High Probability Guarantees for Nonconvex Stochastic Gradient Descent with Heavy Tails

Shaojie Li[1 2] **Yong Liu**[1 2 *]

Abstract

Stochastic gradient descent (SGD) is the workhorse in modern machine learning and data-driven optimization. Despite its popularity, existing theoretical guarantees for SGD are mainly derived in expectation and for convex learning problems. High probability guarantees of nonconvex SGD are scarce, and typically rely on "light-tail" noise assumptions and study the optimization and generalization performance separately. In this paper, we develop high probability bounds for nonconvex SGD with a joint perspective of optimization and generalization performance. Instead of the light tail assumption, we consider the gradient noise following a heavy-tailed sub-Weibull distribution, a novel class generalizing the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Under these complicated settings, we first present high probability bounds with best-known rates in general nonconvex learning, then move to nonconvex learning with a gradient dominance curvature condition, for which we improve the learning guarantees to fast rates. We further obtain sharper learning guarantees by considering a mild Bernstein-type noise condition. Our analysis also reveals the effect of trade-offs between the optimization and generalization performance under different conditions. In the last, we show that gradient clipping can be employed to remove the bounded gradient-type assumptions. Additionally, in this case, the stepsize of SGD is completely oblivious to the knowledge of smoothness.

1. Introduction

Stochastic gradient descent (SGD) has found wide applications in modern statistical and machine learning (Bousquet & Bottou, 2007; Bottou et al., 2018; Lan, 2020). As iterative algorithms, SGD works by querying an oracle for an unbiased gradient estimate built on one or several training examples in place of the exact gradients. Since its simplicity in implementation, low memory requirement, and low computational complexity per iteration, as well as good practical behavior, SGD is becoming ubiquitous in the big data era (Jain & Kar, 2017; Bubeck, 2015; Hazan et al., 2016; Lei & Tang, 2018). The success of SGD motivates the researchers to investigate its theoretical properties (Neu, 2021; Harvey et al., 2019; Madden et al., 2021; Ghadimi & Lan, 2013).

From a theoretical point of view, existing literature provides a quite comprehensive understanding regarding the expected guarantees of SGD (Harvey et al., 2019; Li & Orabona, 2020). However, expectation bounds do not capture the behavior of SGD within a single or few runs, which is related to the probabilistic nature of SGD. In addition, in practical applications such as deep learning, it is often the case that the algorithm is usually run only once since the training process may take a long time. Therefore, obtaining a high probability bound is essential to ensure the performance of the algorithm on single runs (Li & Orabona, 2020; Harvey et al., 2019; Ward et al., 2019; Cutkosky & Mehta, 2021). In particular, many problems of interest (e.g., neural network training) are nonconvex, however, few high probability analyses of SGD in the nonconvex context exist (Ghadimi & Lan, 2013; Li & Orabona, 2020; Zhou et al., 2018; Lei & Tang, 2021; Ward et al., 2019).

Many recent works suggest that SGD exhibits heavier noise than light sub-Gaussian tails (Madden et al., 2021; Gurbuzbalaban et al., 2021; Simsekli et al., 2019; Panigrahi et al., 2019; Şimşekli et al., 2019; Zhang et al., 2020a;b; 2019; Gorbunov et al., 2020; Cutkosky & Mehta, 2021). For example, Zhang et al. (2020b) provide empirical evidence that large natural language processing models based on attention and transformers (Vaswani et al., 2017; Cutkosky & Mehta, 2021) have heavy-tailed gradient noise. It was suggested in (Nguyen et al., 2019; Hodgkinson & Mahoney, 2021) that the heavier-tailed noise of SGD may strongly

[1] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

corrupt its generalization performance to testing data. In this case, many existing theoretical results assuming light sub-Gaussian tails appear to be restrictive (Li & Orabona, 2020; Zhou et al., 2018; Madden et al., 2021). Therefore, it is significant to investigate the high probability theoretical guarantees of nonconvex SGD in a heavy-tailed noise setting since it is towards a more realistic analysis (Cutkosky & Mehta, 2021; Madden et al., 2021; Barsbey et al., 2021).

Moreover, existing learning guarantees of SGD are mainly derived separately either from the point of optimization performance or generalization performance (Lei et al., 2021a). Optimization performance concerns how the learning algorithm minimizes the empirical risk, while generalization performance concerns how the predictive models learned from training samples behave on the testing samples. However, with the development of theoretical studies, it is gradually revealed that the learning performance of models is influenced by both the complexity of models and the optimization algorithms used to train the model (Lei & Tang, 2021; Neyshabur et al., 2017). Bousquet & Bottou (2007) also show that it is the interaction of optimization and generalization that determines the model's final learning performance. In this spirit, to investigate the learning guarantees of SGD, it is necessary to consider both the optimization and generalization properties. However, for nonconvex learning, such theoretical studying is still scarce (Lei & Tang, 2021), since existing works of nonconvex SGD mainly focus on the optimization (Reddi et al., 2016; Li & Orabona, 2020; Ward et al., 2019; Cutkosky & Mehta, 2021; Allen-Zhu & Hazan, 2016; Ghadimi & Lan, 2013; Ghadimi et al., 2016; Zhou et al., 2018; Madden et al., 2021).

Motivated by the above problems, in this paper, we study the high probability guarantees for nonconvex stochastic gradient descent with heavy tails by joint consideration of the optimization properties and generalization properties. To be specific, we consider a novel heavy-tailed distribution, sub-Weibull distribution (Vladimirova et al., 2020), which generalizes the sub-Gaussian and sub-Exponential families to potentially heavier-tailed ones (Camuto et al., 2021; Madden et al., 2021; Vladimirova et al., 2019; 2020; Kuchibhotla & Chakrabortty, 2018). We develop high probability bounds for the optimization and generalization properties under this distribution. Our contributions can be summarized below.

(1.) We first investigate the general nonconvex learning, for which we establish high probability optimization bounds and generalization bounds of gradients with relaxed assumptions. The analysis confirms the trade-off between the optimization and generalization performance.

(2.) We then study the nonconvex stochastic gradient descent with a gradient dominance curvature condition, for which we derive faster rates for the generalization bound of gradients and the optimization error. By balancing the two bounds, we also give fast rates for excess risk.

(3.) We then consider a mild Bernstein-type noise condition, and further provide sharper learning guarantees for the generalization bound of gradients and the excess risk. The analysis reveals that, in this case, optimization will always benefit the generalization, and the over-fitting phenomena would never happen.

(4.) We finally study SGD with clipped gradients, for which we provide a high probability learning guarantee and remove a commonly used bounded gradient assumption. To our knowledge, this is the first high probability bound for nonconvex SGD with clipping.

This paper is organized as follows. We first review the related work in Section 1.1 and then introduce the preliminaries relevant to our discussion in Section 2. Section 3 presents the main results, where we derive a series of learning guarantees for SGD. In Section 4, we conclude this paper. The complete proofs are provided in the Appendix.

1.1. Related Work

High Probability Bounds of SGD. Most of the literature proves bounds of SGD in expectation (Harvey et al., 2019; Lei & Ying, 2021; 2020). The high probability bounds of SGD are mainly provided for convex learning problems, including optimization performance (Kakade & Tewari, 2009; Hazan & Kale, 2014; Rakhlin et al., 2012; Gorbunov et al., 2020; Harvey et al., 2019; Davis & Drusvyatskiy, 2020; Davis et al., 2021; Gorbunov et al., 2021; Jain et al., 2019; Lei & Tang, 2018) and generalization performance (London, 2017; Lei et al., 2021a;b; Feldman & Vondrak, 2019; Bassily et al., 2020). As a comparison, there is relatively less high probability studies on the nonconvex learning (Madden et al., 2021; Li & Orabona, 2020). In the related work of nonconvex learning, Ghadimi & Lan (2013); Li & Orabona (2020); Zhou et al. (2018); Ward et al. (2019); Lei & Tang (2021); Madden et al. (2021) provide high probability bounds for SGD or adaptive SGD. However, most of these works assume the light-tailed sub-Gaussian gradient noise.

Noise in Neural Network. Recently, Simsekli et al. (2019); Panigrahi et al. (2019); Şimşekli et al. (2019); Gurbuzbalaban et al. (2021); Camuto et al. (2021) started the topic of heavy-tailed stochastic gradient noise, mainly focusing on Langevin dynamics and escaping saddle points. Wang et al. (2021) provide expected convergence analysis of optimization for a α-stable distribution with $\alpha \in [1, 2)$ for strongly convex learning problems. Zhang et al. (2020b); Cutkosky & Mehta (2021) instead present convergence rates for a condition that the gradients having bounded p-th moments for some $p \in (1, 2]$. Specifically, Zhang et al. (2020b) provide an in-expectation analysis for nonconvex SGD with clipping. And Cutkosky & Mehta (2021) prove high probability bounds with a combination of gradient clipping, momentum, and normalized gradient descent under the nonconvex setting. It is not clear whether the proof techniques in (Cutkosky & Mehta, 2021) can guarantee the convergence of the vanilla SGD since the momentum brings some exponential terms to accelerate convergence, and the normalized gradient descent operation is coupled with the momentum. Meanwhile, Madden et al. (2021) provide high probability analysis for nonconvex SGD with the heavy-tailed sub-Weibull gradient noise. However, they only derive bounds in the general nonconvex learning regime and focus on the optimization performance of SGD. Therefore, the high probability analysis of nonconvex SGD with heavy tails has not been thoroughly studied, even far from be understood. This paper makes an effort in this direction.

2. Preliminaries

2.1. Notations

Let P be a probability measure defined on a sample space $\mathcal{Z}$, many problems of machine learning can be cast into the following stochastic optimization problem with a hypothesis space indexed by $\mathcal{W} \subseteq \mathbb{R}^d$

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) := \mathbb{E}_{z\sim P}[f(\mathbf{w}; z)],$$

where the objective $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is possibly nonconvex and $\mathbb{E}_{z\sim P}$ denotes the expectation with respect to (w.r.t.) the random variable z drawn form P.

In statistical learning, $F(\mathbf{w})$ is often referred to as population risk (Li & Liu, 2021b). People want to learn a prediction model with small population risk. However, $F(\mathbf{w})$ is typically not accessible since the underlying distribution P is unknown. In practice, we often sample a set of i.i.d. training data $S = \{z_1, ..., z_n\}$ from P and minimize the following empirical risk (Liu, 2021; Yin et al., 2021; Li et al., 2018)

$$F_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i).$$

SGD has shown its powerful efficiency in optimizing the empirical risk $F_S(\mathbf{w})$ (Bottou et al., 2018; Ghadimi & Lan, 2013). The steps of SGD are shown in Algorithm 1.

We then introduce some notations used in this paper. Let $b' = \sup_{z\in\mathcal{Z}} \|\nabla f(0; z)\|$, where $\nabla f(\cdot; z)$ denotes the gradient of f w.r.t. the first argument and $\|\cdot\|$ denotes the Euclidean norm. Let $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}_0\| \leq R\}$ denote a ball with center $\mathbf{w}_0 \in \mathbb{R}^d$ and radius R. In this paper, we mainly assume that the set $\mathcal{W}$ satisfies $\mathcal{W} := B(\mathbf{0}, R)$, denoted by B_R. Let $\mathbf{w}(S) \in \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$ and $\mathbf{w}^* \in \arg\min_{\mathcal{W}} F(\mathbf{w})$. We also denote $A \asymp B$ if there exists universal constants $C_1, C_2 > 0$ such that $C_1 A \leq B \leq C_2 A$.

Algorithm 1 SGD

```
Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$.
for $t = 1, ..., T$ do
    draw $j_t$ from the uniform distribution over the set $\{j : j \in [n]\}$
    update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{j_t})$.
end for
```

2.2. Sub-Weibull Distribution

We now introduce the definition of sub-Weibull random variables, which is characterized by the moment generating function (MGF) (Vershynin, 2018).

Definition 2.1. (Vladimirova et al., 2020) A random variable X, satisfying

$$\mathbb{E}\Big[\exp\big((|X|/K)^{\frac{1}{\theta}}\big)\Big] \leq 2, \tag{1}$$

for some positive K and θ, is called a sub-Weibull random variable with tail parameter θ, which is denoted by $X \sim subW(\theta, K)$.

We provide some important preliminaries of sub-Weibull random variables in Appendix A. This novel class generalizes the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Sub-Weibull distributions are parameterized by a positive tail index θ and reduced to sub-Gaussian distributions for $\theta = 1/2$ and to sub-Exponential distributions for $\theta = 1$. The higher tail parameter θ corresponds to the heavier tails. The light-tailed distributions are often called the sub-Gaussian ones (Vershynin, 2018). To explicitly show the difference between the light-tailed distribution and sub-Weibull distribution, we give the definition of the light-tailed distribution below.

Definition 2.2 (Light-Tailed Distribution)**.** A random variable X, satisfying

$$\mathbb{E}\Big[\exp\big((X/K)^2\big)\Big] \leq 2,$$

for some positive K, is called a light-tailed random variable.

Therefore, in the rest of the paper by stochastic gradient noise with heavy-tailed distribution, we mean such a stochastic gradient noise that satisfies (1) with $\theta > 1/2$. However, to complete the picture of high probability bounds of nonconvex SGD, we also provide theoretical results of $\theta = 1/2$.

Remark 2.3. One of the appearing difficulties in studying sub-Weibull distribution is that when $\theta > 1$, i.e., beyond the sub-Gaussian and sub-Exponential distribution, the MGF of X doesn't exist (Bakhshizadeh et al., 2020). Note that in Definition 2.1, the MGF is defined on $|X|^{1/\theta}$. Therefore, the standard technique, i.e., finding upper bounds for the

MGF, clearly fails for the heavy-tailed sub-Weibull distribution, which also means that it is not easy to establish the concentration of measure inequalities for it. However, the concentration of measure inequalities for martingales, especially the Bernstein-type inequality, plays an essential role in our analysis. With the recent theoretical advances of martingale inequalities for heavy-tailed distributions (Li, 2021; Fan & Giraudo, 2019; Madden et al., 2021; Bakhshizadeh et al., 2020), we use these tools to derive a series of learning guarantees for nonconvex SGD with heavy tails from both points of the optimization and generalization properties.

2.3. Assumptions

We first demonstrate our assumption on the stochastic gradient noise, shown as follows.

Assumption 2.4 (Sub-Weibull Noise)**.** Conditioned on the previous iterates, we assume the gradient noise $\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)$ is centered and $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \sim subW(\theta, K)$ such that $\theta \geq \frac{1}{2}$, i.e.,

$$\mathbb{E}_{j_t}[\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)] = 0,$$

and

$$\mathbb{E}_{j_t}\left[\exp\left((\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|/K)^{\frac{1}{\theta}}\right)\right] \leq 2.$$

Remark 2.5. The clear motivation of our study on the heavy-tailed sub-Weibull stochastic gradient noise is that many recent works suggest that SGD exhibits heavier noise than sub-Gaussian (Madden et al., 2021; Gurbuzbalaban et al., 2021; Simsekli et al., 2019; Panigrahi et al., 2019; Şimşekli et al., 2019; Zhang et al., 2020a;b; 2019; Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Wang et al., 2021). For instance, there is strong empirical evidence that the gradient noise often exhibits a heavy-tailed behavior in fully connected and convolutional neural networks (Şimşekli et al., 2019; Gurbuzbalaban & Hu, 2021) as well as attention-based neural networks (Wang et al., 2021; Zhang et al., 2020b). Moreover, Vladimirova et al. (2019; 2020) show that a Gaussian prior on the weights of a Bayesian neural network produces a sub-Weibull distribution on the weights. Thus, the theoretical analysis of this paper may be helpful to the study of the Bayesian neural network.

Assumption 2.6 (Smoothness)**.** Let $\beta > 0$. For any sample $z \in \mathcal{Z}$ and $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, a differentiable function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is β-smooth if

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

Remark 2.7. This assumption is necessary to have the convergence of the gradients to zero (Li & Orabona, 2020). It is widely used in the optimization and generalization analysis of nonconvex learning problems (Feldman & Vondrak, 2019; Hardt et al., 2016; Reddi et al., 2016; Foster et al., 2018; Li & Liu, 2021c; 2022).

Assumption 2.8. There exists $G > 0$ such that for all $S \in \mathcal{Z}^n$,

$$\eta_t\|\nabla F_S(\mathbf{w}_t)\| \leq G, \forall t \in \mathbb{N}.$$

Remark 2.9. The bounded stochastic gradient assumption, i.e., $\|\nabla f(\mathbf{w}_t; z)\| \leq G$ for $\forall t \in \mathbb{N}$ and $z \in \mathcal{Z}$, is very common in the stochastic optimization literature (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Reddi et al., 2016; Harvey et al., 2019; Li et al., 2021). Compared to this bounded gradient assumption, Assumption 2.8 is mild since the stepsize η_t should goes to zero along with the increase of iterate number t. Typical choices of η_t are $\mathcal{O}(t^{-\frac{1}{2}})$ and $\mathcal{O}(t^{-1})$ (Ghadimi & Lan, 2013; Lei & Tang, 2021), which are also our studied ones in this paper. We highlight that we also show the gradient clipping can be served as a potential tool to remove this assumption, please refer to Section 3.4.

Assumption 2.10 (Noise Condition)**.** There exists $G_* > 0$ such that for all $2 \leq k \leq n$,

$$\mathbb{E}_z\left[\|\nabla f(\mathbf{w}^*, z)\|^k\right] \leq 2^{-1}k!\mathbb{E}_z\left[\|\nabla f(\mathbf{w}^*, z)\|^2\right]G_*^{k-2}.$$

Remark 2.11. Assumption 2.10 is a classical Bernstein condition (Wainwright, 2019) on gradient norms. This assumption is pretty mild since it was assumed at the optima $\mathbf{w}^*$. Moreover, in our theoretical results, G_* will always exist in the $\mathcal{O}(1/n^2)$ term, it, therefore, produces little influence on the learning guarantees.

Assumption 2.12 (PL Condition)**.** Assume that for any $S \in \mathcal{Z}^n$, there exists an $\mu_S > 0$ such that

$$F_S(\mathbf{w}) - F_S(\mathbf{w}(S)) \leq (4\mu_S)^{-1}\|\nabla F_S(\mathbf{w})\|^2, \forall \mathbf{w} \in \mathcal{W}.$$

Remark 2.13. PL condition is also referred to as "gradient dominance condition" (Foster et al., 2018). This condition simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. PL condition is one of the weakest curvature conditions and is widely employed in nonconvex learning, such as (Xu & Zeevi, 2020; Xu & Zeevi, 2020; Lei & Tang, 2021; Lei & Ying, 2021; Lei et al., 2021a; Charles & Papailiopoulos, 2018; Zhou et al., 2018; Reddi et al., 2016; Karimi et al., 2016), to mention but a few. Under suitable assumptions on the input, many popular nonconvex objective functions satisfy the PL condition, including mixture of two Gaussians (Balakrishnan et al., 2017), phase retrieval (Sun et al., 2018), robust regression (Liu et al., 2016), blind deconvolution (Li et al., 2019), matrix factorization (Liu et al., 2016), linear dynamical systems (Hardt et al., 2018), neural networks with one hidden layer (Li & Yuan, 2017), ResNets with linear activations (Hardt & Ma, 2016), etc. Moreover, Liu et al. (2020) recently show that sufficiently over-parameterized systems, including wide neural networks, generally satisfy the PL condition locally around random initialization.

3. Main Results

In this section, we show the main results of this paper. We first consider general non-convex learning setting in Section 3.1 and then non-convex learning satisfying the PL condition in Section 3.2. Section 3.3 further considers the Bernstein-type noise condition (Assumption 2.10). In Section 3.4 we study SGD with clipping.

3.1. General Nonconvex Learning

We study heavy-tailed SGD with joint consideration of optimization and generalization performance, as discussed before. For characterizing this, we first present high probability bounds on the gradients of empirical risks, written as $\|\nabla F_S(\mathbf{w}_t)\|^2$, which is relevant for the optimization performance since it corresponds to that the optimization algorithm minimizes the empirical risk F_S.

Theorem 3.1. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \le 1/(2\beta)$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log(1/\delta)\log T}{\sqrt{T}}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\ =\mathcal{O}\Big(\frac{\log^{\theta-1}(T/\delta)\log(1/\delta)+\log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big).$$

Remark 3.2. In the general nonconvex case, since we cannot guarantee that the algorithm can find a global minimizer, we therefore use the norm of gradients to measure the performance of SGD (Ghadimi & Lan, 2013; Madden et al., 2021). From Theorem 3.1, one can see that when θ increases, the learning guarantees of optimization performance are growing worse. For instance, when $\theta > 1$, an extra term $\log^{\theta-1}(T/\delta)\log(1/\delta)$ appears. Nonetheless, the learning bounds we established are still of the order $\mathcal{O}(1/\sqrt{T})$ when hiding the logarithmic terms. Theorem 3.1 also shows that for the light-tailed sub-Gaussian noise (i.e., $\theta = 1/2$), Assumption 2.8 doesn't required, which is because in this case, the sub-Weibull Freedman inequality in Lemma A.5 is hold for any $\alpha > 0$ (see (3) for details). But for heavy-tailed gradient noise, we need Assumption 2.8 to guarantee $\alpha \ge b\max_{i\in[T]} m_i$ (see (4) and (5) for details). We now compare Theorem 3.1 with the related work of high probability bounds of nonconvex SGD. Ghadimi & Lan (2013) are the first to analyze nonconvex SGD, whose bound's dependency on the confidence parameter $1/\delta$ is linear. As a comparison, Theorem 3.1 presents a logarithm dependency. Lei & Tang (2021) then provide the $\mathcal{O}(\log(1/\delta)/\sqrt{T})$ order bound, but require assuming $\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t; z_{j_t})\| \le G$. The work most relevant to us is (Madden et al., 2021). Motivated by recent research on heavy-tailed phenomena in SGD, Madden et al. (2021) also study the heavy-tailed sub-Weibull gradient noise. They provide upper bounds of the similar order, but require assuming $\|\nabla F_S(\mathbf{w}_t)\| \le G$. By comparison, Assumption 2.8 is milder than the corresponding ones of (Madden et al., 2021; Lei & Tang, 2021), meaning that we establish the comparable bounds by the relaxed assumption. Additionally, Li & Orabona (2020); Zhou et al. (2018); Ward et al. (2019) provide high probability bounds for adaptive SGD under the nonconvex setting. However, Li & Orabona (2020); Zhou et al. (2018) only focus on the case of $\theta = 1/2$ and the bounds in Ward et al. (2019) have linear dependency on the confidence parameter $1/\delta$. Roughly speaking, their bounds are of the order $\mathcal{O}(1/\sqrt{T})$ (Li & Orabona, 2020; Zhou et al., 2018; Ward et al., 2019). Overall, in comparison with related work, our established learning bounds under the heavy-tailed setting in Theorem 3.1 show pretty strong competitiveness. We highlight here we also improve the $\mathcal{O}(1/\sqrt{T})$ order bounds to $\mathcal{O}(1/T)$ order in Section 3.2.

We then provide high probability bounds on the generalization error of gradients, written as $\|\nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|^2$, which is related to the generalization performance since it corresponds to approximating the population gradient by its empirical counterpart based on training samples (Foster et al., 2018; Mei et al., 2018; Lei & Tang, 2021).

Theorem 3.3. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \le 1/(2\beta)$. For any $\delta \in (0,1)$ and uniformly for all $t = 1, ...T$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 \\ =\mathcal{O}\Big(\frac{T^{\frac{1}{2}}\log^2(\frac{1}{\delta})(\log T)\big(d+\log(\frac{1}{\delta})\big)}{n}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 \\ =\mathcal{O}\Big(\frac{T^{\frac{1}{2}}\log^{2\theta+1}(\frac{1}{\delta})(\log T)\big(d+\log(\frac{1}{\delta})\big)}{n}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\big(d + \log(\frac{1}{\delta})\big) \times \frac{T^{\frac{1}{2}}\big(\log^{\theta-1}(T/\delta)\log(1/\delta) + \log^{(2\theta+1)}(\frac{1}{\delta})\log T\big)}{n}\Big).$$

Remark 3.4. Theorem 3.3 shows that when $\theta \in [1/2, 1]$, the generalization error bounds of gradients are of the order $\mathcal{O}(T^{\frac{1}{2}}(d+\log(\frac{1}{\delta}))/n)$ when hiding other logarithmic terms. When the tail $\theta > 1$, the guarantee is clearly worse since an extra term $\log^{\theta-1}(T/\delta)\log(1/\delta)$ appears. Similarly, for light-tailed sub-Gaussian noise, Assumption 2.8 is unnecessary. Lei & Tang (2021) also study the generalization error bound of gradient in the general nonconvex case, but they need assuming $\sqrt{\eta_t}\|\nabla f(\mathbf{w}_t; z_{j_t})\| \leq G$, which is stronger than Assumption 2.8 , as discussed in Remark 3.2. Moreover, Madden et al. (2021); Li & Orabona (2020); Zhou et al. (2018); Ward et al. (2019); Ghadimi & Lan (2013) only study the optimization performance. The generalization bounds in Theorem 3.3 would converge if the sample size $n \geq \mathcal{O}\big(T^{\frac{1}{2}}(d + \log(\frac{1}{\delta}))\big)$. Therefore, it provides novel high probability learning guarantees for the generalization property of nonconvex SGD with heavy tails.

Based on Theorems 3.1 and 3.3, we present high probability generalization bounds on the gradients of population risk.

Theorem 3.5. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \leq 1/(2\beta)$. Selecting $T \asymp n/d$. For any $\delta \in (0, 1)$, with probability $1 - \delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{1}{2}}\log(\frac{n}{d})\log^3(\frac{1}{\delta})\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{1}{2}}\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta})\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\big(\frac{d}{n}\big)^{\frac{1}{2}}\big(\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta}) + \log^{\theta-1}(\frac{n}{d\delta})\log^2(\frac{1}{\delta})\big)\Big).$$

Remark 3.6. The analysis of Theorem 3.5 is based on an error decomposition: $\|\nabla F(\mathbf{w}_t)\|^2 \leq 2(\|\nabla F_S(\mathbf{w}_t)\|^2 + \|\nabla F_S(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|^2)$, where the latter two terms are bounded in Theorems 3.1 and 3.3. Clearly, the optimization performance in Theorem 3.1 improves as the iterate number T increases, while the generalization performance would worsen as shown in Theorem 3.3. Therefore, we need to trade off the optimization and generalization to reach a better balance. By choosing the proper iteration number $T \asymp n/d$, we obtain the stated bounds of Theorem 3.5.

3.2. Nonconvex Learning with PL Condition

Analogous to Section 3.1, we first show high probability bounds for optimization error. Note that we can derive guarantees for optimization error of function values instead of gradients under the PL condition.

Theorem 3.7. *Suppose Assumptions 2.4, 2.6, and 2.12 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log(1/\delta)}{T}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T}{T}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}}T}{T}\Big).$$

Remark 3.8. The learning bounds established in Theorem 3.7 are of the order $\mathcal{O}(1/T)$ when hiding the logarithmic terms, which are faster than the order $\mathcal{O}(1/\sqrt{T})$ in Theorem 3.1. Similar to the analysis in Section 3.1, Theorem 3.7 confirms that when θ increases, the learning guarantees of the optimization error are becoming worse.

We then develop high probability bounds for the generalization error of gradients under the PL condition.

Theorem 3.9. *Suppose Assumptions 2.4, 2.6, and 2.12 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Then for any $\delta \in (0, 1)$ and uniformly for all $t = 1, ...T$, with probability $1 - \delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log T\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log T\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n} \times \big(\log^{(2\theta+1)}(\frac{1}{\delta}) + \log^{\theta-1}(T/\delta)\log(1/\delta)\big)\log T\Big).$$

Remark 3.10. Theorem 3.9 suggests that when the PL condition is satisfied, the generalization error of gradients would have a logarithmic dependency on the iterate number T, which significantly improves the square-root dependency in Theorem 3.3. Therefore, compared to the sample size n in the upper bounds, the increasing optimization process (i.e., increasing T) has little influence on the generalization performance due to the logarithmic dependency of T.

Combined with Theorems 3.7 and 3.9, we show high probability bounds for excess risk $F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)$ (Foster et al., 2018; Feldman & Vondrak, 2019; Bassily et al., 2020). Additionally, we assume the population risk F satisfies the PL condition for some positive constant μ:

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

Theorem 3.11. *Suppose Assumptions 2.4, 2.6, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n$. Then for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log n\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log n\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta})\log n\Big).$$

Remark 3.12. Theorem 3.11 suggests that if F_S and F satisfy the PL condition, the excess risk would be of the order $\mathcal{O}(1/n)$ w.r.t. the sample size n, which significantly improves the $\mathcal{O}(1/\sqrt{n})$ order in Theorem 3.5. The previous Theorem 3.9 also confirms the trade-off between the optimization and generalization even under the PL condition. Specifically, when the generalization error bound of gradients dominates the final excess risk bound, more training processes (i.e., increasing T), although reducing the optimization error of Theorem 3.7, would still increase the excess risk bound due to the logarithmic dependency of T. By choosing the appropriate iteration number $T \asymp n$, we obtain the stated bounds of Theorem 3.11.

3.3. Towards Sharper Learning Guarantees

In this section, we consider the case $\mathbf{w} \in \mathcal{W} := B(\mathbf{w}^*, R)$. When the Bernstein-type noise condition is satisfied, we can further improve the learning guarantees of the generalization error of gradients and the excess risk.

Theorem 3.13. *Suppose Assumptions 2.4, 2.6, 2.10, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, then for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\Big(\frac{\log(1/\delta)}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\Big(\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(\frac{1}{\delta})}{n} + \frac{\log^2(1/\delta)}{n^2} + \frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T}\Big).$$

Remark 3.14. Theorem 3.13 demonstrates that when both the PL condition and the Bernstein-type noise condition are satisfied, the generalization error of gradients would be of the order $\mathcal{O}(1/T)$ w.r.t. the iterate number T when hiding the logarithmic terms, which significantly improves the logarithmic dependency in Theorem 3.9. This means that the optimization (i.e., increasing T) will always benefit the generalization, and the over-fitting phenomena would never happen. Additionally, note that in this case, more training processes will reduce the influence of heavy-tailed gradient noise. For instance, when $\theta > 1$, $\big(\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T\big)/T$ will remain to decrease as the iterate number T increases.

Combined with Theorems 3.13 and 3.7, we further improve the learning guarantees of excess risk, shown as follows.

Theorem 3.15. *Suppose Assumptions 2.4, 2.6, 2.10, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n^2$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, for any $\delta \in (0,1)$, with probability $1-\delta$,* *(a.) if $\theta = \frac{1}{2}$, then we have the following inequality*

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^2(\frac{1}{\delta})}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(\frac{1}{\delta})}{n}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n} + \frac{\log^{\frac{3(\theta-1)}{2}}(n/\delta)\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big).$$

Remark 3.16. Theorem 3.15 suggests that under an extra Bernstein-type noise condition, the excess risk would be of the order $\mathcal{O}\big((\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta))/n\big)$. The term $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2])$ is tiny since it depends on the optima $\mathbf{w}^*$ and involves the expectation operator. Compared to Theorem 3.11, Theorem 3.15 presents sharper learning guarantees, and another distinctive improvement of Theorem 3.15 is that we successfully remove the dimension d. Indeed, from Lemma 4.1 of (Srebro et al., 2010), if f is nonnegative and β-smooth, we have $\|\nabla f(\mathbf{w}^*, z)\|^2 \leq 4\beta\nabla f(\mathbf{w}^*, z)$, implying that $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2] \leq 4\beta F(\mathbf{w}^*)$. Therefore, we can show the following Theorem 3.17.

Theorem 3.17. *Suppose Assumptions 2.4, 2.6, 2.10, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Assume that $F(\mathbf{w}^*) = \mathcal{O}(\frac{1}{n})$. Selecting $T \asymp n^2$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, for any $\delta \in (0, 1)$, with probability $1 - \delta$,*

(a.) if $\theta = \frac{1}{2}$, we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^2(\frac{1}{\delta})}{n^2}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta})\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big).$$

Remark 3.18. The assumption $F(\mathbf{w}^*) = \mathcal{O}(1/n)$ we used just to show that we can get improved bounds under low noise conditions. The term $F(\mathbf{w}^*)$ should be independent of n. This assumption is common and can be found in (Srebro et al., 2010; Zhang et al., 2017; Zhang & Zhou, 2019; Lei et al., 2021a; Liu et al., 2018; Lei & Ying, 2020). Theorem 3.17 presents $\mathcal{O}(1/n^2)$ order high probability generalization bounds for nonconvex SGD with heavy tails when hiding the logarithmic terms. The $\mathcal{O}(1/n^2)$ order bounds significantly improve the $\mathcal{O}\big((d + \log(\frac{1}{\delta}))/n\big)$ order bounds of Theorem 3.11. We highlight the sharper learning bounds in Section 3.3 are novel and have not been derived in the related work of nonconvex SGD (Ghadimi & Lan, 2013; Lei & Tang, 2021; Madden et al., 2021), as well as adaptive SGD (Li & Orabona, 2020; Zhou et al., 2018; Ward et al., 2019).

3.4. SGD with Clipping

The steps of SGD with clipping are shown in Algorithm 2. In this section, we consider gradient clipping to remove Assumption 2.8 for nonconvex heavy-tailed SGD.

Theorem 3.19. *Suppose Assumptions 2.6 and 2.12 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 2. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ for some constant $\eta_1 > 0$ and take $\tau = \max\{20K\log^\theta(\frac{2}{\delta}), 4K\log^\theta\sqrt{T}\}$. Then for any $\delta \in (0, 1)$, with probability $1 - \delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\min\Big\{\|\nabla F_S(\mathbf{w}_t)\|, \|\nabla F_S(\mathbf{w}_t)\|^2\Big\} = \mathcal{O}\Big(\frac{\log^\theta(T/\delta)\log T + \log^{2\theta+1}(T)\log(\frac{T}{\delta})}{\sqrt{T}}\Big).$$

Remark 3.20. The bounded stochastic gradient assumption is very commonly used in previous literature (Hardt et al., 2016; Kuzborskij & Lampert, 2018; Zhou et al., 2018; Reddi et al., 2016), but also pretty strong for deep learning (Li & Orabona, 2020). In Theorem 3.19, we successfully remove Assumption 2.8 and provide the high probability guarantee under smoothness and sub-Weibull noise conditions. We now compare Theorem 3.19 with the related work of SGD with clipping. Gorbunov et al. (2020) focus on convex learning problems, Zhang et al. (2020b) only provide an in-expectation analysis, and Cutkosky & Mehta (2021) prove bounds for momentum, as discussed in Section 1.1. To our knowledge, this is the first high probability bound for nonconvex SGD with clipping. Another improvement of Theorem 3.19 to Theorem 3.1 is that its stepsize η_t does not depend on the smoothness argument β, i.e., completely oblivious to the knowledge of smoothness. Thus, the convergence of clipped SGD is robust to the choice of the stepsize.

4. Conclusions

This paper establishes high probability learning guarantees for nonconvex SGD. In contrast to most theoretical studies, we consider the stochastic gradient noise following a

Algorithm 2 SGD with Clippling

```
Input: initial point $\mathbf{w}_1 = 0$, step sizes $\{\eta_t\}_t$, dataset $S = \{z_1, ..., z_n\}$, and $\tau > 0$.
for $t = 1, ..., T$ do
    draw $j_t$ from the uniform distribution over the set $\{j : j \in [n]\}$
    obtain $\nabla\bar{f}(\mathbf{w}_t; z_{j_t}) = \frac{\nabla f(\mathbf{w}_t; z_{j_t})}{\|\nabla f(\mathbf{w}_t; z_{j_t})\|} \min\{\tau, \|\nabla f(\mathbf{w}_t; z_{j_t})\|\}$
    update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla\bar{f}(\mathbf{w}_t; z_{j_t})$.
end for
```

novel class of heavy-tailed sub-Weibull distribution. Our analysis involves joint consideration of optimization and generalization performance. Under different assumptions, we push the learning guarantees to different orders. We also study clipped SGD to remove a very commonly used assumption. We believe our theoretical findings can provide in-depth insights into the learning guarantees of nonconvex SGD. We also think that further investigating the theoretical properties of clipped algorithms is of great significance.

Acknowledgments

We appreciate all the anonymous reviewers for their invaluable and constructive comments, especially one reviewer's suggestion to add more discussion and restructure. This work is supported in part by the National Natural Science Foundation of China (No. 62076234, No.61703396, No. 62106257), Beijing Outstanding Young Scientist Program NO.BJJWZYJH012019100020098, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" initiative, Renmin University of China, China Unicom Innovation Ecological Cooperation Plan, Public Computing Cloud of Renmin University of China, Beijing Natural Science Foundation (No. 4222029).

References

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

Bakhshizadeh, M., Maleki, A., and de la Pena, V. H. Sharp concentration results for heavy-tailed distributions. *arXiv preprint arXiv:2003.13819*, 2020.

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Şimşekli, U. Heavy tails in sgd and compressibility of overparametrized neural networks. *arXiv preprint arXiv:2106.03795*, 2021.

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

Bastianello, N., Madden, L., Carli, R., and Dall'Anese, E. A stochastic operator framework for inexact static and online optimization. *arXiv preprint arXiv:2105.09884*, 2021.

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

Bousquet, O. and Bottou, L. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pp. 161–168, 2007.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (3-4):231–357, 2015.

Camuto, A., Wang, X., Zhu, L., Holmes, C., Gürbüzbalaban, M., and Şimşekli, U. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *International Conference on Machine Learning*, 2021.

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

Cutkosky, A. and Mehta, H. Momentum improves normalized sgd. In *International Conference on Machine Learning*, pp. 2260–2268, 2020.

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

Davis, D. and Drusvyatskiy, D. High probability guarantees for stochastic convex optimization. In *Conference on Learning Theory*, pp. 1411–1427, 2020.

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22: 49–1, 2021.

Fan, X. and Giraudo, D. Large deviation inequalities for martingales in banach spaces. *arXiv preprint arXiv:1909.05584*, 2019.

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

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pp. 8745–8756, 2018.

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

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

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

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

Gurbuzbalaban, M. and Hu, Y. Fractional moment-preserving initialization schemes for training deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 2233–2241, 2021.

Gurbuzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975, 2021.

Hardt, M. and Ma, T. Identity matters in deep learning. In *International Conference on Learning Representations*, 2016.

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

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Harvey, N. J., Liaw, C., Plan, Y., and Randhawa, S. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613, 2019.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15 (1):2489–2512, 2014.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hodgkinson, L. and Mahoney, M. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274, 2021.

Jain, P. and Kar, P. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

Jain, P., Nagaraj, D., and Netrapalli, P. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pp. 1752–1755, 2019.

Kakade, S. M. and Tewari, A. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pp. 801–808, 2009.

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

Kuchibhotla, A. K. and Chakrabortty, A. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824, 2018.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

Lei, Y. and Tang, K. Stochastic composite mirror descent: Optimal bounds with high probabilities. In *Advances in Neural Information Processing Systems*, pp. 1519–1529, 2018.

Lei, Y. and Tang, K. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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

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

Lei, Y., Hu, T., and Tang, K. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research*, 22: 25–1, 2021a.

Lei, Y., Liu, M., and Ying, Y. Generalization guarantee of sgd for pairwise learning. In *Advances in Neural Information Processing Systems*, 2021b.

Li, C. J. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *arXiv preprint arXiv:1809.02495V3*, 2021.

Li, J., Liu, Y., Yin, R., Zhang, H., Ding, L., and Wang, W. Multi-class learning: From theory to algorithm. *Advances in Neural Information Processing Systems*, 2018.

Li, S. and Liu, Y. Improved learning rates for stochastic optimization: Two theoretical viewpoints. *arXiv preprint arXiv:2107.08686*, 2021a.

Li, S. and Liu, Y. Sharper generalization bounds for clustering. In *International Conference on Machine Learning*, pp. 6392–6402, 2021b.

Li, S. and Liu, Y. Towards sharper generalization bounds for structured prediction. In *Advances in Neural Information Processing Systems*, 2021c.

Li, S. and Liu, Y. High probability generalization bounds with fast rates for minimax problems. In *International Conference on Learning Representations*, 2022.

Li, X. and Orabona, F. A high probability analysis of adaptive sgd with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML*, 2020.

Li, X., Ling, S., Strohmer, T., and Wei, K. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

Li, X., Liu, M., and Orabona, F. On the last iterate convergence of momentum methods. *arXiv preprint arXiv:2102.07002*, 2021.

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

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020.

Liu, H., Wu, W., and So, A. M.-C. Quadratic optimization with orthogonality constraints: Explicit lojasiewicz exponent and linear convergence of line-search methods. In *International Conference on Machine Learning*, pp. 1158–1167, 2016.

Liu, M., Zhang, X., Zhang, L., Jin, R., and Yang, T. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. In *Advances in Neural Information Processing Systems*, pp. 4678–4689, 2018.

Liu, Y. Refined learning bounds for kernel and approximate k-means. In *Advances in Neural Information Processing Systems*, 2021.

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

Madden, L., Dall'Anese, E., and Becker, S. High-probability convergence bounds for non-convex stochastic gradient descent. *arXiv preprint arXiv:2006.05610v4*, 2021.

Mei, S., Bai, Y., Montanari, A., et al. The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774, 2018.

Nesterov, I. E. *Introductory Lectures on Convex Optimization: A Basic Course*. 2014.

Neu, G. Information-theoretic generalization bounds for stochastic gradient descent. *arXiv preprint arXiv:2102.00931*, 2021.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Nguyen, T. H., Simsekli, U., Gurbuzbalaban, M., and RICHARD, G. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, pp. 273–283, 2019.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pp. 1571–1578, 2012.

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

Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837, 2019.

Srebro, N., Sridharan, K., and Tewari, A. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18 (5):1131–1198, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pp. 6458–6467, 2019.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, H., Gürbüzbalaban, M., Zhu, L., Şimşekli, U., and Erdogdu, M. A. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, 2021.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686, 2019.

Wong, K. C., Li, Z., and Tewari, A. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Xu, Y. and Zeevi, A. Towards problem-dependent optimal learning rates. In *Advances in Neural Information Processing Systems*, 2020.

Xu, Y. and Zeevi, A. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

Yin, R., Liu, Y., Wang, W., and Meng, D. Distributed nyström kernel learning with communications. In *International Conference on Machine Learning*, pp. 12019–12028, 2021.

Zhang, B., Jin, J., Fang, C., and Wang, L. Improved analysis of clipping algorithms for non-convex optimization. *arXiv preprint arXiv:2010.02519*, 2020a.

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

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

Zhang, L. and Zhou, Z.-H. Stochastic approximation of smooth and strongly convex functions: Beyond the $\mathcal{O}(1/t)$ convergence rate. In *Conference on Learning Theory*, pp. 3160–3179, 2019.

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

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

Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Zhou, Y., Liang, Y., and Zhang, H. Generalization error bounds with probabilistic guarantee for sgd in nonconvex optimization. *arXiv preprint arXiv:1802.06903*, 2018.

A. Preliminaries of Sub-Weibull Distribution

A.1. Properties

Define the L_p norm of random variable X as $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$, for any $p \geq 1$. A sub-Weibull random variable X can equivalently be characterized using the following properties.

Proposition A.1 (Equivalent definition)**.** *(Vladimirova et al., 2020; Bastianello et al., 2021) Given $\theta \geq 0$, the following properties are equivalent:*

- $\exists K_1 > 0$ *such that* $P(|X| \geq t) \leq 2\exp\left(-(t/K_1)^{1/\theta}\right)$, $\forall t > 0$*;*
- $\exists K_2 > 0$ *such that* $\|X\|_k \leq K_2 k^\theta$, $\forall k \geq 1$*;*
- $\exists K_3 > 0$ *such that* $\mathbb{E}[\exp\left((\lambda|X|)^{1/\theta}\right)] \leq \exp\left((\lambda K_3)^{1/\theta}\right)$, $\forall \lambda \in (0, 1/K_3)$*;*
- $\exists K_4 > 0$ *such that* $\mathbb{E}[\exp((|X|/K_4)^{1/\theta})] \leq 2$*.*

The parameters K_1, K_2, K_3, K_4 differ each by a constant that only depends on θ.

By the tail probabilities in Proposition A.1, we can derive the following high probability bounds for sub-Weibull random variables.

Lemma A.2. *Let $X \sim subW(\theta, K)$ according to Definition 2.1, then for any $\delta \in (0,1)$, with probability $1-\delta$ we have*

$$|X| \leq K \log^\theta(2/\delta).$$

Proof. According to Theorem 2.1 in (Vladimirova et al., 2020), if the forth property in Proposition A.1 hold, then $K_1 = K_4$. By setting the RHS of $P(|X| \geq t) \leq 2\exp\left(-\left(\frac{t}{K}\right)^{1/\theta}\right)$ equal to δ, and solving for t we get $t = K\log^\theta(2/\delta)$. Thus, with probability at least $1-\delta$ there holds $|X| \leq t$. □

A.2. Concentration Inequalities

Lemma A.3. *(Vladimirova et al., 2020; Wong et al., 2020; Madden et al., 2021) Suppose $X_1, \cdots, X_n$ are sub-Weibull(θ) with respective parameters $K_1, \ldots, K_n$. Then, for all $t \geq 0$,*

$$P\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\left(\frac{t}{g(\theta)\sum_{i=1}^n K_i}\right)^{1/\theta}\right),$$

where $g(\theta) = (4e)^\theta$ for $\theta \leq 1$ and $g(\theta) = 2(2e\theta)^\theta$ for $\theta \geq 1$.

The following two lemmas provide results for the concentration of the sum of a sub-Weibull martingale difference sequence.

Lemma A.4. *(Li, 2021; Fan & Giraudo, 2019) Let $\theta \in (0, \infty)$ be given. Assume that $(\mathbf{X}_i, i = 1, \cdots, N)$ is a sequence of $\mathbb{R}^d$-valued martingale differences with respect to filtration $\mathcal{F}_i$, i.e. $\mathbb{E}[\mathbf{X}_i|\mathcal{F}_{i-1}] = 0$, and it satisfies the following weak exponential-type tail condition: for some $\theta > 0$ and all $i = 1, ..., N$ we have for some scalar $0 < K_i$, $\mathbb{E}\left[\exp\left(\left\|\frac{\mathbf{X}_i}{K_i}\right\|^{\frac{1}{\theta}}\right)\right] \leq 2$. Assume that $K_i < \infty$ for each $i = 1, ..., N$. Then for an arbitrary $N \geq 1$ and $t > 0$,*

$$P\left(\max_{n \leq N}\left\|\sum_{i=1}^n \mathbf{X}_i\right\| \geq t\right) \leq 4\left[3 + (3\theta)^{2\theta}\frac{128\sum_{i=1}^N K_i^2}{t^2}\right]\exp\left\{-\left(\frac{t^2}{64\sum_{i=1}^N K_i^2}\right)^{\frac{1}{2\theta+1}}\right\}.$$

Lemma A.5. *(Madden et al., 2021)[Sub-Weibull Freedman inequality] Let $(\Omega, \mathcal{F}, (\mathcal{F}_i), P)$ be a filtered probability space. Let (ξ_i) and (K_i) be adapted to $(\mathcal{F}_i)$. Let $n \in \mathbb{N}$, then for all $i \in [n]$, assume $K_{i-1} \geq 0$, $\mathbb{E}[\xi_i|\mathcal{F}_{i-1}] = 0$, and*

$$\mathbb{E}\Big[\exp\left((|\xi_i|/K_{i-1})^{1/\theta}\right)|\mathcal{F}_{i-1}\Big] \leq 2$$

where $\theta \geq 1/2$. If $\theta > 1/2$, assume there exists (m_i) such that $K_{i-1} \leq m_i$.

If $\theta = 1/2$, let $a = 2$. Then for all $x, \beta' \geq 0$, and $\alpha > 0$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta'\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta'), \tag{3}$$

and for all $x, \beta', \lambda \geq 0$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta'\right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta'\right).$$

If $\theta \in \left(\frac{1}{2}, 1\right]$, let $a = (4\theta)^{2\theta}e^2$ and $b = (4\theta)^\theta e$. For all $x, \beta' \geq 0$, and $\alpha \geq b\max_{i\in[n]} m_i$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta'\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta'), \tag{4}$$

and for all $x, \beta' \geq 0$, and $\lambda \in \left[0, \frac{1}{b\max_{i\in[n]} m_i}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta'\right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta'\right).$$

If $\theta > 1$, let $\delta \in (0,1)$, $a = (2^{2\theta+1}+2)\Gamma(2\theta+1) + \frac{2^{3\theta}\Gamma(3\theta+1)}{3}$ and $b = 2\log^{\theta-1}(n/\delta)$, where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. For all $x, \beta' \geq 0$, and $\alpha \geq b\max_{i\in[n]} m_i$, and $\lambda \in \left[0, \frac{1}{2\alpha}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \alpha\sum_{i=1}^{k}\xi_i + \beta'\right\}\right) \leq \exp(-\lambda x + 2\lambda^2\beta') + 2\delta, \tag{5}$$

and for all $x, \beta' \geq 0$, and $\lambda \in \left[0, \frac{1}{b\max_{i\in[n]} m_i}\right]$,

$$P\left(\bigcup_{k\in[n]}\left\{\sum_{i=1}^{k}\xi_i \geq x \text{ and } \sum_{i=1}^{k} aK_{i-1}^2 \leq \beta'\right\}\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\beta'\right) + 2\delta.$$

B. Some Basic Lemmas

Lemma B.1. *(Lei & Tang, 2021) Let e be the base of the natural logarithm. There holds the following elementary inequalities.*

(a) If $\theta \in (0,1)$, then $\sum_{k=1}^{t} k^{-\theta} \leq t^{1-\theta}/(1-\theta)$;

(b) If $\theta = 1$, then $\sum_{k=1}^{t} k^{-\theta} \leq \log(et)$;

(c) If $\theta > 1$, then $\sum_{k=1}^{t} k^{-\theta} \leq \frac{\theta}{\theta-1}$;

(d) $\sum_{k=1}^{t} \frac{1}{k+k_0} \leq \log(t+1)$, where $k_0 \geq 1$.

Lemma B.2 (Properties of Smoothness)**.** *(Nesterov, 2014; Boyd et al., 2004) If the function f satisfies Assumption 2.6, then we have for any z*

$$f(\mathbf{w}; z) - f(\mathbf{w}'; z) \leq \langle \mathbf{w} - \mathbf{w}', \nabla f(\mathbf{w}'; z)\rangle + \frac{1}{2}\beta\|\mathbf{w} - \mathbf{w}'\|^2$$
$$\text{and} \quad \frac{1}{2\beta}\|\nabla f(\mathbf{w}; z)\|^2 \leq f(\mathbf{w}; z) - \inf_{\mathbf{w}} f(\mathbf{w}; z).$$

Lemma B.3. *(Lei & Tang, 2021) Let $\delta \in (0,1)$, $R > 0$, and $S = \{z_1, ..., z_n\}$ be a set of i.i.d. samples. Suppose Assumption 2.6 holds. Then with probability at least $1-\delta$ we have*

$$\sup_{\mathbf{w}\in B_R} \|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| \leq \frac{(\beta R + b')}{\sqrt{n}} \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right),$$

where e is the base of the natural logarithm.

Lemma B.4. *(Li & Liu, 2021a) Suppose Assumptions 2.6 and 2.10 hold. Assume the population risk F satisfies $F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w})\|$. For all $\mathbf{w} \in \mathcal{W} := B(\mathbf{w}^*, R)$ and any $\delta > 0$, with probability at least $1-\delta$, when $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$,*

$$\|\nabla F(\mathbf{w}) - \nabla F_S(\mathbf{w})\| \leq \|\nabla F_S(\mathbf{w})\| + \frac{\mu}{n} + \frac{2G_* \log(4/\delta)}{n} + 2\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(4/\delta)}{n}},$$

and

$$\|\nabla F(\mathbf{w})\| \leq 2\,\|\nabla F_S(\mathbf{w})\| + \frac{\mu}{n} + \frac{2G_* \log(4/\delta)}{n} + 2\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(4/\delta)}{n}},$$

where c is an absolute constant.

Lemma B.5. *(Zhang, 2005) Let $z_1, ..., z_n$ be a sequence of randoms variables such that z_k may depend the previous variables $z_1, ..., z_{k-1}$ for all $k = 1, ..., n$. Consider a sequence of functionals $\xi_k(z_1, ..., z_k)$, $k = 1, ..., n$. Let $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}[(\xi_k - \mathbb{E}_{z_k}[\xi_k])^2]$ be the conditional variance. Assume $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b$ for each k. Let $\rho \in (0,1]$ and $\delta \in (0,1)$. With probability at least $1-\delta$ we have*

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

Lemma B.6. *(Cutkosky & Mehta, 2020) For any vector $v \in \mathbb{R}^d$, $\langle v/\|v\|, \nabla F_S(\mathbf{w})\rangle \geq \frac{\|\nabla F_S(\mathbf{w})\|}{3} - \frac{8\|v - \nabla F_S(\mathbf{w})\|}{3}$.*

C. Proof of Main Results

For better readability, we restate the theorems in the main paper.

C.1. Proof of Section 3.1

C.1.1. Proof of Theorem 3.1

Theorem C.1. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \leq \frac{1}{2\beta}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log(1/\delta)\log T}{\sqrt{T}}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have the following inequality

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log^{\theta-1}(T/\delta)\log(1/\delta) + \log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big).$$

Proof. According to Assumption 2.6 and Lemma B.2, we have

$$\begin{aligned}
&F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \leq \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla F_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= -\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + \frac{1}{2}\beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t})\|^2 \\
&\leq -\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \Big(\eta_t - \beta\eta_t^2\Big)\|\nabla F_S(\mathbf{w}_t)\|^2 + \beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2 \\
&\leq -\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \frac{1}{2}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + \beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2, \qquad (6)
\end{aligned}$$

where the second inequality follows from that $(a+b)^2 \leq 2(a^2+b^2), \forall a, b \in \mathbb{R}$ and the last inequality follows from that $\beta\eta_t^2 - \eta_t \leq -\frac{\eta_t}{2}$ due to the assumption $\eta_t \leq \frac{1}{2\beta}$.

Then, by a summation form $t = 1$ to T, we get

$$\begin{aligned}
&F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}_1) \\
\leq& -\sum_{t=1}^{T}\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + \sum_{t=1}^{T}\beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2,
\end{aligned}$$

which means that

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\
\leq& 2(F_S(\mathbf{w}_1) - F_S(\mathbf{w}(S))) - \sum_{t=1}^{T}2\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle + \sum_{t=1}^{T}2\beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2. \qquad (7)
\end{aligned}$$

It is clear that $2(F_S(\mathbf{w}_1) - F_S(\mathbf{w}(S)))$ can be seen as a constant.

Now, let us consider the second term $-\sum_{t=1}^{T}2\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$. Since $\mathbb{E}_{j_t}[-\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle] = 0$, thus the sequence $(-\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle, t \in \mathbb{N})$ is a martingale difference sequence. We use the sub-Weibull Freedman inequality in Lemma A.5 to bound this term.

Specifically, we set $\xi_t = -\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$, $K_{t-1} = \eta_t\|\nabla F_S(\mathbf{w}_t)\|K$, $\beta' = 0$, $\lambda = \frac{1}{2\alpha}$, and $x = 2\alpha\log(1/\delta)$.

If $\theta = \frac{1}{2}$, for all $\alpha > 0$, we have the following inequality with probability at least $1-\delta$

$$-\sum_{t=1}^{T}2\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 4\alpha\log(1/\delta) + \frac{2aK^2}{\alpha}\sum_{t=1}^{T}\eta_t^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

If $\theta \in (\frac{1}{2}, 1]$, according to Assumption 2.8, we set $m_t = KG$. Then for all $\alpha \geq bKG$, we have the following inequality with probability at least $1-\delta$

$$-\sum_{t=1}^{T}2\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 4\alpha\log(1/\delta) + \frac{2aK^2}{\alpha}\sum_{t=1}^{T}\eta_t^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

If $\theta > 1$, according to Assumption 2.8, we set $m_t = KG$. We also set $\delta = \delta$. Then, for all $\alpha \geq bKG$, we have the following inequality with probability at least $1-3\delta$

$$-\sum_{t=1}^{T}2\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 4\alpha\log(1/\delta) + \frac{2aK^2}{\alpha}\sum_{t=1}^{T}\eta_t^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

Then, we consider the last term $\sum_{t=1}^T 2\beta\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|^2$. Since $\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|$ is a sub-Weibull random variable, we get

$$\mathbb{E}\left[\exp\left(\frac{\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|^2}{\eta_t^2 K^2}\right)^{\frac{1}{2\theta}}\right]\leq 2,$$

which means that $\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|^2 \sim subW(2\theta,\eta_t^2K^2)$. According to Lemma A.3, we get the following inequality with probability at least $1-\delta$

$$\sum_{t=1}^T 2\beta\eta_t^2\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|^2 \leq 2\beta K^2 g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2,$$

where $g(\theta)=(4e)^\theta$ for $\theta\leq 1$ and $g(\theta)=2(2e\theta)^\theta$ for $\theta\geq 1$.

Taking the above upper bounds into (7), we obtain

$$\begin{aligned}\sum_{t=1}^T\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 &\leq 2(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))\\ &+4\alpha\log(1/\delta)+\frac{2aK^2}{\alpha}\sum_{t=1}^T\eta_t^2\|\nabla F_S(\mathbf{w}_t)\|^2+2\beta K^2g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2,\end{aligned}$$

implying that

$$\sum_{t=1}^T\eta_t\Big(1-\frac{2aK^2\eta_t}{\alpha}\Big)\|\nabla F_S(\mathbf{w}_t)\|^2\leq 2(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+4\alpha\log(1/\delta)+2\beta K^2g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2.$$

To continue the proof, we set $\alpha\geq 4aK^2\eta_1$. Then we obtain $1-\frac{2aK^2\eta_t}{\alpha}\geq\frac{1}{2}$. Thus, we derive that

$$\sum_{t=1}^T\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\leq 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+8\alpha\log(1/\delta)+4\beta K^2g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2.$$

We now use the union bound.

Hence, if $\theta=\frac{1}{2}$, taking $\alpha=4aK^2\eta_1=8K^2\eta_1$, with probability at least $1-2\delta$, we have

$$\sum_{t=1}^T\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\leq 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+64K^2\eta_1\log(1/\delta)+4\beta K^2g(1)\log(2/\delta)\sum_{t=1}^T\eta_t^2. \tag{8}$$

If $\theta\in(\frac{1}{2},1]$, taking $\alpha=\max\{bKG,4aK^2\eta_1\}=\max\{(4\theta)^\theta eKG,4(4\theta)^{2\theta}e^2K^2\eta_1\}$, with probability at least $1-2\delta$, we have

$$\begin{aligned}\sum_{t=1}^T\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2&\leq 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))\\ &+8\max\Big\{(4\theta)^\theta eKG,4(4\theta)^{2\theta}e^2K^2\eta_1\Big\}\log(1/\delta)+4\beta K^2g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2.\end{aligned}\tag{9}$$

If $\theta>1$, taking $\alpha=\max\Big\{bKG,4aK^2\eta_1\Big\}=\max\Big\{2\log^{\theta-1}(T/\delta)KG,4\big((2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}\big)K^2\eta_1\Big\}$, with probability at least $1-4\delta$, we have

$$\begin{aligned}\sum_{t=1}^T\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2&\leq 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+4\beta K^2g(2\theta)\log^{2\theta}(2/\delta)\sum_{t=1}^T\eta_t^2\\ &+8\max\Big\{2\log^{\theta-1}(T/\delta)KG,4\big((2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}\big)K^2\eta_1\Big\}\log(1/\delta).\end{aligned}\tag{10}$$

For brevity, we just focus on parameters δ and T. Moreover, note that the dependence on confidence parameter $1/\delta$ in (8), (9), and (10) is logarithmic. One can replace δ to $\delta/2$ or $\delta/4$. Therefore, (8), (9), and (10) mean that with probability at least $1-\delta$, there holds

$$\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 = \begin{cases}\mathcal{O}\left(\log(1/\delta)\sum_{t=1}^{T}\eta_t^2\right) & \text{if } \theta=\frac{1}{2},\\ \mathcal{O}\left(\log^{2\theta}(1/\delta)\sum_{t=1}^{T}\eta_t^2\right) & \text{if } \theta\in(\frac{1}{2},1],\\ \mathcal{O}\left(\log^{\theta-1}(T/\delta)\log(1/\delta)+\log^{2\theta}(1/\delta)\sum_{t=1}^{T}\eta_t^2\right) & \text{if } \theta>1,\end{cases} \tag{11}$$

$$= \begin{cases}\mathcal{O}\left(\log(1/\delta)\log T\right) & \text{if } \theta=\frac{1}{2},\\ \mathcal{O}\left(\log^{2\theta}(1/\delta)\log T\right) & \text{if } \theta\in(\frac{1}{2},1],\\ \mathcal{O}\left(\log^{\theta-1}(T/\delta)\log(1/\delta)+\log^{2\theta}(1/\delta)\log T\right) & \text{if } \theta>1,\end{cases} \tag{12}$$

where the second equality holds by using Lemma B.1. The proof is complete. □

C.1.2. Proof of Theorem 3.3

Theorem C.2. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t=\eta_1 t^{-\frac{1}{2}}$ with $\eta_1\le\frac{1}{2\beta}$. For any $\delta\in(0,1)$ and uniformly for all $t=1,...T$, with probability $1-\delta$,*

(a.) if $\theta=\frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1})-\nabla F_S(\mathbf{w}_{t+1})\|^2=\mathcal{O}\Big(\frac{T^{\frac{1}{2}}\log^2(\frac{1}{\delta})\log T}{n}\times\big(d+\log(\frac{1}{\delta})\big)\Big);$$

(b.) if $\theta\in(\frac{1}{2},1]$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1})-\nabla F_S(\mathbf{w}_{t+1})\|^2=\mathcal{O}\Big(\frac{T^{\frac{1}{2}}\log^{2\theta+1}(\frac{1}{\delta})\log T}{n}\times\big(d+\log(\frac{1}{\delta})\big)\Big);$$

(c.) if $\theta>1$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1})-\nabla F_S(\mathbf{w}_{t+1})\|^2=\mathcal{O}\Big(\frac{T^{\frac{1}{2}}\big(\log^{\theta-1}(T/\delta)\log(1/\delta)+\log^{(2\theta+1)}(\frac{1}{\delta})\log T\big)}{n}\times\big(d+\log(\frac{1}{\delta})\big)\Big).$$

Proof. For better presentation, we introduce a notation $Y(T,\delta,\theta):=\log^{\theta-1}(T/\delta)\log(1/\delta)\mathbb{I}_{\theta>1}$, where $\mathbb{I}_{\theta>1}$ is an indicate function that is valued 1 when $\theta>1$. Since $\mathbf{w}_{t+1}=\mathbf{w}_t-\eta_t(\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)+\nabla F_S(\mathbf{w}_t))$, by a summation and using $\mathbf{w}_1=0$, we get

$$\mathbf{w}_{t+1}=\sum_{i=1}^{t}-\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))-\sum_{i=1}^{t}\eta_i\nabla F_S(\mathbf{w}_i).$$

This gives that

$$\|\mathbf{w}_{t+1}\|\le\Big\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\Big\|+\Big\|\sum_{i=1}^{t}\eta_i\nabla F_S(\mathbf{w}_i)\Big\|. \tag{13}$$

Let's consider the first term $\Big\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\Big\|$. It is clear that the sequence $(\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i)),i\in\mathbb{N})$ is a martingale difference sequence. Since $\|\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i)\|\sim subW(\theta,K)$, we have $\|\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\|\sim subW(\eta_i\theta,K)$ according to Proposition 2.15 in (Bastianello et al., 2021). Then, we can apply Lemma A.4 to derive the following inequality

$$
\begin{aligned}
&P\left(\max_{1\le t\le T}\left\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\right\|\ge x\right)\\
&\le 4\left[3+(3\theta)^{2\theta}\,\frac{128K^2\sum_{i=1}^{T}\eta_i^2}{x^2}\right]\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=1}^{T}\eta_i^2}\right)^{\frac{1}{2\theta+1}}\right\}.
\end{aligned}
$$

Setting the dominated exponential term $4\exp\left\{-\left(\frac{x^2}{64K^2\sum_{i=1}^{T}\eta_i^2}\right)^{\frac{1}{2\theta+1}}\right\}$ equal to δ, we get $x = 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}$. Thus, with probability at least $1-3\delta-\frac{8(3\theta)^{2\theta}}{\log^{2\theta+1}\frac{4}{\delta}}\delta$, we have

$$
\max_{1\le t\le T}\left\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\right\|\le 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K\Big(\sum_{i=1}^{T}\eta_i^2\Big)^{\frac{1}{2}}. \tag{14}
$$

Since $\theta\ge 1/2$ and $\delta\in(0,1)$, we have $\log^{2\theta+1}\frac{4}{\delta}>1$. Thus, (14) means that with probability at least $1-3\delta-8(3\theta)^{2\theta}\delta$, we have

$$
\max_{1\le t\le T}\left\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\right\|\le 8\log^{(\theta+\frac{1}{2})}(\frac{4}{\delta})K\Big(\sum_{i=1}^{T}\eta_i^2\Big)^{\frac{1}{2}}.
$$

Now, with probability at least $1-\delta$, we get

$$
\max_{1\le t\le T}\left\|\sum_{i=1}^{t}\eta_i(\nabla f(\mathbf{w}_i;z_{j_i})-\nabla F_S(\mathbf{w}_i))\right\|\le 8\log^{(\theta+\frac{1}{2})}\Big(\frac{4(3+8(3\theta)^{2\theta})}{\delta}\Big)K\Big(\sum_{i=1}^{T}\eta_i^2\Big)^{\frac{1}{2}}. \tag{15}
$$

For the second term $\left\|\sum_{i=1}^{t}\eta_i\nabla F_S(\mathbf{w}_i)\right\|$, we have the following inequality with probability at least $1-\delta$ uniformly for all $t=1,...,T$

$$
\begin{aligned}
\Big\|\sum_{i=1}^{t}\eta_i\nabla F_S(\mathbf{w}_i)\Big\|^2&\le\Big(\sum_{i=1}^{t}\eta_i\|\nabla F_S(\mathbf{w}_i)\|\Big)^2\le\Big(\sum_{i=1}^{t}\eta_i\Big)\Big(\sum_{i=1}^{t}\eta_i\|\nabla F_S(\mathbf{w}_i)\|^2\Big)\\
&\le\Big(\sum_{i=1}^{t}\eta_i\Big)\mathcal{O}\left(Y(T,\delta,\theta)+\log^{2\theta}(1/\delta)\sum_{i=1}^{t}\eta_i^2\right), \qquad (16)
\end{aligned}
$$

where the second inequality follows from the Schwartz's inequality and the last inequality follows from (11).

For brevity, in the following proofs, we also just focus on parameters δ and T. Readers can recover the constants by following the analysis.

Plugging (15) and (16) into (13), we have the following inequality with probability at least $1-2\delta$ uniformly for all $t=1,...,T$

$$
\begin{aligned}
\|\mathbf{w}_{t+1}\|&=\mathcal{O}\left(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}\right)+\left(\Big(\sum_{i=1}^{t}\eta_i\Big)\mathcal{O}\Big(Y(T,\delta,\theta)+\log^{2\theta}(1/\delta)\sum_{i=1}^{t}\eta_i^2\Big)\right)^{\frac{1}{2}} \qquad (17)\\
&=\mathcal{O}\left(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T\right)+\Big((t^{\frac{1}{2}})\mathcal{O}\Big(Y(T,\delta,\theta)+\log^{2\theta}(1/\delta)\log t\Big)\Big)^{\frac{1}{2}}\\
&\le\mathcal{O}\left(t^{\frac{1}{4}}\big(Y^{\frac{1}{2}}(T,\delta,\theta)+\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T\big)\right), \qquad (18)
\end{aligned}
$$

where the second equation holds by using Lemma B.1.

Further plugging (18) into Lemma B.3, by the union bound, with probability at least $1-3\delta$, we have the following inequality uniformly for all $t = 1, ...T$

$$\begin{aligned}
&\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\| \leq \frac{(\beta R_{t+1} + b')}{\sqrt{n}} \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right) \\
=&\frac{(\beta\|\mathbf{w}_{t+1}\| + b')}{\sqrt{n}} \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right) \\
\leq&\frac{\mathcal{O}\left(t^{\frac{1}{4}}\left(Y^{\frac{1}{2}}(T,\delta,\theta) + \log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T\right)\right)\beta + b'}{\sqrt{n}} \times \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right), \qquad (19)
\end{aligned}$$

where the first inequality holds due to the fact that the bound of B_R at iterate number $t+1$ is R_{t+1}, that is $\|\mathbf{w}_{t+1}\|$, and where the last inequality follows from (18).

(19) also means that we have the following inequality uniformly for all $t = 1, ...T$ with probability at least $1-\delta$

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\left(\frac{T^{\frac{1}{2}}\left(Y(T,\delta,\theta) + \log^{(2\theta+1)}(\frac{1}{\delta})\log T\right)}{n} \times \left(d + \log(\frac{1}{\delta})\right)\right). \qquad (20)$$

By a transformation of (20), we prove the stated bounds. □

C.1.3. Proof of Theorem 3.5

Theorem C.3. *Suppose Assumptions 2.4 and 2.6 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \leq \frac{1}{2\beta}$. Selecting $T \asymp \frac{n}{d}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}}\log(\frac{n}{d})\log^3(\frac{1}{\delta})\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have the following inequality

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}}\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta})\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}}\big(\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta}) + \log^{\theta-1}(\frac{n}{d\delta})\log^2(\frac{1}{\delta})\big)\Big).$$

Proof. It is clear that

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|^2 &\leq 2\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 + 2\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\
&\leq 2\sum_{t=1}^{T}\eta_t \max_{1\leq t\leq T}\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2 + 2\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\
&= 2\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_T) - \nabla F_S(\mathbf{w}_T)\|^2 + 2\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2, \qquad (21)
\end{aligned}$$

where the last equation follows from the fact that $\|\nabla F(\mathbf{w}_t) - \nabla F_S(\mathbf{w}_t)\|^2$ increases as the iterate number t increases (see (19) for details). Then, we have the following inequality with probability at least $1-2\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 \le \frac{1}{\eta_1\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F(\mathbf{w}_t)\|^2 \le \frac{2}{\sqrt{T}}\Big(\frac{1}{2}T^{1/2}\|\nabla F(\mathbf{w}_T)-\nabla F_S(\mathbf{w}_T)\|^2+\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\Big)$$
$$=\mathcal{O}\left(\frac{T^{1/2}\Big(Y(T,\delta,\theta)+\log^{(2\theta+1)}(\frac{1}{\delta})\log T\Big)}{n}\times\Big(d+\log(\frac{1}{\delta})\Big)\right)+\mathcal{O}\left(Y(T,\delta,\theta)+\log^{2\theta}(1/\delta)\log T\right),$$

where the second inequality follows from that $\sum_{t=1}^{T}\eta_t \le \frac{1}{2}\eta_1 T^{1/2}$ by using Lemma B.1, where the last equality follows from (12) and (20), and where $Y(T,\delta,\theta) := \log^{\theta-1}(T/\delta)\log(1/\delta)\mathbb{I}_{\theta>1}$.

Taking $T \asymp \frac{n}{d}$, then we have the following inequality with probability at least $1-2\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\left(\Big(\frac{d}{n}\Big)^{\frac{1}{2}}\Big(\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta})+Y(\frac{n}{d},\delta,\theta)\log(1/\delta)\Big)\right),$$

which also means that with probability at least $1-\delta$ we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\left(\Big(\frac{d}{n}\Big)^{\frac{1}{2}}\Big(\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta})+Y(\frac{n}{d},\delta,\theta)\log(1/\delta)\Big)\right)$$
$$= \mathcal{O}\left(\Big(\frac{d}{n}\Big)^{\frac{1}{2}}\Big(\log(\frac{n}{d})\log^{(2\theta+2)}(\frac{1}{\delta})+\log^{\theta-1}(\frac{n}{d\delta})\log^{2}(1/\delta)\mathbb{I}_{\theta>1}\Big)\right). \tag{22}$$

By a transformation of (22), we prove the stated bounds. □

C.2. Proof of Section 3.2

We first prove Theorem 3.9.

C.2.1. Proof of Theorem 3.9

Theorem C.4. *Suppose Assumptions 2.4, 2.6, and 2.12 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \ge \max\{\frac{4\beta}{\mu_S}, 1\}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d+\log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log T\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log T\Big);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\Big(\frac{d+\log(\frac{1}{\delta})}{n}\times\big(\log^{(2\theta+1)}(\frac{1}{\delta})+\log^{\theta-1}(T/\delta)\log(1/\delta)\big)\log T\Big).$$

Proof. Since $t_0 \ge \frac{4\beta}{\mu_S}$ and $\eta_t = \frac{2}{\mu_S(t+t_0)}$, we have $\eta_t \le \frac{1}{2\beta}$. Thus, we can use (17) to bound $\|\mathbf{w}_{t+1}\|$ under this stepsize. According to (17), with probability at least $1-\delta$ we have

$$\|\mathbf{w}_{t+1}\| = \mathcal{O}\left(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})(\sum_{i=1}^{T}\eta_i^2)^{\frac{1}{2}}+\Big(\sum_{i=1}^{t}\eta_i\Big)^{\frac{1}{2}}\Big(Y^{\frac{1}{2}}(T,\delta,\theta)+\log^{\theta}(1/\delta)\Big(\sum_{i=1}^{t}\eta_i^2\Big)^{\frac{1}{2}}\Big)\right)$$
$$= \mathcal{O}\left(\Big(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})+Y^{\frac{1}{2}}(T,\delta,\theta)\Big)\log^{\frac{1}{2}} t\right), \tag{23}$$

where $Y(T,\delta,\theta) := \log^{\theta-1}(T/\delta)\log(1/\delta)\mathbb{I}_{\theta>1}$, and where the second equality follows from $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq 1$ and using Lemma B.1.

Plugging (23) into Lemma B.3, by the union bound, with probability at least $1-2\delta$, we have the following inequality uniformly for all $t = 1, ...T$

$$
\begin{aligned}
&\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\| \leq \frac{(\beta R_{t+1} + b')}{\sqrt{n}}\left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right) \\
=&\frac{(\beta\|\mathbf{w}_{t+1}\| + b')}{\sqrt{n}}\left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right) \\
\leq&\frac{\mathcal{O}\left(\left(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta}) + Y^{\frac{1}{2}}(T,\delta,\theta)\right)\log^{\frac{1}{2}} t\right)\beta + b'}{\sqrt{n}} \times \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(\frac{1}{\delta})}\right), \qquad (24)
\end{aligned}
$$

where the first inequality holds due to the fact that the bound of B_R at iterate number $t+1$ is R_{t+1}, that is $\|\mathbf{w}_{t+1}\|$, and where the last inequality follows from (23).

(24) also means that we have the following inequality uniformly for all $t = 1, ...T$ with probability at least $1-\delta$

$$
\|\nabla F(\mathbf{w}_{t+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\left(\frac{d + \log(\frac{1}{\delta})}{n}\left(\log^{(2\theta+1)}(\frac{1}{\delta}) + Y(T,\delta,\theta)\right)\log T\right). \qquad (25)
$$

By a transformation of (25), we prove the stated bounds. □

C.2.2. Proof of Theorem 3.7

Theorem C.5. *Suppose Assumptions 2.4, 2.6, and 2.12 hold. Let* $\mathbf{w}_t$ *be the iterate produced by Algorithm 1. Assume* $\eta_t = \frac{2}{\mu_S(t+t_0)}$ *with* $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$*. Then for any* $\delta \in (0,1)$*, with probability* $1-\delta$*,*

(a.) if $\theta = \frac{1}{2}$*, then we have the following inequality*

$$
F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log(1/\delta)}{T}\Big);
$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ *and Assumption 2.8 holds, then we have the following inequality*

$$
F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T}\Big);
$$

(c.) if $\theta > 1$ *and Assumption 2.8 holds, then we have the following inequality*

$$
F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T}\Big).
$$

Proof. From (6), we know that

$$
\begin{aligned}
F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) &\leq -\eta_t\langle\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \frac{1}{2}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + \beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2 \\
&\leq -\eta_t\langle\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle - \frac{1}{4}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\
&\quad + \eta_t\mu_S(F_S(\mathbf{w}(S)) - F_S(\mathbf{w}_t)) + \beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2,
\end{aligned}
$$

from which we have

$$
\begin{aligned}
&\frac{1}{4}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 + F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}(S)) \leq -\eta_t\langle\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \\
&+ (1 - \eta_t\mu_S)(F_S(\mathbf{w}_t) - F_S(\mathbf{w}(S))) + \beta\eta_t^2\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2.
\end{aligned}
$$

Since $\eta_t = \frac{2}{\mu_S(t+t_0)}$, multiplying both sides by $(t+t_0)(t+t_0-1)$, we get

$$\frac{(t+t_0-1)}{2\mu_S}\|\nabla F_S(\mathbf{w}_t)\|^2 + (t+t_0)(t+t_0-1)(F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}(S))) \leq \frac{4\beta}{\mu_S^2}\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2$$
$$+ (t+t_0-1)(t+t_0-2)(F_S(\mathbf{w}_t) - F_S(\mathbf{w}(S))) - (t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle.$$

Taking a summation from $t=1$ to $t=T$, we derive that

$$\sum_{t=1}^T \frac{(t+t_0-1)}{2\mu_S}\|\nabla F_S(\mathbf{w}_t)\|^2 + (T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S))) \leq \sum_{t=1}^T \frac{4\beta}{\mu_S^2}\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2$$
$$+ t_0(t_0-1)(F_S(\mathbf{w}_1) - F_S(\mathbf{w}(S))) - \sum_{t=1}^T (t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle. \quad (26)$$

Since $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \sim subW(\theta, K)$, we get $\mathbb{E}\left[\exp\left(\frac{\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2}{K^2}\right)^{\frac{1}{2\theta}}\right] \leq 2$. By Lemma A.3, we get the following inequality with probability at least $1-\delta$

$$\frac{4\beta}{\mu_S^2}\sum_{t=1}^T \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2 \leq \frac{4\beta}{\mu_S^2} T K^2 g(2\theta)\log^{2\theta}(2/\delta).$$

Furthermore, since $\mathbb{E}_{j_t}[-(t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle] = 0$, we know that the sequence $(-(t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle, t \in \mathbb{N})$ is a martingale difference sequence. Thus, we can apply Lemma A.5 to bound it.

We set $\xi_t = -(t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$ and

$$K_{t-1} = (t+t_0)(t+t_0-1)\eta_t K\|\nabla F_S(\mathbf{w}_t)\| = 2\mu_S^{-1}(t+t_0-1)K\|\nabla F_S(\mathbf{w}_t)\|.$$

We also set $\beta' = 0$, $\lambda = \frac{1}{2\alpha}$, and $x = 2\alpha\log(1/\delta)$.

Moreover, according to Assumption 2.6, we know

$$\|\nabla F_S(\mathbf{w}_t)\| \leq (\beta\|\mathbf{w}_t\| + \|\nabla F_S(\mathbf{0})\|) \leq (\beta\|\mathbf{w}_t\| + b'). \quad (27)$$

In the next, the bound in (27) will be used to give the bound of m_t.

If $\theta = \frac{1}{2}$, for all $\alpha > 0$, we have the following inequality with probability at least $1-\delta$

$$-\sum_{t=1}^T (t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 2\alpha\log(1/\delta) + \frac{4aK^2}{\mu_S^2\alpha}\sum_{t=1}^T (t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

If $\theta \in (\frac{1}{2}, 1]$, according to (27), we set $m_t = 2\mu_S^{-1}(t+t_0-1)K(\beta\|\mathbf{w}_t\| + b')$. Then for all $\alpha \geq b2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\| + b')$, we have the following inequality with probability at least $1-\delta$

$$-\sum_{t=1}^T (t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 2\alpha\log(1/\delta) + \frac{4aK^2}{\mu_S^2\alpha}\sum_{t=1}^T (t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

If $\theta > 1$, according to (27), we set $m_t = 2\mu_S^{-1}(t+t_0-1)K(\beta\|\mathbf{w}_t\| + b')$ and $\delta = \delta$. Then, for all $\alpha \geq b2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\| + b')$, we have the following inequality with probability at least $1-3\delta$

$$-\sum_{t=1}^T (t+t_0)(t+t_0-1)\eta_t\langle \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq 2\alpha\log(1/\delta) + \frac{4aK^2}{\mu_S^2\alpha}\sum_{t=1}^T (t+t_0-1)^2\|\nabla F_S(\mathbf{w}_t)\|^2.$$

Taking the above bounds into (26), we have

$$\begin{aligned}&\sum_{t=1}^{T}\left(\frac{(t+t_0-1)}{2\mu_S}-\frac{4aK^2}{\mu_S^2\alpha}(t+t_0-1)^2\right)\|\nabla F_S(\mathbf{w}_t)\|^2+(T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1})-F_S(\mathbf{w}(S)))\\ \leq&\frac{4\beta}{\mu_S^2}TK^2g(2\theta)\log^{2\theta}(2/\delta)+t_0(t_0-1)(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+2\alpha\log(1/\delta).\end{aligned}$$

When $\alpha \geq \frac{8aK^2(t+t_0-1)}{\mu_S}$, we have

$$\frac{(t+t_0-1)}{2\mu_S}-\frac{4aK^2}{\mu_S^2\alpha}(t+t_0-1)^2\geq 0.$$

In this case, we derive that

$$\begin{aligned}&(T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1})-F_S(\mathbf{w}(S)))\\ \leq&\frac{4\beta}{\mu_S^2}TK^2g(2\theta)\log^{2\theta}(2/\delta)+t_0(t_0-1)(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+2\alpha\log(1/\delta).\end{aligned}$$

We now use the union bound.

Hence, if $\theta=\frac{1}{2}$, taking $\alpha=\frac{8aK^2(T+t_0-1)}{\mu_S}=\frac{16K^2(T+t_0-1)}{\mu_S}$, with probability at least $1-2\delta$, we have

$$\begin{aligned}&(T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1})-F_S(\mathbf{w}(S)))\\ \leq&\frac{4\beta}{\mu_S^2}TK^2g(1)\log(2/\delta)+t_0(t_0-1)(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))+\frac{32K^2(T+t_0-1)}{\mu_S}\log(1/\delta).\end{aligned}$$

If $\theta \in (\frac{1}{2},1]$, we take $\alpha = \max\left\{\frac{8aK^2(T+t_0-1)}{\mu_S}, b2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\|+b')\right\} = \max\left\{\frac{8(4\theta)^{2\theta}e^2K^2(T+t_0-1)}{\mu_S}, (4\theta)^{\theta}e2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\|+b')\right\}$.

From (23), we know the bound of $\|\mathbf{w}_T\|$. Thus with probability $1-\delta$, $\alpha = \max\left\{\frac{8(4\theta)^{2\theta}e^2K^2(T+t_0-1)}{\mu_S}, (4\theta)^{\theta}e2\mu_S^{-1}(T+t_0-1)K\Big(\beta\mathcal{O}\big((\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})+Y^{\frac{1}{2}}(T,\delta,\theta))\log^{\frac{1}{2}}T\big)+b'\Big)\right\}$.

Thus, with probability at least $1-3\delta$, we have

$$\begin{aligned}&(T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1})-F_S(\mathbf{w}(S)))\leq\frac{4\beta}{\mu_S^2}TK^2g(2\theta)\log^{2\theta}(2/\delta)+t_0(t_0-1)(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S)))\\ &+\max\left\{\frac{8(4\theta)^{2\theta}e^2K^2(T+t_0-1)}{\mu_S},(4\theta)^{\theta}e2\mu_S^{-1}(T+t_0-1)K\Big(\beta\mathcal{O}\big((\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})+Y^{\frac{1}{2}}(T,\delta,\theta))\log^{\frac{1}{2}}T\big)+b'\Big)\right\}\\ &\times 2\log(\frac{1}{\delta}).\end{aligned}$$

If $\theta > 1$, we take $\alpha = \max\left\{\frac{8aK^2(T+t_0-1)}{\mu_S}, b2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\|+b')\right\} = \max\left\{\frac{8\big((2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}\big)K^2(T+t_0-1)}{\mu_S}, 2\log^{\theta-1}(T/\delta)2\mu_S^{-1}(T+t_0-1)K(\beta\|\mathbf{w}_T\|+b')\right\}$.

From (23), we know the bound of $\|\mathbf{w}_T\|$. Thus with probability $1-\delta$,

$$\begin{aligned}\alpha=\max\Bigg\{&\frac{8\big((2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}\big)K^2(T+t_0-1)}{\mu_S},\\ &2\log^{\theta-1}(T/\delta)2\mu_S^{-1}(T+t_0-1)K\Big(\beta\mathcal{O}\big((\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})+Y^{\frac{1}{2}}(T,\delta,\theta)\big)\log^{\frac{1}{2}}T)+b'\Big)\Bigg\}.\end{aligned}$$

Therefore, with probability at least $1-5\delta$, we have

$$
\begin{aligned}
&(T+t_0)(T+t_0-1)(F_S(\mathbf{w}_{T+1})-F_S(\mathbf{w}(S))) \leq \frac{4\beta}{\mu_S^2} T K^2 g(2\theta)\log^{2\theta}(2/\delta) + t_0(t_0-1)(F_S(\mathbf{w}_1)-F_S(\mathbf{w}(S))) \\
&+ 2\max\Bigg\{\frac{8\big((2^{2\theta+1}+2)\Gamma(2\theta+1)+\frac{2^{3\theta}\Gamma(3\theta+1)}{3}\big)K^2(T+t_0-1)}{\mu_S} \\
&, 2\log(T/\delta)^{\theta-1} 2\mu_S^{-1}(T+t_0-1)K\Big(\beta\mathcal{O}\big((\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta}) + Y(T,\delta,\theta)^{\frac{1}{2}}\big)\log^{\frac{1}{2}} T) + b'\Big)\Bigg\}\log(1/\delta).
\end{aligned}
$$

We just focus on parameters δ and T. Finally, the above bounds mean that with probability at least $1-\delta$, there holds

$$
F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \begin{cases}
\mathcal{O}\left(\frac{\log(1/\delta)}{T}\right) & \text{if } \theta = \frac{1}{2}, \\
\mathcal{O}\left(\frac{\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T}\log(\frac{1}{\delta})\right) & \text{if } \theta \in (\frac{1}{2}, 1], \\
\mathcal{O}\left(\frac{\left(\log^{(\theta+\frac{1}{2})}(\frac{1}{\delta}) + Y^{\frac{1}{2}}(T,\delta,\theta)\right)\log^{\frac{1}{2}} T}{T}\log^{\theta-1}(T/\delta)\log(\frac{1}{\delta})\right) & \text{if } \theta > 1,
\end{cases}
$$

which means that with probability at least $1-\delta$ we have

$$
F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \begin{cases}
\mathcal{O}\left(\frac{\log(1/\delta)}{T}\right) & \text{if } \theta = \frac{1}{2}, \\
\mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T}\right) & \text{if } \theta \in (\frac{1}{2}, 1], \\
\mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T}\right) & \text{if } \theta > 1,
\end{cases} \tag{28}
$$

The proof is complete. □

C.2.3. PROOF OF THEOREM 3.11

Theorem C.6. *Suppose Assumptions 2.4, 2.6, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n$. Then for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$
F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d+\log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log n\Big);
$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have the following inequality

$$
F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log n\Big);
$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$
F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta})\log n\right).
$$

Proof. According to (2), we know

$$
F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w}_{T+1})\|^2 \leq \mu^{-1}(\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 + \|\nabla F_S(\mathbf{w}_{T+1})\|^2). \tag{29}
$$

From the smoothness property in Lemma B.2 and the optimization error bound in (28), with probability at least $1-\delta$, we have

$$\begin{aligned}\|\nabla F_S(\mathbf{w}_{T+1})\|^2 &\leq (2\beta)^{-1}(F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)))\\ &= \begin{cases} \mathcal{O}\left(\frac{\log(1/\delta)}{T}\right) & \text{if } \theta = \frac{1}{2},\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T}{T}\right) & \text{if } \theta \in (\frac{1}{2},1],\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}}T}{T}\right) & \text{if } \theta > 1.\end{cases}\end{aligned} \tag{30}$$

Plugging (30) and (25) into (29), we have the following inequality with probability at least $1-2\delta$

$$\begin{aligned}&F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*)\\ &= \begin{cases} \mathcal{O}\left(\frac{\log(1/\delta)}{T} + \frac{d+\log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log T\right) & \text{if } \theta = \frac{1}{2},\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T}{T} + \frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log T\right) & \text{if } \theta \in (\frac{1}{2},1],\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}}T}{T} + \frac{d+\log(\frac{1}{\delta})}{n}\left(\log^{(2\theta+1)}(\frac{1}{\delta}) + \log^{\theta-1}(\frac{T}{\delta})\log(\frac{1}{\delta})\right)\log T\right) & \text{if } \theta > 1.\end{cases}\end{aligned}$$

Selecting $T \asymp n$, the above abounds means that with probability at least $1-\delta$, we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \begin{cases} \mathcal{O}\left(\frac{d+\log(\frac{1}{\delta})}{n}\log^2(\frac{1}{\delta})\log n\right) & \text{if } \theta = \frac{1}{2},\\ \mathcal{O}\left(\frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log n\right) & \text{if } \theta \in (\frac{1}{2},1],\\ \mathcal{O}\left(\frac{d+\log(\frac{1}{\delta})}{n}\log^{(2\theta+1)}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta})\log n\right) & \text{if } \theta > 1.\end{cases}$$

The proof is complete. □

C.3. Proof of Section 3.3

C.3.1. Proof of Theorem 3.13

Theorem C.7. *Suppose Assumptions 2.4, 2.6, 2.10, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, then for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 = \mathcal{O}\left(\frac{\log(1/\delta)}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right),$$

(b.) if $\theta \in (\frac{1}{2},1]$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right),$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{t+1})\|^2 = \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}}T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right).$$

Proof. According to Lemma B.4, by Assumptions 2.6, 2.10, and (2), with probability at least $1-\delta$ we have

$$\begin{aligned}\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 &\leq \Big(\|\nabla F_S(\mathbf{w}_{T+1})\| + \frac{\mu}{n} + 2\frac{G_*\log(4/\delta)}{n} + 2\sqrt{\frac{2\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(4/\delta)}{n}}\Big)^2\\ &\leq 4\Big(\|\nabla F_S(\mathbf{w}_{T+1})\|^2 + 4\frac{G_*^2\log^2(4/\delta)}{n^2} + 8\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(4/\delta)}{n} + \frac{\mu^2}{n^2}\Big).\end{aligned} \tag{31}$$

In (30), we have proved that the following optimization error bound holds with probability at least $1-\delta$

$$\|\nabla F_S(\mathbf{w}_{T+1})\|^2 = \begin{cases} \mathcal{O}\left(\frac{\log(1/\delta)}{T}\right) & \text{if } \theta = \frac{1}{2}, \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T}\right) & \text{if } \theta \in (\frac{1}{2}, 1], \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T}\right) & \text{if } \theta > 1. \end{cases} \tag{32}$$

Plugging (32) into (31), we have the following inequality with probability at least $1-2\delta$

$$\begin{aligned} &\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 \\ &= \begin{cases} \mathcal{O}\left(\frac{\log(1/\delta)}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta = \frac{1}{2}, \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta \in (\frac{1}{2}, 1], \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta > 1, \end{cases} \end{aligned}$$

which also means that with probability at least $1-\delta$ we have

$$\begin{aligned} &\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 \\ &= \begin{cases} \mathcal{O}\left(\frac{\log(1/\delta)}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta = \frac{1}{2}, \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta \in (\frac{1}{2}, 1], \\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T} + \frac{\log^2(1/\delta)}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta > 1. \end{cases} \end{aligned} \tag{33}$$

The proof is complete. □

C.3.2. Proof of Theorem 3.15

Theorem C.8. *Suppose Assumptions 2.4, 2.6, 2.10, 2.12, and (2) hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 1. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n^2$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, for any $\delta \in (0,1)$, with probability $1-\delta$,*

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^2(\frac{1}{\delta})}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(\frac{1}{\delta})}{n}\right);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 2.8 holds, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right);$$

(c.) if $\theta > 1$ and Assumption 2.8 holds, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\left(\frac{\log^{\frac{3(\theta-1)}{2}}(n/\delta)\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right).$$

Proof. According to (2), we know

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu}\|\nabla F(\mathbf{w}_{T+1})\|^2 \leq \mu^{-1}(\|\nabla F(\mathbf{w}_{T+1}) - \nabla F_S(\mathbf{w}_{T+1})\|^2 + \|\nabla F_S(\mathbf{w}_{T+1})\|^2). \tag{34}$$

Plugging (33) and (32) into (34), we derive that the following inequality holds with probability at least $1-2\delta$

$$F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)=\begin{cases}\mathcal{O}\left(\frac{\log(1/\delta)}{T}+\frac{\log^2(1/\delta)}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta=\frac{1}{2},\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}T}{T}+\frac{\log^2(1/\delta)}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta\in(\frac{1}{2},1],\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}}T}{T}+\frac{\log^2(1/\delta)}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta>1.\end{cases}$$

Choosing $T\asymp n^2$, the above abounds means that with probability at least $1-\delta$, we have

$$F(\mathbf{w}_{T+1})-F(\mathbf{w}^*)=\begin{cases}\mathcal{O}\left(\frac{\log^2(1/\delta)}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta=\frac{1}{2},\\ \mathcal{O}\left(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}n}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta\in(\frac{1}{2},1],\\ \mathcal{O}\left(\frac{\log^{\frac{3(\theta-1)}{2}}(n/\delta)\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}}n}{n^2}+\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*,z)\|^2]\log(1/\delta)}{n}\right) & \text{if } \theta>1,\end{cases}$$

The proof is complete. □

C.4. Proof of Section 3.4

Theorem C.9. *Suppose Assumptions 2.6 and 2.12 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm 2. Assume $\eta_t=\eta_1 t^{-\frac{1}{2}}$ for some constant $\eta_1>0$ and take $\tau=\max\left\{20K\log^{\theta}(\frac{2}{\delta}),4K\log^{\theta}\sqrt{T}\right\}$. Then for any $\delta\in(0,1)$, with probability $1-\delta$, we have the following inequality*

$$\frac{1}{T}\sum_{t=1}^{T}\min\left\{\|\nabla F_S(\mathbf{w}_t)\|,\|\nabla F_S(\mathbf{w}_t)\|^2\right\}=\mathcal{O}\left(\frac{\log^{\theta}(T/\delta)\log T+\log^{2\theta+1}(T)\log(\frac{T}{\delta})}{\sqrt{T}}\right).$$

Proof. It is easy to verify that $\|\nabla\bar{f}(\mathbf{w}_t;z_{j_t})\|\le\tau$. We consider two cases: $\|F_S(\mathbf{w}_t)\|\le\tau/2$ and $\|F_S(\mathbf{w}_t)\|\ge\tau/2$.

We first consider the case $\|F_S(\mathbf{w}_t)\|\le\tau/2$. With the smoothness property in Lemma B.2, we have

$$\begin{aligned}&F_S(\mathbf{w}_{t+1})-F_S(\mathbf{w}_t)\le\langle\mathbf{w}_{t+1}-\mathbf{w}_t,\nabla F_S(\mathbf{w}_t)\rangle+\frac{1}{2}\beta\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\\&\le-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle+\frac{1}{2}\beta\eta_t^2\tau^2 \qquad (35)\\&=-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})+\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t),\nabla F_S(\mathbf{w}_t)\rangle-\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2+\frac{1}{2}\beta\eta_t^2\tau^2\\&=-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle-\eta_t\langle\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t),\nabla F_S(\mathbf{w}_t)\rangle\\&-\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2+\frac{1}{2}\beta\eta_t^2\tau^2.\end{aligned}$$

By a summation from $t=1$ to $t=T$, it gives

$$\begin{aligned}&\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\le F_S(\mathbf{w}_1)-F_S(\mathbf{w}_S)+\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\tau^2\\&-\sum_{t=1}^{T}\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle-\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t),\nabla F_S(\mathbf{w}_t)\rangle.\end{aligned}$$

Since $\mathbb{E}_{j_t}[-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle]=0$, thus the sequence $(-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle,t\in\mathbb{N})$ is a martingale difference sequence. Denoted by $\xi_t=-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle$. We have

$$|\xi_t|\le\eta_t(\|\nabla\bar{f}(\mathbf{w}_t;z_{j_t})\|+\|\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})\|)\|\nabla F_S(\mathbf{w}_t)\|\le\eta_1\tau^2.$$

According to the inequality $\mathbb{E}_{j_t}[(\xi_t - \mathbb{E}_{j_t}\xi_t)^2] \leq \mathbb{E}_{j_t}[\xi_t^2]$, we have

$$\sum_{t=1}^{T} \mathbb{E}_{j_t}[(\xi_t - \mathbb{E}_{j_t}\xi_t)^2] \leq \sum_{t=1}^{T} \eta_t^2 \mathbb{E}_{j_t}[\|\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t})\|^2 \|\nabla F_S(\mathbf{w}_t)\|^2] \leq 4\tau^2 \sum_{t=1}^{T} \eta_t^2 \|\nabla F_S(\mathbf{w}_t)\|^2.$$

According to Lemma B.5, with probability $1-\delta$, we have

$$\sum_{t=1}^{T} \xi_t \leq \frac{\rho 4\tau^2 \eta_1 \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2}{\eta_1 \tau^2} + \frac{\eta_1 \tau^2 \log(1/\delta)}{\rho}.$$

Taking $\rho = \frac{1}{16}$, we derive

$$\sum_{t=1}^{T} \xi_t \leq \frac{\sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2}{4} + 16\eta_1 \tau^2 \log(1/\delta).$$

Thus, we have the following inequality with probability $1-\delta$,

$$\begin{aligned} \frac{3}{4}\sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 \leq F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S) + \sum_{t=1}^{T} \frac{1}{2}\beta \eta_t^2 \tau^2 + 16\eta_1 \tau^2 \log(1/\delta) \\ - \sum_{t=1}^{T} \eta_t \langle \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle. \end{aligned} \tag{36}$$

Furthermore, we bound term $-\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle$. We have

$$-\sum_{t=1}^{T} \eta_t \langle \mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t), \nabla F_S(\mathbf{w}_t)\rangle \leq \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\mathbb{E}_{j_t}\nabla \bar{f}(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2. \tag{37}$$

Define $x_t = \mathbb{I}_{\{\|\nabla f(\mathbf{w}_t; z_{j_t})\| > \tau\}}$ and $y_t = \mathbb{I}_{\{\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| > \frac{1}{2}\tau\}}$. Since $\|F_S(\mathbf{w}_t)\| \leq \tau/2$, we have

$$\begin{aligned} \|\nabla f(\mathbf{w}_t; z_{j_t})\| &\leq \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| + \|\nabla F_S(\mathbf{w}_t)\| \\ &\leq \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| + \frac{1}{2}\tau, \end{aligned}$$

which implies that $x_t \leq y_t$.

For the term $\|\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|$, we have

$$\begin{aligned}\|\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\| &= \|\mathbb{E}_{j_t}\Big[(\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla f(\mathbf{w}_t;z_{j_t}))x_t\Big]\| \\ &= \|\mathbb{E}_{j_t}\Big[\nabla f(\mathbf{w}_t;z_{j_t})(\frac{\tau}{\|\nabla f(\mathbf{w}_t;z_{j_t})\|}-1)x_t\Big]\| \\ &= \|\mathbb{E}_{j_t}\Big[\nabla f(\mathbf{w}_t;z_{j_t})(\frac{\tau-\|\nabla f(\mathbf{w}_t;z_{j_t})\|}{\|\nabla f(\mathbf{w}_t;z_{j_t})\|})x_t\Big]\| \\ &\le \mathbb{E}_{j_t}\Big[\|\nabla f(\mathbf{w}_t;z_{j_t})(\frac{\tau-\|\nabla f(\mathbf{w}_t;z_{j_t})\|}{\|\nabla f(\mathbf{w}_t;z_{j_t})\|})x_t\|\Big] \\ &= \mathbb{E}_{j_t}\Big[|\|\nabla f(\mathbf{w}_t;z_{j_t})\|-\tau|x_t\Big] \\ &\le \mathbb{E}_{j_t}\Big[|\|\nabla f(\mathbf{w}_t;z_{j_t})\|-\|\nabla F_S(\mathbf{w}_t)\||x_t\Big] \\ &\le \mathbb{E}_{j_t}\Big[|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\||x_t\Big] \\ &\le \mathbb{E}_{j_t}\Big[\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|y_t\Big] \\ &\le \sqrt{\mathbb{E}_{j_t}\Big[\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|^2\Big]\mathbb{E}_{j_t}y_t^2} \\ &\le \sqrt{2K^2\Gamma(2\theta+1)}\sqrt{\mathbb{E}_{j_t}y_t^2}, \end{aligned} \tag{38}$$

where the first inequality holds due to Jensen's inequality, the second inequality holds due to $\|\nabla F_S(\mathbf{w}_t)\|\le\tau/2$ and $\|\nabla f(\mathbf{w}_t;z_{j_t})\|>\tau$, the fifth inequality follows from the Schwartz's inequality, and the last inequality follows from Lemma 22 of (Madden et al., 2021). Moreover, we can derive the following inequality with probability $1-\delta$

$$\begin{aligned}\mathbb{E}_{j_t}y_t^2 = P_{j_t}(y_t=1) &= P_{j_t}\Big(\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|>\frac{1}{2}\tau\Big) \\ &\le 2\exp\Big(-\Big(\frac{\tau}{4K}\Big)^{\frac{1}{\theta}}\Big),\end{aligned} \tag{39}$$

where the inequality holds due to the tail bound of the subWeibull random variable in Lemma A.1. Thus, Combined (37) with (38) and (39), we have the following inequality with probability $1-T\delta$

$$\begin{aligned}&-\sum_{t=1}^{T}\eta_t\langle\mathbb{E}_{j_t}\nabla\bar{f}(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t),\nabla F_S(\mathbf{w}_t)\rangle \\ \le& 2K^2\Gamma(2\theta+1)\sum_{t=1}^{T}\eta_t\exp\Big(-\Big(\frac{\tau}{4K}\Big)^{\frac{1}{\theta}}\Big)+\frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2.\end{aligned} \tag{40}$$

Plugging (40) into (36), and by a rearrangement, we have the following inequality with probability $1-T\delta-\delta$

$$\begin{aligned}&\frac{1}{4}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 \\ \le& F_S(\mathbf{w}_1)-F_S(\mathbf{w}_S)+\sum_{t=1}^{T}\frac{1}{2}\beta\eta_t^2\tau^2+16\eta_1\tau^2\log(1/\delta)+2K^2\Gamma(2\theta+1)\sum_{t=1}^{T}\eta_t\exp\Big(-\Big(\frac{\tau}{4K}\Big)^{\frac{1}{\theta}}\Big).\end{aligned}$$

To continue the proof, we let

$$\exp\Big(-\Big(\frac{\tau}{4K}\Big)^{\frac{1}{\theta}}\Big)\le\frac{1}{\sqrt{T}}.$$

Then we get

$$\tau\ge 4K\log^{\theta}\sqrt{T}.$$

Thus, when $\tau = 4K\log^{\theta}\sqrt{T}$, with probability $1-T\delta-\delta$ we have

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\\
\leq& 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}_S)) + \sum_{t=1}^{T}2\beta\eta_t^2\tau^2 + 64\eta_1\tau^2\log(1/\delta) + 8K^2\Gamma(2\theta+1)\sum_{t=1}^{T}\eta_t\exp\Big(-\Big(\frac{\tau}{4K}\Big)^{\frac{1}{\theta}}\Big)\\
\leq& 4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}_S)) + 16K^2\log^{2\theta}\sqrt{T}2\beta\sum_{t=1}^{T}\eta_t^2 + 64\eta_1 16K^2\log^{2\theta}\sqrt{T}\log(1/\delta) + 8K^2\Gamma(2\theta+1)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t.
\end{aligned}$$

Since $\eta_t=\eta_1\frac{1}{\sqrt{t}}$, according to Lemma B.1, we can further get the following inequality with probability $1-T\delta-\delta$

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2\\
\leq&\frac{1}{\sqrt{T}}\Big(4(F_S(\mathbf{w}_1)-F_S(\mathbf{w}_S)) + 16K^2\log^{2\theta}\sqrt{T}2\beta\eta_1^2\log(eT) + 64\eta_1 16K^2\log^{2\theta}\sqrt{T}\log(1/\delta) + 4K^2\Gamma(2\theta+1)\eta_1\Big)\\
=&\mathcal{O}\left(\frac{1}{\sqrt{T}}\Big(\log^{2\theta}\sqrt{T}\log(T)\log(1/\delta)\Big)\right),
\end{aligned}$$

which implies that with probability $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(\mathbf{w}_t)\|^2 \leq \frac{1}{\eta_1\sqrt{T}}\sum_{t=1}^{T}\eta_t\|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{T}}\Big(\log^{2\theta+1}(T)\log(\frac{T}{\delta})\Big)\right).$$

Secondly, we consider the case $\|F_S(\mathbf{w}_t)\|\geq\tau/2$. Recall that from (35) we have

$$F_S(\mathbf{w}_{t+1})-F_S(\mathbf{w}_t)\leq -\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle + \frac{1}{2}\beta\eta_t^2\tau^2.$$

We now analyze the term $-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle$. Specifically,

$$\begin{aligned}
&-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle\\
=&-\eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|>\tau} - \eta_t\langle\nabla\bar{f}(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}\\
=&-\eta_t\tau\langle\nabla f(\mathbf{w}_t;z_{j_t})/\|\nabla f(\mathbf{w}_t;z_{j_t})\|,\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|>\tau} - \eta_t\langle\nabla f(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}.
\end{aligned}$$

Then, according to Lemma B.6, we have

$$\begin{aligned}
&-\eta_t\tau\langle\nabla f(\mathbf{w}_t;z_{j_t})/\|\nabla f(\mathbf{w}_t;z_{j_t})\|,\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|>\tau}\\
\leq&-\eta_t\tau\left(\frac{\|\nabla F_S(\mathbf{w}_t)\|}{3}(1-\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}) - \frac{8}{3}\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\right).
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&-\eta_t\langle\nabla f(\mathbf{w}_t;z_{j_t}),\nabla F_S(\mathbf{w}_t)\rangle\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}\\
=&-\eta_t(\|\nabla F_S(\mathbf{w}_t)\|^2 + \langle\nabla F_S(\mathbf{w}_t),\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\rangle)\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}\\
\leq&-\eta_t(\|\nabla F_S(\mathbf{w}_t)\|^2 - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|)\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau}\\
=&-\eta_t\Big(\|\nabla F_S(\mathbf{w}_t)\|^2\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau} - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau,\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\leq\tau/8} -\\
&\|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau,\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\geq\tau/8}\Big)\\
\leq&-\eta_t\Big(\frac{3}{4}\|\nabla F_S(\mathbf{w}_t)\|^2\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau} - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\Big)\\
\leq&-\eta_t\Big(\frac{3}{8}\tau\|\nabla F_S(\mathbf{w}_t)\|\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau} - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\Big)\\
\leq&-\eta_t\Big(\frac{1}{3}\tau\|\nabla F_S(\mathbf{w}_t)\|\mathbb{I}_{\|\nabla f(\mathbf{w}_t;z_{j_t})\|\leq\tau} - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t;z_{j_t})-\nabla F_S(\mathbf{w}_t)\|\Big),
\end{aligned}$$

where the first equation holds due to $\langle \nabla f(\mathbf{w}_t; z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle = \|\nabla F_S(\mathbf{w}_t)\|^2 + \langle \nabla F_S(\mathbf{w}_t), \nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\rangle$, where the second inequality follows from that if $\|F_S(\mathbf{w}_t)\| \geq \tau/2$ and $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \leq \tau/8$, then

$$-\|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \geq -\frac{\tau^2}{16} \geq -\frac{\|\nabla F_S(\mathbf{w}_t)\|^2}{4},$$

and where the third inequality holds due to $\|F_S(\mathbf{w}_t)\| \geq \tau/2$. Thus, we have

$$\begin{aligned}
& -\eta_t \langle \nabla \bar{f}(\mathbf{w}_t; z_{j_t}), \nabla F_S(\mathbf{w}_t)\rangle \\
\leq & -\eta_t \Big(\frac{1}{3}\tau \|\nabla F_S(\mathbf{w}_t)\| \mathbb{I}_{\|\nabla f(\mathbf{w}_t; z_{j_t})\| \leq \tau} - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|\Big) \\
& -\eta_t \tau \left(\frac{\|\nabla F_S(\mathbf{w}_t)\|}{3}(1 - \mathbb{I}_{\|\nabla f(\mathbf{w}_t; z_{j_t})\| \leq \tau}) - \frac{8}{3}\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|\right) \\
= & -\eta_t \Big(\frac{1}{3}\tau \|\nabla F_S(\mathbf{w}_t)\| - \frac{8}{3}\tau \|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| - \|\nabla F_S(\mathbf{w}_t)\|\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|\Big).
\end{aligned}$$

According to Lemma A.2, with probability at least $1 - \delta$, we also have

$$\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \leq K \log^\theta(\frac{2}{\delta}).$$

Thus, with probability at least $1 - \delta$, there holds

$$\begin{aligned}
& F_S(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_t) \leq \frac{1}{2}\beta \eta_t^2 \tau^2 \\
& - \eta_t \Big(\frac{1}{3}\tau \|\nabla F_S(\mathbf{w}_t)\| - \frac{16}{3}\|\nabla F_S(\mathbf{w}_t)\| K \log^\theta(\frac{2}{\delta}) - \|\nabla F_S(\mathbf{w}_t)\| K \log^\theta(\frac{2}{\delta})\Big),
\end{aligned}$$

where the inequality also follows from $\|F_S(\mathbf{w}_t)\| \geq \tau/2$. Taking a Summation form $t = 1$ to $t = T$, with probability at least $1 - T\delta$, we have

$$\begin{aligned}
& F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}_1) \leq \frac{1}{2}\beta \sum_{t=1}^T \eta_t^2 \tau^2 \\
& - \sum_{t=1}^T \eta_t \Big(\frac{1}{3}\tau \|\nabla F_S(\mathbf{w}_t)\| - \frac{16}{3}\|\nabla F_S(\mathbf{w}_t)\| K \log^\theta(\frac{2}{\delta}) - \|\nabla F_S(\mathbf{w}_t)\| K \log^\theta(\frac{2}{\delta})\Big),
\end{aligned}$$

which means

$$\sum_{t=1}^T \Big(\frac{1}{3}\tau - \frac{19}{3}K \log^\theta(\frac{2}{\delta})\Big)\eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S) + \sum_{t=1}^T \frac{1}{2}\beta \eta_t^2 \tau^2.$$

To continue the proof, we take $\tau \geq 20K \log^\theta(\frac{2}{\delta})$. Then we get

$$\sum_{t=1}^T \Big(\frac{1}{3}K \log^\theta(\frac{2}{\delta})\Big)\eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq \sum_{t=1}^T \Big(\frac{1}{3}\tau - \frac{19}{3}K \log^\theta(\frac{2}{\delta})\Big)\eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S) + \sum_{t=1}^T \frac{1}{2}\beta \eta_t^2 \tau^2.$$

It implies that

$$\sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq \frac{3(F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S))}{K \log^\theta(\frac{2}{\delta})} + \sum_{t=1}^T \frac{3}{2K \log^\theta(\frac{2}{\delta})}\beta \eta_t^2 \tau^2.$$

Therefore, with probability at least $1 - T\delta$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq \frac{3(F_S(\mathbf{w}_1) - F_S(\mathbf{w}_S))}{\sqrt{T}K \log^\theta(\frac{2}{\delta})} + \frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{3}{2K \log^\theta(\frac{2}{\delta})}\beta \eta_t^2 \tau^2.$$

Taking $\tau = 20K \log^{\theta}(\frac{2}{\delta})$ and since $\eta_t = \eta_1 \frac{1}{\sqrt{t}}$, we get the following inequality with probability at least $1 - T\delta$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\| = \mathcal{O}\left(\frac{\log^{\theta}(1/\delta) \log T}{\sqrt{T}}\right),$$

implying that with probability at least $1 - \delta$,

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla F_S(\mathbf{w}_t)\| \leq \frac{1}{\eta_1 \sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\| = \mathcal{O}\left(\frac{\log^{\theta}(T/\delta) \log T}{\sqrt{T}}\right).$$

Till here, combined with the two cases and taking $\tau = \max\left\{20K \log^{\theta}(\frac{2}{\delta}), 4K \log^{\theta} \sqrt{T}\right\}$, we finally obtain the following inequality with probability $1 - \delta$

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^{T} \min\left\{\|\nabla F_S(\mathbf{w}_t)\|, \|\nabla F_S(\mathbf{w}_t)\|^2\right\} &= \mathcal{O}\left(\frac{\log^{\theta}(T/\delta) \log T}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\left(\log^{2\theta+1}(T) \log(\frac{T}{\delta})\right)\right) \\ &= \mathcal{O}\left(\frac{\log^{\theta}(T/\delta) \log T + \log^{2\theta+1}(T) \log(\frac{T}{\delta})}{\sqrt{T}}\right).\end{aligned}$$

The proof is complete. □