Journal of Machine Learning Research 27 (2026) 1-55 Submitted 5/24; Revised 11/25; Published 2/26

Stochastic Gradient Methods: Bias, Stability and Generalization

Shuang Zeng ZENGSH9@CONNECT.HKU.HK
Yunwen Lei* LEIYW@HKU.HK
Department of Mathematics
The University of Hong Kong
Hong Kong, China

Editor: Gergely Neu

Abstract

Recent developments of stochastic optimization often suggest *biased gradient estimators* to improve either the robustness, communication efficiency or computational speed. Representative biased stochastic gradient methods (BSGMs) include Zeroth-order stochastic gradient descent (SGD), Clipped-SGD and SGD with delayed gradients. The practical success of BSGMs motivates a lot of convergence analysis to explain their impressive training behaviour. As a comparison, there is far less work on their generalization analysis, which is a central topic in modern machine learning. In this paper, we present the first framework to study the stability and generalization of BSGMs for convex and smooth problems. We introduce a generalized Lipschitz-type condition on gradient estimators and bias, under which we develop a rather general stability bound to show how the bias and the gradient estimators affect the stability. We apply our general result to develop the first stability bound for Zeroth-order SGD with reasonable step size sequences, and the first stability bound for Clipped-SGD. While our stability analysis is developed for general BSGMs, the resulting stability bounds for both Zeroth-order SGD and Clipped-SGD match those of SGD under appropriate smoothing/clipping parameters. We combine the stability and convergence analysis together, and derive excess risk bounds of order $O(1/\sqrt{n})$ for both Zeroth-order SGD and Clipped-SGD, where n is the sample size.

1. Introduction

Stochastic optimization, especially stochastic gradient descent (SGD) and its variants (Johnson and Zhang, 2013; Duchi et al., 2011; Schmidt et al., 2017; Bottou et al., 2018), are the workhorse behind the success of many machine learning applications. Due to their cheap computation cost, simplicity in implementation and impressive generalization behavior, stochastic optimization has achieved a lot of success especially in solving large-scale and complex learning problems (Bottou and Bousquet, 2007; Bottou et al., 2018). This motivates plethora of theoretical work on its convergence analysis (Rakhlin et al., 2012; Bottou et al., 2018; Orabona, 2019; Gower et al., 2020) as well as generalization analysis (Bousquet and Bottou, 2008; Hardt et al., 2016; Yao et al., 2007; Kuzborskij and Lampert, 2018; Lei and Ying, 2020; Zou et al., 2022; Zhang, 2023), which shed lights on understanding the fundamental performance limits of stochastic optimization across broad problem classes.

*. Yunwen Lei is the corresponding author

Stochastic optimization builds gradient estimators to approximate the true gradient by introducing random sampling into the optimization process. Oftentimes, the gradient estimators are unbiased estimators of the true gradient, which simplify both the convergence and generalization analysis. However, in practice, people often build *biased estimators* to either improve the robustness, communication efficiency or computational cost, which lead to a large class of biased stochastic gradient methods (BSGMs) (Ajalloeian and Stich, 2020; Driggs et al., 2022). The framework of BSGMs include popular algorithms such as Zeroth-order SGD (Nesterov and Spokoiny, 2017), Clipped-SGD (Zhang et al., 2020, 2019), SGD with delayed gradients (Stich and Karimireddy, 2020), and stochastic average gradient (SAG) (Schmidt et al., 2017). Below, we give more details on Zeroth-order SGD and Clipped-SGD.

- **Zeroth-order SGD** has been widely used in black-box optimization problems (Nesterov and Spokoiny, 2017; Gasnikov et al., 2022; Sun et al., 2022), where explicit gradients are hard to attain and only the function values can be attained. These include structured-prediction and graphical model inference where the objective is defined variationally (Duchi et al., 2015), as well as modern machine learning such as adversarial learning (Chen et al., 2017; Liu et al., 2019), meta learning (Ruan et al., 2020), reinforcement learning (Vemula et al., 2019; Kumar et al., 2021) and large language models (Tang et al., 2024).

- **Clipped-SGD** has been applied in various applications such as improving training of deep neural networks (Pascanu et al., 2012), diminishing the sensitivity of average gradients in differential privacy (Abadi et al., 2016; Das et al., 2023), and dealing with heavy-tailed gradients (Gorbunov et al., 2020). For instance, large NLP models based on attention and transformers have heavy-tailed gradients, which make SGD unstable to converge (Zhang et al., 2019, 2020). As a comparison, Clipped-SGD is robust to heavy-tailed noises and still guarantees convergence (Zhang et al., 2020).

The practical success of BSGMs motivates increasing interests in their theoretical analysis. Initially, specific BSGMs have been studied. For example, the pioneering work developed optimal convergence rates for Zeroth-order SGD for convex problems (Duchi et al., 2015; Nesterov and Spokoiny, 2017). The benefit of Clipped-SGD over SGD has also been theoretically verified for learning under heavy-tailed noises (Zhang et al., 2020). Recently, several studies consider the convergence of general BSGMs under general assumptions on the bias and variance (Ajalloeian and Stich, 2020; Driggs et al., 2022; Hu et al., 2020). These studies focus on the empirical behavior and the convergence analysis of BSGMs from the perspective of optimization, leaving the issue of generalization untouched, which is a key concern in machine learning. To our knowledge, the existing stability analysis of BSGMs only focused on the specific Zeroth-order SGD (Nikolakakis et al., 2022a; Liu et al., 2024). The stability analyses of Zeroth-order SGD with random gradient estimation (Theorem 5 in Nikolakakis et al. (2022a) and Theorems 10, 12 in Liu et al. (2024)) require a fast-decaying step size $\eta_t \lesssim 1/t$ at the t-th iteration[1], for which the training errors would decay with

1. We say $A \lesssim B$ if there exists some universal constant $C > 0$ (C is independent of A and B) such that $A \leq BC$. We say $A \gtrsim B$ if there exists some universal constant $C > 0$ such that $A \geq BC$. We use the notation $A \asymp B$ if $A \lesssim B$ and $B \lesssim A$.

the slow rate of the order $1/\log(T)$ after T iterations. These limitations motivate us the following question on the issue of generalization: can we develop a general framework for the stability analysis of BSGMs, and if specialized to Zeroth-order SGD, can the framework lead to meaningful stability bounds allowing a moderate decay of step sizes?

In this paper, we provide an affirmative answer to the above question by developing a general analysis of BSGMs for convex and smooth problems. Our contributions are summarized as follows.

- We develop a general framework to study the stability and generalization of BSGMs. Our analysis illustrates how the bias and the sharpness of the gradient estimator affect the stability. We introduce a generalized Lipschitz-type condition on the bias and the gradient estimator, under which we get a simplified bound which is convenient to apply to specific BSGMs such as Zeroth-order SGD and Clipped-SGD.

- We specialize our stability bound to Zeroth-order SGD by verifying the Lipschitz-type condition on the gradient estimator. As compared to existing analysis requiring fast-decaying step size $\eta_t \lesssim 1/t$ (Nikolakakis et al., 2022a), our analysis requires a moderate assumption $\sum_{t=1}^{T} \eta_t^2 \lesssim 1$. This condition holds for $\eta_t \lesssim 1/\sqrt{T}$, which is a typical choice for SGD-type algorithms in the general convex case.

- We also show that the generalized Lipschitz-type condition of the bias and the gradient estimator hold for Clipped-SGD, which, to our knowledge, yields the first stability and generalization bounds for Clipped-SGD. Impressively, our stability bounds for Clipped-SGD match those for SGD under mild assumptions on the clipping parameter.

- Our stability analysis does not require the loss function to be Lipschitz continuous. Moreover, it incorporates the training errors into the stability bounds, which show that a good optimization would be beneficial to improve the stability and generalization. Furthermore, for both Zeroth-order SGD and Clipped-SGD, we combine the stability analysis and convergence analysis to give the excess risk bounds of order $O(1/\sqrt{n})$, where n is the sample size.

We organize the paper as follows. We discuss the related work in Section 2 and formulate the problem in Section 3. The stability bounds for general BSGMs are presented in Section 4 (see detailed proofs in Section 6.1). In Section 5, we apply the general stability bounds in Section 4 to SGD, Zeroth-order SGD and Clipped-SGD (see detailed proofs in Section 7). We conclude the paper in Section 8.

2. Related Work

Algorithmic stability. Algorithmic stability is a fundamental concept in statistical learning theory (SLT), which measures how the output model will change if we either remove or replace a single example. Since its introduction in the 1970s, various algorithmic stability concepts have been proposed to study the generalization gap of learning algorithms (Rogers and Wagner, 1978; Bousquet and Elisseeff, 2002). The most popular stability concept is the uniform stability, which can imply high-probability bounds (Bousquet and Elisseeff, 2002; Feldman and Vondrak, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021; Zhang,

2023; Fan and Lei, 2024). This strong stability measure was weaken to study generalization bounds in expectation, including the on-average stability (Shalev-Shwartz et al., 2010; Koren and Levy, 2015) and hypothesis stability (Bousquet and Elisseeff, 2002; Kuzborskij and Orabona, 2013; Elisseeff et al., 2005). The impressive work (Hardt et al., 2016) gave the first stability bounds for SGD, which motivates a surging of work on the stability analysis of stochastic optimization methods (Lin et al., 2016; Charles and Papailiopoulos, 2018; Kuzborskij and Lampert, 2018; Lei and Ying, 2020; Bassily et al., 2020; Nikolakakis et al., 2022b; Lei, 2023; Chen et al., 2024; Zhang et al., 2024). Recent stability analysis implies optimistic bounds for SGD (Lei and Ying, 2020), which were used to study the implicit bias of gradient methods (Schliserman and Koren, 2022) and the generalization performance of overparameterized models such as neural networks (Richards and Kuzborskij, 2021; Taheri and Thrampoulidis, 2024; Deora et al., 2024). Lower bounds on the stability of gradient methods have also been studied (Bassily et al., 2020; Koren et al., 2022; Amir et al., 2021).

Biased stochastic gradient methods. Now we turn to the analysis of BSGMs. Ajalloeian and Stich (2020) introduced a framework for the convergence analysis of SGD with biased gradients, which applies to several specific instantiations such as Top-k sparsification, Zeroth-order SGD, biased compression operators and delayed gradients. Driggs et al. (2022) presented a systematic analysis for a large class of unbiased and biased stochastic gradient methods, which provides proximal support and identifies the benefit of bias in stochastic gradient estimation. Hu et al. (2020) extended the convergence analysis to biased SGD for conditional stochastic optimization. The above discussions consider general BSGMs. For the specific Zeroth-order SGD, the optimal convergence rates have been established for both smooth and nonsmooth problems (Duchi et al., 2015; Nesterov and Spokoiny, 2017). The behavior of Clipped-SGD has drawn a lot of attention in optimization with heavy-tailed gradient noises (Cutkosky and Mehta, 2021; Zhang et al., 2020, 2019) and differential privacy (Abadi et al., 2016; Das et al., 2023). All the above mentioned work studied BSGMs from an optimization perspective. Recently, the stability of the Zeroth-order SGD has also been studied (Nikolakakis et al., 2022a; Liu et al., 2024). However, the stability analysis there shows that the corresponding gradient operator has the expansive factor $1 + \eta_t$ (even if the loss function is convex)(Nikolakakis et al., 2022a), and therefore the step size η_t has to decay with the order of $1/t$ to get meaningful stability bounds (Nikolakakis et al., 2022a). This step size slows down the optimization process, and the corresponding convergence rates decay logarithmically even if the gradient estimation is accurate, e.g., SGD. In this paper, we go beyond this limitation by developing a framework of stability analysis for BSGMs which handles the step size $\eta_t \lesssim 1/\sqrt{T}$ when applied to Zeroth-order SGD.

3. Background

3.1 Problem Setup

Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y} \subseteq \mathbb{R}$ is an output space. Suppose we are given a training dataset $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ independently drawn from $\mathbb{P}$, based on which we want to build a prediction function $h : \mathcal{X} \mapsto \mathbb{R}$. We consider parametric learning where the prediction function h is parameterized by $\mathbf{w}$ in a parameter space $\mathcal{W} \subset \mathbb{R}^d$. The performance of $\mathbf{w}$ on an example

$\mathbf{z}$ is measured by $f(\mathbf{w};\mathbf{z})$, where $f:\mathcal{W}\times\mathcal{Z}\mapsto\mathbb{R}_+$ is a loss function. The empirical risk and population risk of $\mathbf{w}$ are then defined by

$$F_S(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n f(\mathbf{w};\mathbf{z}_i)\quad\text{and}\quad F(\mathbf{w})=\mathbb{E}_z[f(\mathbf{w};\mathbf{z})],$$

which quantify the behavior of the model on training and testing, respectively. Here $\mathbb{E}_z[\cdot]$ denotes the expectation with respect to (w.r.t.) $\mathbf{z}$.

Upon observing S, we often apply a stochastic learning algorithm $\mathcal{A}$ to build a model as an approximation of $\mathbf{w}^*=\arg\min_{\mathbf{w}\in\mathcal{W}}F(\mathbf{w})$. Then, we use $\mathcal{A}(S)$ to denote the model produced by running $\mathcal{A}$ over the training dataset S. We are interested in studying the excess population risk $\mathbb{E}[F(\mathcal{A}(S))]-F(\mathbf{w}^*)$ to understand how far we are from the best model, which can be decomposed as (note $\mathbb{E}[F_S(\mathbf{w}^*)]=F(\mathbf{w}^*)$) (Bousquet and Bottou, 2008)

$$\mathbb{E}[F(\mathcal{A}(S))]-F(\mathbf{w}^*)=\big(\mathbb{E}[F(\mathcal{A}(S))-F_S(\mathcal{A}(S))]\big)+\big(\mathbb{E}[F_S(\mathcal{A}(S))-F_S(\mathbf{w}^*)]\big).$$

We refer to $\mathbb{E}[F(\mathcal{A}(S))-F_S(\mathcal{A}(S))]$ as the generalization gap, which quantifies the difference between training and testing for the output model. We refer to $\mathbb{E}[F_S(\mathcal{A}(S))-F_S(\mathbf{w}^*)]$ as the optimization error, which measures the training suboptimality of the output model as compared to the best model. Optimization error is a central concept in optimization theory, which has been extensively studied in the literature (Bottou et al., 2018; Orabona, 2019; Gower et al., 2020). Generalization gap is a central concept in SLT (Lugosi and Neu, 2022), which can be studied via several concepts such as Rademacher complexities (Bartlett and Mendelson, 2002), covering numbers (Steinwart and Christmann, 2008; Cucker and Zhou, 2007) and algorithmic stability (Bousquet and Elisseeff, 2002). In this paper, we consider biased stochastic gradient descent methods, and our focus is on their generalization ability.

3.2 Algorithmic Stability

Algorithmic stability is a fundamental concept in SLT to study the generalization of learning algorithms. Various different stability concepts have been introduced in the literature, including the uniform stability (Bousquet and Elisseeff, 2002), on-average stability (Shalev-Shwartz et al., 2010), argument stability (Liu et al., 2017) and on-average model stability (Lei and Ying, 2020). In this paper, we consider the on-average model stability introduced in Lei and Ying (2020), which considers the on-average effect of perturbing each training example and has a benefit of incorporating the empirical risk into the stability bounds (Lei and Ying, 2020; Kuzborskij and Lampert, 2018). Let $\|\cdot\|_2$ denote the Euclidean norm, $\langle\cdot,\cdot\rangle$ denote the dot product and ∇ denote the gradient operator.

Definition 3.1 (Uniform stability) *We say a (randomized) algorithm $\mathcal{A}$ is ϵ-uniformly stable if for all training datasets $S,\widetilde{S}\in\mathcal{Z}^n$ that differ by at most one example, we have* $\sup_{\mathbf{z}}\mathbb{E}_{\mathcal{A}}\big[f(\mathcal{A}(S);\mathbf{z})-f(\mathcal{A}(\widetilde{S});\mathbf{z})\big]\leq\epsilon$.

Definition 3.2 (On-average model stability) *Let $S=\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ and $S'=\{\mathbf{z}'_1,\ldots,\mathbf{z}'_n\}$ be two datasets drawn independently from $\mathbb{P}$. For each $i\in[n]:=\{1,\ldots,n\}$, we denote*

$$S^{(i)}=\{\mathbf{z}_1,\ldots,\mathbf{z}_{i-1},\mathbf{z}'_i,\mathbf{z}_{i+1},\ldots,\mathbf{z}_n\}.\tag{3.1}$$

Let $\epsilon > 0$. We say a (randomized) algorithm $\mathcal{A}$ is on-average ϵ-model stable if

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathcal{A}(S) - \mathcal{A}(S^{(i)})\|_2^2] \leq \epsilon^2.$$

Our stability analysis requires several assumptions on the convexity and smoothness.

Definition 3.3 *Let $g : \mathcal{W} \mapsto \mathbb{R}$, and $\lambda, \widetilde{G}, L \geq 0$.*

- *We say g is λ-strongly convex if $g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}')\rangle + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$. We say g is convex if the above inequality holds with $\lambda = 0$.*
- *We say g is $\widetilde{G}$-Lipschitz if $\|\nabla g(\mathbf{w})\|_2 \leq \widetilde{G}$ for all $\mathbf{w} \in \mathcal{W}$.*
- *We say g is L-smooth if $\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$.*

The following lemma shows that a stable algorithm would automatically enjoy a generalization guarantee. Therefore, it suffices to develop stability bounds to control the generalization behavior of a learning algorithm.

Lemma 3.4 (Lei and Ying 2020) *Let $L, \epsilon > 0$. Let $\mathcal{A}$ be an algorithm which is on-average ϵ-model stable. If for any $\mathbf{z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}, \mathbf{z})$ is L-smooth, then*

$$\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] \leq \frac{L\epsilon^2}{2} + \epsilon\big(2L\mathbb{E}[F_S(\mathcal{A}(S))]\big)^{\frac{1}{2}}.$$

In our stability analysis, we often use the self-bounding property of smooth functions to control gradients by function values.

Lemma 3.5 (Srebro et al. 2010) *Let $L > 0$. If $g : \mathcal{W} \mapsto \mathbb{R}$ is L-smooth and nonnegative, then $\|\nabla g(\mathbf{w})\|_2^2 \leq 2Lg(\mathbf{w})$ for any $\mathbf{w} \in \mathcal{W}$.*

3.3 Biased Stochastic Gradient Method

In this paper we consider biased gradient methods, where we build a possibly biased estimator $g(\mathbf{w}_t; S_{J_t})$ of the gradient and use it to search the next iterate. We consider a generalized bias, where the bias is measured w.r.t. a surrogate loss function $\tilde{f} : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$. The intuition is that while $g(\mathbf{w}_t; S_{J_t})$ may be a biased estimator of $\nabla F_S(\mathbf{w}_t)$, it can be an unbiased estimator (or a biased estimator with a smaller bias) of $\nabla \widetilde{F}_S(\mathbf{w}_t)$, where we define $\widetilde{F}_S$ by

$$\widetilde{F}_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\tilde{f}(\mathbf{w}; \mathbf{z}_i). \tag{3.2}$$

If $\tilde{f}$ is convex, then we may employ the convexity of $\tilde{f}$ and the small bias of $g(\mathbf{w}_t; S_{J_t})$ w.r.t. $\nabla \widetilde{F}_S(\mathbf{w}_t)$ to get better stability bounds. As we will show, this is the case for the Zeroth-order SGD. We can also just simply choose $\tilde{f} = f$, which is the case for SGD and Clipped-SGD. We consider a minibatch version of BSGMs, where at each iteration we randomly draw a batch of training examples to build gradient estimators. This minibatch scheme reduces the variance of the gradient estimator, and does not affect the bias (Bottou et al., 2018; Mücke

et al., 2019). For brevity, we always assume $m \leq n$ in the paper, where m is the batch size. For a set $J = \{j_1, \ldots, j_m\}$, we denote $S_J = \{\mathbf{z}_{j_1}, \ldots, \mathbf{z}_{j_m}\}$. We use the abbreviation $S_j = S_{\{j\}}$. Then our notation system implies (recall $S^{(i)}$ is defined in Eq. (3.1))

$$S_j^{(i)} = \mathbf{z}_j \text{ if } j \neq i \quad \text{and} \quad S_j^{(i)} = \mathbf{z}'_j \text{ if } j = i.$$

Definition 3.6 *Let $\mathbf{w}_1 = 0$ and $\{\eta_t\}$ be a sequence of step sizes. For any $t \in \mathbb{N}$, let $J_t := \{j_{t,1}, \ldots, j_{t,m}\}$ be independently drawn from the uniform distribution on $\{1, \ldots, n\}$ (with replacement). BSGMs update model as follows*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t g(\mathbf{w}_t; S_{J_t}), \tag{3.3}$$

where $g(\mathbf{w}_t; S_{J_t})$ is a (biased) estimator of $\nabla \widetilde{F}_S(\mathbf{w}_t)$ based on J_t.

In this paper, we always assume the estimator $g(\mathbf{w}_t; S_{J_t})$ has a sum structure, i.e., $g(\mathbf{w}_t; S_{J_t}) = \frac{1}{|S_{J_t}|} \sum_{\mathbf{z} \in S_{J_t}} g(\mathbf{w}_t; \mathbf{z})$, where $|S_{J_t}|$ denotes the cardinality of S_{J_t}.

4. Main Results

4.1 Convex Problems

In this section, we present stability bounds for general BSGMs applied to convex problems.

Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be two sequences produced by BSGMs on S and $S^{(i)}$, respectively, i.e., $\{\mathbf{w}_t\}$ is produced by Eq. (3.3) and $\{\mathbf{w}_t^{(i)}\}$ is produced by $\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} - \eta_t g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})$, where $g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})$ is an estimate of $\nabla \widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)})$. Theorem 4.1, to be proved in Section 6.1, provides a general result on the stability of stochastic gradient methods with (biased) gradient estimators. It shows that the stability of BSGMs depends on $\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2$ and the bias of the gradient estimators. For brevity, we introduce the notation

$$\Delta_i = \max_{t \in [T]} \left(\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big]\right)^{\frac{1}{2}}, \quad \forall i \in [n]. \tag{4.1}$$

Theorem 4.1 *Assume for any $\mathbf{z}$, $\mathbf{w} \mapsto \tilde{f}(\mathbf{w}; \mathbf{z})$ is convex. Let*

$$b_t = \mathbb{E}_t[g(\mathbf{w}_t; S_{J_t})] - \nabla \widetilde{F}_S(\mathbf{w}_t) \quad \text{and} \quad b_t^{(i)} = \mathbb{E}_t[g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})] - \nabla \widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)}), \tag{4.2}$$

where $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathbf{w}_t, \mathbf{w}_t^{(1)}, \ldots, \mathbf{w}_t^{(n)}, S, S']$. Then, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \leq \frac{2}{n}\sum_{i=1}^n \sum_{t=1}^T \frac{\eta_t^2}{m} \mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2\big] \\ + \frac{2}{n}\sum_{i=1}^n \sum_{t=1}^T \eta_t^2 \mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2]\Big) + \frac{8}{n}\sum_{i=1}^n \Big(\sum_{t=1}^T \eta_t \mathfrak{C}_{t,i}\Big)^2,\end{aligned}$$

where we introduce

$$\mathfrak{C}_{t,i} := \Big(\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] + 4\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}}. \tag{4.3}$$

In Theorem 4.1, the notation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathbf{w}_t, \mathbf{w}_t^{(1)}, \ldots, \mathbf{w}_t^{(n)}, S, S']$ means the expectation conditioned on $\mathbf{w}_t, \mathbf{w}_t^{(1)}, \ldots, \mathbf{w}_t^{(n)}, S, S'$. Note that b_t does not depend on $\mathbf{w}_t^{(1)}, \ldots, \mathbf{w}_t^{(n)}, S'$. We use $\mathbb{E}_t[\cdot]$ to get a unifying notation applied to all $b_t, b_t^{(i)}, i \in [n]$. As we will see for Zeroth-order SGD, this conditional expectation captures the randomness of both the minibatch J_t and the random direction used to build gradient estimators. Theorem 4.1 shows the stability depends on both the sensitivity of the bias and the sensitivity of gradient estimators.

Remark 4.2 For the upper bound in Theorem 4.1, the term $\mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2\big]$ is related to the variance, the term $\mathfrak{C}_{t,i}$ is related to the bias, and $\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2]$ is closely related to the empirical risk. For example, if we consider SGD, then $\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})] = \nabla F_S(\mathbf{w}_t) - \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})$. One can see that the bias part, as compared to the variance, is more critical to stability. Indeed, if we take $\eta_t = \eta$ and assume $\mathfrak{C}_{t,i} \lesssim \mathfrak{C}_b, \mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2\big] \lesssim \mathfrak{C}_v$ for some $\mathfrak{C}_b, \mathfrak{C}_v \geq 0$. Then, the bias part becomes $8T^2\eta^2\mathfrak{C}_b^2$, while the variance part becomes $2T\eta^2\mathfrak{C}_v/m$. That is, the bias part involves an additional factor of Tm compared to the variance. One can also see how the parameters affect the bias and variance. For example, the use of a batch of size m reduces the variance by a factor of m, and does not affect the bias. This is consistent with the intuition that an average of independent variables decreases the variance but does not change the bias. Then, our stability analysis always benefits from a large batch size. Furthermore, both the bias and the variance benefit from small step sizes. Therefore, one can always choose sufficiently small step sizes to get good stability and generalization. However, small step sizes may affect the convergence. Therefore, one needs to balance the stability and optimization by choosing suitable step sizes.

We now impose a regularity assumption on $\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2$ and the bias, and derive a stability bound which is more convenient to apply. We assume $A, \bar{A}$ and $\widetilde{A}$ are not random variables. The proof is given in Section 6.1.

Theorem 4.3 (Stability bounds) *Let assumptions in Theorem 4.1 hold. Assume there exist $A, \widetilde{A}, \bar{A} \geq 0$ and $B_{t,i}, \widetilde{B}_{t,i}, \bar{B}_{t,i} \geq 0$ such that*

$$\mathbb{E}\big[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2\big] \leq \mathbb{E}\big[A\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + B_{t,i}\big], \tag{4.4}$$

$$\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2] \leq \mathbb{E}\big[\widetilde{A}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \widetilde{B}_{t,i}\big], \tag{4.5}$$

$$\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] \leq \mathbb{E}\big[\bar{A}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{B}_{t,i}\big]. \tag{4.6}$$

If $2A\sum_{t=1}^T \frac{\eta_t^2}{m} + 2\widetilde{A}\sum_{t=1}^T \eta_t^2 + 8\bar{A}T\sum_{t=1}^T \eta_t^2 \leq 1/2$, then

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \leq \frac{4}{n}\sum_{i=1}^n\sum_{t=1}^T \frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m} + \frac{4}{n}\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}[\widetilde{B}_{t,i}] + \frac{16\sum_{t=1}^T \eta_t^2}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbb{E}\Big[\bar{B}_{t,i} + \frac{4\|\nabla\tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\|_2^2}{n^2}\Big].$$

One can directly derive generalization bounds by plugging the stability bounds into Lemma 3.4. We omit the discussion for brevity. The conditions (4.4), (4.5), and (4.6) take a similar form:

the upper bounds involve a constant and a linear function of the distance between the arguments. Intuitively, $\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2$ would be small and therefore these conditions tolerate large $A, \widetilde{A}, \bar{A}$ (as compared to $B_{t,i}, \widetilde{B}_{t,i}, \bar{B}_{t,i}$) to get good stability. We refer to these assumptions as *generalized Lipschitz-type* assumption. To apply Theorem 4.3 to an algorithm $\mathcal{A}$, it suffices to estimate the corresponding $A, \widetilde{A}, \bar{A}$ and $B_{t,i}, \widetilde{B}_{t,i}, \bar{B}_{t,i}$. In Section 5, we will provide explicit estimates of these parameters for SGD, Zeroth-order SGD and Clipped-SGD.

4.2 Strongly Convex Problems

In this subsection, we show that better stability bounds can be derived for strongly convex problems. The proof of Theorem 4.4 and Corollary 4.5 can be found in Section 6.2.

Theorem 4.4 *Assume for any* $\mathbf{z}$, $\mathbf{w} \mapsto \tilde{f}(\mathbf{w}; \mathbf{z})$ *is* λ*-strongly convex. Let Eq.* (4.4), (4.5), (4.6) *hold. If* $A\eta_t/m + \widetilde{A}\eta_t + 2\sqrt{2}\bar{A}^{\frac{1}{2}} + 2\lambda/n \leq \lambda/2$ *for any* $t \in [T]$*, then*

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_{t+1,i}^2 \leq \frac{1}{n}\sum_{k=1}^{t}\eta_k^2\Big(\prod_{k'=k+1}^{t}(1-\lambda\eta_{k'})\Big)\sum_{i=1}^{n}\Big(\frac{\mathbb{E}[B_{k,i}]}{m} + \mathbb{E}[\widetilde{B}_{k,i}]\Big) + \frac{4}{n\lambda}\sum_{k=1}^{t}\eta_k\Big(\prod_{k'=k+1}^{t}(1-\lambda\eta_{k'})\Big)\sum_{i=1}^{n}\Big(\mathbb{E}\big[\bar{B}_{k,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2]/n^2\Big).$$

For strongly convex problems, we often choose step sizes $\eta_t = 1/(\lambda(t+a))$, for which Theorem 4.4 implies the following stability bounds.

Corollary 4.5 *Let conditions in Theorem 4.4 hold and* $\eta_t = 1/(\lambda(t+a))$ *for some* $a \geq 0$*. Then*

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_{t+1,i}^2 \leq \frac{1}{\lambda^2 nt}\sum_{k=1}^{t}\frac{1}{k}\sum_{i=1}^{n}\Big(\frac{\mathbb{E}[B_{k,i}]}{m} + \mathbb{E}[\widetilde{B}_{k,i}]\Big) + \frac{4}{\lambda^2 nt}\sum_{k=1}^{t}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\bar{B}_{k,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2]/n^2\Big).$$

Remark 4.6 If we assume

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[B_{k,i}] \lesssim \frac{1}{n}, \quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\widetilde{B}_{k,i} + \bar{B}_{k,i}] \lesssim \frac{1}{n^2}, \quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2] \lesssim 1, \tag{4.7}$$

then Corollary 4.5 implies that

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_{t+1,i}^2 \lesssim \frac{1}{\lambda^2 t}\sum_{k=1}^{t}\frac{1}{k}\Big(\frac{1}{mn} + \frac{1}{n^2}\Big) + \frac{1}{\lambda^2 t}\sum_{k=1}^{t}\frac{1}{n^2} \lesssim \frac{\log t}{\lambda^2 tmn} + \frac{1}{\lambda^2 n^2}.$$

This shows that BSGM is on-average ϵ-model stable with $\epsilon \lesssim \frac{\log^{1/2} t}{\lambda(tmn)^{1/2}} + \frac{1}{\lambda n}$. Therefore, BSGM is stable no matter how many iterations are taken for strongly convex problems, which is consistent with the existing stability analysis of SGD (Hardt et al., 2016). Note that this result requires assumptions $A\eta_t/m + \widetilde{A}\eta_t + 2\sqrt{2}\bar{A}^{\frac{1}{2}} + 2\lambda/n \leq \lambda/2$ and Eq. (4.7). The assumption $A\eta_t/m + \widetilde{A}\eta_t + 2\sqrt{2}\bar{A}^{\frac{1}{2}} + 2\lambda/n \leq \lambda/2$ holds if we choose a sufficiently small step size and $\bar{A} \lesssim \lambda^2$, which, as we will see, is the case for SGD, Zeroth-order SGD, and Clipped-SGD. Indeed, we can choose $\bar{A} = 0$ for both SGD and Zeroth-order SGD. As we will see

in Lemma 5.1 for SGD, and Lemma 5.14, Lemma 5.15 and Lemma 5.16 for Clipped-SGD, Eq. (4.7) holds if we can show $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] \lesssim 1$, which holds naturally since $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] \leq \frac{2L}{n}\sum_{i=1}^n \mathbb{E}[f(\mathbf{w}_t;\mathbf{z}_i)] = 2L\mathbb{E}[F_S(\mathbf{w}_t)]$ (Lemma 3.5), the latter of which can be bounded by optimization error analysis. According to Lemma 5.5, Lemma 5.6 and Lemma 5.7, Eq. (4.7) also holds for Zeroth-order SGD if the smoothing parameter μ is sufficiently small.

Remark 4.7 Corollary 4.5 shows that BSGMs are resistent to overfitting for strongly convex problems. It is interesting to investigate whether this phenomenon holds if we further relax the strong convexity assumption. Two popular assumptions in this direction include the exp-concavity assumption and the Polyak-Łojasiewicz (PL) condition on F_S. Here, we say F_S is α-exp-concave if $\nabla^2 F_S(\mathbf{w}) \succeq \alpha \nabla F_S(\mathbf{w})\nabla F_S(\mathbf{w})^\top$ ($A \succeq B$ means that $A - B$ is positive semi-definite), and we say F_S satisfies the α-PL condition if $F_S(\mathbf{w}) - \inf_{\mathbf{w}} F_S(\mathbf{w}) \leq \alpha^{-1}\|\nabla F_S(\mathbf{w})\|_2^2$. The stability of the empirical risk minimization (ERM) has been studied for exp-concave problems (Gonen and Shalev-Shwartz, 2017; Koren and Levy, 2015), while the stability of any algorithm converging to global optima has been developed under a PL condition (Charles and Papailiopoulos, 2018; Lei and Ying, 2021). A key property for strongly convex problems is that

$$\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t^{(i)})\rangle \geq \lambda \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2, \tag{4.8}$$

by which we can show the following key inequality (Eq. (6.4)) for the analysis with strongly convex problems

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \leq \big(1 - 2\lambda\eta_t(n-1)/n\big)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + C_t, \tag{4.9}$$

where C_t is a linear function of $\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2$. As a comparison, for exp-concave functions, we only have

$$\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t^{(i)})\rangle \geq \frac{\alpha}{2}\big((\mathbf{w}_t - \mathbf{w}_t^{(i)})^\top \nabla F_S(\mathbf{w}_t)\big)^2 + \frac{\alpha}{2}\big((\mathbf{w}_t - \mathbf{w}_t^{(i)})^\top \nabla F_S(\mathbf{w}_t^{(i)})\big)^2. \tag{4.10}$$

A key difference between strong convexity and exp-concavity is that Eq. (4.8) involves $\|\mathbf{w} - \mathbf{w}_t^{(i)}\|_2^2$, while Eq. (4.10) involves the dot product between $\mathbf{w}_t - \mathbf{w}_t^{(i)}$ and gradients. It is not clear to us how to use Eq. (4.10) to build an inequality similar to Eq. (4.9) for exp-concave problems. Similarly, it is also not clear to us how to build an inequality similar to Eq. (4.9) by using the PL condition. Therefore, it remains an open problem to us on how to use either the exp-concavity or PL condition to improve stability bounds in Theorem 4.3.

On the other hand, it was shown that one can transform the optimization error bounds to excess risk bounds under the PL condition (Charles and Papailiopoulos, 2018; Lei and Ying, 2021). Therefore, one can directly use the existing convergence analysis of BSGMs to derive excess risk bounds under the PL condition.

5. Applications

In this section, we show the effectiveness of our general stability bounds in Section 4 by applying it to SGD, Zeroth-order SGD, and Clipped-SGD. We summarize our stability

Table 1: Comparison of stability bounds for SGD, Zeroth-order SGD and Clipped-SGD for convex and smooth problems. For simplicity, we assume $m = 1, \eta_t = \eta$ and $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t)] \lesssim 1$. For Zeroth-order SGD, we assume $K \asymp d$ and choose sufficiently small μ, i.e., $\mu \lesssim \frac{\widetilde{G}}{nLd^{3/2}}$ in Nikolakakis et al. (2022a) and $\mu \lesssim \min\big\{\frac{F^{1/2}(\mathbf{w}^*)}{(Ld^3)^{1/2}}, \frac{F^{1/2}(\mathbf{w}^*)}{(nL)^{1/2}d}\big\}$ in Eq. (5.8). We take $\tau \asymp GT^{\frac{1}{p}}$ for Clipped-SGD. Here, $\widetilde{G}$ is the Lipscthizness parameter, G is the parameter in either Assumption 5.2 or Assumption 5.3. In the "Type" column, "U" means the uniform stability, while "M" means the on-average model stability.

Algorithm	Reference	Step size	Type	Stability bound
SGD	Hardt et al. (2016)	$\eta_t \lesssim 1/L$	U	$\eta T\widetilde{G}^2/n$
	Lei and Ying (2020)	$\eta_t \lesssim 1/L$	M	$\frac{(LT)^{\frac{1}{2}}\eta}{n^{1/2}} + \frac{L^{\frac{1}{2}}\eta T}{n}$
	Eq. (5.2)	$L^2\sum_{t=1}^{T}\eta_t^2 \lesssim 1$	M	$\frac{(LT)^{\frac{1}{2}}\eta}{n^{1/2}} + \frac{L^{\frac{1}{2}}\eta T}{n}$
Zeroth-order SGD	Nikolakakis et al. (2022a)	$\eta_t = \widetilde{O}(1/(TL))$	U	$\widetilde{G}^2/n$
	Liu et al. (2024)	$\eta_t \lesssim 1/t$	U	$\widetilde{G}^2T/n$
	Eq. (5.8)	$L^2\sum_{t=1}^{T}\eta_t^2 \lesssim 1$	M	$\frac{(LT)^{\frac{1}{2}}\eta}{n^{1/2}} + \frac{L^{\frac{1}{2}}\eta T}{n}$
Clipped SGD	Eq. (5.17)	$L^2\sum_{t=1}^{T}\eta_t^2 \lesssim 1$	M	$\frac{(LT)^{\frac{1}{2}}\eta}{n^{1/2}} + \frac{L^{\frac{1}{2}}\eta T}{n}$
	Eq. (7.48)	$\eta_t \asymp G^{\frac{p}{1-p}}n^{-\frac{1}{2p-2}}$	M	$\frac{(LT)^{\frac{1}{2}}\eta}{n^{1/2}} + \frac{L^{\frac{1}{2}}\eta T}{n}$

Table 2: Excess risk bounds of Zeroth-order SGD and Clipped-SGD for convex and smooth problems. For brevity, we assume $m = 1, \eta_t = \eta$ and indicate the dependency on n and L. For Zeroth-order SGD (ZSGD), we take $K \asymp d$ and a sufficiently small μ, i.e., $\mu \lesssim \min\big\{\frac{1}{d(nL)^{1/2}}, \frac{1}{L^{1/2}d^{3/2}}, \frac{\|\mathbf{w}^*\|_2^{1/2}}{(Ln)^{1/4}d^{1/2}}, \frac{\|\mathbf{w}^*\|_2^{1/2}}{L^{3/4}d^{3/2}\eta^{1/2}n^{1/4}}\big\}$. We take $\tau \asymp GT^{\frac{1}{p}}$ for Clipped-SGD.

Algorithm	Reference	Step size	Iteration number	Excess risk bound
ZSGD	Remark 5.12	$\frac{1}{\sqrt{n}L^{3/2}\|\mathbf{w}^*\|_2}$	$Ln\|\mathbf{w}^*\|_2^2$	$\frac{L^{1/2}}{n^{1/2}}$
Clipped SGD	Remark 5.20	$G^{\frac{p}{1-p}}n^{-\frac{1}{2p-2}}L^{\frac{1}{2p-2}}$	$n^{\frac{p}{2p-2}}G^{\frac{p}{p-1}}L^{\frac{p}{2-2p}}$	$\frac{L^{1/2}}{n^{1/2}}$
	Theorem 5.25	$G^{\frac{p}{1-p}}n^{-\frac{1}{2p-2}}L^{\frac{1}{2p-2}}$	$n^{\frac{p}{2p-2}}G^{\frac{p}{p-1}}L^{\frac{p}{2-2p}}$	$\widetilde{O}\big(\frac{L^{1/2}}{n^{1/2}}\big)$

and excess risk bounds in Table 1 and Table 2, respectively. Note our discussions require $L^2\sum_{t=1}^T \eta_t^2 \lesssim 1$, which is a standard assumption. For example, an assumption $\sum_{t=1}^\infty \eta_t^2 < \infty$ is often imposed for the almost sure convergence of SGD (Bottou et al., 2018).

5.1 Stochastic Gradient Descent

In this section, we apply Theorem 4.3 to derive stability bounds for SGD. The aim is to verify the effectiveness of our general result by showing it recovers existing results on the classical SGD, which has been widely studied in the literature (Hardt et al., 2016; Lei and Ying, 2020; Kuzborskij and Lampert, 2018). For SGD, $\mathbf{w}_{t+1}$ is produced by Eq. (3.3) with

$$g(\mathbf{w}_t; S_{J_t}) = \frac{1}{m}\sum_{k=1}^m \nabla f(\mathbf{w}_t; \mathbf{z}_{j_{t,k}}). \tag{5.1}$$

It is clear that $g(\mathbf{w}_t; S_{J_t})$ is an unbiased estimate of $\nabla F_S(\mathbf{w}_t)$ and therefore we have $b_t = 0$. The following lemma, to be proved in Section 7.1, shows that Eq. (4.4), (4.5) and (4.6) hold.

Lemma 5.1 *Assume for all* $\mathbf{z}$*, the map* $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ *is* L*-smooth. If we choose the gradient estimator in Eq.* (5.1) *and* $\tilde{f} = f$*, then Eq.* (4.4)*,* (4.5) *and* (4.6) *hold with* $\bar{A} = 0, \bar{B}_{t,i} = 0$*,*

$$A = L^2, \quad B_{t,i} = 4\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2/n,$$
$$\widetilde{A} = 3L^2/2, \quad \widetilde{B}_{t,i} = 12\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2/n^2.$$

We now combine Lemma 5.1 and Theorem 4.3 to derive the following corollary.

Corollary 5.2 (Stability of SGD) *Assume for all* $\mathbf{z}$*, the map* $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ *is convex and* L*-smooth. If* $(3 + \frac{2}{m})L^2\sum_{t=1}^T \eta_t^2 \le 1/2$*, then*

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \frac{L}{mn}\sum_{t=1}^T \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)] + \frac{L\sum_{t=1}^T \eta_t^2}{n^2}\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)]. \tag{5.2}$$

Furthermore, if $\eta_t = \eta$*, then*

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \Big(\frac{L\eta^2}{mn} + \frac{L\eta^2 T}{n^2}\Big)\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta} + TF(\mathbf{w}^*)\Big). \tag{5.3}$$

Remark 5.3 If $m = 1$, the stability bound $\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \frac{L+LT/n}{n}\sum_{t=1}^T \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)]$ was developed in Lei and Ying (2020). If $m = 1$ and $\eta_t = \eta$, it is clear that this bound is of the same order of Eq. (5.2). Eq. (5.3) gives explicit stability bounds depending only on problem parameters and step sizes. While Theorem 4.3 is developed for a general stochastic gradient method with a biased estimator, it recovers the existing bound when applied to SGD. This shows the effectiveness of our general approach since we do not incorporate the special update rule of SGD in deriving Theorem 4.3.

5.2 Zeroth-order SGD

In this section, we consider Zeroth-order SGD, which has found wide applications in solving black-box optimization problems. As a derivative-free optimization method, it is especially desirable for applications where explicit gradient calculations are computationally infeasible, expensive, or impossible (Duchi et al., 2015; Nesterov and Spokoiny, 2017). The basic idea of derivative-free optimization is to approximate gradient by a finite difference, which only uses function values. In particular, for Zeroth-order SGD, $\mathbf{w}_{t+1}$ is produced by Eq. (3.3) with ($\mathcal{N}(0, I_d)$ denotes the standard Gaussian random variable in $\mathbb{R}^d$)

$$g(\mathbf{w}_t; S_{J_t}) = \frac{1}{mK}\sum_{k=1}^{m}\sum_{l=1}^{K}\frac{f(\mathbf{w}_t + \mu\mathbf{u}_{t_k,l}; \mathbf{z}_{j_{t,k}}) - f(\mathbf{w}_t; \mathbf{z}_{j_{t,k}})}{\mu}\mathbf{u}_{t_k,l},\ \mathbf{u}_{t_k,l} \sim \mathcal{N}(0, I_d), \quad (5.4)$$

where $\mu \in \mathbb{R}^+$ is a smoothing parameter determining the step size in approximating the gradient of f. For simplicity, we omit the dependency on $\mathbf{u}_{t_k}$ in the notation of $g(\mathbf{w}_t; S_{J_t})$. As indicated in Duchi et al. (2015), $g(\mathbf{w}; S_j)$ is a biased estimator of $\nabla f(\mathbf{w}; \mathbf{z}_j)$. Indeed, the bias satisfies the following inequality if f is L-smooth (note $S_j = \mathbf{z}_j$)

$$\left\|\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}; S_j)] - \nabla f(\mathbf{w}; \mathbf{z}_j)\right\|_2 \lesssim \mu L d^{\frac{3}{2}}, \quad (5.5)$$

which shows that Eq. (4.6) holds with $\bar{A} = 0$ and $\bar{B}_{t,i} \lesssim L^2\mu^2 d^3$. However, directly applying Theorem 4.3 with $\tilde{f} = f$ would lead to a bound involving the following term

$$\frac{\sum_{t=1}^T \eta_t^2}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\big[\bar{B}_{t,i}\big] \lesssim \frac{\sum_{t=1}^T \eta_t^2}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} L^2\mu^2 d^3 = L^2\mu^2 d^3 T\sum_{t=1}^{T}\eta_t^2, \quad (5.6)$$

which is not desirable due to the linear dependency of T and the independence of n. The problem with the above analysis is that it introduces a large bias. Recall that in Theorem 4.1 the bias is defined in terms of a surrogate function $\tilde{f}$. If we can build a convex and smooth surrogate function $\tilde{f}$ such that the bias $\mathbb{E}_t[g(\mathbf{w}_t; S_{J_t})] - \mathbb{E}_{\mathbf{z}}[\nabla\tilde{f}(\mathbf{w}; \mathbf{z})]$ is small, then we can apply Theorem 4.1 with such $\tilde{f}$ to get stronger stability bounds since the bias is a critical term in the stability bound. Our key idea is to show that we can actually get a zero bias by considering the following surrogate function

$$\tilde{f}(\mathbf{w}; \mathbf{z}) = \mathbb{E}_{\mathbf{u}}\big[f(\mathbf{w} + \mu\mathbf{u}; \mathbf{z})\big], \quad (5.7)$$

where $\mathbf{u} \sim \mathcal{N}(0, I_d)$ and we omit the parameter μ in $\tilde{f}$ for brevity. The following lemma shows that $\tilde{f}$ preserves the convexity and smoothness of f. Furthermore, it shows that g is an *unbiased* estimator of $\nabla\widetilde{F}_S(\mathbf{w}_t)$. The last property follows directly from Eq. (21) in Nesterov and Spokoiny (2017).

Lemma 5.4 (Nesterov and Spokoiny 2017) *Let $\tilde{f}$ be defined in Eq. (5.7).*

- *Assume for any $\mathbf{z}$, $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is convex. Then for any $\mathbf{z}$, $\mathbf{w} \mapsto \tilde{f}(\mathbf{w}; \mathbf{z})$ is convex.*
- *Assume for any $\mathbf{z}$, $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth. Then for any $\mathbf{z}$, $\mathbf{w} \mapsto \tilde{f}(\mathbf{w}; \mathbf{z})$ is L-smooth.*

- *Let g be defined in Eq. (5.4). Then, $\mathbb{E}[g(\mathbf{w}_t; S_{J_t})|\mathbf{w}_t] = \nabla \widetilde{F}_S(\mathbf{w}_t)$, where $\widetilde{F}_S$ is defined in Eq. (3.2) with $\tilde{f}$ defined in Eq. (5.7). It also holds that $\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}; S_j)] = \nabla \tilde{f}(\mathbf{w}; S_j)$ for any $\mathbf{w} \in \mathcal{W}$ and $j \in [n]$.*

According to Lemma 5.4, for the above g and $\tilde{f}$, we have $b_t = b_t^{(i)} = 0$, where the bias b_t and $b_t^{(i)}$ are defined in Eq. (4.2). Furthermore, the surrogate function $\tilde{f}$ preserves the smoothness and convexity, which we will exploit in developing a much better stability bound than that in Eq. (5.6).

To apply Theorem 4.3, we will show that Eq. (4.4) and Eq. (4.5) hold with appropriate $A, \widetilde{A}$ and $B_{t,i}, \widetilde{B}_{t,i}$, which are constructed in Lemma 5.5 and Lemma 5.6. The proofs of results in this subsection are given in Section 7.2.

Lemma 5.5 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.4). Assume $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth. Then, for $\tilde{f}$ defined in Eq. (5.7), Eq. (4.4) holds with $A = \big(1 + \frac{2(d+2)}{K}\big)L^2$ and*

$$B_{t,i} = \frac{8(d+2)\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2}{Kn} + \frac{2\mu^2 L^2 d(d+2)(d+4)}{K} + \frac{4}{n}\big\|\nabla \tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\big\|_2^2.$$

Lemma 5.6 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.4). Assume $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth. Then, for $\tilde{f}$ defined in Eq. (5.7), Eq. (4.5) holds with $\widetilde{A} = \big(\frac{3}{2} + \frac{4(d+2)}{K}\big)L^2$ and*

$$\widetilde{B}_{t,i} = \frac{8(d+2)\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2}{Kn^2} + \frac{2\mu^2 L^2 d(d+2)(d+4)}{K} + \frac{12}{n^2}\big\|\nabla \tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\big\|_2^2.$$

The stability bound in Theorem 4.3 involves the term $\|\nabla \tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\|_2^2$, which are related to $\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2$ by the following lemma.

Lemma 5.7 *Let $\tilde{f}(\mathbf{w}; \mathbf{z})$ be defined in Eq. (5.7). Assume $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth. Then*

$$\mathbb{E}\Big[\big\|\nabla \tilde{f}(\mathbf{w}_t; \mathbf{z}_i)\big\|_2^2\Big] \leq 2\mathbb{E}\Big[\,\|\nabla f\,(\mathbf{w}_t; \mathbf{z}_i)\|_2^2\,\Big] + \frac{\mu^2 L^2}{2} d(d+2)(d+4).$$

Now we apply Lemma 5.5, Lemma 5.6, Lemma 5.7 and Theorem 4.3 to derive Corollary 5.8.

Corollary 5.8 (Stability of Zeroth-order SGD) *Assume for all $\mathbf{z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is convex and L-smooth. If $\big(3 + \frac{8(d+2)}{K} + \frac{2}{m} + \frac{4(d+2)}{mK}\big)L^2 \sum_{t=1}^T \eta_t^2 \leq 1/2$, then*

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \frac{L}{mn}\Big(1 + \frac{d}{K}\Big)\sum_{t=1}^T \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)] + \frac{L\sum_{t=1}^T \eta_t^2}{n^2}\sum_{t=1}^T \mathbb{E}\big[F_S(\mathbf{w}_t)\big] + \frac{\mu^2 L^2 d^3}{mn}\sum_{t=1}^T \eta_t^2 + \frac{\mu^2 L^2 d^3}{K}\sum_{t=1}^T \eta_t^2 + \frac{\mu^2 L^2 d^3 T \sum_{t=1}^T \eta_t^2}{n^2}. \quad (5.8)$$

Furthermore, if $\eta_t = \eta \leq (2L(1 + d/K))^{-1}$, we have

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \Big(\frac{L\eta^2}{mn}\Big(1 + \frac{d}{K}\Big) + \frac{LT\eta^2}{n^2}\Big)\Big(TF(\mathbf{w}^*) + \eta^{-1}\|\mathbf{w}^*\|_2^2 + Ld\mu^2 T + \eta T\Big(L(1 + \frac{d}{K})F(\mathbf{w}^*) + \mu^2 L^2 d^3\Big)\Big) + \frac{T\eta^2\mu^2 L^2 d^3}{mn} + \frac{T\eta^2\mu^2 L^2 d^3}{K} + \frac{\mu^2 L^2 d^3 T^2 \eta^2}{n^2}. \quad (5.9)$$

Remark 5.9 To get a clean comparison with SGD, we assume $n \lesssim T$ and $L\eta d/K \lesssim 1$. In this case, Eq. (5.9) implies (note $\eta \lesssim 1/L$)

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2 \lesssim \Big(\frac{L\eta^2 d}{mnK}+\frac{LT\eta^2}{n^2}\Big)\Big(TF(\mathbf{w}^*)+\frac{\|\mathbf{w}^*\|_2^2}{\eta}+Ld\mu^2T+L^2\eta T\mu^2d^3\Big) + \frac{T\eta^2\mu^2L^2d^3}{\min\{mn,K\}}+\frac{\mu^2L^2d^3T^2\eta^2}{n^2}. \quad (5.10)$$

If we choose sufficiently small μ (e.g., $\mu \lesssim \min\big\{\frac{F^{1/2}(\mathbf{w}^*)}{(Ld^3)^{1/2}}, \frac{F^{1/2}(\mathbf{w}^*)}{(mnL)^{1/2}d}\big\}$), then Eq. (5.10) implies

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2 \lesssim \Big(\frac{L\eta^2 d}{mnK}+\frac{LT\eta^2}{n^2}\Big)\Big(TF(\mathbf{w}^*)+\frac{\|\mathbf{w}^*\|_2^2}{\eta}\Big). \quad (5.11)$$

This matches the stability bound of SGD in Eq. (5.3) if $d \asymp K$.

We get this stability bound by taking a surrogate function in Eq. (5.7). If we choose $\tilde{f}=f$, then Eq. (5.6) and Theorem 4.3 shows the stability bound involves $L^2\mu^2d^3\sum_{t=1}^T\eta_t^2T$. As a comparison, our analysis involves a much smaller term $L^2\mu^2d^3\sum_{t=1}^T\eta_t^2\big(\frac{1}{mn}+\frac{1}{K}+\frac{T}{n^2}\big)$. We achieve this improvement by considering the surrogate function $\tilde{f}$, by which we can fully use the zero bias of f as an estimator of $\widetilde{F}_S$ as well as the convexity of $\tilde{f}$. Note a property of our analysis is that the surrogate function is just introduced to simplify the stability analysis, and we do not need to build it in the implementation. This makes our framework flexible since we can try any surrogate function as long as it is convex and smooth.

Remark 5.10 (Comparison with existing analysis) We now compare Corollary 5.8 with existing stability analysis of Zeroth-order SGD (Nikolakakis et al., 2022a). For simplicity, we assume $m=1$. The work (Nikolakakis et al., 2022a) considered two cases to control $\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2$. If $j_{t,1}\neq i$, then the update of $\mathbf{w}_{t+1}$ and $\mathbf{w}_{t+1}^{(i)}$ use the same example $\mathbf{z}_{j_{t,1}}$. Their basic idea is to consider the following decomposition

$$\big\|\mathbf{w}_t-\eta_t g(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-\mathbf{w}_t^{(i)}+\eta_t g(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2 \le \big\|\mathbf{w}_t-\eta_t\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-\mathbf{w}_t^{(i)}+\eta_t\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2 + \eta_t\big\|\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-g(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})+g(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2. \quad (5.12)$$

The discussions in Nikolakakis et al. (2022a) assume

$$\big\|\mathbf{w}_t-\eta_t\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-\mathbf{w}_t^{(i)}+\eta_t\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2 \le \alpha\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2$$

for some $\alpha\ge 1$ (if f is convex and smooth, then $\alpha=1$ (Hardt et al., 2016). Otherwise $\alpha>1$). It was shown that (Nikolakakis et al., 2022a)

$$\big\|\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-g(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})+g(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2 \lesssim \mu L d^{\frac{3}{2}}+\sqrt{d/K}L\|\mathbf{w}_t-\mathbf{w}_t'\|_2.$$

Combining the above bound and Eq. (5.12) yields

$$\big\|\mathbf{w}_t-\eta_t g(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})-\mathbf{w}_t^{(i)}+\eta_t g(\mathbf{w}_t^{(i)};\mathbf{z}_{j_{t,1}})\big\|_2 \le (\alpha+C_1\eta_t\sqrt{d/K}L)\|\mathbf{w}_t-\mathbf{w}_t'\|_2+C_2\mu L\eta_t d^{\frac{3}{2}},$$

where C_1 and C_2 are two universal positive constants. In the second case with $j_{t,1} = i$, the update of $\mathbf{w}_{t+1}$ uses $\mathbf{z}_i$ and the update of $\mathbf{w}_{t+1}^{(i)}$ uses $\mathbf{z}_i'$. Under a $\widetilde{G}$-Lipschitzness assumption, it was shown that (Nikolakakis et al., 2022a)

$$\left\|\mathbf{w}_t - \eta_t g(\mathbf{w}_t; \mathbf{z}_i) - \mathbf{w}_t^{(i)} + \eta_t g(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\right\|_2 \leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + C_1 \eta_t \sqrt{d/K}\widetilde{G} + C_2 \mu L \eta_t d^{\frac{3}{2}}.$$

We combine the above two cases together, and use the fact that the event $j_{t,1} = i$ happens with the probability $1/n$ to get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq (\alpha + C_1 \eta_t \sqrt{d/K} L)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t'\|_2] + C_2 \mu L \eta_t d^{\frac{3}{2}} + C_1 \eta_t \sqrt{d/K}\widetilde{G}/n.$$

This analysis shows that $\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2]$ expands by a factor of $\alpha + C_1 \eta_t \sqrt{d/K} L$ as compared to $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2]$. Even if f is convex, the expansion factor becomes $1 + C_1 \eta_t \sqrt{d/K} L$ for which the step size needs to decay with the rate of $\eta_t \lesssim 1/t$ to get controllable stability bounds (Hardt et al., 2016). To our knowledge, all the existing stability analysis of Zeroth-order SGD with random gradient estimation is based on the decomposition in Eq. (5.12) (Nikolakakis et al., 2022a; Liu et al., 2024), which cannot fully exploit the convexity of f to get a nonexpansiveness of the stability bounds when choosing the same example for updating. Therefore, the existing analysis of Zeroth-order SGD with random gradient estimation requires a fast-decaying step size $\eta_t \lesssim 1/t$. As a comparison, we consider the surrogate loss $\tilde{f}$ and our analysis fully exploits the convexity of $\tilde{f}$ to get an improved stability bound. In this way, we do not need to consider the bias since $g(\mathbf{w}_t; S_j)$ is an *unbiased* estimator of $\nabla \tilde{f}(\mathbf{w}_t; S_j)$ (Nesterov and Spokoiny, 2017). This strategy allows us to get desirable stability bounds for the standard step size $\eta_t \asymp 1/\sqrt{T}$.

It should be mentioned that the work (Theorem 3 in Liu et al. (2024)) also considered Zeroth-order SGD with deterministic coordinate-wise gradient estimation ($g(\mathbf{w}; \mathbf{z}) = \sum_{j=1}^d \frac{f(\mathbf{w}+\mu \mathbf{e}_j; \mathbf{z}) - f(\mathbf{w}; \mathbf{z})}{\mu} \mathbf{e}_j$ with $\mathbf{e}_j$ denoting the j-th elementary basis vector), and derived stability bounds under the condition that $\eta_t \leq 2/L$ and $\mu \lesssim 1/(nd)$. The difference with our work is summarized as follows. First, we consider Zeroth-order SGD with random gradient estimation which is more popular than Zeroth-order SGD with deterministic coordinate-wise gradient estimation. Second, we consider on-average model stability and we do not require a Lipschitzness assumption, while the work (Liu et al., 2024) considered the uniform stability and imposed a Lipschitzness assumption. Furthermore, our stability analysis requires $\mu \lesssim \min\big\{\frac{F^{1/2}(\mathbf{w}^*)}{(Ld^3)^{1/2}}, \frac{F^{1/2}(\mathbf{w}^*)}{(nL)^{1/2}d}\big\}$, while the work (Liu et al., 2024) required $\mu \lesssim 1/(nd)$.

We are now ready to present excess risk bounds for Zeroth-order SGD. For simplicity, we assume $F(\mathbf{w}^*) \lesssim 1$ here. This is a mild assumption and holds for many popular machine learning problems. For example, it holds if we consider margin-based loss $f(\mathbf{w}; \mathbf{z}) = \phi(y\langle \mathbf{w}, \mathbf{x}\rangle)$, where $\mathbf{z} = (\mathbf{x}, y)$ and $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ is a decreasing function. Indeed, in this case, we know

$$F(\mathbf{w}^*) \leq F(\mathbf{0}) = \mathbb{E}_{\mathbf{z}}[\phi(y\langle \mathbf{0}, \mathbf{x}\rangle)] = \phi(0).$$

We also assume $\|\mathbf{w}^* - \mathbf{w}_1\|_2 \lesssim \|\mathbf{w}^*\|_2$ to simplify the notation, which holds if we set $\mathbf{w}_1 = \mathbf{0}$.

Assumption 5.1 *We assume $F(\mathbf{w}^*) \lesssim 1$ and $\|\mathbf{w}^* - \mathbf{w}_1\|_2 \lesssim \|\mathbf{w}^*\|_2$.*

Theorem 5.11 (Excess risk bounds for Zeroth-order SGD) *Assume for all* $\mathbf{z}$, $\mathbf{w} \mapsto f(\mathbf{w};\mathbf{z})$ *is convex and* L*-smooth. Let Assumption 5.1 hold. Let* $\{\mathbf{w}_t\}_t$ *be produced by BSGM with* $g(\mathbf{w}_t; S_{J_t})$ *in Eq. (5.4) and* $\eta_t = \eta \leq (2L(1+d/K))^{-1}$. *Let* $\mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. *If* $\big(3 + \frac{8(d+2)}{K} + \frac{2}{m} + \frac{4(d+2)}{mK}\big)L^2T\eta^2 \leq 1/2, n \lesssim T$ *and*

$$\frac{L(dT)^{\frac{1}{2}}\eta}{(mnK)^{\frac{1}{2}}} + \frac{LT\eta}{n} + \frac{T^{\frac{1}{2}}\eta\mu(Ld)^{\frac{3}{2}}}{\min\{mn,K\}^{\frac{1}{2}}} + \frac{\mu(Ld)^{\frac{3}{2}}T\eta}{n} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\Big(\frac{Ld}{K} + \mu^2L^2d^3\Big) \lesssim 1, \tag{5.13}$$

then

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{L(dT)^{\frac{1}{2}}\eta}{(mnK)^{\frac{1}{2}}} + \frac{LT\eta}{n} + \frac{T^{\frac{1}{2}}\eta\mu(Ld)^{\frac{3}{2}}}{\min\{mn,K\}^{\frac{1}{2}}} + \frac{\mu(Ld)^{\frac{3}{2}}T\eta}{n} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\Big(\frac{Ld}{K} + \mu^2L^2d^3\Big).$$

Remark 5.12 The condition Eq. (5.13) is imposed here to simplify the result. If this assumption does not hold, then it is clear that the above risk bounds are vacuous. Therefore, we always choose parameters to make this condition hold. For simplicity, we assume $m = 1$ and choose $T \gtrsim \max\{\frac{nd}{K}, Ln\|\mathbf{w}^*\|_2^2\big(\frac{d}{K}+1\big)\}$. Then, Theorem 5.11 implies

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{LT\eta}{n} + \frac{T\eta\mu L^{\frac{3}{2}}d}{n^{\frac{1}{2}}} + \frac{\mu(Ld)^{\frac{3}{2}}T\eta}{n} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\mu^2L^2d^3,$$

where we use $K \gtrsim nd/T$ and $n \lesssim T$. If we choose η such that $T\eta \asymp \sqrt{n}\|\mathbf{w}^*\|_2/\sqrt{L}$, then we further get

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{L^{\frac{1}{2}}\|\mathbf{w}^*\|_2}{n^{\frac{1}{2}}} + \mu Ld\|\mathbf{w}^*\|_2 + \frac{\|\mathbf{w}^*\|_2\mu Ld^{\frac{3}{2}}}{n^{\frac{1}{2}}} + Ld\mu^2 + \eta\mu^2L^2d^3.$$

If we choose

$$\mu \lesssim \min\Big\{\frac{1}{d(nL)^{\frac{1}{2}}}, \frac{1}{L^{\frac{1}{2}}d^{\frac{3}{2}}}, \frac{\|\mathbf{w}^*\|_2^{\frac{1}{2}}}{(Ln)^{\frac{1}{4}}d^{\frac{1}{2}}}, \frac{\|\mathbf{w}^*\|_2^{\frac{1}{2}}}{L^{\frac{3}{4}}d^{\frac{3}{2}}\eta^{\frac{1}{2}}n^{\frac{1}{4}}}\Big\},$$

then we get $\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim L^{\frac{1}{2}}\|\mathbf{w}^*\|_2/n^{\frac{1}{2}}$. To our knowledge, this is the first excess risk bounds of order $1/\sqrt{n}$ for Zeroth-order SGD. It is clear that Eq. (5.13) holds under our choices of η and μ. Furthermore, since $T \gtrsim Ln\|\mathbf{w}^*\|_2^2\big(\frac{d}{K}+1\big)$, we know

$$L^2T\eta^2\big(1+\frac{d}{K}\big) = \frac{L^2T^2\eta^2}{T}\big(1+\frac{d}{K}\big) \asymp \frac{L^2(1+d/K)n\|\mathbf{w}^*\|_2^2}{TL} \lesssim 1,$$

which is consistent with the condition $\big(3 + \frac{8(d+2)}{K} + \frac{2}{m} + \frac{4(d+2)}{mK}\big)L^2T\eta^2 \leq 1/2$.

5.3 Clipped-SGD with Bounded Moments

While SGD is still the *de facto* algorithm in machine learning, its performance may deteriorate for problems with heavy-tailed noises, which are often encountered in training deep

learning models such as BERT and convolutional neural networks (Zhang et al., 2020, 2019). To improve the robustness of SGD to tackle heavy-tailed noises, the technique of gradient clipping has been introduced and has found very successful applications. For a vector $\mathbf{v}$, we define the clipping operator with parameter $\tau > 0$ as follows

$$\text{clip}(\mathbf{v}, \tau) := \mathbf{v} \min(1, \tau/\|\mathbf{v}\|_2).$$

The basic idea of Clipped-SGD is to apply the clipping operator to the gradient estimator before the iterate update. At the t-th iteration, Clipped-SGD updates $\mathbf{w}_{t+1}$ according to Eq. (3.3) with

$$g(\mathbf{w}_t; S_{J_t}) = \frac{1}{m} \sum_{k=1}^{m} \text{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_{j_{t,k}}), \tau). \tag{5.14}$$

Other than stabilizing the optimization process, gradient clipping also has found core applications in developing privacy-preserving machine learning algorithms. For example, the gradient clipping controls the magnitude of the gradient, which is essential to apply the gradient perturbation to preserve privacy (Abadi et al., 2016; Das et al., 2023).

Although the gradient clipping operator diminishes the variation of the gradient estimator, it introduces the bias. For example, if $\mathbb{E}_{\mathbf{z}}[\|\nabla f(\mathbf{w}; \mathbf{z})\|_2^p] \leq G^p$ with $p \in (1, 2]$, then it was shown that the bias satisfies the following inequality (Zhang et al., 2020)

$$\big\|\mathbb{E}_{\mathbf{z}}[\text{clip}(\nabla f(\mathbf{w}; \mathbf{z}))] - \mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}; \mathbf{z})]\big\|_2 \leq G^p \tau^{1-p}. \tag{5.15}$$

Therefore, the bias decreases to 0 as τ increases to infinity. According to Eq. (5.15), Eq. (4.6) holds with $\bar{A} = 0$ and $\bar{B}_{t,i} = G^{2p}\tau^{2-2p}$. If we use this crude estimator, Theorem 4.3 implies a stability bound involving the term

$$\frac{\sum_{t=1}^T \eta_t^2}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}\big[\bar{B}_{t,i}\big] = \frac{\sum_{t=1}^T \eta_t^2}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} G^{2p}\tau^{2-2p} = TG^{2p}\tau^{2-2p} \sum_{t=1}^{T} \eta_t^2,$$

which is not desirable due to the factor of T. Our key idea is to consider the difference of bias at two vectors, and show this difference satisfies a much better condition in the sense of Eq. (4.6) with $\bar{A} \lesssim 1/\tau^{2p}$ and $\bar{B}_{t,i} \lesssim 1/n^2$ (Lemma 5.14). The proof of Lemma 5.14 is based on the Lipschitz continuity of the bias established in the following lemma. The proofs of results in this subsection are given in Section 7.3.

Lemma 5.13 *For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{W}$, we have*

$$\|(\text{clip}(\mathbf{v}_1, \tau) - \text{clip}(\mathbf{v}_2, \tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \mathbb{I}\big[\|\mathbf{v}_1\|_2 > \tau \text{ or } \|\mathbf{v}_2\|_2 > \tau\big], \tag{5.16}$$

where $\mathbb{I}[\cdot]$ is the indicator function, i.e., returning 1 if the argument holds and 0 otherwise.

In this subsection, we always assume $\tilde{f} = f$ and impose a bounded moment condition.

Assumption 5.2 *Suppose for any t, i, $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t; S_r)\|_2^p] \leq G^p$, $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2^p] \leq G^p$, where r follows the uniform distribution over $[n]$.*

Assumption 5.2 is standard, which has been widely adopted in the analysis of Clipped-SGD (Zhang et al., 2020; Cutkosky and Mehta, 2021). This assumption is often referred to as a heavy-tailed noise condition if $p < 2$ (Zhang et al., 2020; Cutkosky and Mehta, 2021).

Lemma 5.14 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.14), and b_t, $b_t^{(i)}$ be defined in Eq. (4.2). Let $p \in [1,2]$ and $G > 0$. Let $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ be L-smooth and Assumption 5.2 hold. Then, Eq. (4.6) holds with $\bar{A} = \frac{8L^2G^{2p}}{\tau^{2p}}$ and $\bar{B}_{t,i} = 8\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2/n^2$.*

In the following two lemmas, we show that Eq. (4.4) and Eq. (4.5) hold for Clipped-SGD.

Lemma 5.15 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.14). Assume $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth for any $\mathbf{z}$. Then, Eq. (4.4) holds with $A = L^2$ and $B_{t,i} = \frac{4}{n}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2$.*

Lemma 5.16 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.14). Assume $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ is L-smooth for any $\mathbf{z}$. Then, Eq. (4.5) holds with $\widetilde{A} = 3L^2/2$ and $\widetilde{B}_{t,i} = \frac{12}{n^2}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2$.*

Now we apply Lemma 5.15, Lemma 5.16, Lemma 5.14 and Theorem 4.3 to get Corollary 5.17.

Corollary 5.17 (Stability of Clipped-SGD) *Let Assumptions in Lemma 5.14 hold and f be convex. Let Assumption 5.1 hold. If $(3 + \frac{2}{m} + \frac{64G^{2p}T}{\tau^{2p}})L^2 \sum_{t=1}^T \eta_t^2 \leq 1/2$, then*

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \frac{L}{mn}\sum_{t=1}^T \eta_t^2 \mathbb{E}\big[F_S(\mathbf{w}_t)\big] + \frac{L\sum_{t=1}^T \eta_t^2}{n^2}\sum_{t=1}^T \mathbb{E}\big[F_S(\mathbf{w}_t)\big]. \tag{5.17}$$

Furthermore, if $\eta_t = \eta \leq 1/(4L)$ and $G \lesssim \tau$, then

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \Big(\frac{LT\eta^2}{mn} + \frac{LT^2\eta^2}{n^2}\Big)\Big(F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \eta G^p \tau^{2-p} + G^{2p}\tau^{2-2p}\Big(T\eta + \frac{LT^2\eta^2}{n}\Big)\Big). \tag{5.18}$$

Remark 5.18 If we choose $\tau \asymp GT^{\frac{1}{p}}$ and $\eta \lesssim \min\{1/L, n/(TL), G^{-2}T^{\frac{p-2}{p}}F(\mathbf{w}^*)\}$, then we have $LT^2\eta^2/n \lesssim T\eta$ and

$$G^p\tau^{2-p}\eta + G^{2p}\tau^{2-2p}T\eta \asymp G^2T^{\frac{2-p}{p}}\eta + G^2T^{\frac{2-2p}{p}}T\eta \asymp G^2T^{\frac{2-p}{p}}\eta \lesssim F(\mathbf{w}^*).$$

In this case, Eq. (5.18) implies that

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \Big(\frac{LT\eta^2}{mn} + \frac{LT^2\eta^2}{n^2}\Big)\Big(F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{T\eta}\Big),$$

which is of a similar order as the stability bound for SGD in Eq. (5.3). To our knowledge, our analysis gives the first stability bounds for Clipped-SGD. The choice $\tau \asymp GT^{\frac{1}{p}}$ is also used in the convergence analysis of Clipped-SGD (Nguyen et al., 2023).

We combine the stability and convergence analysis to derive the following excess risk bounds. Recall that Assumption 5.1 gives $F(\mathbf{w}^*) \lesssim 1$.

Theorem 5.19 (Excess risk bounds for Clipped-SGD) *Let assumptions in Corollary 5.17 hold. Let $\eta_t = \eta$ and $\mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. If*

$$\frac{L\eta T^{\frac{1}{2}}}{(mn)^{\frac{1}{2}}} + \frac{LT\eta}{n} + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + G^p\tau^{2-p}\eta + G^{2p}\tau^{2-2p}(T\eta + LT^2\eta^2/n) \lesssim 1, \tag{5.19}$$

then

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + G^p\tau^{2-p}\eta + G^{2p}\tau^{2-2p}T\eta + \frac{L\eta T^{\frac{1}{2}}}{(mn)^{\frac{1}{2}}} + \frac{LT\eta}{n}. \tag{5.20}$$

Remark 5.20 If we take $\tau \asymp GT^{\frac{1}{p}}$ and $T \gtrsim n$, then Eq. (5.20) implies

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + G^2\eta T^{\frac{2-p}{p}} + \frac{LT\eta}{n}.$$

If we further take

$$\eta \asymp G^{\frac{p}{1-p}}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2p-2}} L^{\frac{1}{2p-2}} \quad \text{and} \quad T \asymp n^{\frac{p}{2p-2}} G^{\frac{p}{p-1}} L^{\frac{p}{2-2p}}, \tag{5.21}$$

then $T\eta \asymp \|\mathbf{w}^*\|_2 n^{\frac{p-1}{2p-2}} L^{\frac{1-p}{2p-2}} = \|\mathbf{w}^*\|_2 n^{\frac{1}{2}} L^{-\frac{1}{2}}$ and

$$\begin{aligned}\eta T^{\frac{2-p}{p}} &\asymp G^{\frac{p}{1-p}}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2p-2}} L^{\frac{1}{2p-2}} n^{\frac{p}{2p-2}\frac{2-p}{p}} G^{\frac{p}{p-1}\frac{2-p}{p}} L^{\frac{p}{2-2p}\frac{2-p}{p}} \\ &= G^{\frac{p+(p-2)}{1-p}}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2p-2}+\frac{2-p}{2p-2}} L^{\frac{1}{2p-2}+\frac{p-2}{2p-2}} = G^{-2}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2}} L^{\frac{1}{2}},\end{aligned} \tag{5.22}$$

from which we derive $\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \|\mathbf{w}^*\|_2 L^{\frac{1}{2}} n^{-\frac{1}{2}}$. To our knowledge, this gives the first risk bounds of order $n^{-\frac{1}{2}}$ for Clipped-SGD by stability analysis. The choice $T \asymp n^{\frac{p}{2p-2}}$ means that we require more iterations for a smaller p. For example, for $p = 2$ we can choose $T \asymp n$, while for $p = 3/2$ we require $T \asymp n^{\frac{3}{2}}$ to achieve the excess risk bound $O(1/\sqrt{n})$. Furthermore, we have

$$L^2T\eta^2 \asymp L^2 L^{\frac{p}{2-2p}} L^{\frac{2}{2p-2}} n^{\frac{p-2}{2p-2}} G^{\frac{p}{p-1}} G^{\frac{2p}{1-p}} \|\mathbf{w}^*\|_2^2 = L^{\frac{3p-2}{2p-2}} n^{\frac{p-2}{2p-2}} G^{\frac{-2p}{2p-2}} \|\mathbf{w}^*\|_2^2.$$

Without loss of generality, we always assume $G \gtrsim \|\mathbf{w}^*\|_2^{\frac{2p-2}{p}} n^{\frac{p-2}{2p}} L^{\frac{3p-2}{2p}}$ (otherwise, we can take the maximum between G and $\|\mathbf{w}^*\|_2^{\frac{2p-2}{p}} n^{\frac{p-2}{2p}} L^{\frac{3p-2}{2p}}$). Then, we have $L^2T\eta^2 \lesssim 1$ and therefore our choice of parameters in Eq. (5.21) is consistent with the condition $(3 + \frac{2}{m} + \frac{64G^{2p}T}{\tau^{2p}})L^2 \sum_{t=1}^T \eta_t^2 \leq 1/2$ (note $G^{2p}T/\tau^{2p} \asymp 1/T$). Eq. (5.19) holds directly since the right-hand side of Eq. (5.20) is of order $L^{\frac{1}{2}}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2}}$ (note $\eta \lesssim n/(LT)$ since $T\eta \asymp \|\mathbf{w}^*\|_2 n^{\frac{1}{2}} L^{-\frac{1}{2}}$).

Remark 5.21 An application of Clipped-SGD is to use it to develop differentially private algorithms (Abadi et al., 2016), i.e., we update models by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\big(g(\mathbf{w}_t; S_{J_t}) + G_t\big), \tag{5.23}$$

where $g(\mathbf{w}_t; S_{J_t})$ is defined in Eq. (5.14) and $G_t \sim \mathcal{N}(0, \sigma^2\mathbb{I}_d)$ is a Gaussian noise with variance $\sigma^2\mathbb{I}_d$ introduced to protect the privacy of data. Since the noise G_t does not affect the stability, one can directly apply Corollary 5.17 to show that the method in Eq. (5.23) is on-average ϵ-model stable with

$$\epsilon^2 \lesssim \frac{L}{mn}\sum_{t=1}^T \eta_t^2 \mathbb{E}\big[F_S(\mathbf{w}_t)\big] + \frac{L\sum_{t=1}^T \eta_t^2}{n^2}\sum_{t=1}^T \mathbb{E}\big[F_S(\mathbf{w}_t)\big]. \tag{5.24}$$

Uniform stability was also used to study the utility guarantee of differentially private SGD method (Bassily et al., 2019, 2020)

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\Big(\frac{1}{m}\sum_{k=1}^{m}\nabla f(\mathbf{w}_t; \mathbf{z}_{j_{t,k}}) + G_t\Big), \tag{5.25}$$

Under a $\widetilde{G}$-Lipschitzness, L-smoothness, and convexity assumption, it was shown that the method in Eq. (5.25) is ϵ_{unif}-uniformly stable with $\epsilon_{\text{unif}} \leq T\eta\widetilde{G}^2/n$ (Bassily et al., 2019). This stability analysis considers SGD which requires a Lipschitzness condition to control the sensitivity of the gradient update. As a comparison, we consider Clipped-SGD, which automatically satisfies the sensitivity assumption. Instead, we impose a bounded moment condition to study the stability and generalization. Furthermore, Eq. (5.24) involves training errors, which can benefit from small training errors during the optimization process, while the uniform stability analysis in Bassily et al. (2019) does not capture training errors. Lastly, our result in Eq. (5.24) clarifies the batch size in improving the stability, while the uniform stability analysis in Bassily et al. (2019) does not show the effect of batch size.

5.4 Clipped-SGD with Bounded Central Moments

Section 5.3 studies the stability of Clipped-SGD under a bounded moment condition, which can be strong and implies that the true gradients are bounded (Nguyen et al., 2023). In this section, we consider a weaker assumption called the bounded *central* moment condition (Nguyen et al., 2023; Zhang et al., 2020). If $p = 2$, this becomes a bounded variance assumption.

Assumption 5.3 *Let $G \geq 0$. Suppose for any t, i, $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t; S_r) - \nabla F_S(\mathbf{w}_t)\|_2^p] \leq G^p$, $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)}) - \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2^p] \leq G^p$, where r follows the uniform distribution over $[n]$.*

Remark 5.22 Assumption 5.3 implies that

$$\frac{1}{n}\max_{i\in[n]}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla F_S(\mathbf{w}_t)\|_2^p \leq \mathbb{E}_r[\|\nabla f(\mathbf{w}_t; S_r) - \nabla F_S(\mathbf{w}_t)\|_2^p] \leq G^p$$

and therefore $\max_{i\in[n]}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla F_S(\mathbf{w}_t)\|_2 \leq n^{\frac{1}{p}}G$. This shows that the stochastic gradient noise is bounded by $n^{\frac{1}{p}}G$, which further implies that for any nondecreasing and continuous measurable function φ the following expectation is bounded:

$$\mathbb{E}_r\big[\varphi(\|\nabla f(\mathbf{w}_t; S_r) - \nabla F_S(\mathbf{w}_t)\|_2)\big] \leq \varphi(n^{\frac{1}{p}}G).$$

In particular, this means that the noise is sub-Gaussian, although the sub-Gaussian parameter can be very large, i.e., of the order of $n^{\frac{1}{p}}G$.

The following lemma shows that the condition in Eq. (4.6) still holds under Assumption 5.3. The proofs of results in this subsection are given in Section 7.4.

Lemma 5.23 *Let $g(\mathbf{w}_t; S_{J_t})$ be defined in Eq. (5.14), and b_t, $b_t^{(i)}$ be defined in Eq. (4.2). Let $\mathbf{w} \mapsto f(\mathbf{w}; \mathbf{z})$ be L-smooth and Assumption 5.3 hold with some $p \in [1, 2]$. Then, Eq. (4.6) holds with $\bar{A} = \frac{3 \cdot 2^{2p+2} G^{2p} L^2}{\tau^{2p}}$ and*

$$\bar{B}_{t,i} = \frac{3\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2^2}{n^2} + 3L^2 \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 \chi_{t,i},$$

where we introduce

$$\chi_{t,i} = \mathbb{I}\big[\|\nabla F_S(\mathbf{w}_t)\|_2 > \tau/2 \text{ or } \|\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2 > \tau/2\big]. \tag{5.26}$$

Now we apply Lemma 5.15, Lemma 5.16, Lemma 5.23 and Theorem 4.3 to derive Corollary 5.24. As compared to the analysis under the bounded moment condition, the analysis under the bounded central moment condition yields an additional term involving $\mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2 > \tau/2]$ in Eq (5.28). The stability bound in Eq. (5.28) also involves a weighted summation of training errors, which are removed in Eq. (5.30) by controlling $\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)]$ with high-probability convergence analysis. To this aim, we introduce an assumption on the behavior of $\mathbf{w}^*$. Note that $\|\nabla F_S(\mathbf{w}^*)\|_2 \leq \frac{C_3}{\sqrt{n}} \log^{\frac{1}{2}}(1/\delta)$ is a Sub-Gaussian assumption on $\nabla f(\mathbf{w}^*; \mathbf{z})$ since $\nabla F(\mathbf{w}^*) = 0$.

Assumption 5.4 *Let $C_3 \geq 0$. For any $\delta \in (0, 1)$, with probability at least $1-\delta$: $\|\nabla F_S(\mathbf{w}^*)\|_2 \leq \frac{C_3}{\sqrt{n}} \log^{\frac{1}{2}}(1/\delta)$. Also, assume $\|\mathbf{w}^* - \mathbf{w}_1\|_2 \lesssim \|\mathbf{w}^*\|_2$ and $F(\mathbf{w}^*) \lesssim 1$.*

Let $\delta \in (0, 1)$ be a number to be determined later. We introduce R_T to control the distance between $\mathbf{w}^*$ and $\mathbf{w}_t$ for $t \in [T]$

$$R_T = \sqrt{2}\Big(\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + 4\eta^2\big(32G^p\tau^{2-p}T + 4\tau^2 \log(3/\delta)\big) + 10\eta^2 T^2 \big(\big(12G^p\tau^{1-p}\big)^2 + 4\tau^2 T^{-2} \log^2(3T/\delta) + C_3^2 n^{-1} \log(3/\delta)\big)\Big)^{\frac{1}{2}}. \tag{5.27}$$

Corollary 5.24 (Stability of Clipped-SGD) *Let Assumptions in Lemma 5.23 hold, f be convex and $\delta \in (0, 1)$. If $(3 + \frac{2}{m} + \frac{24 \cdot 2^{2p+2} G^{2p} T}{\tau^{2p}}) L^2 \sum_{t=1}^T \eta_t^2 \leq 1/2$, then*

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \frac{L}{mn}\sum_{t=1}^T \eta_t^2 \mathbb{E}\big[F_S(\mathbf{w}_t)\big] + \frac{L\sum_{t=1}^T \eta_t^2}{n^2}\sum_{t=1}^T \mathbb{E}\big[F_S(\mathbf{w}_t)\big] + \frac{L^2 \sum_{t=1}^T \eta_t^2}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 \mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2 > \tau/2]\big]. \tag{5.28}$$

Furthermore, let Assumption 5.4 hold and

$$\tau \geq 2\Big(\frac{C_3 \log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T\Big), \tag{5.29}$$

where R_T is defined in Eq. (5.27). Then

$$\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \lesssim \Big(\frac{LT\eta^2}{mn} + \frac{LT^2\eta^2}{n^2}\Big)\Big(1 + \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta\big(G^p\tau^{2-p} + \tau^2 T^{-1}\log^2(T/\delta)\big) + \eta T\big(\big(G^p\tau^{1-p}\big)^2 + n^{-1}\log(1/\delta)\big) + L\big(T\eta\tau + \|\mathbf{w}^*\|_2\big)^2\delta\Big) + L^2\tau^2 T^4 \eta^4 \delta. \tag{5.30}$$

We combine the stability and convergence analysis, and derive the following theorem on excess risk bounds under Assumption 5.3. It shows that we can still get excess risk bounds of order $\widetilde{O}(L^{\frac{1}{2}}\|\mathbf{w}^*\|_2/n^{\frac{1}{2}})$, where we use $\widetilde{O}(\cdot)$ to hide logarithmic factors.

Theorem 5.25 (Excess risk bounds for Clipped-SGD) *Let $G \gtrsim n^{\frac{p-2}{2p}} L^{\frac{1}{2}}$ and $\delta = \min\big\{\frac{1}{2}, \frac{1}{L\sqrt{n}(T\eta\tau+\|\mathbf{w}^*\|_2)^2}, (n\tau T\eta)^{-2}/L\big\}$. Let assumptions in Corollary 5.24 hold. Suppose we take $\tau \asymp GT^{\frac{1}{p}}, \eta_t = \eta$ and T in Eq. (5.21). For $\mathcal{A}(S) = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$, we know $\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] = \widetilde{O}(L^{\frac{1}{2}}\|\mathbf{w}^*\|_2/\sqrt{n})$.*

Remark 5.26 Remark 5.20 shows that our parameter choice in Eq. (5.21) is consistent with the assumption $(3 + \frac{2}{m} + \frac{24\cdot 2^{2p+2}G^{2p}T}{\tau^{2p}})L^2\sum_{t=1}^T \eta_t^2 \leq 1/2$. Furthermore, the choice $\tau \asymp GT^{\frac{1}{p}}$ implies

$$\begin{aligned} R_T &\lesssim \|\mathbf{w}^*\|_2 + \eta(G^p\tau^{2-p}T)^{\frac{1}{2}} + \eta T\big(G^p\tau^{1-p} + \tau T^{-1}\log(T/\delta) + n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)\big) \qquad (5.31)\\ &\lesssim \|\mathbf{w}^*\|_2 + \eta(G^pG^{2-p}T^{\frac{2-p}{p}}T)^{\frac{1}{2}} + \eta T\big(G^pG^{1-p}T^{\frac{1-p}{p}} + GT^{\frac{1}{p}}T^{-1}\log(T/\delta) + n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)\big)\\ &\lesssim \|\mathbf{w}^*\|_2 + \eta GT^{\frac{1}{p}}\log(T/\delta) + \eta T n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta) \lesssim \|\mathbf{w}^*\|_2 + \eta GT^{\frac{1}{p}}\log(T/\delta) + \|\mathbf{w}^*\|_2 L^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta), \end{aligned}$$

where we have used $T\eta \asymp \|\mathbf{w}^*\|_2 n^{\frac{1}{2}}L^{-\frac{1}{2}}$ given below Eq. (5.21). Therefore, the choice $\tau \asymp GT^{\frac{1}{p}}$ in Theorem 5.25 is consistent with the assumption in Eq. (5.29).

6. Proofs on General Stability Bounds

In this section, we present the proof for our general stability bounds for BSGMs. We first introduce an elementary result on the solution of a quadratic inequality. We omit the proof.

Lemma 6.1 *Let $a, b \geq 0$. If $x^2 \leq ax + b$, then $x^2 \leq a^2 + 2b$ and $x \leq a + \sqrt{b}$.*

The following lemma relates the difference between two gradient estimates based on a minibatch to the difference based on a single example.

Lemma 6.2 *It holds that*

$$\begin{aligned} \mathbb{E}_{J_t}[\|g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})\|_2^2] \leq \frac{1}{m}\mathbb{E}_{j_{t,1}}\big[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2\big]\\ + \|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2. \end{aligned}$$

Proof Since $\mathbb{E}[X^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X])^2$ for a random variable X, we know

$$\begin{aligned} \mathbb{E}_{J_t}[\|g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})\|_2^2] = \|\mathbb{E}_{J_t}[g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})]\|_2^2\\ + \mathbb{E}_{J_t}\big[\|g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)}) - \mathbb{E}_{J_t}[g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})]\|_2^2\big]. \end{aligned}$$

Recall that g has a sum structure in the sense that $g(\mathbf{w}_t; S_{J_t}) = \frac{1}{|S_{J_t}|}\sum_{\mathbf{z}\in S_{J_t}} g(\mathbf{w}_t; \mathbf{z})$. Since averaging m independent random variables reduces the variance by a factor of m and

$$\mathbb{E}_{J_t}[g(\mathbf{w}_t; S_{J_t}) - g(\mathbf{w}_t^{(i)}; S_{J_t}^{(i)})] = \mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})],$$

we further get

$$\begin{aligned}
&\mathbb{E}_{J_t}[\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2] \le \|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2 \\
&+\frac{1}{m}\mathbb{E}_{j_{t,1}}\big[\|g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})-\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2\big] \\
&=\frac{1}{m}\mathbb{E}_{j_{t,1}}\big[\|g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2\big]+\frac{m-1}{m}\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2.
\end{aligned}$$

The proof is completed. ■

6.1 Proofs on Convex Problems

We now present the proofs of Theorem 4.1 and Theorem 4.3 under a convexity assumption.

Proof of Theorem 4.1 We know

$$\begin{aligned}
&\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2=\|\mathbf{w}_t-\eta_t g(\mathbf{w}_t;S_{J_t})-\mathbf{w}_t^{(i)}+\eta_t g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2 \\
&=\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+\eta_t^2\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2-2\eta_t\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\rangle.
\end{aligned}$$

Taking a conditional expectation implies

$$\begin{aligned}
&\mathbb{E}_t[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2]-\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2 \\
&=\eta_t^2\mathbb{E}_t[\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2]-2\eta_t\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},\mathbb{E}_t[g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})]\rangle \\
&=\eta_t^2\mathbb{E}_t[\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2]-2\eta_t\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},\nabla\widetilde{F}_S(\mathbf{w}_t)+b_t-\nabla\widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)})-b_t^{(i)}\rangle.
\end{aligned}$$

By the convexity of $\tilde{f}$, we know $\langle \mathbf{w}-\mathbf{w}',\nabla\tilde{f}(\mathbf{w};\mathbf{z})-\nabla\tilde{f}(\mathbf{w}';\mathbf{z})\rangle\ge 0$ for any $\mathbf{w},\mathbf{w}'$ and then

$$\begin{aligned}
&\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},\nabla\widetilde{F}_S(\mathbf{w}_t)-\nabla\widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)})\rangle \\
&=\frac{1}{n}\sum_{j:j\neq i}\Big\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_j)-\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_j)\Big\rangle+\frac{1}{n}\langle \mathbf{w}_t-\mathbf{w}_t^{(i)},\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)-\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\rangle \\
&\ge-\frac{\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)-\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2}{n}.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2] &\le \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2]+\eta_t^2\mathbb{E}[\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2] \\
&+2\eta_t\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\|b_t-b_t^{(i)}\|_2]+\frac{2\eta_t}{n}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)-\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2],
\end{aligned}$$

which further yields

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2] &\le \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2]+\eta_t^2\mathbb{E}[\|g(\mathbf{w}_t;S_{J_t})-g(\mathbf{w}_t^{(i)};S_{J_t}^{(i)})\|_2^2] \\
&+2\eta_t\mathbb{E}\Big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\big(\|b_t-b_t^{(i)}\|_2+\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)-\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2/n\big)\Big].
\end{aligned}$$

By the Cauchy–Schwarz inequality and $(a+b)^2 \leq 2(a^2+b^2)$, we know

$$\begin{aligned}
&\mathbb{E}\Big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2\big(\|b_t - b_t^{(i)}\|_2 + \|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla \tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2/n\big)\Big] \\
&\leq \big(\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big]\big)^{\frac{1}{2}}\Big(\mathbb{E}\Big[\big(\|b_t - b_t^{(i)}\|_2 + \|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla \tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2/n\big)^2\Big]\Big)^{\frac{1}{2}} \\
&\leq \sqrt{2}\big(\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big]\big)^{\frac{1}{2}}\Big(\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] + \mathbb{E}\big[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla \tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2\big]/n^2\Big)^{\frac{1}{2}} \\
&\leq \sqrt{2}\big(\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big]\big)^{\frac{1}{2}}\Big(\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] + 4\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}},
\end{aligned}$$

where in the last step we have used the inequality below by the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$

$$\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla \tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2] \leq 2\big(\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] + \mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2]\big) = 4\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2].$$

We combine the above inequalities and Lemma 6.2 to derive that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2] &\leq \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] \\
&+ \frac{\eta_t^2}{m}\mathbb{E}\big[\|g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2 + \eta_t^2\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2] \\
&+ 2\sqrt{2}\eta_t\big(\mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big]\big)^{\frac{1}{2}}\Big(\mathbb{E}\big[\|b_t - b_t^{(i)}\|_2^2\big] + 4\mathbb{E}[\|\nabla \tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

Introduce

$$\Delta_{t,i} = \big(\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2]\big)^{\frac{1}{2}}, \quad \forall t \in [T], i \in [n]. \tag{6.1}$$

Then, $\Delta_i = \max_{t\leq T}\Delta_{t,i}$. The above inequality can be written in terms of $\Delta_{t,i}$ and $\mathfrak{C}_{t,i}$

$$\begin{aligned}
\Delta_{t+1,i}^2 \leq \Delta_{t,i}^2 &+ \frac{\eta_t^2}{m}\mathbb{E}\big[\|g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2] \\
&+ \eta_t^2\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2] + 2\sqrt{2}\eta_t\Delta_{t,i}\mathfrak{C}_{t,i}.
\end{aligned} \tag{6.2}$$

Applying the above inequality recursively and using $\Delta_{1,i} = 0$, we derive

$$\begin{aligned}
\Delta_{t+1,i}^2 \leq \sum_{k=1}^t \eta_k^2\Big(&\frac{1}{m}\mathbb{E}\big[\|g(\mathbf{w}_k;S_{j_{k,1}}) - g(\mathbf{w}_k^{(i)};S_{j_{k,1}}^{(i)})\|_2^2] \\
&+ \mathbb{E}[\|\mathbb{E}_{j_{k,1}}[g(\mathbf{w}_k;S_{j_{k,1}}) - g(\mathbf{w}_k^{(i)};S_{j_{k,1}}^{(i)})]\|_2^2]\Big) + 2\sqrt{2}\sum_{k=1}^t \eta_k\Delta_{k,i}\mathfrak{C}_{k,i}.
\end{aligned}$$

Since the above inequality applies for any $t \in [T]$, we further get

$$\begin{aligned}
\Delta_i^2 \leq \sum_{t=1}^T \eta_t^2\Big(&\frac{1}{m}\mathbb{E}\big[\|g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2] \\
&+ \mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2]\Big) + 2\sqrt{2}\Delta_i\sum_{t=1}^T \eta_t\mathfrak{C}_{t,i}.
\end{aligned}$$

Solving the above quadratic inequality of Δ_i implies that (Lemma 6.1)

$$\Delta_i^2 \le 2\sum_{t=1}^{T}\frac{\eta_t^2}{m}\mathbb{E}\big[\|g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2\big] + 2\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2] + 8\Big(\sum_{t=1}^{T}\eta_t\mathfrak{C}_{t,i}\Big)^2. \quad (6.3)$$

Taking an average over $i\in[n]$ gives the stated bound. The proof is completed. ■

Proof of Theorem 4.3 We plug Eq. (4.4) and (4.5) back into Eq. (6.3), and derive

$$\Delta_i^2 \le 2\sum_{t=1}^{T}\frac{\eta_t^2}{m}\mathbb{E}\big[A\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+B_{t,i}\big] + 2\sum_{t=1}^{T}\eta_t^2\mathbb{E}\big[\widetilde{A}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+\widetilde{B}_{t,i}\big] + 8\Big(\sum_{t=1}^{T}\eta_t\mathfrak{C}_{t,i}\Big)^2.$$

The Cauchy–Schwarz inequality implies that

$$\begin{aligned}\Big(\sum_{t=1}^{T}\eta_t\mathfrak{C}_{t,i}\Big)^2 &= \Big(\sum_{t=1}^{T}\eta_t\Big(\mathbb{E}\big[\|b_t-b_t^{(i)}\|_2^2\big]+4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}}\Big)^2\\ &\le \Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\Big(\mathbb{E}\big[\|b_t-b_t^{(i)}\|_2^2\big]+4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)\Big)\\ &\le \Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\mathbb{E}\Big[\bar{A}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+\bar{B}_{t,i}+4\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\Big]\Big)\\ &\le \bar{A}T\sum_{t=1}^{T}\eta_t^2\Delta_i^2+\Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\mathbb{E}\big[\bar{B}_{t,i}+4\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big]\Big).\end{aligned}$$

We combine the above two inequalities together, and derive

$$\begin{aligned}\Delta_i^2 \le{}& \Delta_i^2\Big(2A\sum_{t=1}^{T}\frac{\eta_t^2}{m}+2\widetilde{A}\sum_{t=1}^{T}\eta_t^2+8\bar{A}T\sum_{t=1}^{T}\eta_t^2\Big)\\ &+2\sum_{t=1}^{T}\frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m}+2\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\widetilde{B}_{t,i}]+8\Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\mathbb{E}\big[\bar{B}_{t,i}+4\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big]\Big).\end{aligned}$$

Since $2A\sum_{t=1}^{T}\frac{\eta_t^2}{m}+2\widetilde{A}\sum_{t=1}^{T}\eta_t^2+8\bar{A}T\sum_{t=1}^{T}\eta_t^2\le 1/2$, we further get

$$\Delta_i^2 \le \frac{\Delta_i^2}{2}+2\sum_{t=1}^{T}\frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m}+2\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\widetilde{B}_{t,i}]+8\Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\mathbb{E}\big[\bar{B}_{t,i}+4\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big]\Big),$$

from which we derive that

$$\Delta_i^2 \le 4\sum_{t=1}^{T}\frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m}+4\sum_{t=1}^{T}\eta_t^2\mathbb{E}[\widetilde{B}_{t,i}]+16\Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\mathbb{E}\big[\bar{B}_{t,i}+4\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big]\Big).$$

We take an average over $i\in[n]$ and get the stated bound. The proof is completed. ■

6.2 Proofs on Strongly Convex Problems

In this subsection, we prove the stability of BSGMs under a strong convexity assumption.

Proof of Theorem 4.4 According to the λ-strong convexity of $\tilde{f}$, we know $\langle \mathbf{w}-\mathbf{w}', \nabla\tilde{f}(\mathbf{w};\mathbf{z})-\nabla\tilde{f}(\mathbf{w}';\mathbf{z})\rangle \geq \lambda\|\mathbf{w}-\mathbf{w}'\|_2^2$ for any $\mathbf{w},\mathbf{w}'$ and therefore

$$\begin{aligned}
&\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla\widetilde{F}_S(\mathbf{w}_t) - \nabla\widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)})\rangle \\
&= \frac{1}{n}\sum_{j:j\neq i}\Big\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_j) - \nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_j)\Big\rangle + \frac{1}{n}\langle \mathbf{w}_t - \mathbf{w}_t^{(i)}, \nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\rangle \\
&\geq \frac{\lambda(n-1)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2}{n} - \frac{\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i) - \nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2}{n}.
\end{aligned}$$

Similar to Eq. (6.2) except using the above inequality, we derive

$$\begin{aligned}
\Delta_{t+1,i}^2 \leq \big(1 - 2\lambda\eta_t(n-1)/n\big)\Delta_{t,i}^2 + \frac{\eta_t^2}{m}\mathbb{E}\big[\|g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2\big] \\
+ \eta_t^2\mathbb{E}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2] + 2\sqrt{2}\eta_t\Delta_{t,i}\mathfrak{C}_{t,i}. \quad (6.4)
\end{aligned}$$

Plugging Eq. (4.4) and Eq. (4.5) into the above inequality, we have

$$\Delta_{t+1,i}^2 \leq \big(1 - 2\lambda\eta_t(n-1)/n + A\eta_t^2/m + \widetilde{A}\eta_t^2\big)\Delta_{t,i}^2 + \frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m} + \eta_t^2\mathbb{E}\big[\widetilde{B}_{t,i}\big] + 2\sqrt{2}\eta_t\Delta_{t,i}\mathfrak{C}_{t,i}.$$

By Eq. (4.6), the definition of $\mathfrak{C}_{t,i}$ in Eq. (4.3) and the definition of $\Delta_{t,i}$ in Eq. (6.1), we further get

$$\begin{aligned}
2\sqrt{2}\eta_t\Delta_{t,i}\mathfrak{C}_{t,i} &\leq 2\sqrt{2}\eta_t\Delta_{t,i}\Big(\mathbb{E}\big[\bar{A}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \bar{B}_{t,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}} \\
&\leq 2\sqrt{2}\bar{A}^{\frac{1}{2}}\eta_t\Delta_{t,i}^2 + 2\sqrt{2}\eta_t\Delta_{t,i}\Big(\mathbb{E}\big[\bar{B}_{t,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big)^{\frac{1}{2}} \\
&\leq 2\sqrt{2}\bar{A}^{\frac{1}{2}}\eta_t\Delta_{t,i}^2 + \frac{\lambda\eta_t}{2}\Delta_{t,i}^2 + \frac{4\eta_t}{\lambda}\Big(\mathbb{E}\big[\bar{B}_{t,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big),
\end{aligned}$$

where we have used the Cauchy's inequality in the last step. It then follows that

$$\begin{aligned}
\Delta_{t+1,i}^2 \leq \big(1 - \frac{3}{2}\lambda\eta_t + A\eta_t^2/m + \widetilde{A}\eta_t^2 + 2\sqrt{2}\bar{A}^{\frac{1}{2}}\eta_t + 2\lambda\eta_t/n\big)\Delta_{t,i}^2 + \frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m} \\
+\eta_t^2\mathbb{E}\big[\widetilde{B}_{t,i}\big] + \frac{4\eta_t}{\lambda}\Big(\mathbb{E}\big[\bar{B}_{t,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big).
\end{aligned}$$

Since $A\eta_t/m + \widetilde{A}\eta_t + 2\sqrt{2}\bar{A}^{\frac{1}{2}} + 2\lambda/n \leq \lambda/2$, the above inequality implies

$$\Delta_{t+1,i}^2 \leq (1-\lambda\eta_t)\Delta_{t,i}^2 + \frac{\eta_t^2\mathbb{E}[B_{t,i}]}{m} + \eta_t^2\mathbb{E}\big[\widetilde{B}_{t,i}\big] + \frac{4\eta_t}{\lambda}\Big(\mathbb{E}\big[\bar{B}_{t,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]/n^2\Big).$$

Applying the above inequality recursively and using $\Delta_{1,i} = 0$, we derive

$$\begin{aligned}
\Delta_{t+1,i}^2 \leq \sum_{k=1}^{t}\eta_k^2\big(\frac{\mathbb{E}[B_{k,i}]}{m} + \mathbb{E}[\widetilde{B}_{k,i}]\big)\prod_{k'=k+1}^{t}\big(1-\lambda\eta_{k'}\big) \\
+ \frac{4}{\lambda}\sum_{k=1}^{t}\eta_k\Big(\mathbb{E}\big[\bar{B}_{k,i}\big] + 4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2]/n^2\Big)\prod_{k'=k+1}^{t}\big(1-\lambda\eta_{k'}\big).
\end{aligned}$$

Taking an average over $i \in [n]$ gives the stated bound. ■

Proof of Corollary 4.5 We take $\eta_t = 1/(\lambda(t+a))$ in Theorem 4.4, and derive

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_{t+1,i}^2 \leq \sum_{k=1}^{t}\frac{1}{n\lambda^2(k+a)^2}\Big(\prod_{k'=k+1}^{t}\frac{k'+a-1}{k'+a}\Big)\sum_{i=1}^{n}\Big(\frac{\mathbb{E}[B_{k,i}]}{m}+\mathbb{E}[\widetilde{B}_{k,i}]\Big)$$
$$+\sum_{k=1}^{t}\frac{4}{n\lambda^2(k+a)}\Big(\prod_{k'=k+1}^{t}\frac{k'+a-1}{k'+a}\Big)\sum_{i=1}^{n}\Big(\mathbb{E}\big[\bar{B}_{k,i}\big]+\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2]/n^2\Big).$$

Since $\prod_{k'=k+1}^{t}\frac{k'+a-1}{k'+a}=\frac{k+a}{t+a}$, we further get that

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_{t+1,i}^2 \leq \sum_{k=1}^{t}\frac{1}{\lambda^2(k+a)^2}\frac{k+a}{(t+a)n}\sum_{i=1}^{n}\Big(\frac{\mathbb{E}[B_{k,i}]}{m}+\mathbb{E}[\widetilde{B}_{k,i}]\Big)$$
$$+\sum_{k=1}^{t}\frac{4}{\lambda^2(k+a)}\frac{k+a}{(t+a)n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\bar{B}_{k,i}\big]+4\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_k;\mathbf{z}_i)\|_2^2]/n^2\Big).$$

The proof is completed by noting $1/(k+a)\leq 1/k$ and $1/(t+a)\leq 1/t$ with $a\geq 0$. ■

7. Proofs on Applications

In this section, we present the stability bounds for specific BSGMs in Section 5.

7.1 Stochastic Gradient Descent

We first prove Lemma 5.1 to show that Eq. (4.4) and Eq. (4.5) hold for SGD.

Proof of Lemma 5.1 Since $b_t = b_t^{(i)} = 0$, it is clear that Eq. (4.6) holds with $\bar{A}=0, \widetilde{B}_{t,i}=0$. Furthermore, we have (note $\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]=\mathbb{E}[\|\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2]$ due to the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$)

$$\mathbb{E}_{j_{t,1}}\big[\|g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2\big]=\mathbb{E}_{j_{t,1}}[\|\nabla f(\mathbf{w}_t;S_{j_{t,1}})-\nabla f(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2]$$
$$=\frac{1}{n}\sum_{k:k\neq i}\|\nabla f(\mathbf{w}_t;\mathbf{z}_k)-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_k)\|_2^2+\frac{1}{n}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2$$
$$\leq L^2\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+\frac{2}{n}\Big(\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2+\|\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2\Big), \tag{7.1}$$

where we have used the fact that $j_{t,1}$ follows the uniform distribution over $[n]$. This shows Eq. (4.4) holds with our choice of A and $B_{t,i}$. Similarly, by $(a+b)^2 \leq 3a^2/2 + 3b^2$ we know

$$\begin{aligned}
&\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2 \\
&= \Big\|\frac{1}{n}\sum_{k:k\neq i}\big(\nabla f(\mathbf{w}_t;\mathbf{z}_k) - \nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_k)\big) + \frac{1}{n}\big(\nabla f(\mathbf{w}_t;\mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\big)\Big\|_2^2 \\
&\leq \frac{3L^2}{2}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{6}{n^2}(\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2 + \|\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i'\|_2^2). \qquad (7.2)
\end{aligned}$$

This, together with $\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] = \mathbb{E}[\|\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2]$, shows that Eq. (4.5) holds with our choice of $\widetilde{A}$ and $\widetilde{B}_{t,i}$. The proof is completed. ■

We now plug these estimates into Theorem 4.3 to prove Corollary 5.2.

Proof of Corollary 5.2 We plug the estimates in Lemma 5.1 to Theorem 4.3, and get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \Delta_i^2 &\leq (\frac{16}{mn^2} + \frac{48}{n^3})\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] + \frac{64\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2] \\
&\lesssim \frac{L}{mn^2}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}[f(\mathbf{w}_t;\mathbf{z}_i)] + \frac{L\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}[f(\mathbf{w}_t;\mathbf{z}_i)] \\
&= \frac{L}{mn}\sum_{t=1}^T\eta_t^2\mathbb{E}[F_S(\mathbf{w}_t)] + \frac{L\sum_{t=1}^T\eta_t^2}{n^2}\sum_{t=1}^T\mathbb{E}[F_S(\mathbf{w}_t)],
\end{aligned}$$

where we have used the self-bounding property (Lemma 3.5). This proves Eq. (5.2). We now turn to Eq. (5.3) for SGD with a constant step size. We have the following standard convergence rates for SGD (see, e.g., Lei and Ying, 2020)

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)\big] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta L F(\mathbf{w}^*), \qquad (7.3)$$

from which we know $\sum_{t=1}^T\mathbb{E}[F_S(\mathbf{w}_t)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\eta} + (1+\eta L)TF(\mathbf{w}^*)$. We can plug this estimator back into Eq. (5.2), and derive

$$\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \lesssim \Big(\frac{L\eta^2}{mn} + \frac{L\eta^2 T}{n^2}\Big)\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta} + (1+\eta L)TF(\mathbf{w}^*)\Big) \lesssim \Big(\frac{L\eta^2}{mn} + \frac{L\eta^2T}{n^2}\Big)\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta} + TF(\mathbf{w}^*)\Big),$$

where we have used the inequality $\eta L \lesssim 1$. ■

7.2 Zeroth-order SGD

In this subsection, we give the proofs for Zeroth-order SGD. Lemma 7.1 is a simple variant of Lemma 1 in Nikolakakis et al. (2022a). We omit the proof for brevity.

Lemma 7.1 *Let $\mathbf{u}_l \in \mathbb{R}^d, l \in \{1, 2 \ldots, K\}$ be i.i.d standard Gaussian. For every random vector $\mathbf{v} \in \mathbb{R}^d$ independent of all $\mathbf{u}_l, l \in \{1, 2 \ldots, K\}$, it is true that*

$$\mathbb{E}_{\mathbf{u}}\Big[\Big\|\frac{1}{K}\sum_{l=1}^{K}\langle \mathbf{v}, \mathbf{u}_l\rangle \mathbf{u}_l\Big\|_2^2\Big] = \Big(1+\frac{d+1}{K}\Big)\|\mathbf{v}\|_2^2,$$

where $\mathbb{E}_{\mathbf{u}}$ denotes the conditional expectation w.r.t. $\mathbf{u}_1, \ldots, \mathbf{u}_K$.

Now we present the proof of Lemma 5.5 and Lemma 5.6, which show that Eq. (4.4) and Eq. (4.5) hold for Zeroth-order SGD.

Proof of Lemma 5.5 Since $\mathbb{E}[X^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X])^2$ for a random variable X, we know

$$\begin{aligned}\mathbb{E}_{\mathbf{u}}[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2] = \mathbb{E}_{\mathbf{u}}[\|(g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)}))- \\ \mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2] + \|\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2. \quad (7.4)\end{aligned}$$

Since an average of K independent variables divides the variance by a factor of K, we know

$$\begin{aligned}&\mathbb{E}_{\mathbf{u}}[\|(g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})) - \mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2] \le \\ &\frac{1}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\frac{f(\mathbf{w}_t + \mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}) - f(\mathbf{w}_t; S_{j_{t,1}})}{\mu}\mathbf{u}_{t_1,1} - \frac{f(\mathbf{w}_t^{(i)} + \mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}^{(i)}) - f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})}{\mu}\mathbf{u}_{t_1,1}\Big\|_2^2\Big],\end{aligned}$$

where we have also used the inequality $\mathbb{E}[(X - \mathbb{E}[X])^2] \le \mathbb{E}[X^2]$. By the Taylor expansion, there exist $\alpha_{\mathbf{u}}$ and $\alpha_{\mathbf{u}}^{(i)}$ in $[0, 1]$ such that

$$\begin{aligned}\frac{f(\mathbf{w}_t + \mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}) - f(\mathbf{w}_t; S_{j_{t,1}})}{\mu} &= \nabla f(\mathbf{w}_t; S_{j_{t,1}})^\top \mathbf{u}_{t_1,1} + \frac{\mu}{2}\mathbf{u}_{t_1,1}^\top \nabla^2 f(\mathbf{w}_t + \alpha_{\mathbf{u}}\mu\mathbf{u}_{t_1,1}; S_{j_{t,1}})\mathbf{u}_{t_1,1}, \\ \frac{f(\mathbf{w}_t^{(i)} + \mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}^{(i)}) - f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})}{\mu} &= \nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})^\top \mathbf{u}_{t_1,1} + \frac{\mu}{2}\mathbf{u}_{t_1,1}^\top \nabla^2 f(\mathbf{w}_t^{(i)} + \alpha_{\mathbf{u}}^{(i)}\mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}^{(i)})\mathbf{u}_{t_1,1}. \quad (7.5)\end{aligned}$$

It then follows that (we use $(a+b)^2 \le 2(a^2+b^2)$)

$$\begin{aligned}&\mathbb{E}_{\mathbf{u}}[\|(g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})) - \mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2] \\ &\le \frac{2}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\big(\nabla f(\mathbf{w}_t; S_{j_{t,1}}) - \nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\big)^\top \mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big] + \\ &\frac{2}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\frac{\mu}{2}\mathbf{u}_{t_1,1}^\top\Big(\nabla^2 f(\mathbf{w}_t + \alpha_{\mathbf{u}}\mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}) - \nabla^2 f(\mathbf{w}_t^{(i)} + \alpha_{\mathbf{u}}^{(i)}\mu\mathbf{u}_{t_1,1}; S_{j_{t,1}}^{(i)})\Big)\mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big]. \quad (7.6)\end{aligned}$$

By Lemma 7.1 with $K = 1$, we derive

$$\begin{aligned}\mathbb{E}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\big(\nabla f(\mathbf{w}_t; S_{j_{t,1}}) - \nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\big)^\top \mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big] &= (d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t; S_{j_{t,1}}) - \nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2] \\ &\le L^2(d+2)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] + 4(d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2]/n, \quad (7.7)\end{aligned}$$

where we have used Eq. (7.1) in the last step. We combine the above discussions together and use the fact that f is L-smooth (eigenvalue of Hessian is smaller than L) to show

$$\mathbb{E}[\|(g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)}))-\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2] \\ \leq \frac{2(d+2)L^2\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2]}{K}+\frac{8(d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]}{Kn}+\frac{2\mu^2L^2\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6]}{K}.$$

Furthermore, according to Lemma 5.4 and similar to Eq. (7.1), we know

$$\mathbb{E}\big[\|\mathbb{E}_{\mathbf{u}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2\big]=\mathbb{E}[\|\nabla\widetilde{f}(\mathbf{w}_t;S_{j_{t,1}})-\nabla\widetilde{f}(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2] \\ \leq L^2\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2]+\frac{4}{n}\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2], \tag{7.8}$$

where we have used $\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]=\mathbb{E}[\|\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2]$. We plug the above two inequalities back into Eq. (7.4), and derive

$$\mathbb{E}[\|g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})\|_2^2]\leq\Big(1+\frac{2(d+2)}{K}\Big)L^2\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2]+ \\ \frac{8(d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]}{Kn}+\frac{2\mu^2L^2\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6]}{K}+\frac{4}{n}\mathbb{E}\big[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big].$$

The stated bound holds by noting $\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6]=d(d+2)(d+4)$ for $\mathbf{u}_{t_1,1}\sim\mathcal{N}(0,I_d)$. ■

Proof of Lemma 5.6 Similar to the proof of Lemma 5.5, we have

$$\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2]=\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]- \\ \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]]\|_2^2]+\|\mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]]\|_2^2. \tag{7.9}$$

Since averaging K independent random variables divides variance by a factor of K, we know

$$\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]-\mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]]\|_2^2] \\ \leq\frac{1}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\mathbb{E}_{j_{t,1}}\Big[\frac{f(\mathbf{w}_t+\mu\mathbf{u}_{t_1,1};S_{j_{t,1}})-f(\mathbf{w}_t;S_{j_{t,1}})}{\mu}\mathbf{u}_{t_1,1}-\frac{f(\mathbf{w}_t^{(i)}+\mu\mathbf{u}_{t_1,1};S_{j_{t,1}}^{(i)})-f(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})}{\mu}\mathbf{u}_{t_1,1}\Big]\Big\|_2^2\Big] \\ =\frac{1}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\frac{F_S(\mathbf{w}_t+\mu\mathbf{u}_{t_1,1})-F_S(\mathbf{w}_t)}{\mu}\mathbf{u}_{t_1,1}-\frac{F_{S^{(i)}}(\mathbf{w}_t^{(i)}+\mu\mathbf{u}_{t_1,1})-F_{S^{(i)}}(\mathbf{w}_t^{(i)})}{\mu}\mathbf{u}_{t_1,1}\Big\|_2^2\Big].$$

By the Taylor expansion, we know there exist $\beta_{\mathbf{u}}$ and $\beta_{\mathbf{u}}^{(i)}$ in $[0,1]$ such that

$$\frac{F_S(\mathbf{w}_t+\mu\mathbf{u}_{t_1,1})-F_S(\mathbf{w}_t)}{\mu}=\nabla F_S(\mathbf{w}_t)^\top\mathbf{u}_{t_1,1}+\frac{\mu}{2}\mathbf{u}_{t_1,1}^\top\nabla F_S^2(\mathbf{w}_t+\beta_{\mathbf{u}}\mu\mathbf{u}_{t_1,1})\mathbf{u}_{t_1,1}, \\ \frac{F_{S^{(i)}}(\mathbf{w}_t^{(i)}+\mu\mathbf{u}_{t_1,1})-F_{S^{(i)}}(\mathbf{w}_t^{(i)})}{\mu}=\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})^\top\mathbf{u}_{t_1,1}+\frac{\mu}{2}\mathbf{u}_{t_1,1}^\top\nabla^2F_{S^{(i)}}(\mathbf{w}_t+\beta_{\mathbf{u}}^{(i)}\mu\mathbf{u}_{t_1,1})\mathbf{u}_{t_1,1}.$$

It then follows from $(a+b)^2 \le 2(a^2+b^2)$ that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})] - \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]]\|_2^2] \\
\le& \frac{2}{K}\mathbb{E}_{\mathbf{u}}\Big[\Big\|\big(\nabla F_S(\mathbf{w}_t) - \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\big)^\top \mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big] \\
&+ \frac{2}{K}\Big\|\frac{\mu}{2}\mathbf{u}_{t,1}^\top\Big(\nabla^2 F_S(\mathbf{w}_t + \beta_{\mathbf{u}}\mu\mathbf{u}_{t_1,1}) - \nabla^2 F_{S^{(i)}}(\mathbf{w}_t + \beta_{\mathbf{u}}^{(i)}\mu\mathbf{u}_{t_1,1}))\Big)\mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big]. \qquad (7.10)
\end{aligned}$$

By Lemma 7.1 with $K=1$, we derive

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}\Big[\Big\|\big(\nabla F_S(\mathbf{w}_t) - \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\big)^\top \mathbf{u}_{t_1,1}\mathbf{u}_{t_1,1}\Big\|_2^2\Big] = (d+2)\mathbb{E}[\|\nabla F_S(\mathbf{w}_t) - \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2^2] \\
=&(d+2)\mathbb{E}\Big[\|\frac{1}{n}\sum_{j:j\neq i}(\nabla f(\mathbf{w}_t;\mathbf{z}_j) - \nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_j)) + \frac{1}{n}(\nabla f(\mathbf{w}_t;\mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i'))\|_2^2\Big] \\
\le&(d+2)\mathbb{E}\Big[2L^2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{2}{n^2}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2\Big] \\
\le&(d+2)\Big(\mathbb{E}[2L^2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] + \frac{4}{n^2}\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]\Big). \qquad (7.11)
\end{aligned}$$

We combine the above discussions together and use the fact that f is L-smooth (eigenvalue of Hessian is smaller than L) to show

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})] - \mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]]\|_2^2] \\
\le&\frac{4(d+2)L^2\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2]}{K} + \frac{8(d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]}{Kn^2} + \frac{2\mu^2L^2\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6]}{K}.
\end{aligned}$$

Furthermore, according to Lemma 5.4 and similar to Eq. (7.2), we know

$$\begin{aligned}
&\|\mathbb{E}_{\mathbf{u}}[\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]]\|_2^2 = \|\nabla\widetilde{F}(\mathbf{w}_t; S) - \nabla\widetilde{F}(\mathbf{w}_t^{(i)}; S^{(i)})\|_2^2 \\
&\le \frac{3L^2}{2}\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{6}{n^2}\Big(\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2 + \|\nabla\tilde{f}(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2\Big). \qquad (7.12)
\end{aligned}$$

We plug the above two inequalities back into Eq. (7.9), and use the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$ to derive

$$\begin{aligned}
\mathbb{E}_{\mathbf{u}}[\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2] \le& \Big(\frac{3}{2} + \frac{4(d+2)}{K}\Big)L^2\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2] + \\
&\frac{8(d+2)\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]}{Kn^2} + \frac{2\mu^2L^2\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6]}{K} + \frac{12}{n^2}\mathbb{E}\big[\|\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big].
\end{aligned}$$

The stated bound holds by noting $\mathbb{E}[\|\mathbf{u}_{t_1,1}\|_2^6] = d(d+2)(d+4)$ for $\mathbf{u}_{t_1,1} \sim \mathcal{N}(0, I_d)$. ■

Next, we prove Lemma 5.7 to control the norm of $\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i)$.

Proof of Lemma 5.7 By Taylor expansion, we know there exists a γ_u in $[0,1]$ such that

$$\begin{aligned}
\nabla\tilde{f}(\mathbf{w}_t;\mathbf{z}_i) =& \mathbb{E}_{\mathbf{u}}\left[\frac{f(\mathbf{w}_t + \mu\mathbf{u};\mathbf{z}_i) - f(\mathbf{w}_t;\mathbf{z}_i)}{\mu}\mathbf{u}\right] \\
=& \mathbb{E}_{\mathbf{u}}\left[\big\langle\nabla f(\mathbf{w}_t;\mathbf{z}_i),\mathbf{u}\big\rangle\mathbf{u} + \frac{\mu}{2}\mathbf{u}^\top\nabla^2 f(\mathbf{w}_t + \gamma_u\mu\mathbf{u};\mathbf{z}_i)\,\mathbf{u}\mathbf{u}\right] \\
=& \nabla f(\mathbf{w}_t;\mathbf{z}_i) + \frac{\mu}{2}\mathbb{E}_{\mathbf{u}}\big[\mathbf{u}^\top\nabla^2 f(\mathbf{w}_t + \gamma_u\mu\mathbf{u};\mathbf{z}_i)\,\mathbf{u}\mathbf{u}\big],
\end{aligned}$$

where we have used the identity $\mathbb{E}_{\mathbf{u}}[\mathbf{u}\mathbf{u}^\top] = I_d$. Since $\mathbb{E}_{\mathbf{u}}[\|\mathbf{u}\|_2^6] \leq d(d+2)(d+4)$, we get

$$\begin{aligned}\mathbb{E}\big[\big\|\nabla \tilde{f}\left(\mathbf{w}_t;\mathbf{z}_i\right)\big\|_2^2\big] =&\mathbb{E}\big[\big\|\nabla f(\mathbf{w}_t;\mathbf{z}_i) + \frac{\mu}{2}\mathbb{E}_{\mathbf{u}}[\mathbf{u}^\top\nabla^2 f(\mathbf{w}_t+\gamma_u\mu\mathbf{u};\mathbf{z}_i)\mathbf{u}\mathbf{u}]\big\|_2^2\big]\\ \leq&2\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big] + \frac{\mu^2L^2}{2}\mathbb{E}_{\mathbf{u}}[\|\mathbf{u}\|_2^6]\\ \leq&2\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big] + \frac{\mu^2L^2}{2}d(d+2)(d+4).\end{aligned}$$

The proof is completed. ■

The following lemma presents convergence rate analysis for Zeroth-order SGD, which is also useful for us to control $\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)]$ in the stability bounds.

Lemma 7.2 (Optimization of Zeroth-order SGD) *Assume for all* $\mathbf{z}$*, the map* $\mathbf{w}\mapsto f(\mathbf{w};\mathbf{z})$ *is convex and* L*-smooth. Let* $\{\mathbf{w}_t\}_t$ *be produced by BSGM with* $g(\mathbf{w}_t;S_{J_t})$ *given in Eq.* (5.4)*. If* $\eta_t=\eta\leq\big(2L\big(1+\frac{d}{K}\big)\big)^{-1}$*, then*

$$\sum_{t=1}^T\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] \lesssim \eta^{-1}\|\mathbf{w}^*\|_2^2 + Ld\mu^2T + \eta T\Big(L(1+\frac{d}{K})\mathbb{E}[F_S(\mathbf{w}^*)] + \mu^2L^2d^3\Big).$$

Proof Duchi et al. (2015) gave the following inequality in the proof of their Corollary 2

$$\mathbb{E}[\|g(\mathbf{w}_t;S_{J_t})\|_2^2] \leq 2(1+\frac{d}{K})\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})\|_2^2] + \frac{1}{2}(1+\frac{1}{K})\mu^2L^2d^3. \tag{7.13}$$

We know

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big] &= \mathbb{E}\big[\|\mathbf{w}_t-\eta_t g(\mathbf{w}_t;S_{J_t})-\mathbf{w}^*\|_2^2\big]\\ &= \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2] - 2\eta_t\mathbb{E}[\langle\mathbf{w}_t-\mathbf{w}^*, g(\mathbf{w}_t,S_{J_t})\rangle] + \eta_t^2\mathbb{E}[\|g(\mathbf{w}_t,S_{J_t})\|_2^2]\\ &= \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2] + 2\eta_t\mathbb{E}[\langle\mathbf{w}^*-\mathbf{w}_t, \nabla\widetilde{F}_S(\mathbf{w}_t)\rangle] + \eta_t^2\mathbb{E}[\|g(\mathbf{w}_t,S_{J_t})\|_2^2]\\ &\leq \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2] + 2\eta_t\mathbb{E}[\widetilde{F}_S(\mathbf{w}^*)-\widetilde{F}_S(\mathbf{w}_t)] + \eta_t^2\mathbb{E}[\|g(\mathbf{w}_t,S_{J_t})\|_2^2],\end{aligned}$$

where we have used the identity $\mathbb{E}[g(\mathbf{w}_t,S_{J_t})|\mathbf{w}_t] = \nabla\widetilde{F}_S(\mathbf{w}_t)$ and the convexity of $\widetilde{F}_S$ established in Lemma 5.4. It was shown in Nesterov and Spokoiny (2017) that

$$F_S(\mathbf{w}) \leq \widetilde{F}_S(\mathbf{w}) \leq F_S(\mathbf{w}) + \frac{\mu^2Ld}{2}, \quad \forall\mathbf{w}\in\mathcal{W}.$$

It then follows that

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big] \leq \mathbb{E}[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2] + 2\eta_t\mathbb{E}[F_S(\mathbf{w}^*)-F_S(\mathbf{w}_t)] + \eta_tLd\mu^2 + \eta_t^2\mathbb{E}[\|g(\mathbf{w}_t,S_{J_t})\|_2^2].$$

Taking a summation of the above inequality and using Eq. (7.13) imply

$$\begin{aligned}\sum_{t=1}^T\eta_t\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] &\lesssim \|\mathbf{w}^*\|_2^2 + Ld\mu^2\sum_{t=1}^T\eta_t + \sum_{t=1}^T\eta_t^2\Big((1+\frac{d}{K})\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})\|_2^2] + \mu^2L^2d^3\Big)\\ &\lesssim \|\mathbf{w}^*\|_2^2 + Ld\mu^2\sum_{t=1}^T\eta_t + \sum_{t=1}^T\eta_t^2\Big(L(1+\frac{d}{K})\mathbb{E}[F_S(\mathbf{w}_t)] + \mu^2L^2d^3\Big),\end{aligned}$$

where we have used the inequality $\|\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})\|_2^2 \le 2Lf(\mathbf{w}_t;\mathbf{z}_{j_{t,1}})$. It then follows that

$$\sum_{t=1}^T \eta_t\Big(1-L\eta_t\big(1+\frac{d}{K}\big)\Big)\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] \lesssim \|\mathbf{w}^*\|_2^2+Ld\mu^2\sum_{t=1}^T\eta_t+\sum_{t=1}^T\eta_t^2\Big(L(1+\frac{d}{K})\mathbb{E}[F_S(\mathbf{w}^*)]+\mu^2L^2d^3\Big).$$

Since $\eta_t=\eta$, we further get

$$\eta\Big(1-L\eta\big(1+\frac{d}{K}\big)\Big)\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] \lesssim \|\mathbf{w}^*\|_2^2+Ld\mu^2T\eta+\eta^2T\Big(L(1+\frac{d}{K})\mathbb{E}[F_S(\mathbf{w}^*)]+\mu^2L^2d^3\Big).$$

The stated bound then follows directly since $L\eta\big(1+\frac{d}{K}\big)\le 1/2$. ■

We are now ready to prove Corollary 5.8 on the stability bounds of Zeroth-order SGD.

Proof of Corollary 5.8 By Lemma 5.5, Lemma 5.6 and Theorem 4.3, we have $\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \le A_1+A_2+A_3$, where

$$A_1=\frac{4}{mn}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}\Big[\frac{8(d+2)\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2}{Kn}+\frac{2\mu^2L^2d(d+2)(d+4)}{K}+\frac{4}{n}\big\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\big\|_2^2\Big],$$

$$A_2=\frac{4}{n}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}\Big[\frac{8(d+2)\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2}{Kn^2}+\frac{2\mu^2L^2d(d+2)(d+4)}{K}+\frac{12}{n^2}\big\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\big\|_2^2\Big],$$

$$A_3=\frac{16\sum_{t=1}^T\eta_t^2}{n}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[4\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big].$$

By combining these three terms, we have

$$\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \lesssim \frac{d}{Kmn^2}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]+\frac{\mu^2L^2d^3}{Kn}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2 + \frac{1}{mn^2}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}[\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]+\frac{\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big].$$

Lemma 5.7 shows the following inequality $\mathbb{E}[\|\nabla\tilde f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]\le 2\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]+\frac{\mu^2L^2}{2}d(d+2)(d+4)$. We combine the above two inequalities together and derive that

$$\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \lesssim \frac{1}{mn^2}\Big(1+\frac{d}{K}\Big)\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]+\frac{\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big] + \frac{\mu^2L^2d^3}{mn^2}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2+\frac{\mu^2L^2d^3}{Kn}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2+\frac{\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mu^2L^2d^3.$$

According to Lemma 3.5, we further get

$$\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \lesssim \frac{L}{mn^2}\Big(1+\frac{d}{K}\Big)\sum_{i=1}^n\sum_{t=1}^T\eta_t^2\mathbb{E}[f(\mathbf{w}_t;\mathbf{z}_i)]+\frac{L\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[f(\mathbf{w}_t;\mathbf{z}_i)\big] + \frac{\mu^2L^2d^3}{mn^2}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2+\frac{\mu^2L^2d^3}{Kn}\sum_{i=1}^n\sum_{t=1}^T\eta_t^2+\frac{\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mu^2L^2d^3.$$

We then get Eq. (5.8) by noting that $\frac{1}{n}\sum_{i=1}^n f(\mathbf{w}_t;\mathbf{z}_i) = F_S(\mathbf{w}_t)$. We now prove Eq. (5.9). Lemma 7.2 shows that

$$\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)] \lesssim TF(\mathbf{w}^*) + \eta^{-1}\|\mathbf{w}^*\|_2^2 + Ld\mu^2 T + \eta T\Big(L(1+\frac{d}{K})\mathbb{E}[F_S(\mathbf{w}^*)] + \mu^2L^2d^3\Big).$$

We plug this inequality back into Eq. (5.8) and derive Eq. (5.9). The proof is completed. ■

Proof of Theorem 5.11 Eq. (5.10) shows that $\mathcal{A}$ is on-average ϵ-model stable with (note $\eta Ld \lesssim K$ by Eq. (5.13))

$$\epsilon^2 \lesssim \Big(\frac{L\eta^2 dT}{mnK} + \frac{LT^2\eta^2}{n^2}\Big)\Big(F(\mathbf{w}^*) + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + L^2\eta\mu^2 d^3\Big) + \frac{T\eta^2\mu^2L^2d^3}{\min\{mn,K\}} + \frac{\mu^2L^2d^3T^2\eta^2}{n^2}.$$

Eq. (5.13) and the assumption $F(\mathbf{w}^*)\lesssim 1$ implies that $F(\mathbf{w}^*)+\frac{\|\mathbf{w}^*\|_2^2}{T\eta}+Ld\mu^2+L^2\eta\mu^2d^3 \lesssim 1$. It then follows that

$$\epsilon^2 \lesssim \frac{L\eta^2dT}{mnK} + \frac{LT^2\eta^2}{n^2} + \frac{T\eta^2\mu^2L^2d^3}{\min\{mn,K\}} + \frac{\mu^2L^2d^3T^2\eta^2}{n^2}$$

and therefore

$$\epsilon \lesssim \frac{(LdT)^{\frac{1}{2}}\eta}{(mnK)^{\frac{1}{2}}} + \frac{L^{\frac{1}{2}}T\eta}{n} + \frac{T^{\frac{1}{2}}\eta\mu Ld^{\frac{3}{2}}}{\min\{mn,K\}^{\frac{1}{2}}} + \frac{\mu Ld^{\frac{3}{2}}T\eta}{n}. \tag{7.14}$$

Lemma 7.2 and the convexity of F_S imply that

$$\mathbb{E}[F_S(\mathcal{A}(S)) - F_S(\mathbf{w}^*)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\Big(\frac{LdF(\mathbf{w}^*)}{K} + LF(\mathbf{w}^*) + \mu^2L^2d^3\Big). \tag{7.15}$$

This together with the assumption $F(\mathbf{w}^*)\lesssim 1$ implies that

$$\mathbb{E}[F_S(\mathcal{A}(S))] \lesssim 1 + \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + Ld\mu^2 + \eta\Big(\frac{Ld}{K} + L + \mu^2L^2d^3\Big) \lesssim 1,$$

where in the last inequality we have used Eq. (5.13) and $\eta \lesssim 1/L$ ($LT\eta/n \lesssim 1$ from Eq. (5.13) implies $\eta \le 1/L$ since $n \lesssim T$). Lemma 3.4 and Eq. (7.14) then show that (note Eq. (5.13) and Eq. (7.14) imply $L^{\frac{1}{2}}\epsilon \lesssim 1$)

$$\begin{aligned}\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] &\lesssim L\epsilon^2 + \epsilon(L\mathbb{E}[F_S(\mathcal{A}(S))])^{\frac{1}{2}} \lesssim L^{\frac{1}{2}}\epsilon \\ &\lesssim \frac{L(dT)^{\frac{1}{2}}\eta}{(mnK)^{\frac{1}{2}}} + \frac{LT\eta}{n} + \frac{T^{\frac{1}{2}}\eta\mu(Ld)^{\frac{3}{2}}}{\min\{mn,K\}^{\frac{1}{2}}} + \frac{\mu(Ld)^{\frac{3}{2}}T\eta}{n}.\end{aligned}$$

We combine the above inequality, $F(\mathbf{w}^*)\lesssim 1$ and Eq. (7.15) together, and derive the stated bound by noting $\eta L \lesssim LT\eta/n$. ■

7.3 Clipped-SGD under Bounded Moment Condition

In this subsection, we present the proofs for Clipped-SGD. We first prove Lemma 5.13 on the Lipschitz continuity of the bias of the clipping operator.

Proof of Lemma 5.13 We proceed with the proof by considering four cases.

Case 1: If $\|\mathbf{v}_1\|_2 \geq \tau$ and $\|\mathbf{v}_2\|_2 < \tau$, then

$$\begin{aligned}\|(\text{clip}(\mathbf{v}_1,\tau) - \text{clip}(\mathbf{v}_2,\tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 &= \|\text{clip}(\mathbf{v}_1,\tau) - \mathbf{v}_1\|_2 = (1 - \frac{\tau}{\|\mathbf{v}_1\|_2})\|\mathbf{v}_1\|_2 \\ &= \|\mathbf{v}_1\|_2 - \tau \leq \|\mathbf{v}_1\|_2 - \|\mathbf{v}_2\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2.\end{aligned}$$

Case 2: If $\|\mathbf{v}_1\|_2 < \tau$ and $\|\mathbf{v}_2\|_2 \geq \tau$, then a similar argument implies directly that

$$\|(\text{clip}(\mathbf{v}_1,\tau) - \text{clip}(\mathbf{v}_2,\tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2.$$

Case 3: If $\|\mathbf{v}_1\|_2 < \tau$ and $\|\mathbf{v}_2\|_2 < \tau$, then

$$\|(\text{clip}(\mathbf{v}_1,\tau) - \text{clip}(\mathbf{v}_2,\tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 = \|(\mathbf{v}_1 - \mathbf{v}_2) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 = 0.$$

Case 4: If $\|\mathbf{v}_1\|_2 \geq \tau$ and $\|\mathbf{v}_2\|_2 \geq \tau$, then

$$\begin{aligned}\|(\text{clip}(\mathbf{v}_1,\tau) - \text{clip}(\mathbf{v}_2,\tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 =& \Big\|\frac{\tau}{\|\mathbf{v}_1\|_2}\mathbf{v}_1 - \frac{\tau}{\|\mathbf{v}_2\|_2}\mathbf{v}_2 - (\mathbf{v}_1 - \mathbf{v}_2)\Big\|_2 \\ =& \Big\|\big(\frac{\tau}{\|\mathbf{v}_1\|_2} - 1\big)(\mathbf{v}_1 - \mathbf{v}_2) - \big(\frac{\tau}{\|\mathbf{v}_2\|_2} - \frac{\tau}{\|\mathbf{v}_1\|_2}\big)\mathbf{v}_2\Big\|_2 \\ \leq& \big(1 - \frac{\tau}{\|\mathbf{v}_1\|_2}\big)\|\mathbf{v}_1 - \mathbf{v}_2\|_2 + \frac{\tau|\|\mathbf{v}_1\|_2 - \|\mathbf{v}_2\|_2|}{\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2}\|\mathbf{v}_2\|_2 \\ =& \big(1 - \frac{\tau}{\|\mathbf{v}_1\|_2}\big)\|\mathbf{v}_1 - \mathbf{v}_2\|_2 + \frac{\tau}{\|\mathbf{v}_1\|_2}|\|\mathbf{v}_1\|_2 - \|\mathbf{v}_2\|_2| \\ \leq& \big(1 - \frac{\tau}{\|\mathbf{v}_1\|_2}\big)\|\mathbf{v}_1 - \mathbf{v}_2\|_2 + \frac{\tau}{\|\mathbf{v}_1\|_2}\|\mathbf{v}_1 - \mathbf{v}_2\|_2 \\ =& \|\mathbf{v}_1 - \mathbf{v}_2\|_2.\end{aligned}$$

We combine the above four cases and derive the stated bound. The proof is completed. ■

To prove Lemma 5.14, we first introduce two lemmas. The following lemma shows that the clipping operator is non-expansive. Actually, one can show that the clipping operator is a projector onto a convex set. We omit the proof for simplicity.

Lemma 7.3 *For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{W}$, we have $\|\text{clip}(\mathbf{v}_1,\tau) - \text{clip}(\mathbf{v}_2,\tau)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2$.*

Lemma 7.4 shows the bias of Clipped-SGD satisfies a generalized Lipschitz continuity.

Lemma 7.4 *Let $\mathbf{w}, \mathbf{v} \in \mathcal{W}$ and r follow the uniform distribution over $[n]$. Then*

$$\begin{aligned}&\big\|\mathbb{E}_r[\text{clip}(\nabla f(\mathbf{w}; S_r),\tau)] - \nabla F_S(\mathbf{w}) - \mathbb{E}_r[\text{clip}(\nabla f(\mathbf{v}; S_r^{(i)}),\tau)] + \nabla F_{S^{(i)}}(\mathbf{v})\big\|_2 \\ &\quad\leq \frac{1}{n}\|\nabla f(\mathbf{w}; \mathbf{z}_i) - \nabla f(\mathbf{v}; \mathbf{z}_i')\|_2 + \frac{L\|\mathbf{w} - \mathbf{v}\|_2\mathbb{E}_r[\|\nabla f(\mathbf{w}; S_r)\|_2^p + \|\nabla f(\mathbf{v}; S_r^{(i)})\|_2^p]}{\tau^p}.\end{aligned}$$

Proof It is clear that

$$\begin{aligned}
&\big\|\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{w};S_r),\tau)]-\nabla F_S(\mathbf{w})-\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{v};S_r^{(i)}),\tau)]+\nabla F_{S^{(i)}}(\mathbf{v})\big\|_2\\
&=\big\|\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{w};S_r),\tau)]-\mathbb{E}_r[\nabla f(\mathbf{w};S_r)]-\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{v};S_r^{(i)}),\tau)]+\mathbb{E}_r[\nabla f(\mathbf{v};S_r^{(i)})]\big\|_2\\
&\leq\mathbb{E}_r\big[\|\mathrm{clip}(\nabla f(\mathbf{w};S_r),\tau)-\nabla f(\mathbf{w};S_r)-\mathrm{clip}(\nabla f(\mathbf{v};S_r^{(i)}),\tau)+\nabla f(\mathbf{v};S_r^{(i)})\|_2\big]\\
&\leq\mathbb{E}_r\big[\|\nabla f(\mathbf{w};S_r)-\nabla f(\mathbf{v};S_r^{(i)})\|_2\mathbb{I}\big[\|\nabla f(\mathbf{w};S_r)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};S_r^{(i)})\|_2>\tau\big]\big],
\end{aligned}$$

where we have used Lemma 5.13. Since r follows the uniform distribution over $[n]$, we know

$$\begin{aligned}
&\mathbb{E}_r\big[\|\nabla f(\mathbf{w};S_r)-\nabla f(\mathbf{v};S_r^{(i)})\|_2\mathbb{I}\big[\|\nabla f(\mathbf{w};S_r)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};S_r^{(i)})\|_2>\tau\big]\big]\\
&\quad=\frac{1}{n}\sum_{j:j\neq i}\|\nabla f(\mathbf{w};\mathbf{z}_j)-\nabla f(\mathbf{v};\mathbf{z}_j)\|_2\mathbb{I}\big[\|\nabla f(\mathbf{w};\mathbf{z}_j)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};\mathbf{z}_j)\|_2>\tau\big]\\
&\qquad+\frac{1}{n}\|\nabla f(\mathbf{w};\mathbf{z}_i)-\nabla f(\mathbf{v};\mathbf{z}_i')\|_2\mathbb{I}\big[\|\nabla f(\mathbf{w};\mathbf{z}_i)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};\mathbf{z}_i')\|_2>\tau\big]. \quad (7.16)
\end{aligned}$$

By the smoothness of $\mathbf{w}\mapsto f(\mathbf{w};\mathbf{z})$, we further get

$$\begin{aligned}
&\mathbb{E}_r\big[\|\nabla f(\mathbf{w};S_r)-\nabla f(\mathbf{v};S_r^{(i)})\|_2\mathbb{I}\big[\|\nabla f(\mathbf{w};S_r)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};S_r^{(i)})\|_2>\tau\big]\big]\\
&\leq\frac{L\|\mathbf{w}-\mathbf{v}\|_2}{n}\sum_{j:j\neq i}\mathbb{I}\big[\|\nabla f(\mathbf{w};\mathbf{z}_j)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};\mathbf{z}_j)\|_2>\tau\big]+\frac{1}{n}\|\nabla f(\mathbf{w};\mathbf{z}_i)-\nabla f(\mathbf{v};\mathbf{z}_i')\|_2,
\end{aligned}$$

where we simply use the inequality $\mathbb{I}\big[\|\nabla f(\mathbf{w};\mathbf{z}_i)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};\mathbf{z}_i')\|_2>\tau\big]\leq 1$. Furthermore, by the Markov's inequality, we know that

$$\begin{aligned}
&\frac{1}{n}\sum_{j:j\neq i}\mathbb{I}\big[\|\nabla f(\mathbf{w};\mathbf{z}_j)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};\mathbf{z}_j)\|_2>\tau\big]\leq\mathbb{E}_r\big[\mathbb{I}\big[\|\nabla f(\mathbf{w};S_r)\|_2>\tau\text{ or }\|\nabla f(\mathbf{v};S_r^{(i)})\|_2>\tau\big]\big]\\
&\leq\Pr\big\{\|\nabla f(\mathbf{w};S_r)\|_2>\tau\big\}+\Pr\big\{\|\nabla f(\mathbf{v};S_r^{(i)})\|_2>\tau\big\}=\Pr\big\{\|\nabla f(\mathbf{w};S_r)\|_2^p>\tau^p\big\}+\Pr\big\{\|\nabla f(\mathbf{v};S_r^{(i)})\|_2^p>\tau^p\big\}\\
&\leq\frac{\mathbb{E}_r[\|\nabla f(\mathbf{w};S_r)\|_2^p+\|\nabla f(\mathbf{v};S_r^{(i)})\|_2^p]}{\tau^p}.
\end{aligned}$$

We combine the above three inequalities together and derive that

$$\begin{aligned}
&\big\|\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{w};S_r),\tau)]-\nabla F_S(\mathbf{w})-\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{v};S_r^{(i)}),\tau)]+\nabla F_{S^{(i)}}(\mathbf{v})\big\|_2\\
&\leq\frac{1}{n}\|\nabla f(\mathbf{w};\mathbf{z}_i)-\nabla f(\mathbf{v};\mathbf{z}_i')\|_2+\frac{L\|\mathbf{w}-\mathbf{v}\|_2\mathbb{E}_r[\|\nabla f(\mathbf{w};S_r)\|_2^p+\|\nabla f(\mathbf{v};S_r^{(i)})\|_2^p]}{\tau^p}.
\end{aligned}$$

The proof is completed. ■

We now apply the above lemmas to show that Eq. (4.6) holds for the Clipped-SGD.

Proof of Lemma 5.14 According to the definition of bias and Lemma 7.4, we know that

$$\begin{aligned}
&\|b_t - b_t^{(i)}\|_2 \\
&= \Big\|\mathbb{E}_{j_{t,1}}[\mathrm{clip}(\nabla f(\mathbf{w}_t; S_{j_{t,1}}), \tau)] - \nabla F_S(\mathbf{w}_t) - \mathbb{E}_{j_{t,1}}[\mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)}), \tau)] + \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\Big\|_2 \\
&\le \frac{1}{n}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2 + \frac{L\|\mathbf{w} - \mathbf{w}_t^{(i)}\|_2 \mathbb{E}_r[\|\nabla f(\mathbf{w}_t; S_r)\|_2^p + \|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2^p]}{\tau^p} \\
&\le \frac{1}{n}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2 + \frac{2LG^p\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2}{\tau^p},
\end{aligned}$$

where we have used the assumption $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t; S_r)\|_2^p] \le G^p$, $\mathbb{E}_r[\|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2^p] \le G^p$ in the last step. It then follows from $(a+b)^2 \le 2(a^2+b^2)$ that

$$\mathbb{E}[\|b_t - b_t^{(i)}\|_2^2] \le \frac{8L^2G^{2p}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2]}{\tau^{2p}} + \frac{2}{n^2}\mathbb{E}\big[\big\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\big\|_2^2\big].$$

The stated bound then follows by the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$. ■

We then show that Eq. (4.4) and Eq. (4.5) hold for the Clipped-SGD.

Proof of Lemma 5.15 According to Lemma 7.3, we know

$$\begin{aligned}
\mathbb{E}_{j_{t,1}}[\|g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2] =& \mathbb{E}_{j_{t,1}}[\|\mathrm{clip}(\nabla f(\mathbf{w}_t; S_{j_{t,1}}), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)}), \tau)\|_2^2] \\
\le& \mathbb{E}_{j_{t,1}}[\|\nabla f(\mathbf{w}_t; S_{j_{t,1}}) - \nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})\|_2^2] \\
\le& L^2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + \frac{2}{n}\Big(\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2 + \|\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2^2\Big),
\end{aligned}$$

where we have used Eq. (7.1) in the last step. Since $\mathbb{E}[\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2] = \mathbb{E}[\|\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2^2]$, we can choose $A = L^2$ and $B_{t,i} = \frac{4}{n}\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2$. The proof is completed. ■

Proof of Lemma 5.16 By $(a+b)^2 \le 3a^2/2 + 3b^2$ we know

$$\begin{aligned}
&\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t; S_{j_{t,1}}) - g(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)})]\|_2^2 = \|\mathbb{E}_{j_{t,1}}[\mathrm{clip}(\nabla f(\mathbf{w}_t; S_{j_{t,1}}), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; S_{j_{t,1}}^{(i)}), \tau)]\|_2^2 \\
=& \Big\|\frac{1}{n}\sum_{j:j\neq i}(\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_j), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j), \tau)) + \frac{1}{n}\big(\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_i), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i'), \tau)\big)\Big\|_2^2 \\
\le& \frac{3}{2n}\sum_{j:j\neq i}\|\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_j), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j), \tau)\|_2^2 + \frac{3}{n^2}\|\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_i), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i'), \tau)\|_2^2.
\end{aligned}$$

From Lemma 7.3 and using the fact that f is L-smooth, we have

$$\|\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_j), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j), \tau)\|_2^2 \le \|\nabla f(\mathbf{w}_t; \mathbf{z}_j) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j)\|_2^2 \le L^2\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2,$$

and

$$\begin{aligned}
\|\mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_i), \tau) - \mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i'), \tau)\|_2^2 &\le \|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2^2 \\
&\le 2(\|\nabla f(\mathbf{w}_t; \mathbf{z}_i)\|_2^2 + \|\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2^2).
\end{aligned}$$

We combine the above three inequalities together and derive

$$\|\mathbb{E}_{j_{t,1}}[g(\mathbf{w}_t;S_{j_{t,1}})-g(\mathbf{w}_t^{(i)};S_{j_{t,1}}^{(i)})]\|_2^2 \leq \frac{3L^2}{2}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+\frac{6}{n^2}(\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2+\|\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2).$$

Therefore, Eq. (4.5) holds with the stated $\widetilde{A}$ and $\widetilde{B}_{t,i}$. ■

We are now in the position to prove Corollary 5.17 on the stability of Clipped-SGD.

Proof of Corollary 5.17 By Lemmas 5.15, 5.16, 5.14 and Theorem 4.3, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \Delta_i^2 \leq \frac{4}{mn}\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}\Big[\frac{4}{n}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\Big] + \frac{4}{n}\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}\Big[\frac{12}{n^2}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\Big] \\ + \frac{16\sum_{t=1}^T\eta_t^2}{n}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[8\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2+4\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2/n^2\big],\end{aligned}$$

which can be simplified as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \Delta_i^2 &\lesssim \frac{1}{mn^2}\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big] + \frac{\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big] \\ &\lesssim \frac{L}{mn^2}\sum_{i=1}^n\sum_{t=1}^T \eta_t^2\mathbb{E}\big[f(\mathbf{w}_t;\mathbf{z}_i)\big] + \frac{L\sum_{t=1}^T\eta_t^2}{n^3}\sum_{i=1}^n\sum_{t=1}^T\mathbb{E}\big[f(\mathbf{w}_t;\mathbf{z}_i)\big],\end{aligned}$$

where we have used Lemma 3.5. Eq. (5.17) follows by the definition of empirical risk.

We now turn to Eq. (5.18). By Lemma 7.5 below, we have the following rate

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)\big] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \big(G^p\tau^{2-p}+LF(\mathbf{w}^*)/n\big)\eta \\ + \Big(G^{2p}\tau^{2-2p}+G^p\tau^{1-p}(LF(\mathbf{w}^*)/n)^{\frac{1}{2}}\Big)T\eta. \quad (7.17)\end{aligned}$$

Since $\eta \lesssim 1/L$ and $2G^p\tau^{1-p}(LF(\mathbf{w}^*)/n)^{\frac{1}{2}}T\eta \leq F(\mathbf{w}^*)+G^{2p}\tau^{2-2p}LT^2\eta^2/n$, we further get

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}[F_S(\mathbf{w}_t)] \lesssim F(\mathbf{w}^*)+\frac{\|\mathbf{w}^*\|_2^2}{T\eta}+\eta G^p\tau^{2-p}+G^{2p}\tau^{2-2p}\Big(T\eta+\frac{LT^2\eta^2}{n}\Big). \quad (7.18)$$

Since $\eta_t=\eta$, Eq. (5.17) implies $\frac{1}{n}\sum_{i=1}^n\Delta_i^2 \lesssim \big(\frac{LT\eta^2}{mn}+\frac{LT^2\eta^2}{n^2}\big)\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[F_S(\mathbf{w}_t)\big]$. We combine these two inequalities together, and derive the stated bound. ■

The following lemma presents optimization error bounds for Clipped-SGD under Assumption 5.2. Recall that if g is convex and L-smooth, then it satisfies the coercivity, i.e., (Hardt et al., 2016)

$$g(\mathbf{w}) \geq g(\mathbf{w}')+\langle \mathbf{w}-\mathbf{w}', \nabla g(\mathbf{w}')\rangle + \frac{1}{2L}\|\nabla g(\mathbf{w})-\nabla g(\mathbf{w}')\|_2^2, \quad (7.19)$$

$$\langle \mathbf{w}-\mathbf{w}', \nabla g(\mathbf{w})-\nabla g(\mathbf{w}')\rangle \geq \frac{1}{L}\|\nabla g(\mathbf{w})-\nabla g(\mathbf{w}')\|_2^2. \quad (7.20)$$

Lemma 7.5 (Optimization error for Clipped-SGD) *Let Assumptions in Lemma 5.14 hold and f be convex. Let Assumption 5.1 hold. If $\eta_t \leq 1/(4L)$ and $G \lesssim \tau$, then*

$$\sum_{k=1}^{t} \eta_k \mathbb{E}\big[F_S(\mathbf{w}_k) - F_S(\mathbf{w}^*)\big] \lesssim \|\mathbf{w}^*\|_2^2 + \big(G^p \tau^{2-p} + LF(\mathbf{w}^*)/n\big) \sum_{k=1}^{t} \eta_k^2 + \Big(G^{2p}\tau^{2-2p} + G^p \tau^{1-p} (LF(\mathbf{w}^*)/n)^{\frac{1}{2}}\Big)\Big(\sum_{k=1}^{t} \eta_k\Big)^2.$$

Proof For simplicity, we assume $m = 1$ and denote $\hat{g}_t = \mathrm{clip}(\nabla f(\mathbf{w}_t; \mathbf{z}_{j_{t,1}}), \tau)$. By the update of Clipped-SGD, we know that

$$\begin{aligned} \mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\big] &= \mathbb{E}\big[\big\|\mathbf{w}_t - \eta_t \hat{g}_t - \mathbf{w}^*\big\|_2^2\big] \\ &= \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2\big] + \eta_t^2 \mathbb{E}[\|\hat{g}_t\|_2^2] + 2\eta_t \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \hat{g}_t\rangle\big]. \end{aligned} \tag{7.21}$$

By $(a+b+c+d)^2 \leq 4a^2 + 4b^2 + 4c^2 + 4d^2$, we know

$$\begin{aligned} \mathbb{E}[\|\hat{g}_t\|_2^2] \leq 4\mathbb{E}[\|\hat{g}_t - \mathbb{E}_t[\hat{g}_t]\|_2^2] + 4\mathbb{E}[\|\mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\|_2^2] \\ + 4\mathbb{E}[\|\nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}^*)\|_2^2] + 4\mathbb{E}[\|\nabla F_S(\mathbf{w}^*)\|_2^2]. \end{aligned} \tag{7.22}$$

By the coercivity property (i.e., Eq. (7.19) and Eq. (7.20)), we know

$$\mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle\big] \leq \mathbb{E}\big[F_S(\mathbf{w}^*) - F_S(\mathbf{w}_t) - (2L)^{-1}\|\nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}^*)\|_2^2\big]$$

and

$$\begin{aligned} \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle\big] &= \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}^*)\rangle\big] - \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}^*) - \nabla F_S(\mathbf{w}_t)\rangle\big] \\ &\leq \mathbb{E}\big[\|\mathbf{w}^* - \mathbf{w}_t\|_2 \|\nabla F_S(\mathbf{w}^*)\|_2\big] - L^{-1}\mathbb{E}[\|\nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}^*)\|_2^2]. \end{aligned}$$

The above two bounds of $\mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle\big]$ then imply the following two inequalities

$$\begin{aligned} &\mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \hat{g}_t\rangle\big] = \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\rangle\big] + \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle\big] \\ &\leq \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\rangle\big] + \mathbb{E}\big[F_S(\mathbf{w}^*) - F_S(\mathbf{w}_t) - (2L)^{-1}\|\nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}^*)\|_2^2\big] \end{aligned} \tag{7.23}$$

and

$$\begin{aligned} \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \hat{g}_t\rangle\big] \leq \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\rangle\big] \\ + \mathbb{E}\big[\|\mathbf{w}^* - \mathbf{w}_t\|_2 \|\nabla F_S(\mathbf{w}^*)\|_2\big] - L^{-1}\mathbb{E}[\|\nabla F_S(\mathbf{w}_t) - \nabla F_S(\mathbf{w}^*)\|_2^2]. \end{aligned} \tag{7.24}$$

We first estimate $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2]$. We combine Eq. (7.21), (7.22), (7.24) together and get

$$\begin{aligned} &\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2\big] \leq \mathbb{E}\big[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2\big] + 4\eta_t^2 \mathbb{E}[\|\hat{g}_t - \mathbb{E}_t[\hat{g}_t]\|_2^2] + 4\eta_t^2 \mathbb{E}[\|\mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\|_2^2] \\ &+4\eta_t^2 \mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big] + 2\eta_t \mathbb{E}\big[\langle \mathbf{w}^* - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_t] - \nabla F_S(\mathbf{w}_t)\rangle\big] + 2\eta_t \mathbb{E}\big[\|\mathbf{w}^* - \mathbf{w}_t\|_2 \|\nabla F_S(\mathbf{w}^*)\|_2\big], \end{aligned} \tag{7.25}$$

where we have used the inequality $4\eta_t^2\mathbb{E}[\|\nabla F_S(\mathbf{w}_t)-\nabla F_S(\mathbf{w}^*)\|_2^2]-2\eta_t L^{-1}\mathbb{E}[\|\nabla F_S(\mathbf{w}_t)-\nabla F_S(\mathbf{w}^*)\|_2^2]\leq 0$. It was shown that (Zhang et al., 2020)

$$\mathbb{E}[\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2]\leq G^p\tau^{2-p}\quad\text{and}\quad\big\|\mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\big\|_2\leq G^p\tau^{1-p}. \tag{7.26}$$

We combine the above two inequalities together, and derive that

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big]&\leq\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2\big]+4\eta_t^2G^p\tau^{2-p}+4\eta_t^2G^{2p}\tau^{2-2p}\\&\quad+4\eta_t^2\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big]+2G^p\tau^{1-p}\eta_t\mathbb{E}\big[\|\mathbf{w}^*-\mathbf{w}_t\|_2\big]+2\eta_t\mathbb{E}\big[\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla F_S(\mathbf{w}^*)\|_2\big].\end{aligned}$$

We take a summation of the above inequality, and derive

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big]&\leq\mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}^*\|_2^2\big]+4\big(G^p\tau^{2-p}+G^{2p}\tau^{2-2p}\big)\sum_{k=1}^t\eta_k^2\\&\quad+4\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big]\sum_{k=1}^t\eta_k^2+2\Big(G^p\tau^{1-p}+\big(\mathbb{E}[\|\nabla F_S(\mathbf{w}^*)\|_2^2]\big)^{\frac{1}{2}}\Big)\sum_{k=1}^t\eta_k\big(\mathbb{E}[\|\mathbf{w}^*-\mathbf{w}_k\|_2^2\big)^{\frac{1}{2}},\end{aligned}$$

where we use $\mathbb{E}[XY]\leq(\mathbb{E}[X^2])^{\frac{1}{2}}(\mathbb{E}[Y^2])^{\frac{1}{2}}$. Note the above inequality holds for any $t\in[T]$. If we denote $\bar{\triangle}=\big(\max_{k\in[t+1]}\mathbb{E}\big[\|\mathbf{w}_k-\mathbf{w}^*\|_2^2\big]\big)^{\frac{1}{2}}$, then it follows that

$$\begin{aligned}\bar{\triangle}^2\leq\mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}^*\|_2^2\big]+4\big(G^p\tau^{2-p}+G^{2p}\tau^{2-2p}\big)\sum_{k=1}^t\eta_k^2+\frac{8LF(\mathbf{w}^*)}{n}\sum_{k=1}^t\eta_k^2\\+2\Big(G^p\tau^{1-p}+\big(\frac{2LF(\mathbf{w}^*)}{n}\big)^{\frac{1}{2}}\Big)\bar{\triangle}\sum_{k=1}^t\eta_k,\end{aligned} \tag{7.27}$$

where we have used the following inequality due to $\nabla F(\mathbf{w}^*)=0$ and Lemma 3.5

$$\begin{aligned}\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big]&=\mathbb{E}\Big[\big\langle\nabla F_S(\mathbf{w}^*)-\nabla F(\mathbf{w}^*),\nabla F_S(\mathbf{w}^*)-\nabla F(\mathbf{w}^*)\big\rangle\Big]\\&=\mathbb{E}\Big[\frac{1}{n^2}\sum_{i,j=1}^n\Big\langle\nabla f(\mathbf{w}^*;\mathbf{z}_i)-\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}^*;\mathbf{z})],\nabla f(\mathbf{w}^*;\mathbf{z}_j)-\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}^*;\mathbf{z})]\Big\rangle\Big]\\&=\mathbb{E}\Big[\frac{1}{n^2}\sum_{i=1}^n\Big\langle\nabla f(\mathbf{w}^*;\mathbf{z}_i)-\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}^*;\mathbf{z})],\nabla f(\mathbf{w}^*;\mathbf{z}_i)-\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}^*;\mathbf{z})]\Big\rangle\Big]\\&=\frac{1}{n}\mathbb{E}_{\mathbf{z}}\big[\|\nabla f(\mathbf{w}^*;\mathbf{z})-\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{w}^*;\mathbf{z})]\|_2^2\big]\leq\frac{1}{n}\mathbb{E}_{\mathbf{z}}[\|\nabla f(\mathbf{w}^*;\mathbf{z})\|_2^2]\\&\leq\frac{2L}{n}\mathbb{E}_{\mathbf{z}}[f(\mathbf{w}^*;\mathbf{z})]=\frac{2LF(\mathbf{w}^*)}{n}.\end{aligned} \tag{7.28}$$

Solving the quadratic inequality of $\bar{\triangle}$ in Eq. (7.27) (by Lemma 6.1) implies that

$$\begin{aligned}\bar{\triangle}\leq 2\Big(G^p\tau^{1-p}+(2LF(\mathbf{w}^*))^{\frac{1}{2}}/n^{\frac{1}{2}}\Big)\sum_{k=1}^t\eta_k+\big(\mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}^*\|_2^2\big]\big)^{\frac{1}{2}}\\+2\Big(G^p\tau^{2-p}+G^{2p}\tau^{2-2p}+2LF(\mathbf{w}^*)/n\Big)^{\frac{1}{2}}\big(\sum_{k=1}^t\eta_k^2\big)^{\frac{1}{2}}.\end{aligned} \tag{7.29}$$

We now give the optimization error bounds. We combine Eq. (7.21), (7.22), (7.23) together and get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big] \le \mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2\big] + 4\eta_t^2\mathbb{E}[\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2] + 4\eta_t^2\mathbb{E}[\|\mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\|_2^2] + 4\eta_t^2\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big] + 2\eta_t\mathbb{E}\big[\langle \mathbf{w}^*-\mathbf{w}_t, \mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\rangle\big] + 2\eta_t\mathbb{E}\big[F_S(\mathbf{w}^*)-F_S(\mathbf{w}_t)\big], \tag{7.30}$$

where we have used the inequality $4\eta_t^2\mathbb{E}[\|\nabla F_S(\mathbf{w}_t)-\nabla F_S(\mathbf{w}^*)\|_2^2] - \eta_t L^{-1}\mathbb{E}[\|\nabla F_S(\mathbf{w}_t)-\nabla F_S(\mathbf{w}^*)\|_2^2] \le 0$. We combine this inequality and Eq. (7.26) together, and derive that

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big] \le{}& \mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}^*\|_2^2\big] + 4\eta_t^2G^p\tau^{2-p} + 4\eta_t^2G^{2p}\tau^{2-2p} \\ &+ 4\eta_t^2\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big] + 2G^p\tau^{1-p}\eta_t\mathbb{E}\big[\|\mathbf{w}^*-\mathbf{w}_t\|_2\big] + 2\eta_t\mathbb{E}\big[F_S(\mathbf{w}^*)-F_S(\mathbf{w}_t)\big].\end{aligned}$$

We take a summation of the above inequality, and derive

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2\big] + 2\sum_{k=1}^t \eta_k\mathbb{E}\big[F_S(\mathbf{w}_k)-F_S(\mathbf{w}^*)\big] \le \mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}^*\|_2^2\big] + 4\big(G^p\tau^{2-p}+G^{2p}\tau^{2-2p}+\mathbb{E}\big[\|\nabla F_S(\mathbf{w}^*)\|_2^2\big]\big)\sum_{k=1}^t\eta_k^2 + 2G^p\tau^{1-p}\sum_{k=1}^t\eta_k\mathbb{E}\big[\|\mathbf{w}^*-\mathbf{w}_k\|_2\big]. \tag{7.31}$$

We plug Eq. (7.28) and Eq. (7.29) into Eq. (7.31) and derive (noting $\tau \gtrsim G$)

$$\sum_{k=1}^t \eta_k\mathbb{E}\big[F_S(\mathbf{w}_k)-F_S(\mathbf{w}^*)\big] \lesssim \|\mathbf{w}^*\|_2^2 + \Big(G^p\tau^{2-p}+LF(\mathbf{w}^*)/n\Big)\sum_{k=1}^t\eta_k^2 + G^p\tau^{1-p}\sum_{k=1}^t\eta_k\Big(\Big(G^p\tau^{1-p}+(LF(\mathbf{w}^*))^{\frac{1}{2}}/n^{\frac{1}{2}}\Big)\sum_{k=1}^t\eta_k + \|\mathbf{w}^*\|_2 + \Big(G^p\tau^{2-p}+LF(\mathbf{w}^*)/n\Big)^{\frac{1}{2}}\Big(\sum_{k=1}^t\eta_k^2\Big)^{\frac{1}{2}}\Big),$$

where we use the assumption that $\|\mathbf{w}_1-\mathbf{w}^*\|_2^2 \lesssim \|\mathbf{w}^*\|_2^2$. It then follows that

$$\begin{aligned}\sum_{k=1}^t \eta_k\mathbb{E}\big[F_S(\mathbf{w}_k)-F_S(\mathbf{w}^*)\big] \lesssim{}& \|\mathbf{w}^*\|_2^2 + \big(G^p\tau^{2-p}+LF(\mathbf{w}^*)/n\big)\sum_{k=1}^t\eta_k^2 \\ &+ \Big(G^{2p}\tau^{2-2p}+G^p\tau^{1-p}(LF(\mathbf{w}^*)/n)^{\frac{1}{2}}\Big)\Big(\sum_{k=1}^t\eta_k\Big)^2 + G^p\tau^{1-p}\|\mathbf{w}^*\|_2\sum_{k=1}^t\eta_k \\ &+ \Big(G^{\frac{3p}{2}}\tau^{2-\frac{3p}{2}}+G^p\tau^{1-p}(LF(\mathbf{w}^*)/n)^{\frac{1}{2}}\Big)\Big(\sum_{k=1}^t\eta_k\Big)\Big(\sum_{k=1}^t\eta_k^2\Big)^{\frac{1}{2}}.\end{aligned}$$

It is clear that $2G^p\tau^{1-p}\|\mathbf{w}^*\|_2\sum_{k=1}^t\eta_k \le \|\mathbf{w}^*\|_2^2 + G^{2p}\tau^{2-2p}\big(\sum_{k=1}^t\eta_k\big)^2$ and

$$2G^{\frac{3p}{2}}\tau^{2-\frac{3p}{2}}\Big(\sum_{k=1}^t\eta_k\Big)\Big(\sum_{k=1}^t\eta_k^2\Big)^{\frac{1}{2}} \le G^p\tau^{2-p}\sum_{k=1}^t\eta_k^2 + G^{2p}\tau^{2-2p}\Big(\sum_{k=1}^t\eta_k\Big)^2,$$

$$2G^p\tau^{1-p}(LF(\mathbf{w}^*)/n)^{\frac{1}{2}}\Big(\sum_{k=1}^t\eta_k\Big)\Big(\sum_{k=1}^t\eta_k^2\Big)^{\frac{1}{2}} \le \frac{LF(\mathbf{w}^*)}{n}\sum_{k=1}^t\eta_k^2 + G^{2p}\tau^{2-2p}\Big(\sum_{k=1}^t\eta_k\Big)^2.$$

The proof is completed by combining the above inequalities together. ■

Proof of Theorem 5.19 By Assumption 5.1, Eq. (5.18) and Eq. (5.19), we know that $\mathcal{A}(S)$ is on-average ϵ-model stable with $\epsilon^2 \lesssim \frac{LT\eta^2}{mn} + \frac{LT^2\eta^2}{n^2}$. Furthermore, Assumption 5.1, Eq. (7.18) and Eq. (5.19) imply that $\mathbb{E}[F_S(\mathcal{A}(S))] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t)] \lesssim 1$. Lemma 3.4 then implies that

$$\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] \lesssim \frac{L^2T\eta^2}{mn} + \frac{L^2T^2\eta^2}{n^2} + \frac{LT^{\frac{1}{2}}\eta}{(mn)^{\frac{1}{2}}} + \frac{LT\eta}{n}.$$

We combine this generalization bound, Assumption 5.1 and Eq. (7.17) to derive

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] \lesssim \frac{\|\mathbf{w}^*\|_2^2}{T\eta} + \big(G^p\tau^{2-p} + L/n\big)\eta + \Big(G^{2p}\tau^{2-2p} + G^p\tau^{1-p}L^{\frac{1}{2}}/n^{\frac{1}{2}}\Big)T\eta + \frac{L\eta T^{\frac{1}{2}}}{(mn)^{\frac{1}{2}}} + \frac{LT\eta}{n},$$

where we have used $L\eta T^{\frac{1}{2}}/(mn)^{\frac{1}{2}} + LT\eta/n \lesssim 1$ due to Eq. (5.19). The proof is completed by noting that $2\big(G^p\tau^{1-p}L^{\frac{1}{2}}/n^{\frac{1}{2}}\big)T\eta \leq G^{2p}\tau^{2-2p}T\eta + LT\eta/n$. ■

7.4 Clipped-SGD under Bounded Central Moment Condition

We first verify the generalized Lipschitzness condition for the bias of Clipped-SGD.

Proof of Lemma 5.23 By the definition of bias, we know

$$\|b_t - b_t^{(i)}\|_2 = \big\|\mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{w}_t; S_r), \tau)] - \nabla F_S(\mathbf{w}_t) - \mathbb{E}_r[\mathrm{clip}(\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)}), \tau)] + \nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\big\|_2,$$

where r follows the uniform distribution in $[n]$. By Eq. (7.16), we know

$$\|b_t - b_t^{(i)}\|_2 \leq \frac{\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2}{n} + \frac{1}{n}\sum_{j:j\neq i}\|\nabla f(\mathbf{w}_t; \mathbf{z}_j) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j)\|_2 \mathbb{I}\big[\|\nabla f(\mathbf{w}_t; \mathbf{z}_j)\|_2 > \tau \text{ or } \|\nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_j)\|_2 > \tau\big].$$

By the smoothness of f, we know

$$\|b_t - b_t^{(i)}\|_2 \leq \frac{\|\nabla f(\mathbf{w}_t; \mathbf{z}_i) - \nabla f(\mathbf{w}_t^{(i)}; \mathbf{z}_i')\|_2}{n} + L\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 \mathbb{E}_r\big[\mathbb{I}\big[\|\nabla f(\mathbf{w}_t; S_r)\|_2 > \tau \text{ or } \|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2 > \tau\big]\big]. \quad (7.32)$$

It is clear that

$$\mathbb{I}\big[\|\nabla f(\mathbf{w}_t; S_r)\|_2 > \tau \text{ or } \|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2 > \tau\big] \leq \mathbb{I}\big[\|\nabla F_S(\mathbf{w}_t)\|_2 > \tau/2 \text{ or } \|\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2 > \tau/2\big] + \mathbb{I}\big[\|\nabla f(\mathbf{w}_t; S_r)\|_2 > \tau \text{ and } \|\nabla F_S(\mathbf{w}_t)\|_2 \leq \tau/2\big] + \mathbb{I}\big[\|\nabla f(\mathbf{w}_t^{(i)}; S_r^{(i)})\|_2 > \tau \text{ and } \|\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2 \leq \tau/2\big].$$

It then follows that

$$\mathbb{I}\big[\|\nabla f(\mathbf{w}_t;S_r)\|_2>\tau \text{ or } \|\nabla f(\mathbf{w}_t^{(i)};S_r^{(i)})\|_2>\tau\big]\le \mathbb{I}\big[\|\nabla F_S(\mathbf{w}_t)\|_2>\tau/2 \text{ or } \|\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2>\tau/2\big]$$
$$+\mathbb{I}\big[\|\nabla f(\mathbf{w}_t;S_r)-\nabla F_S(\mathbf{w}_t)\|_2>\tau/2\big]+\mathbb{I}\big[\|\nabla f(\mathbf{w}_t^{(i)};S_r^{(i)})-\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2>\tau/2\big].$$

By Markov's inequality and Assumption 5.3, we further get

$$\mathbb{E}_r\Big[\mathbb{I}\big[\|\nabla f(\mathbf{w}_t;S_r)-\nabla F_S(\mathbf{w}_t)\|_2>\tau/2\big]+\mathbb{I}\big[\|\nabla f(\mathbf{w}_t^{(i)};S_r^{(i)})-\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2>\tau/2\big]\Big]$$
$$\le\frac{\mathbb{E}_r\big[\|\nabla f(\mathbf{w}_t;S_r)-\nabla F_S(\mathbf{w}_t)\|_2^p\big]}{(\tau/2)^p}+\frac{\mathbb{E}_r\big[\|\nabla f(\mathbf{w}_t^{(i)};S_r^{(i)})-\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2^p\big]}{(\tau/2)^p}\le\frac{2^{p+1}G^p}{\tau^p}.$$

We combine the above two inequalities together and derive

$$\mathbb{E}_r\big[\mathbb{I}\big[\|\nabla f(\mathbf{w}_t;S_r)\|_2>\tau \text{ or } \|\nabla f(\mathbf{w}_t^{(i)};S_r^{(i)})\|_2>\tau\big]\big]\le\chi_{t,i}+\frac{2^{p+1}G^p}{\tau^p}.$$

We plug the above inequality back into Eq. (7.32), and get

$$\|b_t-b_t^{(i)}\|_2\le\frac{\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2}{n}+L\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\Big(\chi_{t,i}+\frac{2^{p+1}G^p}{\tau^p}\Big).$$

By the standard inequality $(a+b+c)^2\le 3a^2+3b^2+3c^2$, we further get

$$\|b_t-b_t^{(i)}\|_2^2\le\frac{3\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2}{n^2}+3L^2\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\chi_{t,i}+\frac{3L^2 2^{2p+2}G^{2p}}{\tau^{2p}}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2.$$

This gives the stated result and finishes the proof. ■

We then present high-probability convergence rates for Clipped-SGD under Assumption 5.3. To this aim, we introduce some lemmas. The following lemma give bias and variance estimates under Assumption 5.3.

Lemma 7.6 (Nguyen et al. 2023) *Let Assumption 5.3 hold and* $\hat{g}_t=\mathrm{clip}(\nabla f(\mathbf{w}_t;\mathbf{z}_{j_{t,1}}),\tau)$. *Then* $\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2\le 2\tau$. *Furthermore, if* $\|\nabla F_S(\mathbf{w}_t)\|_2\le\frac{\tau}{2}$, *then*

$$\mathbb{E}_t[\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2]\le 16G^p\tau^{2-p}\quad \textit{and}\quad \big\|\mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\big\|_2\le 4G^p\tau^{1-p}. \tag{7.33}$$

The following lemma gives a Bernstein inequality for martingale difference sequences.

Lemma 7.7 (Zhang 2005) *Let* $Z_1,\ldots,Z_t$ *be a sequence of independent random variables. Consider a sequence of functionals* $\xi_k(Z_1,\ldots,Z_k),k\in[t]$. *Let* $\sigma_t^2=\sum_{k=1}^t\mathbb{E}_{Z_k}\big[\big(\xi_k-\mathbb{E}_{Z_k}[\xi_k]\big)^2\big]$ *be the conditional variance. Assume that* $\xi_k-\mathbb{E}_{Z_k}[\xi_k]\le b$ *for each* $k\in[t]$ *and some* $b\ge 0$. *Let* $\rho\in(0,1]$ *and* $\delta\in(0,1)$. *With probability at least* $1-\delta$ *we have*

$$\sum_{k=1}^t\xi_k-\sum_{k=1}^t\mathbb{E}_{Z_k}[\xi_k]\le\frac{\rho\sigma_t^2}{b}+\frac{b\log\frac{1}{\delta}}{\rho}. \tag{7.34}$$

Lemma 7.8 (High-probability Rates of Clipped-SGD) *Let assumptions in Lemma 5.23 hold, f be convex, $\delta \in (0,1)$. Let Assumption 5.4 hold and R_T be defined in Eq. (5.27). If $\tau \geq 2\big(\frac{C_3 \log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T\big)$. Then, with probability at least $1-\delta$ we have $\|\nabla F_S(\mathbf{w}_t)\|_2 \leq \tau/2$ simultaneously for all $t \in [T]$ and*

$$\frac{1}{T}\sum_{t=1}^{T}\big(F_S(\mathbf{w}_t)-F_S(\mathbf{w}_S)\big) \lesssim \frac{\|\mathbf{w}^*\|_2^2}{\eta T}+\eta\big(G^p\tau^{2-p}+\tau^2 T^{-1}\log^2(T/\delta)\big)+\eta T\Big(\big(G^p\tau^{1-p}\big)^2+n^{-1}\log(1/\delta)\Big).$$

Proof Similar to Eq. (7.25) and Eq. (7.30) (just ignore the expectation), we know

$$\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t-\mathbf{w}^*\|_2^2+4\eta^2\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2+4\eta^2\|\mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\|_2^2+4\eta^2\|\nabla F_S(\mathbf{w}^*)\|_2^2+ 2\eta\langle \mathbf{w}^*-\mathbf{w}_t, \hat{g}_t-\mathbb{E}_t[\hat{g}_t]\rangle + 2\eta\langle \mathbf{w}^*-\mathbf{w}_t, \mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\rangle + 2\eta\|\mathbf{w}^*-\mathbf{w}_t\|_2\|\nabla F_S(\mathbf{w}^*)\|_2$$

and

$$\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2 \leq \|\mathbf{w}_t-\mathbf{w}^*\|_2^2+4\eta^2\|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2+4\eta^2\|\mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\|_2^2+4\eta^2\|\nabla F_S(\mathbf{w}^*)\|_2^2+ 2\eta\langle \mathbf{w}^*-\mathbf{w}_t, \hat{g}_t-\mathbb{E}_t[\hat{g}_t]\rangle + 2\eta\langle \mathbf{w}^*-\mathbf{w}_t, \mathbb{E}_t[\hat{g}_t]-\nabla F_S(\mathbf{w}_t)\rangle + 2\eta\big(F_S(\mathbf{w}^*)-F_S(\mathbf{w}_t)\big).$$

Taking summations over both sides of the above two inequalities, we get

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}_1-\mathbf{w}^*\|_2^2+4\eta^2\sum_{k=1}^{t}\|\hat{g}_k-\mathbb{E}_k[\hat{g}_k]\|_2^2+4\eta^2\sum_{k=1}^{t}\|\mathbb{E}_k[\hat{g}_k]-\nabla F_S(\mathbf{w}_k)\|_2^2\\ &+4\eta^2 t\|\nabla F_S(\mathbf{w}^*)\|_2^2+2\eta\sum_{k=1}^{t}\langle \mathbf{w}^*-\mathbf{w}_k, \hat{g}_k-\mathbb{E}_k[\hat{g}_k]\rangle\\ &+2\eta\sum_{k=1}^{t}\|\mathbf{w}^*-\mathbf{w}_k\|_2\|\mathbb{E}_k[\hat{g}_k]-\nabla F_S(\mathbf{w}_k)\|_2+2\eta\|\nabla F_S(\mathbf{w}^*)\|_2\sum_{k=1}^{t}\|\mathbf{w}^*-\mathbf{w}_k\|_2 \qquad (7.35)\end{aligned}$$

and

$$\begin{aligned}2\eta\sum_{k=1}^{t}\big(F_S(\mathbf{w}_k)-F_S(\mathbf{w}^*)\big) &\leq \|\mathbf{w}_1-\mathbf{w}^*\|_2^2+4\eta^2\sum_{k=1}^{t}\|\hat{g}_k-\mathbb{E}_k[\hat{g}_k]\|_2^2+4\eta^2\sum_{k=1}^{t}\|\mathbb{E}_k[\hat{g}_k]-\nabla F_S(\mathbf{w}_k)\|_2^2\\ &+4\eta^2 t\|\nabla F_S(\mathbf{w}^*)\|_2^2+2\eta\sum_{k=1}^{t}\langle \mathbf{w}^*-\mathbf{w}_k, \hat{g}_k-\mathbb{E}_k[\hat{g}_k]\rangle+2\eta\sum_{k=1}^{t}\|\mathbf{w}^*-\mathbf{w}_k\|_2\|\mathbb{E}_k[\hat{g}_k]-\nabla F_S(\mathbf{w}_k)\|_2. \qquad (7.36)\end{aligned}$$

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

$$\xi_t = \|\hat{g}_t-\mathbb{E}_t[\hat{g}_t]\|_2^2\mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2 \leq \tau/2], \quad \xi_t' = \langle \mathbf{w}^*-\mathbf{w}_t, \hat{g}_t-\mathbb{E}_t[\hat{g}_t]\rangle\mathbb{I}[\|\mathbf{w}_t-\mathbf{w}^*\|_2 \leq R_T].$$

Lemma 7.6 shows that $\xi_t \in [0, 4\tau^2]$ and $\mathbb{E}_t[\xi_t] \leq 16G^p\tau^{2-p}$. Furthermore, we know

$$\mathbb{E}_t[\xi_t^2] \leq 4\tau^2\mathbb{E}_t[\xi_t] \leq 64G^p\tau^{4-p}.$$

Lemma 7.7 with $b = 4\tau^2, \sigma_T^2 = 64G^p\tau^{4-p}T, \rho = 1$ then shows the following inequality with probability at least $1 - \delta/3$

$$\sum_{t=1}^{T}\xi_t \le \sum_{t=1}^{T}\mathbb{E}_t[\xi_t] + \frac{64G^p\tau^{4-p}T}{4\tau^2} + 4\tau^2\log\frac{3}{\delta} \le 16G^p\tau^{2-p}T + 16G^p\tau^{2-p}T + 4\tau^2\log(3/\delta)$$
$$= 32G^p\tau^{2-p}T + 4\tau^2\log(3/\delta), \tag{7.37}$$

where we have used $\mathbb{E}_t[\xi_t] \le 16G^p\tau^{2-p}$. Suppose that $\|\nabla F_S(\mathbf{w}^*)\|_2 \le \frac{C_3\log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}}$. By the L-smoothness of F_S, we know if $\|\mathbf{w} - \mathbf{w}^*\|_2 \le R_T$, then

$$\|\nabla F_S(\mathbf{w})\|_2 \le \|\nabla F_S(\mathbf{w}^*)\|_2 + \|\nabla F_S(\mathbf{w}) - \nabla F_S(\mathbf{w}^*)\|_2$$
$$\le \|\nabla F_S(\mathbf{w}^*)\|_2 + L\|\mathbf{w} - \mathbf{w}^*\|_2 \le \frac{C_3\log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T. \tag{7.38}$$

Lemma 7.6 then shows that

$$\xi_t' - \mathbb{E}_t[\xi_t'] = \xi_t' \le \|\mathbf{w}^* - \mathbf{w}_t\|_2\|\hat{g}_t - \mathbb{E}_t[\hat{g}_t]\|_2\mathbb{I}[\|\mathbf{w}_t - \mathbf{w}^*\|_2 \le R_T] \le 2R_T\tau$$

and

$$\mathbb{E}_t\big[\big(\xi_t' - \mathbb{E}_t[\xi_t']\big)^2\big] = \mathbb{E}_t[(\xi_t')^2] \le \|\mathbf{w}^* - \mathbf{w}_t\|_2^2\mathbb{I}[\|\mathbf{w}_t - \mathbf{w}^*\|_2 \le R_T]\mathbb{E}_t\big[\|\hat{g}_t - \mathbb{E}_t[\hat{g}_t]\|_2^2\big] \le 16G^p\tau^{2-p}R_T^2,$$

where in the last step we have used Lemma 7.6 (note $\|\nabla F_S(\mathbf{w}_t)\|_2 \le \tau/2$ due to $\tau \ge 2\big(\frac{C_3\log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T\big)$ and Eq. (7.38)). According to Lemma 7.7 with $b = 2R_T\tau, \sigma_t^2 = 16G^p\tau^{2-p}R_T^2 t, \rho = 1$ (with the confidence parameter being $\delta/(3T)$) and the union bounds of probability, we get the following inequality with probability at least $1 - \delta/3$ simultaneously for all $t \in [T]$

$$\sum_{k=1}^{t}\xi_k' \le \frac{16G^p\tau^{2-p}R_T^2 t}{2R_T\tau} + 2R_T\tau\log\frac{3T}{\delta} = 2R_T(4G^p\tau^{1-p}t + \tau\log(3T/\delta)). \tag{7.39}$$

From now on, we assume that $\|\nabla F_S(\mathbf{w}^*)\|_2 \le \frac{C_3\log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}}$, Eq. (7.37) and Eq. (7.39) hold simultaneously for all $t \in [T]$, which according to Assumption 5.4 and the above analyses, happens with probability at least $1 - \delta$.

We now show by induction that $\|\mathbf{w}_t - \mathbf{w}^*\|_2 \le R_T$ for all $t \in [T]$. This inequality holds with $k = 1$. Assume that it holds for $k \le t$. We now show that it holds for $k = t + 1$. Since $\tau \ge 2\big(\frac{C_3\log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T\big)$ and Eq. (7.38), Lemma 7.6 shows that

$$\mathbb{E}_k[\|\hat{g}_k - \mathbb{E}_k[\hat{g}_k]\|_2^2] \le 16G^p\tau^{2-p} \ \text{ and } \ \big\|\mathbb{E}_k[\hat{g}_k] - \nabla F_S(\mathbf{w}_k)\big\|_2 \le 4G^p\tau^{1-p}, \ \ \forall k \in [t]. \tag{7.40}$$

We plug this inequality into Eq. (7.35), and get

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \le \|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + 4\eta^2\sum_{k=1}^{t}\|\hat{g}_k - \mathbb{E}_k[\hat{g}_k]\|_2^2 + 4\eta^2 t\big(4G^p\tau^{1-p}\big)^2 + 4\eta^2 t\|\nabla F_S(\mathbf{w}^*)\|_2^2$$
$$+ 2\eta\sum_{k=1}^{t}\langle\mathbf{w}^* - \mathbf{w}_k, \hat{g}_k - \mathbb{E}_k[\hat{g}_k]\rangle + 2\eta t R_T\cdot 4G^p\tau^{1-p} + 2\eta t R_T\|\nabla F_S(\mathbf{w}^*)\|_2. \tag{7.41}$$

According to Eq. (7.38), the induction assumption $\|\mathbf{w}_k - \mathbf{w}^*\|_2 \leq R_T, k \in [t]$ and $\tau \geq 2\big(\frac{C_3 \log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}} + LR_T\big)$, we know $\mathbb{I}[\|\nabla F_S(\mathbf{w}_k)\|_2 \leq \tau/2] = 1$ and therefore

$$\|\hat{g}_k - \mathbb{E}_k[\hat{g}_k]\|_2^2 = \xi_k, \quad \text{and} \quad \langle \mathbf{w}^* - \mathbf{w}_k, \hat{g}_k - \mathbb{E}_k[\hat{g}_k]\rangle = \xi_k', \quad \forall k \in [t].$$

This, together with Eq. (7.37), Eq. (7.39) and Eq. (7.41), implies that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 \\
&\leq 4\eta^2 \sum_{k=1}^t \xi_t + 4\eta^2 t\big(\big(4G^p\tau^{1-p}\big)^2 + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big) + 2\eta \sum_{k=1}^t \xi_k' + 2\eta t R_T\big(4G^p\tau^{1-p} + \|\nabla F_S(\mathbf{w}^*)\|_2\big) \\
&\leq 4\eta^2\big(32G^p\tau^{2-p}T + 4\tau^2\log(3/\delta)\big) + 4\eta^2 t\big(4G^p\tau^{1-p}\big)^2 + 4\eta^2 t\|\nabla F_S(\mathbf{w}^*)\|_2^2 \\
&+ 2\eta t R_T\Big(8G^p\tau^{1-p} + 2\tau t^{-1}\log(3T/\delta) + 4G^p\tau^{1-p} + \|\nabla F_S(\mathbf{w}^*)\|_2\Big).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + 4\eta^2\big(32G^p\tau^{2-p}T + 4\tau^2\log(3/\delta)\big) \\
&+ \big(4\eta^2 t + 6\eta^2 t^2\big)\big(\big(12G^p\tau^{1-p}\big)^2 + 4\tau^2 t^{-2}\log^2(3T/\delta) + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big) + 2^{-1}R_T^2, \quad (7.42)
\end{aligned}$$

where we have used (note $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$)

$$\begin{aligned}
&2\eta t R_T\big(12G^p\tau^{1-p} + 2\tau t^{-1}\log(3T/\delta) + \|\nabla F_S(\mathbf{w}^*)\|_2\big) \\
&\leq 2^{-1}R_T^2 + 2\eta^2 t^2\big(12G^p\tau^{1-p} + 2\tau t^{-1}\log(3T/\delta) + \|\nabla F_S(\mathbf{w}^*)\|_2\big)^2 \\
&\leq 2^{-1}R_T^2 + 6\eta^2 t^2\big(\big(12G^p\tau^{1-p}\big)^2 + 4\tau^2 t^{-2}\log^2(3T/\delta) + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big).
\end{aligned}$$

By $\|\nabla F_S(\mathbf{w}^*)\|_2 \leq \frac{C_3 \log^{\frac{1}{2}}(3/\delta)}{\sqrt{n}}$ and the definition of R_T in Eq. (5.27), Eq. (7.42) implies that $\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 \leq R_T^2$. This completes the induction hypothesis. Therefore, with probability at least $1 - \delta$ we have

$$\|\mathbf{w}_t - \mathbf{w}^*\|_2 \leq R_T \quad \text{and} \quad \|\nabla F_S(\mathbf{w}_t)\|_2 \leq \tau/2, \quad \forall t \in [T]. \tag{7.43}$$

We now assume that Eq. (7.43) holds, for which we know $\|\hat{g}_t - \mathbb{E}_t[\hat{g}_t]\|_2^2 = \xi_t, \langle \mathbf{w}^* - \mathbf{w}_t, \hat{g}_t - \mathbb{E}_t[\hat{g}_t]\rangle = \xi_t', \forall t \in [T]$. Then, Eq. (7.36), Eq. (7.37), Eq. (7.39) and Eq. (7.40) imply

$$\begin{aligned}
&2\eta \sum_{t=1}^T \big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)\big) - \|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 \\
&\leq 4\eta^2 \sum_{t=1}^T \xi_t + 4\eta^2 T\big(\big(4G^p\tau^{1-p}\big)^2 + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big) + 2\eta \sum_{t=1}^T \xi_t' + 2\eta T R_T \cdot 4G^p\tau^{1-p} \\
&\lesssim \eta^2\big(G^p\tau^{2-p}T + \tau^2\log(1/\delta)\big) + \eta^2 T\big(\big(G^p\tau^{1-p}\big)^2 + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big) + \eta T R_T\Big(G^p\tau^{1-p} + \tau T^{-1}\log(T/\delta)\Big) \\
&\lesssim \eta^2\big(G^p\tau^{2-p}T + \tau^2\log(1/\delta)\big) + \eta^2 T\big(\big(G^p\tau^{1-p}\big)^2 + \|\nabla F_S(\mathbf{w}^*)\|_2^2\big) + \eta^2 T^2\Big(G^p\tau^{1-p} + \tau T^{-1}\log(T/\delta)\Big)^2 + R_T^2.
\end{aligned}$$

By Eq. (5.31), we further get

$$\eta\sum_{t=1}^{T}\big(F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)\big)\lesssim\|\mathbf{w}^*\|_2^2+\eta^2\big(G^p\tau^{2-p}T+\tau^2\log(1/\delta)\big)\\+\eta^2T^2\Big(G^p\tau^{1-p}+\tau T^{-1}\log(T/\delta)\Big)^2+n^{-1}\eta^2T^2\log(1/\delta).$$

The proof is completed by dividing both sides by ηT. ■

Now, we are ready to prove Corollary 5.24 on stability bounds of Clipped-SGD.

Proof of Corollary 5.24 By Lemma 5.15, Lemma 5.16, Lemma 5.23 and Theorem 4.3, we have

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2\lesssim\frac{1}{mn}\sum_{i=1}^{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\Big[\frac{1}{n}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\Big]+\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\Big[\frac{1}{n^2}\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\Big]\\+\frac{\sum_{t=1}^{T}\eta_t^2}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2}{n^2}+L^2\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\chi_{t,i}\Big],$$

where we have used $\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)-\nabla f(\mathbf{w}_t^{(i)};\mathbf{z}_i')\|_2^2]\le 4\mathbb{E}[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2]$. By the symmetry between $\mathbf{z}_i$ and $\mathbf{z}_i'$, we know

$$\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\chi_{t,i}\big]\le\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2>\tau/2]\big]+\\\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\mathbb{I}[\|\nabla F_{S^{(i)}}(\mathbf{w}_t^{(i)})\|_2>\tau/2]\big]=2\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2>\tau/2]\big].$$

Combining the above two inequalities together gives (note $m\lesssim n$)

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2\lesssim\frac{1}{mn^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big]+\frac{\sum_{t=1}^{T}\eta_t^2}{n^3}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\mathbf{w}_t;\mathbf{z}_i)\|_2^2\big]+\\+\frac{L^2\sum_{t=1}^{T}\eta_t^2}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\mathbb{I}[\|\nabla F_S(\mathbf{w}_t)\|_2>\tau/2]\big].$$

Eq. (5.28) then follows directly from Lemma 3.5.

We now turn to Eq. (5.30). Let E be the event that the inequalities in Lemma 7.8 hold, which happens with probability at least $1-\delta$. By the update of clipped SGD, we know $\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\le\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+2\tau\eta_t$, from which we get $\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\le 2\tau\sum_{t=1}^{T}\eta_t$. We plug the above inequality back into Eq. (5.28), and derive

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i^2\lesssim\Big(\frac{L\eta^2}{mn}+\frac{LT\eta^2}{n^2}\Big)\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t)\big]+L^2\tau^2\Big(\sum_{t=1}^{T}\eta_t^2\Big)\Big(\sum_{t=1}^{T}\eta_t\Big)^2\sum_{t=1}^{T}\Pr\big\{\|\nabla F_S(\mathbf{w}_t)\|_2>\tau/2\big\}\\\lesssim\Big(\frac{L\eta^2}{mn}+\frac{LT\eta^2}{n^2}\Big)\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t)\big]+L^2\tau^2T^4\eta^4\delta,\tag{7.44}$$

where we have used $\Pr\big\{\|\nabla F_S(\mathbf{w}_t)\|_2 > \tau/2\big\} \leq \delta$. The update of Clipped-SGD implies

$$\|\mathbf{w}_t - \mathbf{w}^*\|_2 \leq \|\mathbf{w}_t - \mathbf{w}_1\|_2 + \|\mathbf{w}_1 - \mathbf{w}^*\|_2 \leq T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2, \quad \forall t \in [T],$$

from which and the smoothness of F_S we know

$$\begin{aligned} F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*) &\leq \langle \mathbf{w}_t - \mathbf{w}^*, \nabla F_S(\mathbf{w}^*)\rangle + \frac{L\|\mathbf{w}_t - \mathbf{w}^*\|_2^2}{2} \leq \|\mathbf{w}_t - \mathbf{w}^*\|_2 \|\nabla F_S(\mathbf{w}^*)\|_2 + \frac{L\|\mathbf{w}_t - \mathbf{w}^*\|_2^2}{2} \\ &\leq \big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)\|\nabla F_S(\mathbf{w}^*)\|_2 + \frac{L}{2}\big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)^2. \end{aligned}$$

It then follows from $\mathbb{E}[XY] \leq (\mathbb{E}[X^2])^{\frac{1}{2}}(\mathbb{E}[Y^2])^{\frac{1}{2}}$ that ($\bar{E}$ denotes the complement of E)

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)\big)\mathbb{I}[\bar{E}]] \\ &\leq \big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)\mathbb{E}[\|\nabla F_S(\mathbf{w}^*)\|_2 \mathbb{I}[\bar{E}]] + \frac{L}{2}\big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)^2 \mathbb{E}[\mathbb{I}[\bar{E}]] \\ &\leq \big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)\big(\mathbb{E}[\|\nabla F_S(\mathbf{w}^*)\|_2^2]\big)^{\frac{1}{2}}\big(\mathbb{E}[\mathbb{I}[\bar{E}]]\big)^{\frac{1}{2}} + \frac{L}{2}\big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)^2 \Pr\{\bar{E}\} \\ &\leq \big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)\Big(\frac{2L\delta F(\mathbf{w}^*)}{n}\Big)^{\frac{1}{2}} + \frac{L}{2}\big(T\eta\tau + \|\mathbf{w}_1 - \mathbf{w}^*\|_2\big)^2\delta \lesssim L\big(T\eta\tau + \|\mathbf{w}^*\|_2\big)^2\delta + n^{-1}, \end{aligned}$$

where we have used Eq. (7.28) in the last second step, and $2ab \leq a^2 + b^2, F(\mathbf{w}^*) \lesssim 1$ in the last step. By the law of total expectation, we know

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)] &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)|E]\Pr\{E\} + \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\big(F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)\big)\mathbb{I}[\bar{E}]] \\ &\lesssim \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta\big(G^p\tau^{2-p} + \tau^2 T^{-1}\log^2(T/\delta)\big) + \eta T\big(\big(G^p\tau^{1-p}\big)^2 + n^{-1}\log(1/\delta)\big) \\ &\quad + L\big(T\eta\tau + \|\mathbf{w}^*\|_2\big)^2\delta + n^{-1}, \end{aligned} \tag{7.45}$$

where we have used Lemma 7.8. It then follows from the assumption $F(\mathbf{w}^*) \lesssim 1$ that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^T \mathbb{E}[F_S(\mathbf{w}_t)] \lesssim 1 + \frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta\big(G^p\tau^{2-p} + \tau^2 T^{-1}\log^2(T/\delta)\big) + \\ \eta T\big(\big(G^p\tau^{1-p}\big)^2 + n^{-1}\log(1/\delta)\big) + L\big(T\eta\tau + \|\mathbf{w}^*\|_2\big)^2\delta. \end{aligned}$$

We plug this inequality back into Eq. (7.44), and derive Eq. (5.30). ■

Finally, we prove Theorem 5.25 on excess risk bounds by combining the stability and convergence analyses together.

Proof of Theorem 5.25 Since $\delta \le \frac{1}{L\sqrt{n}(T\eta\tau+\|\mathbf{w}^*\|_2)^2}$, Eq. (7.45) shows

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta\big(G^p\tau^{2-p}+\tau^2T^{-1}\big) + \eta T\big(\big(G^p\tau^{1-p}\big)^2+n^{-1}\big) + \frac{1}{\sqrt{n}}\Big)\\
&= \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta\big(G^p(GT^{\frac{1}{p}})^{2-p}+(GT^{\frac{1}{p}})^2T^{-1}\big) + \eta T\big(\big(G^p(GT^{\frac{1}{p}})^{1-p}\big)^2+n^{-1}\big) + \frac{1}{\sqrt{n}}\Big)\\
&= \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta G^2T^{\frac{2-p}{p}} + \eta T\big(G^2T^{\frac{2-2p}{p}}+n^{-1}\big) + \frac{1}{\sqrt{n}}\Big) = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2^2}{\eta T} + \eta G^2T^{\frac{2-p}{p}} + \frac{\eta T}{n} + \frac{1}{\sqrt{n}}\Big),
\end{aligned}$$

where we use the choice $\tau \asymp GT^{\frac{1}{p}}$ in the second step. By (5.22) and the analysis below Eq. (5.21), we know

$$T\eta \asymp \|\mathbf{w}^*\|_2 n^{\frac{1}{2}}L^{-\frac{1}{2}} \quad \text{and} \quad \eta T^{\frac{2-p}{p}} \asymp G^{-2}\|\mathbf{w}^*\|_2 n^{-\frac{1}{2}}L^{\frac{1}{2}}. \tag{7.46}$$

It then follows that

$$\mathbb{E}[F_S(\mathcal{A}(S)) - F_S(\mathbf{w}^*)] \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)] = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2 L^{\frac{1}{2}}}{n^{\frac{1}{2}}}\Big), \tag{7.47}$$

where we assume $1 \lesssim L^{\frac{1}{2}}\|\mathbf{w}^*\|_2$ for simplicity. It then follows from the assumption $F(\mathbf{w}^*) \lesssim 1$ and $\|\mathbf{w}^*\|_2 L^{\frac{1}{2}} \lesssim n^{\frac{1}{2}}$ that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t)] \lesssim 1$. We plug this inequality back into Eq. (7.44) and get that $\mathcal{A}(S)$ is on-average ϵ-model stable with

$$\epsilon^2 \lesssim \frac{LT\eta^2}{n} + \frac{LT^2\eta^2}{n^2} + L^2\tau^2T^4\eta^4\delta \lesssim \frac{LT\eta^2}{n} + \frac{LT^2\eta^2}{n^2}, \tag{7.48}$$

where the last step uses $\delta \le (n\tau T\eta)^{-2}/L$. By our choice of parameter, we know $L^{\frac{1}{2}}\epsilon \lesssim 1$ and therefore Lemma 3.4 implies that

$$\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] \lesssim L^{\frac{1}{2}}\epsilon \lesssim \frac{LT^{\frac{1}{2}}\eta}{n^{\frac{1}{2}}} + \frac{LT\eta}{n} \lesssim \frac{LT\eta}{n}, \tag{7.49}$$

where in the last step we have used the following inequality (i.e., $T^{\frac{1}{2}}/n^{\frac{1}{2}} \lesssim T/n$) due to Eq. (5.21) and the condition $G \gtrsim L^{\frac{1}{2}}n^{\frac{p-2}{2p}}$:

$$T \asymp n^{\frac{p}{2p-2}}G^{\frac{p}{p-1}}L^{\frac{p}{2-2p}} \ge n^{\frac{p}{2p-2}}L^{\frac{p}{2p-2}}n^{\frac{p-2}{2p}\frac{p}{p-1}}L^{\frac{p}{2-2p}} = n.$$

We combine Eq. (7.47) and Eq. (7.49) together and get

$$\mathbb{E}[F(\mathcal{A}(S)) - F(\mathbf{w}^*)] = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2 L^{\frac{1}{2}}}{n^{\frac{1}{2}}} + \frac{LT\eta}{n}\Big) = \widetilde{O}\Big(\frac{\|\mathbf{w}^*\|_2 L^{\frac{1}{2}}}{n^{\frac{1}{2}}}\Big),$$

where in the last step we have used Eq. (7.46). ■

8. Conclusions

In this paper, we develop a general framework to study the stability and generalization of BSGMs. We introduce a generalized Lipschitz-type condition on the bias and the gradient estimator, under which we present a unifying generalization bound which is convenient to apply: one just needs to check the generalized Lipschitz-type condition, after which the stability bounds can be directly derived. We apply our general results to both Zeroth-order SGD and Clipped-SGD. For Zeroth-order SGD, we notice that the corresponding gradient estimator is an *unbiased* estimator of a surrogate gradient, based on which we develop the first stability bounds under a mild condition on step size sequence. For Clipped-SGD, we develop the first stability bounds under a heavy-tailed noise condition. Remarkably, our stability bounds for both Zeroth-order SGD and Clipped-SGD match those for the vanilla SGD under appropriate choice of hyperparameters. We build our analysis on the on-average model stability, which incorporates the training errors in the stability bounds to show the benefits of optimization in generalization.

There are several interesting problems for further investigation. This paper only considers applications of our framework to Zeroth-order SGD and Clipped-SGD. In the future, it would be interesting to apply our framework to study stability bounds for other BSGMs, e.g., top-k and random-k sparsification, biased compression operators, and delayed gradients. Second, it is interesting to incorporate the exp-concavity and the PL condition to the stability analysis of BSGMs.

Acknowledgement

We are grateful to the action editor and reviewers for their thoughtful comments and constructive suggestions. The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project No. 22303723, 17305425].

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

A. Ajalloeian and S. U. Stich. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

I. Amir, Y. Carmon, T. Koren, and R. Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34:25033–25043, 2021.

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

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.

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.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems*, 20, 2007.

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

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

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

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

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.

L. Chen, H. Fernando, Y. Ying, and T. Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *Advances in Neural Information Processing Systems*, 36, 2024.

P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

F. Cucker and D.-X. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

A. Cutkosky and H. Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

R. Das, S. Kale, Z. Xu, T. Zhang, and S. Sanghavi. Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*, pages 7066–7101. PMLR, 2023.

P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

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

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

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.

A. Gasnikov, D. Dvinskikh, P. Dvurechensky, E. Gorbunov, A. Beznosikov, and A. Lobanov. Randomized gradient-free methods in convex optimization. *arXiv preprint arXiv:2211.13566*, 2022.

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

E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

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.

Y. Hu, S. Zhang, X. Chen, and N. He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33:2759–2770, 2020.

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

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 34, 2021.

T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. *Advances in Neural Information Processing Systems*, 28, 2015.

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

H. Kumar, D. S. Kalogerias, G. J. Pappas, and A. Ribeiro. Actor-only deterministic policy gradient via zeroth-order gradient oracles in action space. In *IEEE International Symposium on Information Theory*, pages 1676–1681, 2021.

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

I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

Y. Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Annual Conference on Learning Theory*, pages 191–227, 2023.

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.

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

J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

S. Liu, P.-Y. Chen, X. Chen, and M. Hong. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167, 2017.

X. Liu, H. Zhang, B. Gu, and H. Chen. General stability analysis for zeroth-order optimization algorithms. In *International Conference on Learning Representations*, 2024.

G. Lugosi and G. Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546. PMLR, 2022.

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.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

K. Nikolakakis, F. Haddadpour, D. Kalogerias, and A. Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems*, 35:31525–31541, 2022a.

K. Nikolakakis, F. Haddadpour, A. Karbasi, and D. Kalogerias. Beyond Lipschitz: Sharp generalization and excess risk bounds for full-batch GD. In *International Conference on Learning Representations*, 2022b.

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

R. Pascanu, T. Mikolov, and Y. Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2(417):1, 2012.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

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

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Y. Ruan, Y. Xiong, S. Reddi, S. Kumar, and C.-J. Hsieh. Learning to learn by zeroth-order oracle. In *International Conference on Learning Representations*, 2020.

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, 2022.

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

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.

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

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

S. U. Stich and S. P. Karimireddy. The error-feedback framework: SGD with delayed gradients. *Journal of Machine Learning Research*, 21(237):1–36, 2020.

T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855, 2022.

H. Taheri and C. Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

X. Tang, A. Panda, M. Nasr, S. Mahloujifar, and P. Mittal. Private fine-tuning of large language models with zeroth-order optimization. *arXiv preprint arXiv:2401.04343*, 2024.

A. Vemula, W. Sun, and J. Bagnell. Contrasting exploration in parameter and action space: A zeroth-order optimization perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 2926–2935, 2019.

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

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

J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

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

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

T. Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

D. Zou, Y. Cao, Y. Li, and Q. Gu. Understanding the generalization of adam in learning neural networks with proper regularization. In *International Conference on Learning Representations*, 2022.