{(m)}\|_2, \qquad \mathfrak{C}_{t,m} = \|\nabla f(\mathbf{w}_t; z_m) - \nabla f(\mathbf{w}_t^{(m)}; z_m')\|_2. \tag{6.7}$$

Since f is L-smooth and $\sum_{k: k \neq m} \alpha_{t,k} \leq b$, we know the function $\mathbf{w} \mapsto \frac{1}{b} \sum_{k: k \neq m} \alpha_{t,k} f(\mathbf{w}; z_k)$ is L-smooth. By Lemma 14 and the assumption $\eta_t \leq 1/L$, we know

$$\begin{aligned} \Delta_{t+1,m} &\leq \big\|\mathbf{w}_t - \frac{\eta_t}{b} \sum_{k: k \neq m} \alpha_{t,k} \nabla f(\mathbf{w}_t; z_k) - \mathbf{w}_t^{(m)} + \frac{\eta_t}{b} \sum_{k: k \neq m} \alpha_{t,k} \nabla f(\mathbf{w}_t^{(m)}; z_k)\big\|_2 + \frac{\eta_t \alpha_{t,m} \mathfrak{C}_{t,m}}{b} \\ &\leq \Delta_{t,m} + \frac{\eta_t \alpha_{t,m} \mathfrak{C}_{t,m}}{b}. \end{aligned}$$

We can apply the above inequality recursively and derive (note $\mathbf{w}_1 = \mathbf{w}_1^{(m)}$)

$$\Delta_{t+1,m} \leq \frac{1}{b} \sum_{k=1}^t \eta_k \alpha_{k,m} \mathfrak{C}_{k,m}. \tag{6.8}$$

According to the definition of $\alpha_{t,k}$, we know that $\alpha_{t,k}$ is a random variable following from the binomial distribution $B(b, 1/n)$ with parameters b and $1/n$, from which we know

$$\mathbb{E}[\alpha_{t,k}] = b/n, \quad \mathrm{Var}(\alpha_{t,k}) = b(1 - 1/n) \cdot (1/n) \le b/n. \tag{6.9}$$

Furthermore, Lemma 12 implies

$$\mathfrak{C}_{k,m} \le \|\nabla f(\mathbf{w}_k; z_m)\|_2 + \|\nabla f(\mathbf{w}_k^{(m)}; z'_m)\|_2 \le \sqrt{2Lf(\mathbf{w}_k; z_m)} + \sqrt{2Lf(\mathbf{w}_k^{(m)}; z'_m)}. \tag{6.10}$$

We can combine the above inequality, Eq. (6.9) and Eq. (6.8) together to derive

$$\mathbb{E}[\Delta_{t+1,m}] \le \frac{1}{b}\sum_{k=1}^{t} \eta_k \mathbb{E}\big[\alpha_{k,m}\mathfrak{C}_{k,m}\big] = \frac{1}{b}\sum_{k=1}^{t} \eta_k \mathbb{E}\big[\mathbb{E}_{J_k}[\alpha_{k,m}]\mathfrak{C}_{k,m}\big] = \frac{1}{n}\sum_{k=1}^{t} \eta_k \mathbb{E}\big[\mathfrak{C}_{k,m}\big], \tag{6.11}$$

where we have used the fact that $\mathfrak{C}_{k,m}$ is independent of J_k. According to the symmetry between z_m and z'_m, we know $\mathbb{E}[f(\mathbf{w}_t^{(m)}; z'_m)] = \mathbb{E}[f(\mathbf{w}_t; z_m)]$ and therefore Eq. (6.10) implies $\mathbb{E}\big[\mathfrak{C}_{k,m}\big] \le 2\sqrt{2L}\mathbb{E}\big[\sqrt{f(\mathbf{w}_k; z_m)}\big]$. It then follows that

$$\mathbb{E}[\Delta_{t+1,m}] \le \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t} \eta_k \mathbb{E}\big[\sqrt{f(\mathbf{w}_k; z_m)}\big].$$

It then follows from the concavity of $x \mapsto \sqrt{x}$ that

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^{n} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] &\le \frac{2}{n}\sum_{m=1}^{n}\sum_{k=1}^{t} \frac{\eta_k}{n}\mathbb{E}\big[\sqrt{2Lf(\mathbf{w}_k; z_m)}\big] \\ &\le \sum_{k=1}^{t} \frac{2\eta_k}{n}\sqrt{\frac{2L}{n}\sum_{m=1}^{n}\mathbb{E}[f(\mathbf{w}_k; z_m)]} = \sum_{k=1}^{t} \frac{2\eta_k\sqrt{2L\mathbb{E}[F_S(\mathbf{w}_k)]}}{n}. \end{aligned}\tag{6.12}$$

This establishes the stated bound (4.2).

We now prove Eq. (4.3). We introduce an expectation-variance decomposition in (6.8) as follows

$$\Delta_{t+1,m} \le \frac{1}{b}\sum_{k=1}^{t} \eta_k\big(\alpha_{k,m} - b/n\big)\mathfrak{C}_{k,m} + \frac{1}{n}\sum_{k=1}^{t} \eta_k \mathfrak{C}_{k,m}.$$

We take square on both sides followed with an expectation (w.r.t. S and $J_1, \ldots, J_t$) and use $(a+b)^2 \le 2(a^2 + b^2)$ to show

$$\begin{aligned}\mathbb{E}\big[\Delta_{t+1,m}^2\big] &\le \frac{2}{b^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{t} \eta_k\big(\alpha_{k,m} - b/n\big)\mathfrak{C}_{k,m}\Big)^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}\Big)^2\Big] \\ &= \frac{2}{b^2}\mathbb{E}\Big[\sum_{k,k'=1}^{t} \eta_k\eta_{k'}\big(\alpha_{k,m} - b/n\big)\big(\alpha_{k',m} - b/n\big)\mathfrak{C}_{k,m}\mathfrak{C}_{k',m}\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}\Big)^2\Big].\end{aligned}$$

For any $k \neq k'$, it follows from $\mathbb{E}_{J_{k'}}[\alpha_{k',m}] = b/n$ (we can assume $k < k'$ without loss of generality)

$$\begin{aligned}\mathbb{E}\Big[\big(\alpha_{k,m} - b/n\big)\big(\alpha_{k',m} - b/n\big)\mathfrak{C}_{k,m}\mathfrak{C}_{k',m}\Big] &= \mathbb{E}\mathbb{E}_{J_{k'}}\Big[\big(\alpha_{k,m} - b/n\big)\big(\alpha_{k',m} - b/n\big)\mathfrak{C}_{k,m}\mathfrak{C}_{k',m}\Big] \\ &= \mathbb{E}\Big[\big(\alpha_{k,m} - b/n\big)\mathbb{E}_{J_{k'}}\big[\alpha_{k',m} - b/n\big]\mathfrak{C}_{k,m}\mathfrak{C}_{k',m}\Big] = 0,\end{aligned} \tag{6.13}$$

where we have used the fact that $\alpha_{k,m}, \mathfrak{C}_{k,m}$ and $\mathfrak{C}_{k',m}$ are independent of $J_{k'}$. It then follows from Eq. (6.9) that

$$\begin{aligned}\mathbb{E}\big[\Delta_{t+1,m}^2\big] &\leq \frac{2}{b^2}\mathbb{E}\Big[\sum_{k=1}^t \eta_k^2\big(\alpha_{k,m}-b/n\big)^2\mathfrak{C}_{k,m}^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \mathfrak{C}_{k,m}\Big)^2\Big]\\ &= \frac{2}{b^2}\mathbb{E}\Big[\sum_{k=1}^t \eta_k^2 \mathrm{Var}(\alpha_{k,m})\mathfrak{C}_{k,m}^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \mathfrak{C}_{k,m}\Big)^2\Big]\\ &\leq \frac{2}{nb}\mathbb{E}\Big[\sum_{k=1}^t \eta_k^2 \mathfrak{C}_{k,m}^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \mathfrak{C}_{k,m}\Big)^2\Big]\\ &\leq \frac{2}{nb}\mathbb{E}\Big[\sum_{k=1}^t \eta_k^2 \mathfrak{C}_{k,m}^2\Big] + \frac{8}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big],\end{aligned}$$

where we have used the following inequality in the last step

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \mathfrak{C}_{k,m}\Big)^2\Big] &\leq 2\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big] + 2\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k^{(m)};z_m')\|_2\Big)^2\Big]\\ &= 4\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big]. \end{aligned} \tag{6.14}$$

Analogous to Eq. (6.10), we have

$$\begin{aligned}\mathbb{E}[\mathfrak{C}_{k,m}^2] &\leq 2\mathbb{E}[\|\nabla f(\mathbf{w}_k;z_m)\|_2^2] + 2\mathbb{E}[\|\nabla f(\mathbf{w}_k^{(m)};z_m')\|_2^2]\\ &\leq 4L\mathbb{E}\big[f(\mathbf{w}_k;z_m)+f(\mathbf{w}_k^{(m)};z_m')\big] = 8L\mathbb{E}\big[f(\mathbf{w}_k;z_m)\big]. \end{aligned} \tag{6.15}$$

It then follows that

$$\mathbb{E}\big[\Delta_{t+1,m}^2\big] \leq \frac{16L}{nb}\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[f(\mathbf{w}_k;z_m)\big] + \frac{8}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big].$$

We take an average over all $m \in [n]$ and get

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^n \mathbb{E}\big[\Delta_{t+1,m}^2\big] &\leq \frac{16L}{n^2 b}\sum_{k=1}^t\sum_{m=1}^n \eta_k^2 \mathbb{E}\big[f(\mathbf{w}_k;z_m)\big] + \frac{8}{n^3}\sum_{m=1}^n \mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big]\\ &= \frac{16L}{nb}\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[F_S(\mathbf{w}_k)\big] + \frac{8}{n^3}\sum_{m=1}^n \mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k \|\nabla f(\mathbf{w}_k;z_m)\|_2\Big)^2\Big].\end{aligned}$$

The proof is completed. □

Remark 8 (Novelty in the analysis)**.** Similar stability bounds involving $F_S(\mathbf{w}_k)$ were developed for the vanilla SGD [25]. Their argument needs to distinguish two cases according to whether the algorithm chooses a particular example at each iteration. This argument does not work for the minibatch SGD since we draw b examples per iteration and we can draw the particular example several times. We bypass this difficulty by introducing the *expectation-variance decomposition* and self-bounding property [24, 25] into the stability analysis based on a reformulation of minibatch SGD with binomial variables. Indeed, the paper [25] considers SGD with $\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t - \eta_t \nabla f(\tilde{\mathbf{w}}_t; z_{i_t})$. Their discussion controls $\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_{t+1}^{(m)}\|_2^2$ by considering two cases: $i_t = m$ or $i_t \neq m$. If $i_t = m$, they use

$$\|\mathbf{v}_1 + \mathbf{v}_2\|_2^2 \leq (1+p)\|\mathbf{v}_1\|_2^2 + (1+1/p)\|\mathbf{v}_2\|_2^2 \tag{6.16}$$

and get $\|\tilde{\mathbf{w}}_{t+1}-\tilde{\mathbf{w}}_{t+1}^{(m)}\|_2^2 \le (1+p)\|\tilde{\mathbf{w}}_t-\tilde{\mathbf{w}}_t^{(m)}\|_2^2+(1+1/p)\eta_t^2\mathfrak{C}_{t,m}^2$. Since $i_t = m$ happens with probability $1/n$, they derive

$$\mathbb{E}[\|\tilde{\mathbf{w}}_{t+1} - \tilde{\mathbf{w}}_{t+1}^{(m)}\|_2^2] \le (1 + p/n)\mathbb{E}[\|\tilde{\mathbf{w}}_t - \tilde{\mathbf{w}}_t^{(m)}\|_2^2] + O\Big(\big(\frac{1}{n} + \frac{1}{np}\big)\eta_t^2\mathbb{E}[\mathfrak{C}_{t,m}^2]\Big). \tag{6.17}$$

For minibatch SGD, we may select i_m several times and cannot divide the discussions into two cases as in [25]. Instead, we reformulate SGD as Eq. (6.5) with $\alpha_{t,k}$ being a binomial random variable. Furthermore, even with the formulation, the existing techniques [25] would imply suboptimal bounds. Indeed, applying (6.16) to Eq. (6.6) would imply

$$\begin{aligned}\mathbb{E}[\Delta_{t+1,m}^2] &\le (1 + p)\mathbb{E}[\Delta_{t,m}^2] + \eta_t^2 b^{-2}(1 + 1/p)\mathbb{E}[\alpha_{t,m}^2\mathfrak{C}_{t,m}^2]\\ &\le (1 + p)\mathbb{E}[\Delta_{t,m}^2] + 2\eta_t^2 b^{-1}n^{-1}(1 + 1/p)\mathbb{E}[\mathfrak{C}_{t,m}^2],\end{aligned} \tag{6.18}$$

where we have used $\mathbb{E}_{J_t}[\alpha_{t,m}^2] \le 2b/n$. The key difference is we have a factor of $1 + p/n$ for SGD and $1 + p$ for minibatch SGD. To see how Eq. (6.18) implies sub-optimal bounds, we continue the deduction as follows. We apply Eq. (6.18) recursively and get

$$\begin{aligned}\mathbb{E}[\Delta_{t+1,m}^2] &\le 2b^{-1}n^{-1}(1 + 1/p)\sum_{k=1}^{t}(1 + p)^{t+1-k}\eta_k^2\mathbb{E}[\mathfrak{C}_{k,m}^2] \le 2b^{-1}n^{-1}(1 + 1/p)(1 + p)^t\sum_{k=1}^{t}\eta_k^2\mathbb{E}[\mathfrak{C}_{k,m}^2]\\ &\le 2b^{-1}n^{-1}(1 + t)e\sum_{k=1}^{t}\eta_k^2\mathbb{E}[\mathfrak{C}_{k,m}^2] \le 16Lb^{-1}n^{-1}(1 + t)e\sum_{k=1}^{t}\eta_k^2\mathbb{E}[f(\mathbf{w}_k; z_m)],\end{aligned}$$

where we choose $p = 1/t$ and use $(1 + 1/t)^t \le e$ in the last second inequality, and use Eq. (6.15) in the last inequality. An average over all $m \in [n]$ implies

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\Delta_{t+1,m}^2] \le \frac{16L(1 + t)e\eta^2}{nb}\sum_{k=1}^{t}\mathbb{E}[F_S(\mathbf{w}_k)], \tag{6.19}$$

which is much worse than Eq. (4.3). Indeed, if $\mathbb{E}[F_S(\mathbf{w}_k)] \lesssim 1$, then Eq. (6.19) implies $\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\Delta_{t+1,m}^2] \lesssim t^2\eta^2/(nb)$. As a comparison, Eq. (4.3) implies $\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\Delta_{t+1,m}^2] \lesssim t\eta^2/(nb)+t^2\eta^2/n^2$. Note $t\eta^2/(nb)$ outperforms $t^2\eta^2/(nb)$ by a factor of t, and $t^2\eta^2/n^2$ outperforms $t^2\eta^2/(nb)$ by a factor of n/b.

We significantly improve the analysis in [25] by introducing new techniques in the analysis with ℓ_2 on-average model stability. Our idea is to use an expectation-variance decomposition $\Delta_{t+1,m} \le \frac{1}{b}\sum_{k=1}^{t}\eta_k\big(\alpha_{k,m}-b/n\big)\mathfrak{C}_{k,m}+\frac{1}{n}\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}$. The key observation is that $\mathbb{E}\big[\big(\alpha_{k,m}-b/n\big)\mathfrak{C}_{k,m}\big(\alpha_{k',m}-b/n\big)\mathfrak{C}_{k',m}\big] = 0$ if $k \ne k'$. This removes the cross-over terms when taking a square followed by an expectation, and implies

$$\mathbb{E}\big[\Delta_{t+1,m}^2\big] \le \frac{2}{b^2}\mathbb{E}\Big[\sum_{k=1}^{t}\eta_k^2\big(\alpha_{k,m} - b/n\big)^2\mathfrak{C}_{k,m}^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}\Big)^2\Big].$$

It is also possible to derive high-order stability bounds under a Lipschitzness assumption. We omit the discussions for simplicity.

Remark 9 (Lower bounds)**.** Recently, lower bounds on the uniform stability have also received increasing attention. Let ϵ_{unif} be the uniform stability of SGD with t iterations and a constant step size η. For nonsmooth and Lipschitz loss functions, it was shown $\epsilon_{\text{unif}} \gtrsim \min\{1, t/n\}\eta\sqrt{t} + \eta t/n$ for convex problems [3], $\epsilon_{\text{unif}} \gtrsim 1/\mu\sqrt{n}$ for μ-strongly convex problems ($\mu \geq 1/\sqrt{n}$) and $\epsilon_{\text{unif}} \gtrsim \eta^2 n$ for nonconvex problems ($\eta \leq 1/\sqrt{n}$) [23]. For smooth loss functions, it was shown $\epsilon_{\text{unif}} \gtrsim \eta t/n$ for convex and Lipschitz problems, and $\epsilon_{\text{unif}} \gtrsim 1/(\mu n)$ for μ-strongly convex problems [51]. It is clear that our on-average stability bounds in Eq. (4.2) match the existing lower bounds on uniform stability in the convex and smooth case.

Finally, we give some direct corollaries of Theorem 2. By the Cauchy-Schwarz's inequality $(\sum_{k=1}^t a_k)^2 \leq t\sum_{k=1}^t a_k^2$, Eq. (4.3) further implies

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^n \mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2^2\big] &\leq \frac{16L}{nb}\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[F_S(\mathbf{w}_k)\big] + \frac{8t}{n^3}\sum_{m=1}^n\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[\|\nabla f(\mathbf{w}_k; z_m)\|_2^2\big] \\ &\leq \Big(\frac{16L}{nb} + \frac{16Lt}{n^2}\Big)\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[F_S(\mathbf{w}_k)\big], \end{aligned} \tag{6.20}$$

where we use $\|\nabla f(\mathbf{w}_k; z_m)\|_2^2 \leq 2Lf(\mathbf{w}_k; z_m)$ due to the self-bounding property. If $b = 1$, our analysis implies stability bounds of order $L\big(\frac{1}{n} + \frac{t}{n^2}\big)\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[F_S(\mathbf{w}_k)\big]$, which match the stability bounds for SGD [25]. Furthermore, under a stronger self-bounding property $\|\nabla f(\mathbf{w}_k; z_m)\|_2 \leq f(\mathbf{w}_k; z_m)$ (e.g., logistic loss) [36], Eq. (4.3) implies

$$\frac{1}{n}\sum_{m=1}^n \mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(m)}\|_2^2\big] \leq \frac{16L}{nb}\sum_{k=1}^t \eta_k^2 \mathbb{E}\big[F_S(\mathbf{w}_k)\big] + \frac{8}{n^3}\sum_{m=1}^n \mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k f(\mathbf{w}_k; z_m)\Big)^2\Big]. \tag{6.21}$$

6.2. Proof of Theorem 3

In this section, we present the proof of Theorem 3 on excess population risk bounds of minibatch SGD. We first introduce the following optimization error bounds.

Lemma 15 (Optimization Errors of Minibatch SGD: Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_t\}$ be produced by Eq. (4.1) with $\eta \leq 1/L$. Then the following inequality holds for all $\mathbf{w}$*

$$\frac{1}{R}\sum_{t=1}^R \mathbb{E}_A\big[F_S(\mathbf{w}_t)\big] - F_S(\mathbf{w}) \leq \frac{2\eta L}{bR}\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)] + \frac{\|\mathbf{w}\|_2^2}{2\eta R} + \frac{F_S(\mathbf{w}_1)}{R}. \tag{6.22}$$

Proof. Denote $B_t = \{z_{i_{t,1}}, \ldots, z_{i_{t,b}}\}$ and $f(\mathbf{w}; B_t) = \frac{1}{b}\sum_{j=1}^b f(\mathbf{w}; z_{i_{t,j}})$. Then the update of minibatch SGD can be written as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla f(\mathbf{w}_t; B_t).$$

Since $\mathbb{E}_{B_t}[f(\mathbf{w}_t;B_t)] = F_S(\mathbf{w}_t)$ we know

$$\begin{aligned}\mathbb{E}_A[\|\nabla f(\mathbf{w}_t;B_t)\|_2^2] &= \mathbb{E}_A[\|\nabla f(\mathbf{w}_t;B_t) - \nabla F_S(\mathbf{w}_t)\|_2^2] + \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] \\ &= \frac{1}{b}\mathbb{E}_A[\|\nabla f(\mathbf{w}_t;z_{i_{t,1}}) - \nabla F_S(\mathbf{w}_t)\|_2^2] + \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] \\ &= \frac{\mathbb{E}_A[\|\nabla f(\mathbf{w}_t;z_{i_{t,1}})\|_2^2]}{b} - \frac{\mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2]}{b} + \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] \\ &\le \frac{2L\mathbb{E}_A[f(\mathbf{w}_t;z_{i_{t,1}})]}{b} - \frac{\mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2]}{b} + \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] \\ &\le \frac{2L\mathbb{E}_A[F_S(\mathbf{w}_t)]}{b} + \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2], \end{aligned} \tag{6.23}$$

where we have used the self-bounding property of smooth functions. Furthermore, by the convexity of f we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta^2\|\nabla f(\mathbf{w}_t;B_t)\|_2^2 + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t;B_t)\rangle \\ &\le \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta^2\|\nabla f(\mathbf{w}_t;B_t)\|_2^2 + 2\eta(f(\mathbf{w};B_t) - f(\mathbf{w}_t;B_t)).\end{aligned}$$

It then follows that

$$\mathbb{E}_A[\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2] \le \mathbb{E}_A[\|\mathbf{w}_t - \mathbf{w}\|_2^2] + \frac{2L\eta^2\mathbb{E}_A[F_S(\mathbf{w}_t)]}{b} + \eta^2\mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] + 2\eta\mathbb{E}_A[F_S(\mathbf{w}) - F_S(\mathbf{w}_t)].$$

Taking a summation of the above inequality gives ($\mathbf{w}_1 = 0$)

$$2\eta\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t) - F_S(\mathbf{w})] \le \|\mathbf{w}\|_2^2 + \frac{2L\eta^2}{b}\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)] + \eta^2\sum_{t=1}^R \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2]. \tag{6.24}$$

By the L-smoothness of F_S and Eq. (6.23) we have

$$\begin{aligned}\mathbb{E}_A[F_S(\mathbf{w}_{t+1})] &\le \mathbb{E}_A[F_S(\mathbf{w}_t)] + \mathbb{E}_A[\langle \nabla F_S(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle] + \frac{L\mathbb{E}_A[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2]}{2} \\ &= \mathbb{E}_A[F_S(\mathbf{w}_t)] - \eta\mathbb{E}_A[\langle \nabla F_S(\mathbf{w}_t), \nabla f(\mathbf{w}_t;B_t)\rangle] + \frac{L\eta^2\mathbb{E}_A[\|\nabla f(\mathbf{w}_t;B_t)\|_2^2]}{2} \\ &\le \mathbb{E}_A[F_S(\mathbf{w}_t)] - \eta\mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] + \frac{L^2\eta^2\mathbb{E}_A[F_S(\mathbf{w}_t)]}{b} + \frac{L\eta^2\mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2]}{2}.\end{aligned}$$

It then follows from $\eta \le 1/L$ that

$$\frac{\eta}{2}\sum_{t=1}^R \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2] \le \mathbb{E}_A[F_S(\mathbf{w}_1)] + \frac{L^2\eta^2\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)]}{b}.$$

We combine the above inequality and Eq. (6.24) to derive (note $\eta \le 1/L$)

$$\begin{aligned}2\eta\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t) - F_S(\mathbf{w})] &\le \|\mathbf{w}\|_2^2 + \frac{2L\eta^2}{b}\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)] + 2\eta F_S(\mathbf{w}_1) + \frac{2L^2\eta^3\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)]}{b} \\ &\le \|\mathbf{w}\|_2^2 + \frac{4L\eta^2}{b}\sum_{t=1}^R \mathbb{E}_A[F_S(\mathbf{w}_t)] + 2\eta F_S(\mathbf{w}_1).\end{aligned}$$

The proof is completed. $\square$

Proof of Theorem 3. We choose $\mathbf{w} = \mathbf{w}^*$ and take expectations w.r.t. S over both sides of Eq. (6.22) to get

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\big[F_S(\mathbf{w}_t)\big] - F(\mathbf{w}^*) \leq \frac{2\eta L}{bR}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t)] + \frac{\|\mathbf{w}^*\|_2^2}{2\eta R} + \frac{F(\mathbf{w}_1)}{R}. \tag{6.25}$$

We consider two cases. If $\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t)] \leq F(\mathbf{w}^*)$, then this means that the optimization error in Eq. (3.1) is non-positive (this is the easier case since one does not need to consider the optimization error)

$$\mathbb{E}[F_S(\bar{\mathbf{w}}_R)] \leq \frac{1}{R}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t)] \leq F(\mathbf{w}^*) = \mathbb{E}[F_S(\mathbf{w}^*)].$$

We now consider the case $\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t)] \geq F(\mathbf{w}^*)$. Then it follows from Eq. (6.25) that

$$\begin{aligned}\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\big[F_S(\mathbf{w}_t)\big] - F(\mathbf{w}^*) &\leq \frac{2\eta L}{bR}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t) - F(\mathbf{w}^*)] + \frac{2\eta L}{bR}\sum_{t=1}^{R}F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{2\eta R} + \frac{F(\mathbf{w}_1)}{R}\\ &\leq \frac{1}{2R}\sum_{t=1}^{R}\mathbb{E}[F_S(\mathbf{w}_t) - F(\mathbf{w}^*)] + \frac{2\eta L}{bR}\sum_{t=1}^{R}F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{2\eta R} + \frac{F(\mathbf{w}_1)}{R},\end{aligned}$$

where we have used $\eta \leq b/(4L)$ due to $b \geq 2$. It then follows that

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\big[F_S(\mathbf{w}_t)\big] - F(\mathbf{w}^*) \leq \frac{4\eta LF(\mathbf{w}^*)}{b} + \frac{\|\mathbf{w}^*\|_2^2}{\eta R} + \frac{2F(\mathbf{w}_1)}{R}. \tag{6.26}$$

By Lemma 1 (Part (b)) and Eq. (6.20), we know

$$\mathbb{E}[F(\bar{\mathbf{w}}_R) - F_S(\bar{\mathbf{w}}_R)] \leq \frac{L}{\gamma}\mathbb{E}[F_S(\bar{\mathbf{w}}_R)] + (L+\gamma)\Big(\frac{8L}{nb} + \frac{8LR}{n^2}\Big)\sum_{t=1}^{R}\eta_t^2\mathbb{E}\big[F_S(\mathbf{w}_t)\big].$$

Eq. (6.26) implies that

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\big[F_S(\mathbf{w}_t)\big] \lesssim F(\mathbf{w}^*) + \|\mathbf{w}^*\|_2^2/(\eta R).$$

We combine the above two inequalities together and derive (note $F_S(\bar{\mathbf{w}}_R) \leq \frac{1}{R}\sum_{t=1}^{R}F_S(\mathbf{w}_t)$)

$$\mathbb{E}[F(\bar{\mathbf{w}}_R) - F_S(\bar{\mathbf{w}}_R)] \lesssim L\Big(\frac{F(\mathbf{w}^*) + \|\mathbf{w}^*\|_2^2/(\eta R)}{\gamma}\Big) + L(L+\gamma)\eta^2\Big(\frac{1}{nb} + \frac{R}{n^2}\Big)\big(RF(\mathbf{w}^*) + \|\mathbf{w}^*\|_2^2/\eta\big).$$

We plug the above generalization error bounds, the optimization error bounds in Eq. (6.26) back into Eq. (3.1), and derive

$$\begin{aligned}\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{\eta LF(\mathbf{w}^*)}{b} + \frac{\|\mathbf{w}^*\|_2^2}{\eta R} + \frac{LF(\mathbf{w}^*) + L\|\mathbf{w}^*\|_2^2/(\eta R)}{\gamma}&\\ + L(L+\gamma)\eta^2\Big(\frac{1}{nb} + \frac{R}{n^2}\Big)\big(RF(\mathbf{w}^*) + \|\mathbf{w}^*\|_2^2/\eta\big)&.\end{aligned}$$

The proof is completed. □

Proof of Corollary 4. We first consider the case that $F(\mathbf{w}^*) \geq 4Lb^2\|\mathbf{w}^*\|_2^2/n$. In this case, we have $\frac{\|\mathbf{w}^*\|_2 b}{\sqrt{LnF(\mathbf{w}^*)}} \leq \frac{1}{2L}$ and therefore $\eta = \frac{\|\mathbf{w}^*\|_2 b}{\sqrt{LnF(\mathbf{w}^*)}}$. We have

$$\eta R \asymp \frac{\|\mathbf{w}^*\|_2 b}{\sqrt{LnF(\mathbf{w}^*)}}\frac{n}{b} = \frac{\sqrt{n}\|\mathbf{w}^*\|_2}{\sqrt{LF(\mathbf{w}^*)}} \tag{6.27}$$

and therefore

$$F(\mathbf{w}^*)\eta R \asymp \frac{\sqrt{nF(\mathbf{w}^*)}\|\mathbf{w}^*\|_2}{\sqrt{L}} \geq \frac{2\sqrt{L}b\|\mathbf{w}^*\|_2^2}{\sqrt{L}} = 2b\|\mathbf{w}^*\|_2^2.$$

Theorem 3 together with $R \asymp n/b$ then implies

$$\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{\eta LF(\mathbf{w}^*)}{b} + \frac{\|\mathbf{w}^*\|_2^2}{\eta R} + LF(\mathbf{w}^*)\Big(\frac{1}{\gamma} + (L+\gamma)\eta^2\frac{R^2}{n^2}\Big).$$

Since $\eta = \frac{\|\mathbf{w}^*\|_2 b}{\sqrt{nLF(\mathbf{w}^*)}}$, $R \asymp \frac{n}{b}$ and $\gamma = \sqrt{LnF(\mathbf{w}^*)}/\|\mathbf{w}^*\|_2$, we know

$$\frac{\eta LF(\mathbf{w}^*)}{b} \asymp \frac{Lb\|\mathbf{w}^*\|}{\sqrt{LnF(\mathbf{w}^*)}}\frac{F(\mathbf{w}^*)}{b} \asymp \frac{\|\mathbf{w}^*\|_2(LF(\mathbf{w}^*))^{\frac{1}{2}}}{\sqrt{n}},$$
$$\frac{LF(\mathbf{w}^*)}{\gamma} \asymp \frac{LF(\mathbf{w}^*)\|\mathbf{w}^*\|_2}{\sqrt{LnF(\mathbf{w}^*)}} \asymp \frac{(LF(\mathbf{w}^*))^{\frac{1}{2}}\|\mathbf{w}^*\|_2}{\sqrt{n}}$$

and

$$\frac{L(L+\gamma)\eta^2R^2F(\mathbf{w}^*)}{n^2} \asymp \frac{L(L+(LnF(\mathbf{w}^*))^{\frac{1}{2}}/\|\mathbf{w}^*\|_2)\|\mathbf{w}^*\|_2^2b^2R^2F(\mathbf{w}^*)}{n^2LnF(\mathbf{w}^*)}$$
$$\asymp \frac{(L+(LnF(\mathbf{w}^*))^{\frac{1}{2}}/\|\mathbf{w}^*\|_2)\|\mathbf{w}^*\|_2^2}{n} \lesssim \frac{(LF(\mathbf{w}^*))^{\frac{1}{2}}\|\mathbf{w}^*\|_2}{\sqrt{n}}.$$

We plug the above inequalities back into Eq. (6.27) and get $\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{(LF(\mathbf{w}^*))^{\frac{1}{2}}\|\mathbf{w}^*\|_2}{\sqrt{n}}$.

We now consider the case $F(\mathbf{w}^*) \leq 4Lb^2\|\mathbf{w}^*\|_2^2/n$. In this case, we have $\eta = 1/(2L)$, $R \asymp n$ and choose $\gamma \asymp L$. Theorem 3 implies

$$\mathbb{E}[F(\bar{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{F(\mathbf{w}^*)}{b} + \frac{L\|\mathbf{w}^*\|_2^2}{n} + L\Big(F(\mathbf{w}^*) + \frac{L\|\mathbf{w}^*\|_2^2}{n}\Big)\big(L^{-1} + LL^{-2}\big) \lesssim F(\mathbf{w}^*) + \frac{L\|\mathbf{w}^*\|_2^2}{n}.$$

The proof is completed. □

6.3. Proof of Theorem 5 and Theorem 6

In this section, we prove stability and risk bounds for minibatch SGD applied to strongly convex problems.

Proof of Theorem 5. For simplicity, we assume $f(\mathbf{w}; z) = g(\mathbf{w}; z) + r(\mathbf{w})$ with $r : \mathcal{W} \mapsto \mathbb{R}^+$ being μ-strongly convex and $g : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ being convex (this is a typical form for strongly convex problems in machine learning). According to Eq. (6.5) and the sub-additivity of $\|\cdot\|_2$, we know

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2 \leq \frac{\eta_t\alpha_{t,m}}{b}\|\nabla g(\mathbf{w}_t; z_m) - \nabla g(\mathbf{w}_t^{(m)}; z'_m)\|_2 +$$
$$\Big\|\mathbf{w}_t - \frac{\eta_t}{b}\sum_{k:k\neq m}\alpha_{t,k}\nabla g(\mathbf{w}_t; z_k) - \eta_t\nabla r(\mathbf{w}_t) - \mathbf{w}_t^{(m)} + \frac{\eta_t}{b}\sum_{k:k\neq m}\alpha_{t,k}\nabla g(\mathbf{w}_t^{(m)}; z_k) + \eta_t\nabla r(\mathbf{w}_t^{(m)})\Big\|_2.$$

Since f is L-smooth and $\sum_{k:k\neq m}\alpha_{t,k} \leq b$, we know the function $\mathbf{w} \mapsto \frac{1}{b}\sum_{k:k\neq m}\alpha_{t,k}f(\mathbf{w}; z_k) + r(\mathbf{w})$ is L-smooth and μ-strongly convex. By Lemma 14 and the assumption $\eta_t \leq 1/L$, we know

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2 \leq (1 - \mu\eta_t/2)\|\mathbf{w}_t - \mathbf{w}_t^{(m)}\|_2 + \frac{\eta_t\alpha_{t,m}}{b}\|\nabla g(\mathbf{w}_t; z_m) - \nabla g(\mathbf{w}_t^{(m)}; z'_m)\|_2. \quad (6.28)$$

Taking an expectation over both sides yields (note $\mathbf{w}_t, \mathbf{w}_t^{(m)}$ are independent of J_t)

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] \leq (1 - \mu\eta_t/2)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(m)}\|_2] + \frac{2\eta_t\sqrt{2L\mathbb{E}[f(\mathbf{w}_t; z_m)]}}{n},$$

where we have used Eq. (6.10) and Eq. (6.9). It then follows that

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] \leq \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t}\eta_k\sqrt{\mathbb{E}[f(\mathbf{w}_k; z_m)]}\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2).$$

We take an average over m to derive

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] &\leq \frac{2\sqrt{2L}}{n^2}\sum_{k=1}^{t}\sum_{m=1}^{n}\eta_k\sqrt{\mathbb{E}[f(\mathbf{w}_k; z_m)]}\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\\ &\leq \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t}\eta_k\Big(\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}[f(\mathbf{w}_k; z_m)]\Big)^{\frac{1}{2}}\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\\ &= \frac{2\sqrt{2L}}{n}\sum_{k=1}^{t}\eta_k\sqrt{\mathbb{E}[F_S(\mathbf{w}_k)]}\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2),\end{aligned}$$

where we have used the concavity of $x \mapsto \sqrt{x}$. This proves Eq. (4.4).

We now turn to Eq. (4.5). Let $\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. The following inequality was established in [45]

$$\mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}_S\|_2^2] \leq (1 - \mu\eta_k)\mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_S\|_2^2] - \eta_k\mathbb{E}[F_S(\mathbf{w}_k) - F_S(\mathbf{w}_S)] + \frac{2\eta_k^2\sigma_S^2}{b},$$

where $\sigma_S^2 = \mathbb{E}_{i_t}[\|\nabla f(\mathbf{w}_S; z_{i_t}) - \nabla F_S(\mathbf{w}_S)\|_2^2]$. We multiply both sides by $\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)$ and derive

$$\begin{aligned}\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\mathbb{E}[\|\mathbf{w}_{k+1} - \mathbf{w}_S\|_2^2] \leq \prod_{k'=k}^{t}(1 - \mu\eta_{k'}/2)\mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_S\|_2^2] -\\ \eta_k\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\mathbb{E}[F_S(\mathbf{w}_k) - F_S(\mathbf{w}_S)] + \frac{2\sigma_S^2\eta_k^2\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)}{b}.\end{aligned}$$

We take a summation of the above inequality and derive

$$\begin{aligned}\sum_{k=1}^{t}\eta_k\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\mathbb{E}[F_S(\mathbf{w}_k) - F_S(\mathbf{w}_S)] \leq \mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_S\|_2^2]\prod_{k'=1}^{t}(1 - \mu\eta_{k'}/2) +\\ \frac{2\sigma_S^2}{b}\sum_{k=1}^{t}\eta_k^2\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2). \qquad (6.29)\end{aligned}$$

There holds

$$\begin{aligned}\frac{\mu}{2}\sum_{k=1}^{t}\eta_k\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2) &= \sum_{k=1}^{t}\big(1 - (1 - \mu\eta_k/2)\big)\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2)\\ &= \sum_{k=1}^{t}\Big(\prod_{k'=k+1}^{t}(1 - \mu\eta_{k'}/2) - \prod_{k'=k}^{t}(1 - \mu\eta_{k'}/2)\Big)\\ &= 1 - \prod_{k'=1}^{t}(1 - \mu\eta_{k'}/2) \leq 1. \qquad (6.30)\end{aligned}$$

By the strong convexity of F_S and $\nabla F_S(\mathbf{w}_S) = 0$, we know

$$F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S) = F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S) - \langle \mathbf{w}_1 - \mathbf{w}_S, \nabla F_S(\mathbf{w}_S)\rangle \geq \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_S\|_2^2$$

and therefore

$$\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_S\|_2^2] \leq \frac{2}{\mu}\mathbb{E}[F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S)] \lesssim 1/\mu.$$

We can plug the above inequality and Eq. (6.30) back into Eq. (6.29) to derive (note $\eta_k \leq 1/L$ and $\eta_k\mu \leq \mu/L \leq 1$)

$$\sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\mathbb{E}[F_S(\mathbf{w}_k)-F_S(\mathbf{w}_S)] \leq \mathbb{E}[\|\mathbf{w}_1-\mathbf{w}_S\|_2^2]+\frac{2\sigma_S^2}{bL}\sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2) \lesssim 1/\mu.$$

We combine the above inequality and Eq. (6.30) together and derive

$$\begin{aligned}\sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\mathbb{E}[F_S(\mathbf{w}_k)] &= \mathbb{E}[F_S(\mathbf{w}_S)]\sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2) \\ &\quad + \sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\mathbb{E}[F_S(\mathbf{w}_k) - F_S(\mathbf{w}_S)] \lesssim 1/\mu. \quad (6.31)\end{aligned}$$

This together with Eq. (6.30) implies that

$$\sum_{k=1}^t \eta_k \sqrt{\mathbb{E}[F_S(\mathbf{w}_k)]} \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2) \leq \frac{1}{2}\sum_{k=1}^t \eta_k \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\big(1+\mathbb{E}[F_S(\mathbf{w}_k)]\big) \lesssim 1/\mu.$$

We plug the above inequality back into Eq. (4.4) to derive Eq. (4.5).

Finally, we prove Eq. (4.6). Recall the notations in Eq. (6.7). Then, Eq. (6.28) implies $\Delta_{t+1,m} \leq (1-\mu\eta_t/2)\Delta_{t,m} + \eta_t\alpha_{t,m}\mathfrak{C}_{t,m}/b$. We apply this inequality recursively, and get

$$\begin{aligned}\Delta_{t+1,m} &\leq \frac{1}{b}\sum_{k=1}^t \eta_k\alpha_{k,m}\mathfrak{C}_{k,m} \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2) \\ &= \frac{1}{b}\sum_{k=1}^t \eta_k\big(\alpha_{k,m} - b/n\big)\mathfrak{C}_{k,m} \prod_{k'=k+1}^t (1-\mu\eta_{k'}/2) + \frac{1}{n}\sum_{k=1}^t \eta_k\mathfrak{C}_{k,m}\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2).\end{aligned}$$

We take a square and an expectation over both sides, and get

$$\begin{aligned}&\mathbb{E}[\Delta_{t+1,m}^2] \\ &\leq \frac{2}{b^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k\big(\alpha_{k,m}-b/n\big)\mathfrak{C}_{k,m}\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\Big)^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k\mathfrak{C}_{k,m}\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\Big)^2\Big] \\ &= \frac{2}{b^2}\sum_{k=1}^t \eta_k^2\mathbb{E}\Big[\big(\alpha_{k,m}-b/n\big)^2\mathfrak{C}_{k,m}^2\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)^2\Big] + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k\mathfrak{C}_{k,m}\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\Big)^2\Big] \\ &\leq \frac{2}{nb}\sum_{k=1}^t \eta_k^2\mathbb{E}\big[\mathfrak{C}_{k,m}^2\big]\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)^2 + \frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{k=1}^t \eta_k\mathfrak{C}_{k,m}\prod_{k'=k+1}^t (1-\mu\eta_{k'}/2)\Big)^2\Big],\end{aligned}$$

where we have used Eq. (6.13) and $\mathbb{E}_{J_k}\big[\big(\alpha_{k,m}-b/n\big)^2\big] = \mathrm{Var}(\alpha_{k,m}) \leq b/n$. Furthermore, by the

Schwarz's inequality and Eq. (6.30), we know

$$\Big(\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2)\Big)^2 \leq \Big(\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}^2\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2)\Big)\Big(\sum_{k=1}^{t}\eta_k\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2)\Big)$$
$$\leq \frac{2}{\mu}\sum_{k=1}^{t}\eta_k\mathfrak{C}_{k,m}^2\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2).$$

We can combine the above two inequalities together and derive

$$\mathbb{E}[\Delta_{t+1,m}^2] \leq \sum_{k=1}^{t}\Big(\frac{2\eta_k^2}{nb}+\frac{4\eta_k}{n^2\mu}\Big)\mathbb{E}[\mathfrak{C}_{k,m}^2]\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2).$$

Analogous to Eq. (6.10), we know $\mathbb{E}[\mathfrak{C}_{k,m}^2] \leq 8L\mathbb{E}[f(\mathbf{w}_k;z_m)]$ and therefore

$$\mathbb{E}[\Delta_{t+1,m}^2] \leq \sum_{k=1}^{t}\Big(\frac{16L\eta_k^2}{nb}+\frac{32L\eta_k}{n^2\mu}\Big)\mathbb{E}[f(\mathbf{w}_k;z_m)]\prod_{k'=k+1}^{t}(1-\mu\eta_{k'}/2).$$

We can take an average over $m \in [n]$ to get the stated bound. The proof is completed.

□

Proof of Theorem 6. Since $F_S(\mathbf{w}_S) \leq F_S(\mathbf{w}^*)$, an upper bound on $F_S(A(S)) - F_S(\mathbf{w}_S)$ is also an upper bound on $F_S(A(S)) - F_S(\mathbf{w}^*)$. Then, according to [45], there exists an average $\hat{\mathbf{w}}_R$ of $\{\mathbf{w}_t\}$ such that

$$\mathbb{E}_A[F_S(\hat{\mathbf{w}}_R)] - F_S(\mathbf{w}^*) \lesssim \frac{L}{\mu}\exp\big(-\mu R/L\big) + \frac{\sigma_*^2}{\mu bR}. \tag{6.32}$$

Theorem 5 shows that an algorithm outputting any iterate produced by Eq. (4.1) would be ℓ_1 on-average model $O(1/(n\mu))$-stable. It then follows that the output model $\hat{\mathbf{w}}_R$ would also be ℓ_1 on-average model $O(1/(n\mu))$-stable. Lemma 1 (Part (a)) then implies

$$\mathbb{E}[F(\hat{\mathbf{w}}_R) - F_S(\hat{\mathbf{w}}_R)] \lesssim G/(n\mu).$$

We plug the above two inequalities back to Eq. (3.1) and derive

$$\mathbb{E}[F(\hat{\mathbf{w}}_R)] - F(\mathbf{w}^*) \lesssim \frac{L}{\mu}\exp\big(-\mu R/L\big) + \frac{\sigma_*^2}{\mu bR} + \frac{G}{n\mu}.$$

If we choose $R > \frac{L}{\mu}\log\frac{nL}{G}$ and $b > \frac{n\sigma_*^2}{GR}$, we get

$$\frac{L}{\mu}\exp\big(-\mu R/L\big) \lesssim G/n\mu \quad\text{and}\quad \frac{\sigma_*^2}{\mu bR} \lesssim G/n\mu.$$

The proof is completed. □

6.4. Proof of Theorem 7 and Theorem 8

In this section, we present the proof of minibatch SGD for nonconvex problems. We first prove Theorem 7.

Proof of Theorem 7. According to Eq. (6.6) and the smoothness of f, we know $\Delta_{t+1,m} \leq (1+\eta_t L)\Delta_{t,m} + \frac{\eta_t \alpha_{t,m}\mathfrak{C}_{t,m}}{b}$. We apply the above inequality recursively, and derive

$$\mathbb{E}[\Delta_{t+1,m}] \leq \sum_{k=1}^{t} \frac{\eta_k \mathbb{E}[\alpha_{k,m}\mathfrak{C}_{k,m}]}{b} \prod_{k'=k+1}^{t} (1+\eta_{k'}L) = \sum_{k=1}^{t} \frac{\eta_k \mathbb{E}[\mathfrak{C}_{k,m}]}{n} \prod_{k'=k+1}^{t} (1+\eta_{k'}L).$$

Analogous to Eq. (6.12), we then get

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^{n} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(m)}\|_2] &\leq \frac{2\sqrt{2L}}{n} \sum_{m=1}^{n}\sum_{k=1}^{t} \frac{\eta_k \mathbb{E}\big[\sqrt{f(\mathbf{w}_k; z_m)}\big]}{n} \prod_{k'=k+1}^{t} (1+\eta_{k'}L) \\ &\leq \frac{2\sqrt{2L}}{n} \sum_{k=1}^{t} \eta_k \mathbb{E}\big[\sqrt{F_S(\mathbf{w}_k)}\big] \prod_{k'=k+1}^{t} (1+\eta_{k'}L).\end{aligned}$$

The proof is completed. □

We now prove Theorem 8 on risk bounds of minibatch SGD under the PL condition. We first introduce the following lemma relating generalization to optimization for problems under the PL condition [26].

Lemma 16 (Generalization Bounds under PL Condition)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. Let A be an algorithm. If Assumption 1 holds and $L \leq n\mu/4$, then*

$$\mathbb{E}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{16L\mathbb{E}[F_S(A(S))]}{n\mu} + \frac{L\mathbb{E}\big[F_S(A(S)) - F_S(\mathbf{w}_S)\big]}{2\mu}. \tag{6.33}$$

The following lemma gives the optimization error bounds for minibatch SGD under the PL condition.

Lemma 17 (Optimization Errors for Minibatch SGD: PL condition)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. Let Assumption 1 hold and $\mathbb{E}_{i_k}\big[\|\nabla f(\mathbf{w}_t; z_{i_k}) - \nabla F_S(\mathbf{w}_t)\|_2^2\big] \leq \sigma^2$, where i_k follows from the uniform distribution over $[n]$. Let $\{\mathbf{w}_t\}$ be produced by the algorithm A defined in (4.1) with $\eta_t = 2/(\mu(t+a))$ and $a \geq 4L/\mu$. Then*

$$\mathbb{E}_A[F_S(\mathbf{w}_{R+1})] - F_S(\mathbf{w}_S) \lesssim \frac{L^2}{\mu^2 R^2} + \frac{L\sigma^2}{b\mu^2 R}. \tag{6.34}$$

Proof. Note the assumption $a \geq 4L/\mu$ implies $\eta_t \leq 1/(2L)$. For simplicity, we denote $g_t = \frac{1}{b}\sum_{j=1}^{b} \nabla f(\mathbf{w}_t; z_{i_{t,j}})$. Then Eq. (4.1) becomes $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t g_t$. By the L-smoothness of F_S, we have

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

We take a conditional expectation over both sides and derive

$$\begin{aligned}\mathbb{E}_{J_t}[F_S(\mathbf{w}_{t+1})] &\leq F_S(\mathbf{w}_t) - \eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2 + L\eta_t^2\big(\mathbb{E}_{J_t}[\|g_t - \nabla F_S(\mathbf{w}_t)\|_2^2] + \|\nabla F_S(\mathbf{w}_t)\|_2^2\big) \\ &= F_S(\mathbf{w}_t) - \eta_t\|\nabla F_S(\mathbf{w}_t)\|_2^2/2 + L\eta_t^2\mathbb{E}_{J_t}[\|g_t - \nabla F_S(\mathbf{w}_t)\|_2^2],\end{aligned}$$

where we have used the assumption $\eta_t \le 1/(2L)$. Note the variance reduces by a factor of b with a minibatch, i.e.,

$$\mathbb{E}_{J_t}[\|g_t - \nabla F_S(\mathbf{w}_t)\|_2^2] = \frac{1}{b}\mathbb{E}_{i_k}[\|\nabla f(\mathbf{w}_t; z_{i_k}) - \nabla F_S(\mathbf{w}_t)\|_2^2] \le \frac{\sigma^2}{b}.$$

We combine the above two inequalities together and take an expectation w.r.t. the remaining random variables to get

$$\mathbb{E}_A[F_S(\mathbf{w}_{t+1})] \le \mathbb{E}_A[F_S(\mathbf{w}_t)] - \eta_t \mathbb{E}_A[\|\nabla F_S(\mathbf{w}_t)\|_2^2]/2 + \frac{L\eta_t^2\sigma^2}{b}.$$

We subtract both sides by $F_S(\mathbf{w}_S)$ and use the PL condition to derive

$$\begin{aligned}\mathbb{E}_A[F_S(\mathbf{w}_{t+1})] - F_S(\mathbf{w}_S) &\le \mathbb{E}_A[F_S(\mathbf{w}_t)] - F_S(\mathbf{w}_S) - \mu\eta_t\big(\mathbb{E}_A[F_S(\mathbf{w}_t)] - F_S(\mathbf{w}_S)\big) + \frac{L\eta_t^2\sigma^2}{b} \\ &= (1-\mu\eta_t)\big(\mathbb{E}_A[F_S(\mathbf{w}_t)] - F_S(\mathbf{w}_S)\big) + \frac{L\eta_t^2\sigma^2}{b}.\end{aligned}$$

Since $\eta_t = \frac{2}{\mu(a+t)}$, we know

$$\mathbb{E}_A[F_S(\mathbf{w}_{t+1})] - F_S(\mathbf{w}_S) \le \Big(1 - \frac{2}{a+t}\Big)\big(\mathbb{E}_A[F_S(\mathbf{w}_t)] - F_S(\mathbf{w}_S)\big) + \frac{4L\sigma^2}{b\mu^2(a+t)^2}.$$

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

$$(t+a)(t+a-1)\big(\mathbb{E}_A[F_S(\mathbf{w}_{t+1})] - F_S(\mathbf{w}_S)\big) \le (t+a-1)(t+a-2)\big(\mathbb{E}_A[F_S(\mathbf{w}_t)] - F_S(\mathbf{w}_S)\big) + \frac{4L\sigma^2}{b\mu^2}.$$

We take a summation of the above inequality from $t=1$ to R and get

$$(R+a)(R+a-1)\big(\mathbb{E}_A[F_S(\mathbf{w}_{R+1})] - F_S(\mathbf{w}_S)\big) \le a(a-1)\big(\mathbb{E}_A[F_S(\mathbf{w}_1)] - F_S(\mathbf{w}_S)\big) + \frac{4LR\sigma^2}{b\mu^2}.$$

The stated bound then follows directly since $a \ge 4L/\mu$. The proof is completed. □

Now we are ready to prove Theorem 8 for nonconvex problems.

Proof of Theorem 8. According to Lemma 16 and Lemma 17, we know

$$\mathbb{E}[F(\mathbf{w}_R) - F_S(\mathbf{w}_S)] \lesssim \frac{L}{n\mu} + \frac{L\mathbb{E}[F_S(\mathbf{w}_R) - F_S(\mathbf{w}_S)]}{\mu} \lesssim \frac{L}{n\mu} + \frac{L^3}{\mu^3 R^2} + \frac{L^2\sigma^2}{b\mu^3 R}.$$

Since $F_S(\mathbf{w}_S) \le F_S(\mathbf{w}^*)$, we then derive

$$\mathbb{E}[F(\mathbf{w}_R)] - F(\mathbf{w}^*) = \mathbb{E}[F(\mathbf{w}_R) - F_S(\mathbf{w}^*)] \le \mathbb{E}[F(\mathbf{w}_R) - F_S(\mathbf{w}_S)] \lesssim \frac{L}{n\mu} + \frac{L^3}{\mu^3 R^2} + \frac{L^2\sigma^2}{b\mu^3 R}.$$

Since $R \ge \max\big\{L\sqrt{n}/\mu, nL\sigma^2/(b\mu^2)\big\}$, we know

$$\mu^2 R^2 \ge n \quad \text{and} \quad b\mu^2 R \ge n.$$

It then follows that $\mathbb{E}[F(\mathbf{w}_R)] - F(\mathbf{w}^*) \lesssim L/(n\mu)$. The proof is completed. □

7. Proofs on Local SGD

7.1. Proof of Theorem 9

In this section, we prove stability bounds on local SGD.

Proof of Theorem 9. Let $\{\mathbf{w}^{(k)}_{m,r,t+1}\}$ be the sequence produced by Eq. (5.1) on $S^{(k)}$. We introduce the notations

$$\Delta_{m,r,t,k} = \big\|\mathbf{w}_{m,r,t} - \mathbf{w}^{(k)}_{m,r,t}\big\|_2, \quad \mathfrak{C}_{m,r,t,k} = \|\nabla f(\mathbf{w}_{m,r,t}; z_k) - \nabla f(\mathbf{w}^{(k)}_{m,r,t}; z'_k)\|_2.$$

If $i_{m,r,t} \neq k$, we can use Lemma 14 to derive

$$\Delta_{m,r,t+1,k} = \big\|\mathbf{w}_{m,r,t} - \eta_{r,t}\nabla f(\mathbf{w}_{m,r,t}; z_{i_{m,r,t}}) - \mathbf{w}^{(k)}_{m,r,t} + \eta_{r,t}\nabla f(\mathbf{w}^{(k)}_{m,r,t}; z_{i_{m,r,t}})\big\|_2 \leq \|\mathbf{w}_{m,r,t} - \mathbf{w}^{(k)}_{m,r,t}\|_2.$$

If $i_{m,r,t} = k$, we have

$$\Delta_{m,r,t+1,k} = \big\|\mathbf{w}_{m,r,t} - \eta_{r,t}\nabla f(\mathbf{w}_{m,r,t}; z_k) - \mathbf{w}^{(k)}_{m,r,t} + \eta_{r,t}\nabla f(\mathbf{w}^{(k)}_{m,r,t}; z'_k)\big\|_2 \leq \Delta_{m,r,t,k} + \eta_{r,t}\mathfrak{C}_{m,r,t,k}.$$

We combine the above two cases together and derive

$$\Delta_{m,r,t+1,k} \leq \Delta_{m,r,t,k} + \eta_{r,t}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}, \tag{7.1}$$

where $\mathbb{I}_{[i_{m,r,t}=k]}$ denotes the indicator function of the event $\{i_{m,r,t} = k\}$, i.e., $\mathbb{I}_{[i_{m,r,t}=k]} = 1$ if $i_{m,r,t} = k$, and 0 otherwise. We apply the above inequality recursively and get

$$\Delta_{m,r,K+1,k} \leq \Delta_{m,r,1,k} + \sum_{t=1}^{K}\eta_{r,t}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}.$$

We take an average over $m \in [M]$ and use $\mathbf{w}_{r+1} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,K+1}$ to derive

$$\big\|\mathbf{w}_{r+1} - \mathbf{w}^{(k)}_{r+1}\big\|_2 \leq \frac{1}{M}\sum_{m=1}^{M}\|\mathbf{w}_{m,r,K+1} - \mathbf{w}^{(k)}_{m,r,K+1}\|_2 \leq \big\|\mathbf{w}_r - \mathbf{w}^{(k)}_r\big\|_2 + \sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}, \tag{7.2}$$

where we have used $\mathbf{w}_{m,r,1} = \mathbf{w}_r$. We can apply the above inequality recursively, and derive

$$\big\|\mathbf{w}_{R+1} - \mathbf{w}^{(k)}_{R+1}\big\|_2 \leq \sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\mathbb{I}_{[i_{m,r,t}=k]}. \tag{7.3}$$

We first consider the ℓ_1 on-average model stability. We know that $i_{m,r,t}$ takes the value k with probability $1/n$, and other values with probability $1 - 1/n$. We take expectation w.r.t. $i_{m,r,t}$ and note $\mathfrak{C}_{m,r,t,k}$ is independent of $i_{m,r,t}$, which implies

$$\mathbb{E}\big[\|\mathbf{w}_{R+1} - \mathbf{w}^{(k)}_{R+1}\|_2\big] \leq \sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{nM}\mathbb{E}[\mathfrak{C}_{m,r,t,k}] \leq \frac{2\sqrt{2L}}{nM}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathbb{E}\Big[\sqrt{f(\mathbf{w}_{m,r,t}; z_k)}\Big], \tag{7.4}$$

where we have used the self-bounding property and the symmetry between z_k and z'_k (analogous to Eq. (6.10)). It then follows from the concavity of $x \mapsto \sqrt{x}$ that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{R+1}-\mathbf{w}^{(k)}_{R+1}\|_2\big] &\leq \frac{2\sqrt{2L}}{n^2M}\sum_{k=1}^{n}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathbb{E}\Big[\sqrt{f(\mathbf{w}_{m,r,t};z_k)}\Big]\\ &\leq \frac{2\sqrt{2L}}{nM}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{k=1}^{n}f(\mathbf{w}_{m,r,t};z_k)\Big)^{\frac{1}{2}}\Big]\\ &= \frac{2\sqrt{2L}}{nM}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathbb{E}\Big[\Big(F_S(\mathbf{w}_{m,r,t})\Big)^{\frac{1}{2}}\Big].\end{aligned}$$

This proves Eq. (5.2). We now consider the ℓ_2 on-average model stability. We take an expectation-variance decomposition in Eq. (7.3) and derive

$$\big\|\mathbf{w}_{R+1}-\mathbf{w}^{(k)}_{R+1}\big\|_2 \leq \sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\big(\mathbb{I}_{[i_{m,r,t}=k]}-1/n\big)+n^{-1}\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}. \quad (7.5)$$

Analogous to Eq. (6.13), we have (note $i_{m,r,t}$ is independent of $i_{m',r',t'}$ if $(m,r,t)\neq(m',r',t')$, $\mathfrak{C}_{m,r,t,k}$ is independent of $i_{m,r,t}$, and $\mathfrak{C}_{m',r',t',k}$ is independent of $i_{m',r',t'}$)

$$\mathbb{E}\Big[\mathfrak{C}_{m,r,t,k}\big(\mathbb{I}_{[i_{m,r,t}=k]}-1/n\big)\mathfrak{C}_{m',r',t',k}\big(\mathbb{I}_{[i_{m',r',t'}=k]}-1/n\big)\Big]=0 \quad \text{if either } t\neq t', m\neq m', \text{ or } r\neq r'.$$

Then, we take a square on both sides of Eq. (7.5) followed by expectation, and analyze analogously to the proof of Eq. (4.3):

$$\begin{aligned}&\mathbb{E}\big[\big\|\mathbf{w}_{R+1}-\mathbf{w}^{(k)}_{R+1}\big\|_2^2\big]\\ &\leq 2\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\big(\mathbb{I}_{[i_{m,r,t}=k]}-1/n\big)\Big)^2\Big]+\frac{2}{n^2}\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\frac{\eta_{r,t}}{M}\mathfrak{C}_{m,r,t,k}\Big)^2\Big]\\ &= \frac{2}{M^2}\mathbb{E}\Big[\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}^2\mathfrak{C}_{m,r,t,k}^2\mathrm{Var}(\mathbb{I}_{[i_{m,r,t}=k]})\Big]+\frac{2}{n^2M^2}\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathfrak{C}_{m,r,t,k}\Big)^2\Big]\\ &\leq \frac{2}{nM^2}\mathbb{E}\Big[\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}^2\mathfrak{C}_{m,r,t,k}^2)\Big]+\frac{2}{n^2M^2}\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathfrak{C}_{m,r,t,k}\Big)^2\Big],\end{aligned}$$

where we have used $\mathrm{Var}(\mathbb{I}_{[i_{m,r,t}=k]})\leq 1/n$. By the self-bounding property of f we know

$$\mathbb{E}\big[\mathfrak{C}^2_{m,r,t,k}\big]\leq 4L\mathbb{E}\big[f(\mathbf{w}_{m,r,t};z_k)+f(\mathbf{w}^{(k)}_{m,r,t};z'_k)\big]=8L\mathbb{E}[f(\mathbf{w}_{m,r,t};z_k)]. \quad (7.6)$$

It then follows that

$$\mathbb{E}\big[\big\|\mathbf{w}_{R+1}-\mathbf{w}^{(k)}_{R+1}\big\|_2^2\big]\leq\frac{16L}{nM^2}\mathbb{E}\Big[\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}^2f(\mathbf{w}_{m,r,t};z_k)\Big]+\frac{2}{n^2M^2}\mathbb{E}\Big[\Big(\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\eta_{r,t}\mathfrak{C}_{m,r,t,k}\Big)^2\Big]. \quad (7.7)$$

The stated bound then follows by taking an average over $k\in[n]$ and noting $F_S(\mathbf{w})=\frac{1}{n}\sum_{k=1}^{n}f(\mathbf{w};z_k)$. The proof is completed. □

7.2. *Proof of Theorem 10*

In this section, we prove Theorem 10 on excess population risk bounds of local SGD for convex problems. To this aim, we require the following lemma on the optimization error bounds [44]. Note $F_S(\mathbf{w}^*)\geq F_S(\mathbf{w}_S)$.

Lemma 18 (Optimization Errors of Local SGD: Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_{m,r,t}\}$ be produced by the algorithm A defined in (5.1) with $\eta \leq 1/(4L)$. Assume for all $r \in [R], t \in [K]$, $\mathbb{E}_{i_{m,r,t}}[\|\nabla f(\mathbf{w}_{r,t}; z_{i_{m,r,t}}) - \nabla F_S(\mathbf{w}_{r,t})\|_2^2] \leq \sigma^2$. Then the following inequality holds*

$$\mathbb{E}_A[F_S(\bar{\mathbf{w}}_{R,1})] - F_S(\mathbf{w}^*) \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\eta K R} + \frac{\eta \sigma^2}{M} + L(K-1)\eta^2\sigma^2. \tag{7.8}$$

We are now ready to prove Theorem 10. For simplicity, we assume $\mathbb{E}\big[\sqrt{F_S(\mathbf{w}_{m',r',t'})}\big] \lesssim 1$, which is reasonable since we are minimizing F_S by local SGD. Note this assumption is used to bound the stability and can be removed if we assume f is Lipschitz continuous ($F_S(\mathbf{w}_{m',r',t'})$ appears in the stability analysis since we control the gradient norm by function values).

Proof of Theorem 10. Analogous to Eq. (7.7), one can show that

$$\begin{aligned}\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,t} - \sum_{m=1}^{M}\mathbf{w}_{m,r,t}^{(k)}\Big\|_2^2\Big] &\leq \frac{16L}{nM^2}\mathbb{E}\Big[\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}^2 f(\mathbf{w}_{m,r',t'}; z_k)\Big] \\ &\quad + \frac{2}{n^2M^2}\mathbb{E}\Big[\Big(\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}\mathfrak{C}_{m,r',t',k}\Big)^2\Big].\end{aligned}$$

We take an average over $k \in [n]$, and derive

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,t} - \sum_{m=1}^{M}\mathbf{w}_{m,r,t}^{(k)}\Big\|_2^2\Big] &\leq \frac{16L}{n^2M^2}\sum_{k=1}^{n}\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}^2\mathbb{E}\big[f(\mathbf{w}_{m,r',t'}; z_k)\big] \\ &\quad + \frac{2}{n^3M^2}\sum_{k=1}^{n} rMK\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}^2\mathbb{E}[\mathfrak{C}_{m,r',t',k}^2].\end{aligned}$$

By the self-bounding property and the symmetry between z_k and z_k', we further know

$$\mathbb{E}[\mathfrak{C}_{m,r',t',k}^2] \leq 2\mathbb{E}\big[\|\nabla f(\mathbf{w}_{m,r',t'}; z_k)\|_2^2\big] + \mathbb{E}\big[\|\nabla f(\mathbf{w}_{m,r',t'}^{(k)}; z_k')\|_2^2\big] \leq 8L\mathbb{E}[f(\mathbf{w}_{m,r',t'}; z_k)].$$

It then follows that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,t} - \sum_{m=1}^{M}\mathbf{w}_{m,r,t}^{(k)}\Big\|_2^2\Big] &\leq \frac{16L}{nM^2}\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}^2\mathbb{E}\big[F_S(\mathbf{w}_{m,r',t'})\big] \\ &\quad + \frac{16LrK}{n^2M}\sum_{r'=1}^{r}\sum_{m=1}^{M}\sum_{t'=1}^{K}\eta_{r',t'}^2\mathbb{E}\big[F_S(\mathbf{w}_{m,r',t'})\big].\end{aligned}$$

It then follows the convexity of $\|\cdot\|^2$ that

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\|\bar{\mathbf{w}}_{R,1} - \bar{\mathbf{w}}_{R,1}^{(k)}\|_2^2\big] \leq \frac{1}{KRn}\sum_{k=1}^{n}\sum_{r=1}^{R}\sum_{t=1}^{K}\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_{m,r,t} - \sum_{m=1}^{M}\mathbf{w}_{m,r,t}^{(k)}\Big\|_2^2\Big] \\ &\lesssim \frac{1}{KR}\sum_{r=1}^{R}\sum_{t=1}^{K}\frac{L\eta^2}{nM}\Big(\frac{1}{M} + \frac{rK}{n}\Big)\sum_{r'=1}^{r}\sum_{m'=1}^{M}\sum_{t'=1}^{K}\mathbb{E}\big[F_S(\mathbf{w}_{m',r',t'})\big] \\ &\lesssim \frac{L\eta^2}{KRnM}\Big(\frac{KR}{M} + \frac{K^2R^2}{n}\Big)\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{t=1}^{K}\mathbb{E}\big[F_S(\mathbf{w}_{m,r,t})\big].\end{aligned}$$

According to Lemma 1 (Part (b)) and using the assumption $\mathbb{E}\big[F_S(\mathbf{w}_{m,r,t})\big] \lesssim 1$, we know

$$\mathbb{E}[F(\bar{\mathbf{w}}_{R,1}) - F_S(\bar{\mathbf{w}}_{R,1})] \lesssim \frac{L}{\gamma}\mathbb{E}[F_S(\bar{\mathbf{w}}_{R,1})] + \frac{L(L+\gamma)\eta^2}{n}\Big(\frac{KR}{M} + \frac{K^2R^2}{n}\Big).$$

We combine the above inequality and Lemma 18 together, and derive

$$\mathbb{E}[F(\bar{\mathbf{w}}_{R,1})] - F(\mathbf{w}^*) \lesssim \frac{L}{\gamma}\mathbb{E}[F_S(\bar{\mathbf{w}}_{R,1})] + \frac{L(L+\gamma)\eta^2}{n}\Big(\frac{KR}{M} + \frac{K^2R^2}{n}\Big) + \frac{\|\mathbf{w}^*\|_2^2}{\eta KR} + \frac{\eta\sigma^2}{M} + L(K-1)\eta^2\sigma^2.$$

We can minimize γ and use $KRM \asymp n$ to get

$$\mathbb{E}[F(\bar{\mathbf{w}}_{R,1})] - F(\mathbf{w}^*) \lesssim \frac{LKR\eta}{n} + \frac{L^2\eta^2K^2R^2}{n^2} + \frac{\|\mathbf{w}^*\|_2^2}{\eta KR} + \frac{\eta\sigma^2}{M} + L(K-1)\eta^2\sigma^2.$$

Since $\eta \asymp \|\mathbf{w}^*\|_2\sqrt{n}/(KR\sqrt{L})$, we know

$$\begin{aligned}
\frac{LKR\eta}{n} &\asymp \frac{\|\mathbf{w}^*\|_2^2}{\eta KR} \asymp \frac{\sqrt{L}\|\mathbf{w}^*\|_2}{\sqrt{n}},\\
\frac{L^2\eta^2K^2R^2}{n^2} &\asymp \frac{L^2\|\mathbf{w}^*\|_2^2 nK^2R^2}{n^2K^2R^2L} = \frac{L\|\mathbf{w}^*\|_2^2}{n},\\
\frac{\eta\sigma^2}{M} &\asymp \frac{\|\mathbf{w}^*\|_2\sqrt{n}\sigma^2}{MKR\sqrt{L}} \asymp \frac{\|\mathbf{w}^*\|_2\sigma^2}{\sqrt{nL}}.
\end{aligned}$$

Since $\eta \lesssim (K-1)^{-\frac{1}{2}}\|\mathbf{w}^*\|_2^{\frac{1}{2}}/(nL)^{\frac{1}{4}}$, we further know

$$L(K-1)\eta^2\sigma^2 \asymp \frac{\sqrt{L}\|\mathbf{w}^*\|_2}{\sqrt{n}}.$$

The stated bound then follows by combining the above discussions together. □

7.3. Proof of Theorem 11

To prove Theorem 11, we require the following lemma on optimization errors [41, 21].

Lemma 19 (Optimization Errors of Local SGD: Strongly Convex Case)**.** *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, μ-strongly convex and L-smooth. Let $\{\mathbf{w}_{m,r,t}\}$ be produced by the algorithm A defined in* (5.1) *with $\eta_{r,t} = \frac{4}{\mu(a+(r-1)K+t)} \leq 2/L$ with $a > 2L/\mu$. Assume for all $r \in [R], t \in [K]$, $\mathbb{E}_{i_{m,r,t}}[\|\nabla f(\mathbf{w}_{r,t}; z_{i_{m,r,t}}) - \nabla F_S(\mathbf{w}_{r,t})\|_2^2] \leq \sigma^2$. Then the following inequality holds*

$$\mathbb{E}_A[F_S(\bar{\mathbf{w}}_{R,2})] - F_S(\mathbf{w}^*) \lesssim \frac{\sigma^2}{\mu MKR} + \frac{L\log(RK)}{\mu^2KR^2}.$$

Proof of Theorem 11. By the analysis in the proof of Theorem 9 (e.g. Eq. (7.4)), we know

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^n \mathbb{E}\big[\|\mathbf{w}_{m,r,t} - \mathbf{w}_{m,r,t}^{(k)}\|_2\big] &\leq \frac{2\sqrt{2L}}{nM}\sum_{r'=1}^R\sum_{m'=1}^M\sum_{t'=1}^K \eta_{r',t'}\mathbb{E}\Big[\sqrt{F_S(\mathbf{w}_{m',r',t'})}\Big]\\
&\lesssim \frac{\sqrt{L}}{nM}\sum_{r'=1}^R\sum_{m'=1}^M\sum_{t'=1}^K \frac{1}{\mu(a+(r'-1)K+t')}\\
&\lesssim \frac{\sqrt{L}}{nM}\sum_{m'=1}^M \frac{\log(KR)}{\mu} \lesssim \frac{\sqrt{L}\log(KR)}{n\mu}, \quad \forall r \in [R], t \in [K].
\end{aligned}$$

Since the above inequality holds for all $r \in [R], t \in [K]$ and $\bar{\mathbf{w}}_{R,2}$ is a weighted average of $\mathbf{w}_{m,r,t}$, we then get

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\|\bar{\mathbf{w}}_{R,2} - \bar{\mathbf{w}}_{R,2}^{(k)}\|_2\big] \lesssim \frac{\sqrt{L}\log(KR)}{n\mu}$$

and therefore $\mathbb{E}[F(\bar{\mathbf{w}}_{R,2}) - F_S(\bar{\mathbf{w}}_{R,2})] \lesssim \frac{\sqrt{L}G\log(KR)}{n\mu}$. We combine this generalization error bound and the optimization error bound in Lemma 19 to derive

$$\mathbb{E}[F(\bar{\mathbf{w}}_{R,2})] - F(\mathbf{w}^*) \lesssim \frac{G\sqrt{L}\log(KR)}{n\mu} + \frac{\sigma^2}{\mu MKR} + \frac{L\log(RK)}{\mu^2 KR^2} \lesssim \frac{G\sqrt{L}\log(KR)}{n\mu},$$

where we have used $KR \gtrsim \frac{n\sigma^2}{MG\sqrt{L}}$ and $\mu KR^2 \gtrsim n\sqrt{L}/G$ in the last inequality. □

8. Conclusion

We investigate the stability and generalization of minibatch SGD and local SGD, which are widely used for large-scale learning problems. While there are many discussions on the speedup of these methods for optimization, we study the linear speedup in generalization. We develop on-average stability bounds for convex, strongly convex and nonconvex problems, and show how small training errors can improve stability. For strongly convex problems, our stability bounds are independent of the iteration number, which is new for the vanilla SGD in the sense of removing the Lipschitzness assumption. Our stability analysis implies optimal excess population risk bounds with both a linear speedup w.r.t. the batch size for minibatch SGD and a linear speedup w.r.t. the number of machines for local SGD.

There are several limitations of our work. A limitation of our work is that we do not get optimistic bounds for local SGD which are important to show the benefit of low noises. Another limitation is that we only consider homogeneous setups in local SGD. It would be very interesting to extend the analysis to heterogeneous setups, i.e., where different local machines have different sets of examples. We will study these limitations in our future work.

Acknowledgements

The authors are grateful to the associate editor and the anonymous reviewers for their thoughtful comments and constructive suggestions.

References

[1] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[2] R. Bassily, V. Feldman, K. Talwar, and A. G. Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, pages 11279–11288, 2019.

[3] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[4] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2 (Mar):499–526, 2002.

[5] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[6] L. Carratino, A. Rudi, and L. Rosasco. Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pages 10213–10224, 2018.

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

[8] S. Chen, Q. Zheng, Q. Long, and W. J. Su. Minimax estimation for personalized federated learning: An alternative between fedavg and local training? *Journal of Machine Learning Research*, 24(262):1–59, 2023.

[9] Y. Chen, C. Jin, and B. Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

[10] A. Christmann, D. Xiang, and D.-X. Zhou. Total stability of kernel methods. *Neurocomputing*, 289: 101–118, 2018.

[11] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in Neural Information Processing Systems*, 24:1647–1655, 2011.

[12] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

[13] J. Fan and Y. Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

[14] V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

[15] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[16] R. Gower, O. Sebbouh, and N. Loizou. SGD for structured nonconvex functions: Learning rates, mini-batching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

[17] X. Guo, Z.-C. Guo, and L. Shi. Capacity dependent analysis for functional online learning algorithms. *Applied and Computational Harmonic Analysis*, 67:101567, 2023.

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

[19] T. Hu, Q. Wu, and D.-X. Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Applied and Computational Harmonic Analysis*, 49(1):229–256, 2020.

[20] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[21] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[22] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[23] T. Koren, R. Livni, Y. Mansour, and U. Sherman. Benign underfitting of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2022.

[24] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

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

[26] Y. Lei and Y. Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

[27] B. Li, Z. Li, and Y. Chi. Destress: Computation-optimal and communication-efficient decentralized nonconvex finite-sum optimization. *SIAM Journal on Mathematics of Data Science*, 4(3):1031–1051, 2022.

[28] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

[29] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

[30] S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

[31] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations*, 2019.

[32] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[33] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[34] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

[35] N. Mücke, G. Neu, and L. Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

[36] M. Schliserman and T. Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394. PMLR, 2022.

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

[38] O. Shamir and N. Srebro. Distributed stochastic optimization and learning. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 850–857. IEEE, 2014.

[39] A. Spiridonoff, A. Olshevsky, and Y. Paschalidis. Communication-efficient SGD: From local SGD to one-shot averaging. In *Advances in Neural Information Processing Systems*, volume 34, pages 24313–24326, 2021.

[40] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural*

Information Processing Systems, pages 2199–2207, 2010.

[41] S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

[42] Z. Sun, X. Niu, and E. Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684, 2024.

[43] P. Wang, Y. Lei, Y. Ying, and H. Zhang. Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022.

[44] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. McMahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[45] B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292, 2020.

[46] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[47] D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1998–2007. PMLR, 2018.

[48] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[49] H. Yu, S. Yang, and S. Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *AAAI Conference on Artificial Intelligence*, pages 5693–5700, 2019.

[50] C. Yun, S. Rajput, and S. Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022.

[51] Y. Zhang, W. Zhang, S. Bald, V. Pingali, C. Chen, and M. Goswami. Stability of SGD: Tightness analysis and improved bounds. In *Uncertainty in Artificial Intelligence*, pages 2364–2373. PMLR, 2022.

[52] M. A. Zinkevich, M. Weimer, A. Smola, and L. Li. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.