Tight High-Probability Bounds for Nonconvex Heavy-Tailed Scenario under Weaker Assumptions

Weixin An[1], Yuanyuan Liu[1]*, Fanhua Shang[2]*, Han Yu[3], Junkang Liu[2], Hongying Liu[4,5]*
[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, China
[2]School of Computer Science and Technology, Tianjin University, China
[3]College of Computing and Data Science, Nanyang Technological University, Singapore
[4]Medical School, Tianjin University, China
[5]Peng Cheng Lab, Shenzhen, China
`weixinanut@163.com, yyliu@xidian.edu.cn, fhshang@tju.edu.cn`
`han.yu@ntu.edu.sg, junkangliukk@gmail.com, hyliu2009@tju.edu.cn`

Abstract

Gradient clipping is increasingly important in centralized learning (CL) and federated learning (FL). Many works focus on its optimization properties under strong assumptions involving Gaussian noise and standard smoothness. However, practical machine learning tasks often only satisfy weaker conditions, such as heavy-tailed noise and (L_0, L_1)-smoothness. To bridge this gap, we propose a high-probability analysis for clipped Stochastic Gradient Descent (SGD) under these weaker assumptions. Our findings show a better convergence rate than existing ones can be achieved, and our high-probability analysis does not rely on the bounded gradient assumption. Moreover, we extend our analysis to FL, where a gap remains between expected and high-probability convergence, which the naive clipped SGD can not bridge. Thus, we design a new Federated Clipped Batched Gradient (`FedCBG`) algorithm, and prove the convergence and generalization bounds with high probability for the first time. Our analysis reveals the trade-offs between the optimization and generalization performance. Extensive experiments demonstrate that `FedCBG` can generalize better to unseen client distributions than state-of-the-art baselines.

1 Introduction

Gradient clipping has proven effective in training vision and language models [55, 58]. Many studies demonstrated an optimal convergence rate of $\mathcal{O}(T^{-\frac{1}{2}})$ under a finite-variance assumption, where T is the number of iterations or communication rounds. However, recent studies [56, 16] pointed out that assuming finite-variance noise is overly optimistic for modern machine learning tasks. Instead, it is more appropriate to assume that the noise has a bounded p-th moment, as stated in Assumption 3 below (**the first weaker assumption**), which is called heavy-tailed regime. This assumption brings significant challenges for theoretical analysis. Attempts to establish the convergence rate under this assumption have been made. For example, Zhang et al. [56] showed that clipped SGD achieves the state-of-the-art convergence rate in expectation. In practice, models are usually trained only once due to the long training process. Thus, Cutkosky and Mehta [7], Nguyen et al. [37], Puchkin et al. [39] studied high-probability convergence, offering a stronger guarantee for each individual run.

However, the above high-probability results are achievable only under standard smoothness. Works have demonstrated that some language and vision models [55, 48] can not satisfy the standard

*Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

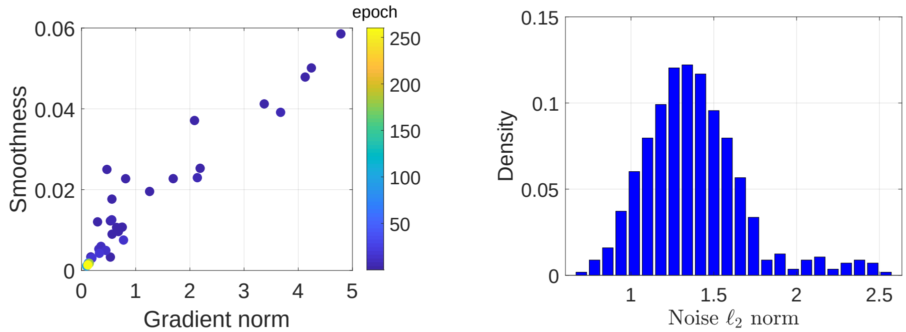

Figure 1: Gradient norm vs estimated Lipschitz smoothness (left) and gradient noise distribution (right) during training for AWD-LSTM [36] on the PTB dataset. Local smoothness positively correlates with gradient norm instead of a constant, which satisfies Assumption 2. The gradient noise exhibits a heavy-tailed behavior and its norm can be as large as 2.5. Similar phenomena also appeared in the Shakespeare dataset, as shown in Fig. 2.

smoothness assumption. Instead, Assumption 2 (**the second weaker assumption**) applies. Zhang et al. [55] first analyzed the convergence properties of clipped SGD under the (L_0, L_1)-smoothness assumption, which covers many large language models [6, 20].

The two types of works focus on the single weaker condition, but both conditions can appear in the same model, which can be verified in Fig. 1. Thus, there is an urgent need for analysis under both weaker conditions. Besides, these methods mentioned above focus on optimization, not including the generalization properties. Li and Liu [26, 27] first analyzed the generalization of clipped SGD, but both depend on the bounded gradient assumption, which is stronger than Assumption 3 and can not hold even when f is quadratic.

As for federated learning (FL) [10, 42, 50, 40, 35, 41], heavy-tailed noise also exists. The works [47, 46, 49] focuses on this issue. The most related work to our paper is [49]. Their analysis is in expectation, which offers a weak guarantee for each single run. Besides, their analysis depends on the restrictive assumption that local gradients are bounded, which may not hold even for the quadratic function and the independent Gaussian random variables. In addition, they focus on the optimization performance under the standard smoothness assumption. In summary, there is a lack of studies jointly considering the optimization and generalization properties under weaker conditions in the FL setting.

The above analysis naturally raises the following questions:

Q1: Can we analyze the clipped methods under only weaker conditions such as heavy-tailed noise and (L_0, L_1)-smoothness assumptions in high probability?

Q2: Can the analysis in CL inspire to design an FL algorithm to achieve the convergence rate matching the lower bound under weaker conditions?

Q3: Can we analyze the high-probability generalization properties under weaker assumptions for both CL and FL?

1.1 Contributions

To answer these questions, we summarize our contributions as follows:

• By induction, we prove a faster convergence rate of the clipped SGD under the weaker conditions. By carefully choosing clipping parameter, we obtain a convergence rate $\mathcal{O}(T^{\frac{2-2p}{3p-2}} \log^{\frac{2p-2}{2p-1}} \frac{1}{\delta})$ with a probability of at least $1-\delta$, $\delta \in (0,1)$, which improves existing high-probability bound $\mathcal{O}(T^{\frac{2-2p}{3p-2}} \log^2 \frac{1}{\delta})$ $(p \in (1,2])$. Interestingly, our analysis does not rely on the bounded gradient assumption used in [7, 26, 27]. Besides, we provide the generalization analysis for the first time.

• We design a new Federated Clipped Batched Gradient (`FedCBG`) algorithm for FL under weaker assumptions. We creatively prove the bounded variance of batch gradients, which opens the door to analyzing batch gradients under the heavy-tailed scenario. Then, we prove that `FedCBG` can achieve the advanced convergence rate of $\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}} \log^{\frac{p-1}{p}} \frac{T}{\delta})$, where m and K is the number of clients and local iterations, respectively. Finally, we provide a generalization upper bound for the federated setting for the first time. Our analysis reveals the trade-off between optimization and generalization. A summary of our theoretical contributions can be found in Tables 1 and 2 below.

2 Related Work

2.1 Existing analysis under weaker conditions

Many works such as [56, 49] have shown that there exists heavy-tailed noise in many applications. Analyzing convergence under this scenario is more challenging than for light-tailed noise (e.g., Gaussian noise) due to the unbounded variance, which makes most existing proof techniques inapplicable. The first type of analysis addressed this issue by assuming the bounded gradient $\mathbb{E}_t[\|\nabla f(x_t;\xi_{j_t})\|^p] \leq G^p$ [56, 49, 27]. However, this assumption is strong and cannot hold even when the loss is quadratic. As a comparison, our inductive analysis only needs the weakened Assumption 3, which is also used in works [37, 32]. Besides, the loss functions of many language and vision models can not satisfy the standard smoothness but rather a weaker (L_0, L_1)-smoothness [55, 25]. (L_0, L_1)-smoothness assumption is applied by the works [55, 53, 6, 20, 25] under the light-tailed noise. As for the heavy-tailed noise, the existing works [56, 31] only analyzed the expected rather than high-probability convergence rates as shown in Table 4 in the Appendix. In contrast, we propose tight high-probability analysis under both weaker conditions, offering a stronger guarantee for each individual training.

2.2 Optimization properties for clipped methods

In centralized learning settings, existing studies [56, 31] focused on clipping for the heavy-tailed scenario and analyzed the convergence bounds in expectation. Besides, works such as [26, 27, 37, 32] provide high-probability analysis, which matches the lower bound in expectation. However, the order of $\log \frac{T}{\delta}$ is at least 2 as shown in Table 1. We reduce this order to $\frac{2p-2}{2p-1}$ as shown in the same table. In FL, to the best of our knowledge, there is only one work [49] analyzing convergence rates under the heavy-tailed noise. However, they focus on expected rather than high-probability analysis, and optimization properties rather than generalization aspects.

2.3 Generalization for nonconvex problems

Existing generalization analysis contains three types: 1) in expectation [5, 9, 17, 23, 24, 38, 45, 52], 2) high probability [12, 34, 13, 22, 18], and 3) information theory [1, 4, 54]. For example, Hardt et al. [17] pioneered generalization analysis in expectation based on stability. However, their analysis requires a very small step size, which leads to an exponential number of iterations. As a comparison, the high-probability analysis allows the constant step size, which controls the generalization error and makes the optimization error decay faster. Besides, it can provide a stronger guarantee for each single run and is a tighter criterion for bounded losses [1]. As for the information-theoretic analysis, they are usually algorithm-independent [3]. In this paper, we provide the high-probability upper bounds for optimization and generalization and focus on their joint perspective.

3 Preliminaries

Problem setting: In this paper, we focus on the clipped methods for solving the problems in both CL and FL settings. For the CL setting, the population loss is defined as:

$$F(x) = \mathbb{E}_{\xi\sim P_\xi} f(x;\xi), \tag{1}$$

where the loss function f is nonconvex w.r.t. x, x is network weights, and one sample ξ is sampled from the distribution P_ξ. Normally, the population risk $F(x)$ is used for generalization but is computationally invisible and it can only be estimated using the empirical risk $F_S(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;\xi_i)$.

FL [33] allows multiple participants to share model training results but not data, reducing the risk of data leakage. FL usually addresses the following problem:

$$F(x) := \mathbb{E}_{i\sim\mathcal{P}}\{F_i(x) := \mathbb{E}_{\xi_j\sim P_i} f(x;\xi_j)\}, \tag{2}$$

where $f(x;\xi_j)$ is the loss at sample ξ_j, ξ_j is sampled from the local distribution P_i, each client i is sampled from a meta-distribution $\mathcal{P}$. We define the client empirical risk by $f_i(x) := \frac{1}{n_i}\sum_{j=1}^{n_i} f_i(x;\xi_j)$

and the empirical risk on the participating training client data is defined by $F_S(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$, where n_i is the number of samples of the i-th client and m is the number of participating clients. Generalization research in FL includes the performance gap on unseen client data and unseen client distributions. For the former, the CL can provide help. In this paper, we focus on the latter. For example, Problem (2) is common in the cross-device FL setting, where m is generally large and it is reasonable to sample from a meta-distribution to model local distributions of clients [51], which makes it clear the generalization to non-participating clients.

Notation: We use lower-case letters to denote vectors. For a differentiable function f, $\nabla f(x)$ is the gradient of f at x. We let $\mathcal{F}_t$ be the natural filtration for the algorithms. $\mathbb{E}_t$ is used to denote $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ for brevity.

Assumption 1 (Bounded Function). *F admits a finite lower bound, i.e., $F^* = \inf_x F(x) > -\infty$.*

Assumption 2 ((L_0, L_1)-Smoothness). *The smoothness of the function F_S means that for $\forall x, y$ satisfying $\|x - y\| \le \frac{1}{L_1}$, $\|\nabla F_S(x) - \nabla F_S(y)\| \le (L_0 + L_1\|\nabla F_S(x)\|)\|x - y\|$ holding with smoothness parameter $\ell = L_0 + L_1\|\nabla F_S(x)\|$.*

For the federated setting, the local (L_0, L_1)-Lipschitz continuous gradient for each client means $\|\nabla f_i(x) - \nabla f_i(y)\| \le (L_0 + L_1\|\nabla f_i(x)\|)\|x - y\|$. When $L_1 = 0$, they become standard smoothness.

Assumption 3 (Heavy-tailed Noise). *For the centralized setting, the stochastic gradient estimator is unbiased, i.e., $\mathbb{E}[\nabla f(x;\xi)] = \nabla F_S(x)$. Besides, the gradient noise satisfies the heavy-tailed condition $\mathbb{E}_\xi\|\nabla F_S(x) - \nabla f(x;\xi)\|^p \le \sigma^p, p \in (1, 2]$.*

For the federated setting, the local gradient estimator is unbiased, i.e., $\mathbb{E}[\nabla f_i(x;\xi)] = \nabla f_i(x)$. Besides, the local stochastic gradient noise in the i-th client follows the heavy-tailed distribution, i.e., $\mathbb{E}_\xi\|\nabla f_i(x) - \nabla f_i(x;\xi)\|^p \le \sigma^p, p \in (1, 2]$.

Many works such as image classification [44], training the large language models [56] and FL [49] have shown that stochastic gradient noise usually follows the heavy-tailed distribution, which is also corroborated by Fig. 1. Some works [7, 27] have made this assumption concrete to that the stochastic gradients are bounded in p-th moment, i.e., $\mathbb{E}_t\|\nabla f(x_t, \xi_{j_t})\|^p \le G^p$ (or $\mathbb{E}_t\|\nabla f_i(x_t, \xi_{j_t})\|^p \le G^p$ in FL), for some $G > 0$. However, it does not hold even when f is quadratic and $\nabla f(x;\xi) - \nabla F_S(x)$ (or $\nabla f_i(x;\xi) - \nabla f_i(x)$ in FL) is an independent centered Gaussian random variable. In contrast, Assumption 3 is weaker. In this paper, we focus on the analysis under Assumptions 1-3.

4 Tighter High-probability Bounds in the Centralized Setting

To answer Q1, we first consider the optimization properties of clipped SGD in Subsection 4.1. Besides, we prove its generalization bound in Subsection 4.2.

4.1 Tighter high-probability convergence under weaker conditions

The pseudocode of clipped SGD is shown in Algorithm 1. In each iteration, clipped SGD performs gradient descent along the clipped gradient $\widetilde{\nabla} f(x_t;\xi_{j_t})$.

We extend the analysis in [37] to the (L_0, L_1)-smoothness assumption, which can cover more applications. In Theorem 1, we propose better parameter choices and prove a faster convergence rate than existing analyses such as [37, 32].

Algorithm 1 Clipped SGD

Initialize: x_0, step size η and clipping parameter λ.

```
for t = 0, 1, ..., T − 1 do
    Draw i.i.d. ξ_{j_t} stochastic sample;
    ∇̃f(x_t; ξ_{j_t}) = min{1, λ / ‖∇̃f(x_t; ξ_{j_t})‖} ∇f(x_t; ξ_{j_t});
    x_{t+1} = x_t − η ∇̃f(x_t; ξ_{j_t});
end for
Randomly draw x̂ from x_1, ..., x_T at uniform;
```

Output: $\hat{x}$.

Theorem 1. *We assume that Assumptions 1, 2 and 3 hold. If we choose λ and η satisfying $\lambda = \mathcal{O}(T^{\frac{1}{3p-2}}(\log\frac{T}{\delta})^{\frac{1}{1-2p}})$, $\eta = \mathcal{O}(T^{\frac{-p}{3p-2}}(\log\frac{T}{\delta})^{\frac{2-2p}{2p-1}})$, where $\mathcal{R} = L_0 + 2(L_1 + 1)R$, the constant $R \ge 4\Delta_1 L_1 + 4\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1}$, $\Delta_1 = F_S(x_1) - F^*$, $\rho = \max\{\log\frac{4T}{\delta}, 1\}$, the clipped SGD (Algorithm 1) can achieve the convergence rate of $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 = \mathcal{O}(T^{\frac{2-2p}{3p-2}}(\log\frac{T}{\delta})^{\frac{2p-2}{2p-1}})$ with the probability at least $1 - \delta$ for any $\delta \in (0, 1)$.*

Theorem 1 offers a new high-probability optimization bound for clipped SGD. According to Jensen's inequality, the bound implies that $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\| \leq \mathcal{O}(\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}/T^{\frac{p-1}{3p-2}})$, which matches the lower bound $\Omega(T^{\frac{p-1}{3p-2}})$ in [56] up to a logarithmic factor. Compared with existing results, our bound has the following advantages.

Table 1: Comparison of existing high-probability (HP) analysis in centralized learning (CL). We use $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2$ and $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2$ as the criterion in high probability. Abbreviation: Standard smoothness (SS), (L_0, L_1)-smoothness ((L_0, L_1)), Heavy-tailed (HT), Theorem (Th.). Here, G is a constant, $\delta \in (0,1)$, and $p \in (1,2]$. It can be seen that our Theorems 1 and 2 achieve better optimization convergence and the state-of-the-art generalization bound under weaker assumptions.

Methods	Assumptions		Additional Assumptions	Bounds	
	Smooth	Noise		Optimization	Generalization
[26]	SS	HT	$\eta\|\nabla F_S(x_t)\| \leq G$	$\mathcal{O}(\frac{\log T}{T^{1/2}}\log^2\frac{1}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$
[27]	SS	HT	$\mathbb{E}_t[\|\nabla f(x_t;\xi_{j_t})\|^p] \leq G^p$	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log\frac{1}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$
[37]	SS	HT	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{p}{p-1}}\frac{T}{\delta})$	——
[32]	SS	HT	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^2\frac{T}{\delta})$	——
Th. 1, 2	(L_0, L_1)	HT	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$

• **Addressed the challenges under the weaker assumption.** In our analysis, the clipped SGD can deal with the (L_0, L_1)-smoothness rather than only standard smoothness. The generalized smoothness increased analysis difficulty due to extra $\|\nabla F_S(x)\|$ in the upper bound ((L_0, L_1)-smoothness makes the gradient upper bound implicit in an inequality, which complicates the analysis). Specifically, it can lead to a high-order term containing $\|\nabla F_S(x)\|$. The previous works like [53] keep it till the end and use the boundedness of clipped gradients to choose step size η. Instead, Zhang et al. [55] chooses carefully clipped step size $\min\{\eta, \frac{\eta\lambda}{\|\nabla F_S(x)\|}\}$ to achieve convergence. But they focus on noise with bounded variance or need additional assumption $\|\nabla f(x;\xi) - \nabla F_S(x)\| \leq \sigma$, which are not practical even when the loss is quadratic. However, the noise variance can not be easily bounded under the heavy-tailed scenario, and the boundedness of clipped gradients and clipped step sizes can not be used for high-order terms. Thus, this paper still faces the challenge of high-order terms.

Inspired by the analysis in [11, 25] for bounded variance, we prove Theorem 1 by induction. In Appendix B.1, we show how to use induction arguments to handle (L_0, L_1)-smoothness and remove the bounded gradient assumption, and here we give a proof sketch.

Proof sketch. The key in the convergence rate analysis is to show that $\|\nabla F_S(x_t)\| \leq \frac{\lambda}{2}$. By induction hypothesis at l ($l \leq t$), we creatively solve a quadratic inequality w.r.t. $\|\nabla F_S(x)\|$ so that the gradient $\|\nabla F_S(x)\|$ can be controlled under the (L_0, L_1)-smoothness when λ is greater than a constant. Based on these, we construct a new martingale difference sequence $\sum_{t=0}^{l-1}(L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle\nabla F_S(x_t), \theta_t^a\rangle$ produced by (L_0, L_1)-smoothness, which does not appear in standard analysis like in [37], where $\theta_t^a = \widetilde{\nabla} f(x_t;\xi_{j_t}) - \mathbb{E}_t[\widetilde{\nabla} f(x_t;\xi_{j_t})]$. Next, by carefully choosing λ and η, we can obtain the following induction $\Delta_{T+1} + \frac{\eta}{4}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 \leq 2\Delta_1$ with the probability at least $1-\delta$, thereby achieving the desired convergence rate.

• **Tighter convergence bound.** We analyzed the parameter selection in [37] and found that $\mathcal{O}(T^{\frac{2-2p}{3p-2}})$ is already tight but the order of $\log\frac{T}{\delta}$ can be reduced. By analyzing the inequalities that λ satisfies in our induction, we set $\lambda = \mathcal{O}(T^{\frac{1}{3p-2}}(\log\frac{T}{\delta})^{\frac{1}{1-2p}})$ and $\eta = \mathcal{O}(T^{\frac{-p}{3p-2}}(\log\frac{T}{\delta})^{\frac{2-2p}{2p-1}})$, which yields a tighter convergence rate on the logarithmic factor compared with [37, 32], as shown in Table 1.

4.2 Generalization bound under weaker assumptions

In addition to optimization, we analyze the generalization bound of clipped SGD to answer Q3. We use the term $\|\nabla F(x_t)\|^2$ to estimate this bound. Similar criteria can be found in [27, 22].

Theorem 2. *We assume that Assumptions 1, 2 and 3 hold. We set the same step size and clipping parameter as Theorem 1. If we choose $T = \mathcal{O}\left(\sqrt{\frac{n}{d}}\right)$, then with probability at least $1-\delta$, Algorithm 1 can achieve $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 \le \mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}} \log^{\frac{2p-2}{2p-1}}(\frac{1}{\delta}\sqrt{\frac{n}{d}}))$.*

Theorem 2 shows that clipped SGD can guarantee the generalization bound of the order $\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}} \log^{\frac{2p-2}{2p-1}}(\frac{1}{\delta}\sqrt{\frac{n}{d}}))$ under weaker assumptions, such as heavy-tailed noise and (L_0, L_1)-smoothness. Besides, Theorem 2 is the first high-probability generalization analysis without bounded gradient assumption. For clarification, we provide a proof sketch.

Proof sketch. We estimate the term $\|\nabla F(x_t)\|^2$ as follows $\|\nabla F(x_t)\|^2 \le 2\|\nabla F_S(x_t)\|^2 + 2\|\nabla F(x_t) - \nabla F_S(x_t)\|^2$. The first term is optimization error and we can bound it by Theorem 1. The second term is generalization error due to approximating the true gradient with its empirical counterpart and we bound it by generalized uniform convergence as shown in Lemma 5. In Lemma 5, the value of R needs to be quantified. We prove that the generalization error increases as training progresses and we can choose $R = \max_{1\le t\le T}\|x_t\|$. Next, we decompose $\|x_t\|$ into A_1, A_2, A_3 by Triangle Inequality and combine them with our inductive Lemma 6 to get $\|x_{t+1}\| = \mathcal{O}(T^{\frac{2p-1}{3p-2}} / \log^{\frac{p-1}{2p-1}}\frac{T}{\delta})$. Along this line of thought, we successfully bounded $\|\nabla F(x_t)\|^2$.

Advantages compared with existing analysis

• We offer generalized uniform convergence for (L_0, L_1)-smooth objective as shown in Lemma 5, which generalizes the results in [22, 27].

• Remove the bounded gradient assumption. We use our induction Lemma 6, which can guarantee that $\|\nabla F_S(x_t)\| \le \frac{\lambda}{2}$ with the probability at least $1-\delta$. This analysis removes the bounded gradient assumption in [27], i.e., $\mathbb{E}_t[\|\nabla F_S(x_t;\xi_t)\|^p] \le G^p$ and achieves the first generalization upper bound.

5 The Proposed `FedCBG` Algorithm for the Federated Setting

As we discussed above, heavy-tailed noise also exists in FL. To answer Q2, we extend Theorems 1 and 2 to FL, which inspires us to design an `FedCBG` algorithm to match the optimization lower bound. Besides, we provide the high-probability generalization bound for the first time.

5.1 Federated clipped batch gradient algorithm

To handle the heavy-tailed noise, we design a Federated Clipped Batch Gradient (`FedCBG`) algorithm as shown in Algorithm 2, which mainly contains two parts:

• In the client, we use clipped batch gradient $\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)$ to perform gradient descent, where $\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k) = \min\{1, \frac{\lambda}{\|\nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)\|}\}\nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)$ and $\boldsymbol{\xi}_{t,i}^k = \{(\xi_{t,i}^k)_j\}_{j=1}^b$, which is different from the existing methods such as [49], where they use a single sample in each local update. This difference is one of the key reasons why we obtain a convergence rate matching the lower bound under the more difficult criterion, i.e., in high probability.

• In the server, we design a "sum-aggregation" paradigm $x_{t+1} = x_t - \gamma\sum_{i=1}^m \widetilde{\Delta}_t$, which is the other reason why our `FedCBG` can achieve the convergence rate in high probability matching the lower bound.

Algorithm 2 `FedCBG` Algorithm

Initialize: Initial point x_0, local step size η, global learning rate γ and clipping parameter λ.
for $t = 0, 1, \ldots, T-1$ (communication round) **do**
  for each client $i \in [m]$ in parallel **do**
    Update local model: $x_{t,i}^0 = x_t$.
    for $k = 0, \cdots, K-1$ (local update step) **do**
      Draw i.i.d. stochastic samples $\boldsymbol{\xi}_{t,i}^k$;
      $x_{t,i}^{k+1} = x_{t,i}^k - \eta\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)$;
    end for
    Send $\widetilde{\Delta}_t^i = \sum_{k=0}^{K-1}\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)$ to the server.
  end for
  Global sum-aggregation at server:
  Server update: $x_{t+1} = x_t - \gamma\sum_{i=1}^m \widetilde{\Delta}_t$;
  Broadcasting x_{t+1} to clients.
end for
Output: x_T.

5.2 Convergence rate of our FedCBG algorithm

To prove the convergence rate of FedCBG, we need to bound the variance of the batch gradient under the heavy-tailed noise assumption. Compared to the Gaussian noise (light-tailed) assumption, such analysis is more difficult. Fortunately, by Hölder Inequality and Markov's Inequality, we have proved an upper bound on this variance for the first time in Lemma 1.

Lemma 1 (**Batch gradient variance bound**). *If Assumptions 3 holds and $\|\nabla f_i(x_t)\| \leq \frac{\lambda}{2}, \forall i \in [m]$, for batch gradient $\nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k) = \frac{1}{b}\sum_{\xi_{t,i}^k \in \boldsymbol{\xi}_{t,i}^k} \nabla f_i(x_{t,i}^k; \xi_{t,i}^k)$, we have the batch gradient variance bound*

$$\mathbb{E}_t[\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)\|^2] \leq \frac{3\sigma^p\lambda^{2-p}}{b}. \tag{3}$$

Remark 1. *In Lemma 1, we provide the* first upper bound *for batched gradient variance under the heavy-tailed noise. In our high-probability analysis,* $b > 1$ *provides one parameter of freedom for choosing the clipping parameter λ, thereby achieving the convergence rate matching the lower bound in expectation. Specifically, we set $b = \mathcal{O}((mKT)^{\frac{2p-2}{3(3p-2)}})$, which allows us to choose $\lambda = \mathcal{O}((mKT)^{\frac{1}{2(3p-2)}})$, thereby achieving the desired convergence rate.*

Inspired by the definitions of θ_t^a and θ_t^b ($\theta_t^b = \mathbb{E}_t[\widetilde{\nabla} f(x_t; \xi_{j_t})] - \nabla F_S(x_t)$) in the centralized setting, we construct three errors in the federated setting: stochastic batch error ϵ_t, the clipped batch gradient deviation ϵ_t^a, and the bias ϵ_t^b between the expected clipped batch gradient and full gradient, where $\epsilon_t^a = \frac{1}{mK}\sum_{i=1}^m \sum_{k=1}^K (\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k) - \mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)])$, $\epsilon_t^b = \frac{1}{mK}\sum_{i=1}^m \sum_{k=1}^K \mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)] - \nabla F_S(x_t)$, and $\epsilon_t = \epsilon_t^a + \epsilon_t^b$. Based on Lemma 1, we analyze their upper bounds in Lemma 2.

Lemma 2. *For Algorithm 2, $\forall t \in [T]$, we have $\|\epsilon_t^a\| \leq 2\lambda$. Besides, if $\|\nabla f_i(x_t)\| \leq \frac{\lambda}{2}$, there is $\|\epsilon_t^b\| \leq \frac{12\sigma^p\lambda^{1-p}}{b}$ and $\mathbb{E}_t[\|\epsilon_t^a\|^2] \leq \frac{100\sigma^p\lambda^{2-p}}{mKb}$.*

Lemma 2 shows that the clipped batch gradient can add a parameter of freedom b compared with [37, 39], which relaxes the conditions for choosing hyperparameters. Now, we begin to prove the convergence rate of our FedCBG algorithm. The key to our derivation lies in Lemma 3.

Lemma 3. *For $1 \leq N \leq T+1$, let E_N' be the event that for all $l = 1, \cdots, N$,*

$$\begin{aligned}
&\Delta_l' + \frac{\gamma mK}{2}\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^2 \leq \Delta_1' + \gamma mK\sum_{t=1}^{l-1}[(1 + L_1\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]) \\
&+ L_1\|\nabla F_S(x_t)\|(\langle\epsilon_t^a, \nabla F_S(x_t)\rangle + \|\nabla F_S(x_t)\|\|\epsilon_t^b\|)] + \frac{L_1\gamma^2m^2K^2}{2}\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^3 \\
&+ \gamma mK\sum_{t=1}^{l-1}(1 + L_1\|\nabla F_S(x_t)\|)(\|\epsilon_t^b\|^2 + \mathbb{E}_t[\|\epsilon_t^a\|^2]) \leq 2\Delta_1'.
\end{aligned} \tag{4}$$

Then E_N' happens with probability at least $1 - \frac{(N-1)\delta}{T}$ for each $N \in [T]$.

Lemma 3 explains why our Algorithm 2 can achieve a convergence rate matching the lower bound. Specifically, in our inductive analysis, we focus on constructing the martingale difference sequences $\{\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]\}$ and $\{\langle\epsilon_t^a, \nabla F_S(x_t)\rangle\}$ and bound them in high probability by Freedman's inequality. Besides, by induction hypothesis at l ($l \leq t$), we also creatively solve a quadratic inequality w.r.t. $\|\nabla F_S(x)\|$ and $\|\nabla f_i(x)\|$ so that they can be controlled under the (L_0, L_1)-smoothness when λ is greater than a constant. Then, we leverage Lemma 2 and choose appropriate η, γ, b and λ to balance all the terms to achieve the desired convergence rate. Based on Lemma 3, we prove the convergence rate of Algorithm 2 as shown in Theorem 3.

Theorem 3. *We assume that Assumptions 1, 2 and 3 hold. If we choose $b = \mathcal{O}((mKT)^{\frac{2p-2}{3(3p-2)}})$, $\lambda = \mathcal{O}((mKT)^{\frac{1}{3(3p-2)}}/\rho^{\frac{1}{2p}})$, $\gamma = \mathcal{O}((mKT)^{\frac{-p}{3p-2}}/\rho^{\frac{p-1}{p}})$, $\eta = \mathcal{O}(\frac{\log^{\frac{1}{p}}\frac{T}{\delta}}{K^{\frac{17p-8}{4(3p-2)}}(mT)^{\frac{5p}{4(3p-2)}}})$, where $\mathcal{R}' = 1 + 2(\frac{L_1}{L_0} + 1)R'$, $R' \geq 4\Delta_1'L_1 + 4\sqrt{(\Delta_1')^2L_1^2 + L_0\Delta_1'}$, $\rho = \max\{\log\frac{4T}{\delta}, 1\}$, $\Delta_t' = F_S(x_t) - F^*$, and $\beta = \min\{\frac{32\mathcal{R}\rho}{L_0}, \frac{3}{4}, \frac{3L_0}{8L_1R'}\}$, Algorithm 2 can achieve the convergence rate* *of* $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 = \mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{T}{\delta})$ *with the probability at least* $1-\delta$ *for any* $\delta \in (0,1)$*.*

Table 2: Comparison of the existing analysis in FL. "AA" indicates whether the additional gradient boundedness assumption $\mathbb{E}_t[\|\nabla f_i(x_t;\xi_{j_t})\|^p] \le G^p$ are required, and "LB" refers to the lower bound.

Methods	Assumptions		Criteria	AA	Bounds	
	Smooth.	Noise			Optimization	Generalization
[49]	SS	HT	Exp	✓	$\mathcal{O}((mT)^{\frac{2-2p}{3p-2}}K^{\frac{4-2p}{3p-2}})$	——
[49]	SS	HT	Exp	✓	$\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}})$	——
LB	SS	HT	Exp	✗	$\Omega((mKT)^{\frac{2-2p}{3p-2}})$	——
Th. 3, 4	(L_0, L_1)	HT	HP	✗	$\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{1}{\delta})$	$\mathcal{O}\big((\frac{d}{n})^{\frac{p-1}{7p-6}}\log(\frac{1}{\delta}(\frac{n}{d})^{\frac{3p-2}{2(7p-6)}})\big)$

Theorem 3 shows that our `FedCBG` algorithm can achieve the desired convergence rate for the heavy-tailed noise setting. The size of b is consistent with our intuition that the smaller p is, the more sensitive the algorithm is to noise, the gradient differences between different samples may be large, thus a small b can achieve an ideal convergence rate. This convergence rate $\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{T}{\delta})$ matches the lower bound proposed in [49] up to a logarithmic factor as shown in Table 2. Thus, `FedCBG` effectively reduces the number of communication rounds. Besides, our clipping parameter λ is smaller than that of [49]. Small λ is typically used and often leads to good performance [36, 57], as stated in [19]. The specific parameter choices and the inequalities they need to satisfy can be found in Appendix C.1. Compared with the state-of-the-art methods in [49], our analysis has the following advantages. **1)** Our analysis is in high probability and provides a stronger guarantee for a single run. **2)** Our analysis use the weaker Assumption 3 rather than the bounded gradient assumption, i.e., $\mathbb{E}_t[\|\nabla f_i(x_t;\xi_{j_t})\|^p] \le G^p$. **3)** Our analysis use the weaker Assumption 2 rather than standard smoothness. Thus, our analysis can apply to a wider range of applications than existing methods.

5.2.1 Challenges and techniques for our analysis

In our analysis, we attempt to extend our Theorem 1 to the federated setting, but we find it is very difficult or even impossible to match the lower bound. The reasons are the following: a faster convergence rate requires a larger step size, but the inductive property requires a smaller step size. Thus, a contradiction arises. We balance the contradiction by addressing the following challenges.

Construct high-probability criteria. Starting from the smoothness of the function $F_S(x)$, we use our proposed "sum-aggregation" paradigm and $-\langle a,b\rangle = \frac{1}{2}\|a-b\|^2 - \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2$ to handle the tricky inner product term $\langle \nabla F_S(x_t), x_{t+1}-x_t\rangle$. It helps to produce the term $\frac{\gamma m K}{2}\|\nabla F_S(x_t)\|^2$ and construct martingale difference sequences in Lemma 9, which constructs the high-probability criteria and relaxes the restrictions on parameter selection in the induction.

Difficulty of the analysis in high probability. Many upper bounds in expectation are usually tighter and more concise than those of high probability. For example, there is the bound $\mathbb{E}_t\|\theta_t^a\|^2 \le 10\sigma^p\lambda^{2-p}$ but only the bound $\|\theta_t^a\|^2 \le 4\lambda^2$ in the centralized setting, where λ is usually of the order $\mathcal{O}(T^\alpha)$, $\alpha > 0$. A similar phenomenon also appears in federated learning, which makes the analysis difficult. Fortunately, we prove that our clipped batch gradient can provide one extra parameter of freedom to handle these rough upper bounds in Lemma 1.

Weaker assumptions. In FL, the difficulties caused by heavy-tailed noise and (L_0, L_1)-smoothness were addressed by induction and our martingale difference sequence, just like in CL.

5.3 Generalization bound for our `FedCBG` algorithm

In FL, there is no work jointly considering the optimization and generalization. To answer Q3, we analyze the generalization upper bound for our `FedCBG` in Theorem 4.

Theorem 4. *We assume that Assumptions 1, 2 and 3 hold. We choose the same parameter setting as in Theorem 3. If we choose* $T = \mathcal{O}\big((\frac{n}{d})^{\frac{3p-2}{2(7p-6)}}/m^{\frac{6p-5}{7p-6}}\big)$ *and* $K = \mathcal{O}\big((\frac{n}{d})^{\frac{3p-2}{2(6p-5)}}\big)$*, then with* *probability at least* $1-\delta$*, Algorithm 2 can achieve the generalization bound* $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 = \mathcal{O}\big((\frac{d}{n})^{\frac{p-1}{7p-6}}\log(\frac{1}{\delta}(\frac{n}{d})^{\frac{3p-2}{2(7p-6)}})\big)$.

To the best of our knowledge, this generalization bound is the first upper bound for FL with heavy-tailed noise. If we set a small global learning rate γ to obtain better generalization, the convergence speed will be slower, which reflects the trade-off between optimization and generalization.

6 Experiments

In this section, we evaluate our `FedCBG` algorithm against only FL algorithms FAT-clipping-PR (PR) and FAT-clipping-PI (PI) [49] that can theoretically handle heavy-tailed noise. We also compare the well-known FedAvg algorithm [33]. We test these methods on the CIFAR-10, CIFAR-100 [21] and Shakespeare [43] datasets. By the way, our goal is to compare the relative performance of `FedCBG` and baselines and larger models can achieve better performance on these datasets. All the experiments were performed on the GeForce RTX 2080Ti platform with the PyTorch framework.

Training an LSTM only satisfying the weaker Assumptions 2 and 3. Firstly, to verify that the federated scenarios may only meet weaker assumptions (i.e., Assumptions 2 and 3), we evaluate the smoothness and gradient noise distribution of a stacked LSTM as in [28] training on the Shakespeare dataset. We show smoothness and the histograms of gradient noise probability density for two randomly selected clients i in Fig. 2. More results are shown in the Appendix. It can be seen that local smoothness $\frac{\|\nabla f_i(x_t)-\nabla f_i(x_{t-1})\|}{\|x_t-x_{t-1}\|}$ positively correlates with gradient norm $\|\nabla f_i(x_t)\|$ instead of a constant, which satisfies weaker Assumption 2. Besides, the gradient noise meets heavy-tailed distribution, i.e., Assumption 3, rather than light-tailed Gaussian distribution.

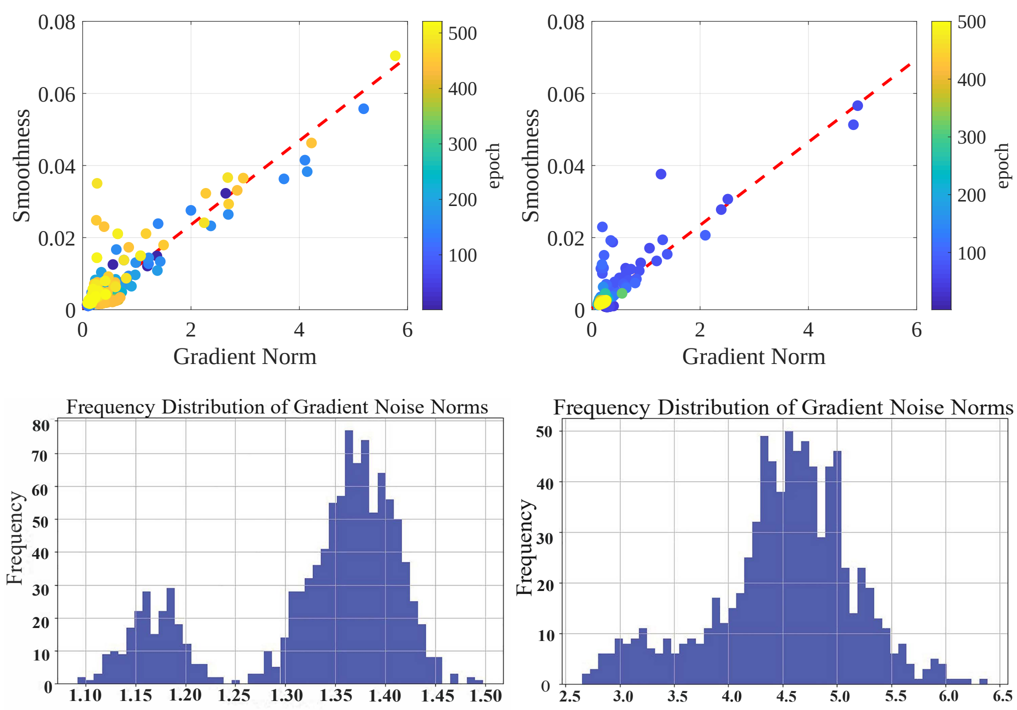

Figure 2: Gradient norm vs estimated Lipschitz smoothness (left) and distributions of the gradient noises (right) during training a stacked LSTM [28] on the Shakespeare dataset.

Hyperparameter selection. We conducted ablation experiments on hyperparameters γ, λ, b and K as shown in Fig. 3 and Fig. 6 in the Appendix. When global learning rate $\gamma = 0.2$ or 0.3 and $\lambda = 3.0$, our `FedCBG` algorithm performs better than other choices. The performance of $b = 100$ and $K = 10 \times n_i/b$ exceeds that of other values.

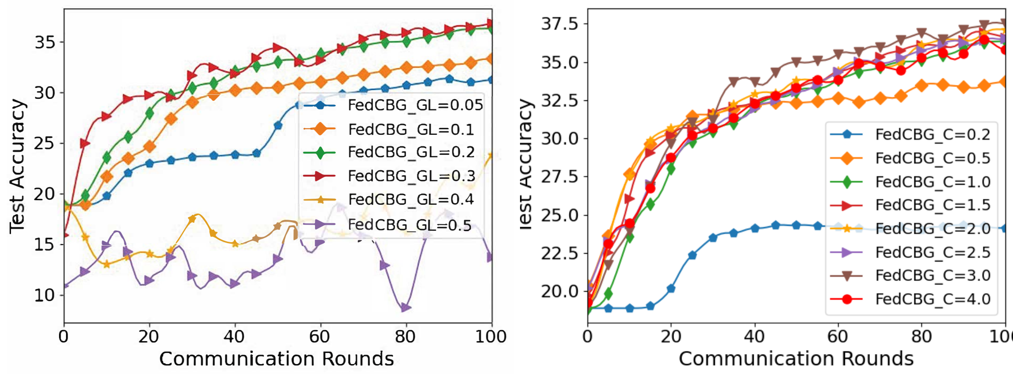

Figure 3: Test accuracy with different global learning (GL) rate (left) and clipping (C) parameter (right) on the Shakespeare dataset.

Table 3: Comparison of the training loss (TLoss.), testing classification accuracy (TAcc.) and the number of communication rounds (Round) to reach target test accuracy (84.5% for CIFAR-10, 45.0% for CIFAR-100 and 35.5% for Shakespeare datasets) in FL with heavy-tailed noise on various datasets.

Datasets	Evaluation	CIFAR-10	CIFAR-100	Shakespeare
PR	TLoss	0.16	3.25	3.17
	TAcc. (%)	83.1	42.2	34.8
	Round	282 (3.1×)	412 (1.9×)	219 (2.2×)
PI	TLoss	0.10	3.16	3.18
	TAcc. (%)	84.0	42.8	35.2
	Round	189 (2.1×)	346 (1.6×)	178 (1.8×)
FedAvg	TLoss	0.12	3.20	3.52
	TAcc. (%)	83.8	42.0	32.0
	Round	201 (2.3×)	409 (1.9×)	268(2.7×)
`FedCBG`	TLoss	**0.07**	**3.00**	**3.04**
	TAcc. (%)	**85.6**	**44.2**	**36.5**
	Round	**89**	**221**	**98**

Experimental details. Firstly, we choose $\eta = 1$, $\lambda = 3.0$, $\gamma = 0.3$, $K = n_i/b$ and $b = 100$ to train all the methods. Device distributions are non-IID. We use 100 randomly selected clients to train the model and the remaining 39 clients to test the model performance, which can quantify the performance gap on unseen client distributions. We report the average experimental results of 10 random initializations in Table 3. Secondly, we also conducted an experimental comparison based on the well-chosen hyperparameters and more results are shown in the Appendix. Table 3 shows that `FedCBG` can achieve 1.6-3.1 times gains over the competitors on all the tasks including vision and text models, which verifies the validity of our analysis: our `FedCBG` converges faster and performs better generalization ability than baselines on unseen client distributions.

7 Conclusions and Future Work

In this paper, we study clipped SGD for heavy-tailed noise. We prove a tighter optimization upper bound and the advanced generalization bound in high probability under weaker conditions. We extend our analysis to the federated setting based on our batch gradient variance bound and propose a `FedCBG` algorithm, which first achieves the high-probability convergence rate matching the lower bound and the first high-probability generalization bound. In future work, we will explore the role of recursive momentum [8] and minimax optimization [29, 30, 2] in heavy-tailed scenarios.

8 Acknowledgements

We want to thank the anonymous reviewers for their valuable suggestions and comments. This work was supported by the National Key Research and Development Program of China (No. 2023YFF0906204), National Natural Science Foundation of China (No. 62276182), Peng Cheng Lab Program (No. PCL2023A08), and the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG101/24).

References

[1] Ibrahim Alabdulmohsin. An information-theoretic route from generalization in expectation to generalization in probability. In *Artificial intelligence and statistics*, pages 92–100. PMLR, 2017.

[2] Weixin An, Yuanyuan Liu, Fanhua Shang, and Hongying Liu. Robust and faster zeroth-order minimax optimization: Complexity and applications. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] Idan Attias, Gintare Karolina Dziugaite, Mahdi Haghifam, Roi Livni, and Daniel M Roy. Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing. In *ICML*, 2024.

[4] Leighton Pate Barnes, Alex Dytso, and Harold Vincent Poor. Improved information-theoretic generalization bounds for distributed, federated, and iterative learning. *Entropy*, 24(9):1178, 2022.

[5] Aurélien Bellet, Marc Tommasi, Kevin Scaman, Giovanni Neglia, et al. Improved stability and generalization guarantees of the decentralized sgd algorithm. In *ICML*, 2024.

[6] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

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

[8] Ashok Cutkosky and Francesco Orabona. *Momentum-based variance reduction in non-convex SGD*. 2019.

[9] Xiaoge Deng, Tao Sun, Shengwei Li, and Dongsheng Li. Stability-based generalization analysis of the asynchronous decentralized sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7340–7348, 2023.

[10] Tao Fan, Hanlin Gu, et al. Ten challenging problems in federated foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[11] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 89–160. PMLR, 2023.

[12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

[13] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in neural information processing systems*, 31, 2018.

[14] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[15] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

[16] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pages 3964–3975. PMLR, 2021.

[17] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[18] Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

[19] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. In *The 28th International Conference on Artificial Intelligence and Statistics*.

[20] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR, 2024.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Yunwen Lei and Ke Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4505–4511, 2021.

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

[24] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

[25] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pages 12931–12963. PMLR, 2022.

[27] Shaojie Li and Yong Liu. High probability analysis for non-convex stochastic optimization with clipping. In *ECAI 2023*, pages 1406–1413. IOS Press, 2023.

[28] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[29] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6083–6093. PMLR, 13–18 Jul 2020.

[30] Yuanyuan Liu, Fanhua Shang, Weixin An, Junhao Liu, Hongying Liu, and Zhouchen Lin. A single-loop accelerated extra-gradient difference algorithm with improved complexity bounds for constrained minimax optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=wIlmx4bHrO`.

[31] Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. *arXiv preprint arXiv:2412.19529*, 2024.

[32] Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2266–2290. PMLR, 2023.

[33] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[34] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[35] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36, 2024.

[36] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. In *International Conference on Learning Representations*, 2018.

[37] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

[38] Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? *Advances in Neural Information Processing Systems*, 35:38190–38203, 2022.

[39] Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864. PMLR, 2024.

[40] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3099–3107, 2023.

[41] Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 19986–19994, 2025.

[42] Chao Ren, Han Yu, et al. Advances and open challenges in federated foundation models. *IEEE Communications Surveys and Tutorials*, 2025.

[43] William Shakespeare. *The complete works of William Shakespeare*. Wordsworth Editions, 2007.

[44] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[45] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024.

[46] Youming Tao, Sijia Cui, Wenlu Xu, Haofei Yin, Dongxiao Yu, Weifa Liang, and Xiuzhen Cheng. Byzantine-resilient federated learning at edge. *IEEE Transactions on Computers*, 72(9): 2600–2614, 2023.

[47] Chenhao Wang, Zihan Chen, Nikolaos Pappas, Howard H. Yang, Tony Q. S. Quek, and H. Vincent Poor. Adaptive federated learning over the air. *IEEE Transactions on Signal Processing*, 73:3187–3202, 2025.

[48] Chenghan Xie, Chenxi Li, Chuwen Zhang, Qi Deng, Dongdong Ge, and Yinyu Ye. Trust region methods for nonconvex stochastic optimization beyond lipschitz smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16049–16057, 2024.

[49] Haibo Yang, Peiwen Qiu, and Jia Liu. Taming fat-tailed ("heavier-tailed" with potentially infinite variance) noise in federated learning. *Advances in Neural Information Processing Systems*, 35:17017–17029, 2022.

[50] Qiang Yang, Lixin Fan, and Han Yu. *Federated Learning: Privacy and Incentive*. Springer, Cham, 2020.

[51] Honglin Yuan, Warren Richard Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=VimqQq-i_Q`.

[52] Dun Zeng, Zheshun Wu, Shiyu Liu, Yu Pan, Xiaoying Tang, and Zenglin Xu. Understanding generalization of federated learning: the trade-off between model stability and optimization. *arXiv preprint arXiv:2411.16303*, 2024.

[53] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

[54] Hao Zhang, Chenglin Li, Nuowen Kan, Ziyang Zheng, Wenrui Dai, Junni Zou, and Hongkai Xiong. Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach. In *NeurIPS*, 2024.

[55] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.

[56] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[58] Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *ICML*, 2022.

Appendix

In Section A, we provide more related works in detail and useful lemmas. In Section B, we provide the proof details for Clipped SGD about high-probability optimization and generalization bounds. In Section C, we provide the proof details for the `FedCBG` algorithm about high-probability optimization and generalization bounds. In section D, we provide more experimental results.

A Some related works and useful lemmas

A.1 Optimization properties for clipped methods

In the centralized setting, studies like [56, 31] focused on the clipping technology solving the heavy-tailed scenario and have successfully analyzed convergence bounds in expectation. Besides, works such as [26, 27, 37, 32] provide high-probability analysis for the convergence rates of clipped methods, which matches the lower bound in expectation up to a logarithmic factor. However, the order of $\log \frac{T}{\delta}$ is at least 2 in existing high-probability analysis as shown in Table 1. Our analysis reduces this order to $\frac{2p-2}{2p-1}$ as shown in the same table.

In federated learning, to the best of our knowledge, there is only one work [49] analyzing the first convergence rate result under the heavy-tailed noise assumption. However, their analysis focuses on convergence in expectation rather than high probability, and optimization properties rather than generalization aspects.

A.2 Generalization for nonconvex problems

The model generalization ability is always a concern in machine learning. Existing generalization analysis contains three types: in expectation [5, 17, 23, 38, 45, 52], in high probability [12, 34, 13, 22, 18], and based on information theory [1, 4, 54]. For example, Hardt et al. [17] pioneered generalization analysis in expectation based on algorithm stability. This type of method has been successfully used for asynchronous decentralized SGD [9], SGDA [24], and FedAvg and SCAFFOLD [45]. However, their analysis requires a very small step size to enjoy good stability, which leads to an exponential number of iterations for good convergence performance. As a comparison, the high-probability analysis can allow the constant step size, which will control the generalization error and make the optimization error decay faster. Besides, it can provide a stronger guarantee for each single run and is a tighter criterion for the case of bounded losses [1]. As for the information-theoretic analysis, they are usually algorithm-independent [3]. In this paper, we focus on the joint perspective of optimization and generalization and provide the high-probability upper bounds for both.

Lemma 4 (Freedman's inequality [14])**.** *Let $\{X_t\}_{t\geq 1}$ be a martingale difference sequence. Assume that there exists a constant $c > 0$ such that $|X_t| \leq c$ almost surely for all $t \geq 1$ and define $\sigma_t^2 = \mathbb{E}[X_t^2 | X_{t-1}, \cdots, X_1]$. Then, for all $b > 0$, $M > 0$ and $T \geq 1$,*

$$\Pr\left[\left|\sum_{t=1}^{T} X_t\right| > b \text{ and } \sum_{t=1}^{T} \sigma_t^2 \leq M\right] \leq 2\exp\left(-\frac{b^2}{2M + 2cb/3}\right). \tag{5}$$

Lemma 5 (Rademacher chaos complexities, generalized version of Corollary 2 in [22])**.** *Let $\delta \in (0,1)$, $R > 0$ and $S = \{z_1, z_2, \cdots, z_n\}$ be examples drawn independently form probability distribution P. Suppose Assumption 2 holds, with the probability at least $1-\delta$, we have*

$$\begin{aligned}
&\sup_{x\in B_R} \|\nabla F(x) - \nabla F_S(x)\| \\
&\leq ((L_0 + L_1\widetilde{b})R + \widetilde{b})n^{-\frac{1}{2}} \times \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log\frac{1}{\delta}}\right) \\
&= (L_0 R + (L_1 R + 1)\widetilde{b})n^{-\frac{1}{2}} \times \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log\frac{1}{\delta}}\right),
\end{aligned} \tag{6}$$

where $\widetilde{b} = \sup_{\xi} \|\nabla f(0,\xi)\|$ and $B_R = \{x \in \mathbb{R}^d : \|x\| \leq R\}$.

This lemma is a generalized version of Corollary 2 in [22] and can be used for the (L_0, L_1)-smoothness condition.

Table 4: Comparison of existing analysis of clipped methods for solving nonconvex problems in centralized learning (CL). We use $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_S(x_t)\|^2$ and $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\|^2$ as the criterion in expectation and use $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2$ and $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2$ as the criterion in high probability. Abbreviation: Standard smoothness (SS), (L_0, L_1)-smoothness ((L_0, L_1)), Heavy-tailed (HT), Light-tailed (LT), High-probability (HP), In expectation (Exp). Here, G is a constant, $\delta \in (0, 1)$, and $p \in (1, 2]$.

	Assumptions		Criteria	Additional Assumptions	Bounds	
	Smooth.	Noise			Optimization	Generalization
[53]	(L_0, L_1)	LT	Exp	$\|\nabla f(x;\xi) - \nabla F_S(x)\| \leq \sigma$	$\mathcal{O}(1/T^{\frac{1}{2}})$	——
[56]	(L_0, L_1)	HT	Exp	$\mathbb{E}_t[\|\nabla f(x_t;\xi_{j_t})\|^p] \leq G^p$	$\mathcal{O}(T^{\frac{2-2p}{3p-2}})$	——
[31]	(L_0, L_1)	HT	Exp	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}})$	——
[26]	SS	HT[1]	HP	$\eta\|\nabla F_S(x_t)\| \leq G$	$\mathcal{O}(\frac{\log T}{T^{\frac{1}{2}}}\log^2\frac{1}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$
[27]	SS	HT	HP	$\mathbb{E}_t[\|\nabla f(x_t;\xi_{j_t})\|^p] \leq G^p$	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log\frac{1}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$
[37]	SS	HT	HP	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{p}{p-1}}\frac{T}{\delta})$	——
[32]	SS[2]	HT	HP	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^2\frac{T}{\delta})$	——
Th. 1, 2	(L_0, L_1)	HT	HP	——	$\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta})$	$\mathcal{O}((\frac{d}{n})^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}(\sqrt{\frac{n}{\delta^2 d}}))$

[1] Li & Liu (2022) analyzed a special heavy-tailed distribution, i.e., sub-Weibull distribution. They obtain the optimal convergence rate but under assumption $\eta\|\nabla F_S(x_t)\| \leq G$, which is orthogonal to our analysis.

[2] Under the standard smoothness condition, the work [32] achieves the convergence rate $\mathcal{O}(T^{(2-2p)/(3p-2)}\log^2\frac{T}{\delta})$ with the probability at least $1-\delta$. Besides, under the smoothness condition of individual loss function, i.e., $\|\nabla f(x;\xi_i) - \nabla f(y;\xi_i)\| \leq L_i\|x-y\|$, $\forall i \in [n]$, it breaks the lower bound and achieves the convergence rate of $\mathcal{O}(T^{(2-2p)/(2p-1)}\log^2\frac{T}{\delta})$.

Table 5: Comparison of the existing analysis for nonconvex problems in federated learning (FL). "AA" indicates whether the additional gradient boundedness assumption $\mathbb{E}_t[\|\nabla f_i(x_t;\xi_{j_t})\|^p] \leq G^p$ are required, and "LB" refers to the lower bound.

Algorithms	Assumptions		Criteria	AA	Bounds	
	Smooth.	Noise			Optimization	Generalization
[49][1]	SS	HT	Exp	✓	$\mathcal{O}((mT)^{\frac{2-2p}{3p-2}}K^{\frac{4-2p}{3p-2}})$	——
[49]	SS	HT	Exp	✓	$\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}})$	——
LB	SS	HT	Exp	✗	$\Omega((mKT)^{\frac{2-2p}{3p-2}})$	——
Theorems 3, 4	(L_0, L_1)	HT	HP	✗	$\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{1}{\delta})$	$\mathcal{O}(\frac{(\frac{d}{n})^{\frac{2p-2}{10p-5}}\log(\frac{1}{\delta}(\frac{n}{d})^{\frac{1}{8}})}{m^{\frac{3-8p}{3p-2}}})$

[1] Yang et al. [49] proposed the FAT-Clipping-PR and FAT-Clipping-PI algorithms, where the former can achieve the convergence rate of $\mathcal{O}((mT)^{\frac{2-2p}{3p-2}}K^{\frac{4-2p}{3p-2}})$ and the latter can improve it to $\mathcal{O}((mKT)^{\frac{2-2p}{3p-2}})$, which matches their proposed lower bound.

B Centralized Setting

B.1 High Probability Optimization Upper Bound

Lemma 6. *For $1 \leq N \leq T+1$, let E_N be the event that for all $l = 1, \cdots, N$,*

$$\Delta_l + \frac{\eta}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^2 \leq \Delta_1 + (L_0\eta^2 - \eta)\sum_{t=0}^{l-1}\langle\nabla F_S(x_t), \theta_t^a\rangle + \sum_{t=0}^{l-1}(L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle\nabla F_S(x_t), \theta_t^a\rangle$$
$$+ \ell\eta^2\sum_{t=0}^{l-1}(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2]) + 2\eta\sum_{t=0}^{l-1}\|\theta_t^b\|^2 + \ell\eta^2\sum_{t=0}^{l-1}\mathbb{E}_t[\|\theta_t^a\|^2] + \frac{L_1\eta^2}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^3 \leq 2\Delta_1. \tag{7}$$

Then E_N happens with probability at least $1 - \frac{(N-1)\delta}{T}$ for each $N \in [T+1]$.

Proof. We choose $\lambda = \mathcal{O}(T^{\frac{1}{3p-2}} \log^{\frac{1}{1-2p}} \frac{T}{\delta})$ and $\eta = \mathcal{O}(T^{\frac{-p}{3p-2}} \log^{\frac{2-2p}{2p-1}} \frac{T}{\delta})$ such that we have $\|x_{t+1} - x_t\| \le \eta\lambda \le \frac{1}{L_1}$ for any $t \le T$. By the smoothness of function f, we have

$$\begin{aligned}
&F_S(x_{t+1}) - F_S(x_t) \\
&\le \langle \nabla F_S(x_t), x_{t+1} - x_t\rangle + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\
&= -\eta\langle \nabla F_S(x_t), \widetilde{\nabla} f(x_t;\xi_{j_t})\rangle + \frac{\ell\eta^2}{2}\|\widetilde{\nabla} f(x_t;\xi_{j_t})\| \\
&= -\eta\langle \nabla F_S(x_t), \theta_t + \nabla F_S(x_t)\rangle + \frac{\ell\eta^2}{2}\|\theta_t + \nabla F_S(x_t)\|^2 \\
&= -\eta\|\nabla F_S(x_t)\|^2 - \eta\langle \nabla F_S(x_t), \theta_t\rangle + \frac{\ell\eta^2}{2}\|\theta_t\|^2 + \frac{\ell\eta^2}{2}\|\nabla F_S(x_t)\|^2 + \ell\eta^2\langle \nabla F_S(x_t), \theta_t\rangle \\
&= -\left(\eta - \frac{L_0\eta^2}{2}\right)\|\nabla F_S(x_t)\|^2 + \frac{\ell\eta^2}{2}\|\theta_t\|^2 + (\ell\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t\rangle + \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 \\
&= -\left(\eta - \frac{L_0\eta^2}{2}\right)\|\nabla F_S(x_t)\|^2 + \frac{\ell\eta^2}{2}\|\theta_t\|^2 + (L_0\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t^a + \theta_t^b\rangle \\
&\quad + \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a + \theta_t^b\rangle
\end{aligned} \tag{8}$$

where $\ell = L_0 + L_1\|\nabla F_S(x_t)\|$, $\theta_t = \widetilde{\nabla} f(x_t;\xi_{j_t}) - \mathbb{E}_t[\widetilde{\nabla} f(x_t;\xi_{j_t})] + \mathbb{E}_t[\widetilde{\nabla} f(x_t;\xi_{j_t})] - \nabla F_S(x_t)$, $\theta_t = \theta_t^a + \theta_t^b$.

Using Cauchy-Schwarz, we have $\langle \nabla F_S(x_t), \theta_t^b\rangle \le \frac{1}{2}\|\nabla F_S(x_t)\|^2 + \frac{1}{2}\|\theta_t^b\|^2$. Thus, we have

$$\begin{aligned}
&\Delta_{t+1} - \Delta_t \\
\le& -\left(\eta - \frac{L_0\eta^2}{2}\right)\|\nabla F_S(x_t)\|^2 + \frac{\ell\eta^2}{2}\|\theta_t\|^2 + (L_0\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + \frac{\eta - L_0\eta^2}{2}\|\nabla F_S(x_t)\|^2 \\
&+ \frac{\eta - L_0\eta^2}{2}\|\theta_t^b\|^2 + \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a + \theta_t^b\rangle
\end{aligned} \tag{9}$$

where we use the setting $\eta \le \frac{1}{L_0}$. Thus, we have

$$\begin{aligned}
&\Delta_{t+1} - \Delta_t \\
\le& -\frac{\eta}{2}\|\nabla F_S(x_t)\|^2 + \ell\eta^2\|\theta_t^a\|^2 + (L_0\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + \frac{\eta + L_0\eta^2}{2}\|\theta_t^b\|^2 \\
&+ L_1\eta^2\|\nabla F_S(x_t)\|\|\theta_t^b\|^2 + \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a + \theta_t^b\rangle \\
\le& -\frac{\eta}{2}\|\nabla F_S(x_t)\|^2 + \ell\eta^2\|\theta_t^a\|^2 + (L_0\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + \eta\|\theta_t^b\|^2 + L_1\eta^2\|\nabla F_S(x_t)\|\|\theta_t^b\|^2 \\
&+ \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^b\rangle,
\end{aligned} \tag{10}$$

where the last inequality holds due $\eta \le \frac{1}{L_0}$. Rearranging, adding and subtracting $\mathbb{E}[\|\theta_t^a\|^2|\mathcal{F}_{t-1}]$ (to construct a martingale difference sequence), we have

$$\begin{aligned}
\Delta_{t+1} - \Delta_t \le& -\frac{\eta}{2}\|\nabla F_S(x_t)\|^2 + \ell\eta^2\left(\|\theta_t^a\|^2 - \mathbb{E}[\|\theta_t^a\|^2|\mathcal{F}_{t-1}]\right) + (L_0\eta^2 - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle \\
&+ \eta\|\theta_t^b\|^2 + \ell\eta^2\mathbb{E}[\|\theta_t^a\|^2|\mathcal{F}_{t-1}] + L_1\eta^2\|\nabla F_S(x_t)\|\|\theta_t^b\|^2 + \frac{L_1\eta^2}{2}\|\nabla F_S(x_t)\|^3 \\
&+ (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + (L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^b\rangle.
\end{aligned} \tag{11}$$

The term $\ell\eta^2\left(\|\theta_t^a\|^2 - \mathbb{E}[\|\theta_t^a\|^2|\mathcal{F}_{t-1}]\right)$ is a martingale difference sequence. We will show by induction that the probability of event E_N occurring is at least $1 - \frac{(N-1)\delta}{T}$.

For $N = 1$, the event happens with probability 1. Suppose that for some $N \le T$, $\Pr[E_N] \ge 1 - \frac{(N-1)\delta}{T}$. We will prove that $\Pr[E_{N+1}] \ge 1 - \frac{N\delta}{T}$.

For $l \leq N$, under the event E_N, the following inequality

$$
\begin{aligned}
&\Delta_l + \frac{\eta}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^2 \\
&\leq \Delta_1 + (L_0\eta^2 - \eta)\sum_{t=0}^{l-1}\langle \nabla F_S(x_t), \theta_t^a\rangle + \ell\eta^2\sum_{t=0}^{l-1}(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2]) \\
&\quad + 2\eta\sum_{t=0}^{l-1}\|\theta_t^b\|^2 + \ell\eta^2\sum_{t=0}^{l-1}\mathbb{E}_t[\|\theta_t^a\|^2] + \frac{L_1\eta^2}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^3 + \sum_{t=0}^{l-1}(L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle
\end{aligned} \tag{12}
$$

holds with the probability at least $1 - \frac{(N-1)\delta}{T}$. From the induction hypothesis, we have $\Delta_l \leq 2\Delta_1$, $\forall l \leq N$.

By the smoothness of f and f is bounded below [32], we have

$$
\|\nabla F_S(x_t)\| \leq \sqrt{2\ell\Delta_t} = \sqrt{2(L_0 + L_1\|\nabla F_S(x_t)\|)\Delta_t}. \tag{13}
$$

Let $y = x - \frac{\nabla F_S(x)}{L_0 + L_1\|\nabla F_S(x)\|}$, we have $\|x - y\| \leq \frac{1}{L_1}$. Thus, we have

$$
\begin{aligned}
F_S(y) &\leq F_S(x) + \langle \nabla F_S(x), y - x\rangle + \frac{L_0 + L_1\|\nabla F_S(x)\|}{2}\|x - y\|^2 \\
&= F_S(x) - \frac{\|\nabla F_S(x)\|^2}{2(L_0 + L_1\|\nabla F_S(x)\|)} \\
\Rightarrow \|\nabla F_S(x)\|^2 &\leq 2(L_0 + L_1\|\nabla F_S(x)\|)(F_S(x) - F_S(y)) \\
&\leq 2(L_0 + L_1\|\nabla F_S(x)\|)\Delta_1.
\end{aligned} \tag{14}
$$

Thus, there is

$$
\|\nabla F_S(x_t)\| \leq \frac{2\Delta_t L_1 + \sqrt{4(\Delta_t)^2 L_1^2 + 8L_0\Delta_t}}{2} \leq \frac{\lambda}{2}, \tag{15}
$$

where we use the $\lambda \geq 4\Delta_1 L_1 + 4\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1}$, and under the event E_N, we have $\Delta_t \leq 2\Delta_1$ $\forall t \leq N$. We define $R := \frac{1}{2}(L_0 + (4L_1 + 1)\Delta_1)$, (due to $\Delta_1 L_1 + \sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1} \leq 2\Delta_1 L_1 + \frac{\Delta_1}{2} + \frac{L_0}{2} \leq \frac{1}{2}(L_0 + (4L_1 + 1)\Delta_1)$). Thus, $R \geq \frac{\Delta_1}{2}$.

Besides, according to [32], we can use its **Lemma 5** and we have

$$
\|\theta_t^b\| \leq 2\sigma^p\lambda^{1-p}, \mathbb{E}_t[\|\theta_t^a\|^2] \leq 10\sigma^p\lambda^{2-p}. \tag{16}
$$

Now, we can analyze the E_{N+1}. By summing over t from 0 to N regarding inequality (11) and applying the Cauchy-Schwarz and Young's inequalities we have

$$
\begin{aligned}
&\Delta_{N+1} + \frac{\eta}{2}\sum_{t=0}^{N}\|\nabla F_S(x_t)\|^2 \\
&\leq \Delta_1 + (L_0\eta^2 - \eta)\sum_{t=0}^{N}\langle \nabla F_S(x_t), \theta_t^a\rangle + \sum_{t=0}^{N}(L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle + \ell\eta^2\sum_{t=0}^{N}(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2]) \\
&\quad + \eta\sum_{t=0}^{N}\|\theta_t^b\|^2 + L_1\eta^2\sum_{t=0}^{N}\|\nabla F_S(x_t)\|\|\theta_t^b\|^2 + \sum_{t=0}^{N}\frac{|\eta - L_1\eta^2\|\nabla F_S(x_t)\||}{4}\|\nabla F_S(x_t)\|^2 \\
&\quad + |\eta - L_1\eta^2\|\nabla F_S(x_t)\||\|\theta_t^b\|^2 + \ell\eta^2\sum_{t=0}^{N}\mathbb{E}_t[\|\theta_t^a\|^2] + \frac{L_1\eta^2}{2}\sum_{t=0}^{N}\|\nabla F_S(x_t)\|^3.
\end{aligned} \tag{17}
$$

We set $\eta \le \frac{1}{2L_1 R}$ and we have $\eta - L_1\eta^2\|\nabla F_S(x_t)\| \ge 0$. Thus

$$\Delta_{N+1} + \frac{\eta}{4}\sum_{t=0}^{N}\|\nabla F_S(x_t)\|^2 \le \Delta_1 + \underbrace{(L_0\eta^2-\eta)\sum_{t=0}^{N}\langle \nabla F_S(x_t), \theta_t^a\rangle + \sum_{t=0}^{N}(L_1\eta^2\|\nabla F_S(x_t)\| - \eta)\langle \nabla F_S(x_t), \theta_t^a\rangle}_{A}$$
$$+ \underbrace{\ell\eta^2\sum_{t=0}^{N}(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2])}_{B} + \underbrace{2\eta\sum_{t=0}^{N}\|\theta_t^b\|^2}_{C} + \underbrace{\ell\eta^2\sum_{t=0}^{N}\mathbb{E}_t[\|\theta_t^a\|^2]}_{D} + \underbrace{\frac{L_1\eta^2}{4}\sum_{t=0}^{N}\|\nabla F_S(x_t)\|^3}_{H}. \tag{18}$$

Thus, we have $E_{N+1} = E_A \cap E_B \cap E_C \cap E_D \cap E_H$.

Upper bound for C

When the event E_N happens,

$$C = 2\eta\sum_{t=0}^{N}\|\theta_t^b\|^2 \le 8\eta\sum_{t=0}^{N}\sigma^{2p}\lambda^{2-2p} = 8\sigma^{2p}\lambda^{2-2p}\eta N \le \frac{\Delta_1}{256}, \tag{19}$$

where we choose $\eta = \frac{\Delta_1 T^{\frac{1-p}{3p-2}}}{32(L_0+2(L_1+1)R)\lambda\max\{\log\frac{4T}{\delta},1\}}$, and the last inequality holds due to $\lambda \ge \frac{64^{\frac{1}{2p-1}}\sigma^{\frac{2p}{2p-1}}T^{\frac{1}{3p-2}}}{(\mathcal{R}\rho)^{\frac{1}{2p-1}}}$, where $\mathcal{R} = L_0 + 2(L_1+1)R$.

Upper bound for D

When the event E_N happens,

$$D = \ell\eta^2\sum_{t=1}^{N}\mathbb{E}_t[\|\theta_t^a\|^2] \le 10\sigma^p\lambda^{2-p}L_0\eta^2 N + 20R\sigma^p\lambda^{2-p}L_1\eta^2 N \le 10(L_0+2L_1R)\left(\frac{\sigma}{\lambda}\right)^p\lambda^2\eta^2 T \overset{(b)}{\le} \frac{\Delta_1}{256}, \tag{20}$$

where in the inequality (b) we use $\lambda \ge \frac{(3\Delta_1)^{\frac{1}{p}}\sigma T^{\frac{1}{3p-2}}}{(\mathcal{R}\rho^2)^{\frac{1}{p}}}$.

Upper bound for H

Under the event E_N, we have

$$H = \frac{L_1\eta^2}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^3 \le 2L_1\eta^2 R^3 T \le \frac{\Delta_1}{32}, \tag{21}$$

where we use the setting $\eta \le \mathcal{O}(T^{\frac{-p}{3p-2}}) \le \frac{\sqrt{\Delta_1}}{\sqrt{4L_1R^3T}}$. This condition is naturally satisfied if we choose the step size in Theorem 1.

Upper bound for A

To bound A and B, we use the Freedman's inequality.

To construct the martingale difference sequence, we define the following random variables

$$X_t = \begin{cases} -\nabla F_S(x) & \text{if } \Delta_t \le 2\Delta_1 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Thus $\|X_t\| \le \|\nabla F_S(x_t)\| \le 2\Delta_1 L_1 + 2\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1}$ for all t.

Instead of bounding A, we bound $A' = (2\eta - L_0\eta^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\sum_{t=1}^{N}\langle X_t, \theta_t^a\rangle$. We check the conditions to apply Freedman's inequality. First $\mathbb{E}_t[(2\eta_t - L_0\eta_t^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle] = 0$. Furthermore, with probability 1, $\|\theta_t^a\| \le 2\lambda$, and $\|X_t\| \le 2\Delta_1 L_1 + 2\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1}$, thus $|(2\eta - L_0\eta^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle| \le (2\eta - L_0\eta^2)\|X_t\|\|\theta_t^a\| + L_1\eta^2\|\nabla F_S(x_t)\|\|X_t\|\|\theta_t^a\| \le (4\lambda\Delta_1 L_1 + 4\lambda\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1})(2\eta - L_0\eta^2 + L_1\eta^2\|\nabla F_S(x_t)\|) \le (\eta + \frac{L_1\eta^2\lambda}{2})4\lambda(\Delta_1 L_1 +

$\sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1})$. Hence, $\{(2\eta - L_0\eta^2 - \eta^2 L_1\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle\}$ is a bounded martingale difference sequence. Therefore, for constant a and M to be chosen we have

$$\begin{aligned}
&\Pr\left[\left|\sum_{t=1}^{N}(2\eta - L_0\eta^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle\right| > a \text{ and}\right.\\
&\left.\sum_{t=1}^{N}\mathbb{E}_t\left[((2\eta - L_0\eta^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle)^2\right] \le M\ln\frac{4T}{\delta}\right]\\
&\le 2\exp\left(-\frac{a^2}{2M\ln\frac{4T}{\delta} + \frac{8}{3}(\Delta_1 L_1 + \sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1})(\eta + \frac{L_1\eta^2\lambda}{2})\lambda a}\right).
\end{aligned}$$

We choose a such that

$$2\exp\left(-\frac{a^2}{2M\ln\frac{4T}{\delta} + \frac{8}{3}(\Delta_1 L_1 + \sqrt{\Delta_1^2 L_1^2 + L_0\Delta_1})(\eta + \frac{L_1\eta^2\lambda}{2})\lambda a}\right) = \frac{\delta}{2T}, \tag{23}$$

which gives

$$a = \left(\frac{4}{3}R(1 + \frac{L_1\eta\lambda}{2})\eta\lambda + \sqrt{\frac{16R^2(1+\frac{L_1\eta\lambda}{2})^2\eta^2\lambda^2}{9} + 2M}\right)\ln\frac{4T}{\delta}. \tag{24}$$

If we choose $M = 40R^2\sigma^p\lambda^{2-p}\eta^2(1+\frac{L_1\eta\lambda}{2})^2 T$ and set $\eta = \frac{\Delta_1 T^{\frac{1-p}{3p-2}}}{32(L_0+2(L_1+1)R)\lambda\max\{\log\frac{4T}{\delta},1\}}$ and $\lambda \ge 40^{\frac{1}{p}}T^{\frac{1}{3p-2}}\rho^{-\frac{2}{p}}\sigma(1+\frac{L_1\Delta_1}{32\mathcal{R}})^{\frac{2}{p}}$, we can obtain that $a \le \frac{\Delta_1}{16}$. Thus, we have

$$E_A = \left\{\text{either } |A'| \le \frac{\Delta_1}{16} \text{ or } \sum_{t=1}^{N}\mathbb{E}_t\left[((\eta - L_0\eta^2 - \frac{1}{2}L_1\eta^2\lambda)\langle X_t, \theta_t^a\rangle)^2\right] \ge M\ln\frac{4T}{\delta}\right\} \tag{25}$$

holds with probability at least $1 - \frac{\delta}{2T}$.

Also notice that under the event E_N, we have

$$\begin{aligned}
&\sum_{t=1}^{N}\mathbb{E}_t\left[((2\eta - L_0\eta^2 - L_1\eta^2\|\nabla F_S(x_t)\|)\langle X_t, \theta_t^a\rangle)^2\right]\\
&\le \eta^2(1+\frac{L_1\eta\lambda}{2})^2\sum_{t=1}^{N}\mathbb{E}_t\left[\|X_t\|^2\|\theta_t^a\|^2\right] \le \eta^2 R^2(1+\frac{L_1\eta\lambda}{2})^2\sum_{t=1}^{N}\mathbb{E}_t[\|\theta_t^a\|^2]\\
&\le 40R^2\sigma^p\lambda^{2-p}\eta^2(1+\frac{L_1\eta\lambda}{2})^2 N \le 40R^2 T\left(\frac{\sigma}{\lambda}\right)^p\lambda^2\eta^2(1+\frac{L_1\eta\lambda}{2})^2 = M \le M\ln\frac{4T}{\delta}.
\end{aligned} \tag{26}$$

Under E_N, we have that $X_t = -\nabla F_S(x_t)$ for all $t \le N$. Thus, when $E_N \cap E_A$ happens, i.e., with the probability $(1 - \frac{(N-1)\delta}{T})(1 - \frac{\delta}{2T})$, we have $A = A' \le \frac{\Delta_1}{16}$.

Upper bound for B

We check the conditions to apply Freedman's inequality. First, $\mathbb{E}_t\left[\ell\eta^2(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2])\right] = 0$. Further, with probability 1, $\|\theta_t^a\| \le 2\lambda$, thus $|\ell\eta^2(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2])| \le 4(L_0 + 2L_1R)\eta^2\lambda^2$. Thus, the sequence $\{\ell\eta^2(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2])\}$ is a bounded martingale difference sequence. Applying Freedman's inequality for constants c and G to be chosen, we have

$$\begin{aligned}
&\Pr\left[\left|\sum_{t=1}^{N}\ell\eta^2\left(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2]\right)\right| > c \text{ and } \sum_{t=1}^{N}\mathbb{E}_t\left[\left(\ell\eta^2(\|\theta_t^a\|^2 - \mathbb{E}_t[\|\theta_t^a\|^2])\right)^2\right] \le G\ln\frac{4T}{\delta}\right]\\
&\le 2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta} + \frac{8}{3}(L_0 + 2L_1R)\lambda^2\eta^2 c}\right)
\end{aligned} \tag{27}$$

We choose c such that

$$2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta}+\frac{8}{3}(L_0+2L_1R)\lambda^2\eta^2 c}\right)=\frac{\delta}{2T} \tag{28}$$

which gives

$$c=\left(\underbrace{\frac{4}{3}(L_0+2L_1R)\lambda^2\eta^2}_{\leq\frac{1}{64}\Delta_1}+\sqrt{\frac{16(L_0+2L_1R)^2\lambda^4\eta^4}{9}+\underbrace{2G}_{\leq\frac{1}{64^2}\Delta_1^2}}\right)\ln\frac{4T}{\delta} \tag{29}$$

In we choose $G=80(L_0+2L_1R)^2\sigma^p\lambda^{4-p}\eta^4 T$ and $\eta=\frac{\Delta_1 T^{\frac{1-p}{3p-2}}}{32(L_0+2(L_1+1)R)\lambda\max\{\ln\frac{4T}{\delta},1\}}$, we can obtain $c\leq\frac{14\Delta_1}{768}\leq\frac{2}{64}\Delta_1$. Thus, the following event happens

$$E_C=\left\{\text{either } C\leq|\ell\eta^2\sum_{t=1}^{N}(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^2\|^2])|\leq\frac{\Delta_1}{32}\text{ or }\sum_{t=1}^{N}\mathbb{E}_t\left[(\ell\eta^2(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^a\|^2]))^2\right]\geq G\ln\frac{4T}{\delta}\right\}. \tag{30}$$

Under the event E_N, we have

$$\begin{aligned}
&\sum_{t=1}^{N}\mathbb{E}_t\left[(\ell\eta^2(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^a\|^2]))^2\right]\\
&\leq 4(L_0+2L_1R)\lambda^2\eta^2\sum_{t=1}^{N}\mathbb{E}_t[|(L_0+2L_1R)\eta^2(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^a\|^2])|]\\
&\leq 8(L_0+2L_1R)^2\lambda^2\eta^4\sum_{t=1}^{N}\mathbb{E}_t[\|\theta_t^a\|^2]\\
&\leq 80(L_0+2L_1R)^2\sigma^p\lambda^{4-p}\eta^4 N\leq G\leq G\ln\frac{4T}{\delta}
\end{aligned} \tag{31}$$

Thus, when $E_N\cap E_C$ happens, we have $\Pr(C\leq\frac{\Delta_1}{32})\geq 1-\frac{\delta}{2T}$. □

Proof of Theorem 1.

Proof. In summary, we choose

$$\begin{aligned}
\lambda=&\ \max\{(\frac{64}{\mathcal{R}\rho})^{\frac{1}{2p-1}}\sigma^{\frac{2p}{2p-1}}T^{\frac{1}{3p-2}},\frac{(2\Delta_1)^{\frac{1}{p}}\sigma T^{\frac{1}{3p-2}}}{(\mathcal{R}\rho^2)^{\frac{1}{p}}},40^{\frac{1}{p}}T^{\frac{1}{3p-2}}\rho^{-\frac{2}{p}}\sigma(1+\frac{L_1\Delta_1}{32\mathcal{R}})^{\frac{2}{p}},4R\}\\
&=\mathcal{O}(T^{\frac{1}{3p-2}}(\log\frac{T}{\delta})^{\frac{1}{1-2p}}),\\
\eta=&\ \frac{\Delta_1 T^{\frac{1-p}{3p-2}}}{32\mathcal{R}\lambda\rho}=\frac{\Delta_1}{32\mathcal{R}\rho}\min\{\frac{T^{\frac{-p}{3p-2}}}{\sigma^{\frac{2p}{2p-1}}}(\frac{\mathcal{R}\rho}{64})^{1-p},\frac{T^{\frac{1-p}{3p-2}}}{4R},\frac{(\mathcal{R}\rho^2)^{\frac{1}{p}}T^{\frac{-p}{3p-2}}}{\sigma(2\Delta_1)^{\frac{1}{p}}},\frac{T^{\frac{-p}{3p-2}}}{\sigma(40\rho^{-2})^{\frac{1}{p}}(1+\frac{L_1\Delta_1}{32\mathcal{R}})^{\frac{2}{p}}}\}\\
&=\mathcal{O}(T^{\frac{-p}{3p-2}}(\log\frac{T}{\delta})^{\frac{2-2p}{2p-1}}),
\end{aligned} \tag{32}$$

Summing the descent lemma from 1 to T, we have

$$\sum_{t=1}^{T}\frac{\eta}{4}\|\nabla F_S(x_t)\|^2+\Delta_{T+1}$$
$$\leq \Delta_1+(L_0\eta^2-\eta)\sum_{t=1}^{T}\langle\nabla F_S(x_t),\theta_t^a\rangle+\ell\eta^2\sum_{t=1}^{T}(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^a\|^2])+2\eta\sum_{t=1}^{T}\|\theta_t^b\|^2$$
$$+\ell\eta^2\sum_{t=1}^{T}\mathbb{E}_t[\|\theta_t^a\|^2]+\frac{L_1\eta^2}{4}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^3+\sum_{t=1}^{T}(L_1\eta^2\|\nabla F_S(x_t)\|-\eta)\langle\nabla F_S(x_t),\theta_t^a\rangle\leq 2\Delta_1 \tag{33}$$

holding with the probability at least $\Pr(E_N\cap E_A\cap E_B\cap E_C\cap E_D\cap E_H)=(1-\frac{(N-1)\delta}{T})(1-\frac{\delta}{2T})\cdot 1\cdot 1\cdot 1\geq 1-\frac{N\delta}{T}$. Thus, the clipped SGD can achieve the convergence rate

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2\leq\mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta}) \tag{34}$$

with the probability at least $1-\delta$. □

B.2 High Probability Generalization Bound

Theorem 5. *Suppose Assumptions 1, 2 and 3 hold. Let $\{x_t\}$ be the sequence produced by Algorithm 1 with the step size η and clipping parameter λ set to be the same as in Theorem 1. If we choose $T=\mathcal{O}(\sqrt{\frac{n}{d}})$, then with probability at least $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2\leq\mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}\left(\frac{1}{\delta}\sqrt{\frac{n}{d}}\right)\right) \tag{35}$$

for any $\delta\in(0,1)$.

Proof. Since $x_{t+1}-x_t=\eta\widetilde{\nabla}f(x_t;\xi_{j_t})$, using $x_0=0$, we have

$$\|x_{t+1}\|$$
$$=\eta\|\sum_{l=1}^{t}\widetilde{\nabla}f(x_l;\xi_{j_l})\|=\eta\|\sum_{l=1}^{t}\widetilde{\nabla}f(x_l;\xi_{j_l})-\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})+\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})-\nabla F_S(x_l)+\nabla F_S(x_l)\|$$
$$\leq\underbrace{\eta\|\sum_{l=1}^{t}\widetilde{\nabla}f(x_l;\xi_{j_l})-\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})\|}_{A_1}+\underbrace{\eta\|\sum_{l=0}^{t}\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})-\nabla F_S(x_l)\|}_{A_2}+\underbrace{\eta\|\sum_{l=0}^{t}\nabla F_S(x_l)\|}_{A_3}. \tag{36}$$

Upper Bound for A_1

For A_1, it is clear that the sequence $\{\widetilde{\nabla}f(x_l;\xi_{j_l})-\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})\}$ is a martingale difference sequence. With probability 1, there is $|\widetilde{\nabla}f(x_l;\xi_{j_l})-\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})|\leq 2\lambda$. Thus, $A_1\leq\eta\sum_{l=0}^{t}\|\widetilde{\nabla}f(x_l;\xi_{j_l})-\mathbb{E}_l\widetilde{\nabla}f(x_l;\xi_{j_l})\|\leq 2\eta\lambda t$.

Upper Bound for A_2

According to Lemma 6, we know that

$$(L_0\eta^2-\eta)\sum_{t=0}^{l-1}\langle\nabla F_S(x_t),\theta_t^a\rangle+\sum_{t=0}^{l-1}(L_1\eta^2\|\nabla F_S(x_t)\|-\eta)\langle\nabla F_S(x_t),\theta_t^a\rangle$$
$$+\ell\eta^2\sum_{t=0}^{l-1}(\|\theta_t^a\|^2-\mathbb{E}_t[\|\theta_t^a\|^2])+2\eta\sum_{t=0}^{l-1}\|\theta_t^b\|^2+\ell\eta^2\sum_{t=0}^{l-1}\mathbb{E}_t[\|\theta_t^a\|^2]+\frac{L_1\eta^2}{4}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^3\leq\Delta_1 \tag{37}$$

holds with the probability of $1-\frac{(t-1)\delta}{T}$. Furthermore, we know that $\Delta_t \leq 2\Delta_1$ holds for $0 \leq t \leq T$ with the probability $1-\frac{(t-1)\delta}{T}$. According to our setting for λ in Theorem 1, we know that $\|\nabla F_S(x_t)\| \leq \frac{\lambda}{2}$ also holds with the probability $1-\frac{(t-1)\delta}{T}$. Thus, according to Lemma 5 in [32], $A_2 \leq \eta \sum_{l=0}^{t} \| \sum_{l=0}^{t} \mathbb{E}_l \widetilde{\nabla} f(x_l;\xi_{j_l}) - \nabla F_S(x_l)\| \leq 2\eta t \sigma^p \lambda^{1-p}$

Upper Bound for A_3

Besides, we have the following inequality with probability at least $1-\delta$

$$\begin{aligned}\|\sum_{l=0}^{t}\nabla F_S(x_l)\|^2 \leq \left(\sum_{l=0}^{t}\|\nabla F_S(x_l)\|\right)^2 \leq \left(\sum_{l=1}^{t}\right)\left(\sum_{l=0}^{t}\|\nabla F_S(x_l)\|^2\right) \\ \leq t^2\mathcal{O}\left(t^{\frac{p}{3p-2}}\log^{\frac{2p-2}{2p-1}}\frac{t}{\delta}\right) \leq \mathcal{O}\left(t^{\frac{7p-4}{3p-2}}(\log\frac{t}{\delta})^{\frac{2p-2}{2p-1}}\right),\end{aligned} \tag{38}$$

where the second inequality holds due to the Cauchy-Schwarz inequality and the third inequality holds due to the Theorem 1. Thus, with the probability $1-\delta$,

$$\|\sum_{l=0}^{t}\nabla F_S(x_l)\| = \mathcal{O}\left(t^{\frac{3.5p-2}{3p-2}}\log^{\frac{p-1}{2p-1}}\frac{t}{\delta}\right). \tag{39}$$

Thus, combining these bounds, we have the following inequality

$$\|x_{t+1}\| \leq A_1 + A_2 + A_3 \leq 2\eta\lambda t + 2\eta t\sigma^p\lambda^{1-p} + \eta\mathcal{O}\left(t^{\frac{3.5p-2}{3p-2}}\log^{\frac{p-1}{2p-1}}\frac{t}{\delta}\right) \tag{40}$$

holding with the probability of $(1-\frac{(t-1)\delta}{T})(1-\delta) \geq 1-2\delta$. Since we choose the step size $\eta = \mathcal{O}(T^{\frac{-p}{3p-2}}/\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta})$, $\lambda = \mathcal{O}(T^{\frac{1}{3p-2}}\log^{\frac{1}{1-2p}}\frac{T}{\delta})$, we have

$$\|x_{t+1}\| \leq \mathcal{O}(T^{\frac{2p-1}{3p-2}}/\log\frac{T}{\delta}) + \mathcal{O}(T^{\frac{p-1}{3p-2}}/\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}) + \mathcal{O}(T^{\frac{2.5p-2}{3p-2}}/\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}) \leq \mathcal{O}(T^{\frac{2p-1}{3p-2}}/\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}). \tag{41}$$

Thus, we know that x_t traverses a sphere, and as t increases, the radius of the sphere increases. We need to replace R in Lemma 5 with $\|x_{t+1}\|$. Plugging the bound of $\|x_{t+1}\|$ into Lemma 5, we have the following inequality

$$\begin{aligned}\|\nabla F(x_{t+1}) - \nabla F_S(x_{t+1})\| &\leq (L_0\|x_{t+1}\| + (L_1\|x_{t+1}\|+1)\widetilde{b})n^{-\frac{1}{2}}\mu \\ &\leq \mu\frac{L_0\mathcal{O}(\frac{T^{\frac{2p-1}{3p-2}}}{\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}}) + (L_1\mathcal{O}(\frac{T^{\frac{2p-1}{3p-2}}}{\log^{\frac{p-1}{2p-1}}\frac{T}{\delta}})+1)\widetilde{b}}{\sqrt{n}}\end{aligned} \tag{42}$$

holding with the probability of $1-2\delta$, where $\mu = \left(2+2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log\frac{1}{\delta}}\right)$. This bound also means that the following inequality

$$\|\nabla F(x_{t+1}) - \nabla F_S(x_{t+1})\|^2 = \mathcal{O}\left(\frac{T^{\frac{4p-2}{3p-2}}}{n\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta}}\left(d+\log\frac{1}{\delta}\right)\right) \tag{43}$$

holding with the probability $1-2\delta$. Finally, we begin to bound $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2$. Firstly,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 &\leq 2\sum_{t=1}^{T}\|\nabla F(x_t) - \nabla F_S(x_t)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 \\ &\leq 2T\max_{1\leq t\leq T}\|\nabla F(x_t) - \nabla F_S(x_t)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 \\ &= 2T\|\nabla F(x_T) - \nabla F_S(x_T)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2,\end{aligned} \tag{44}$$

where the last inequality holds due to the fact that $\|\nabla F(x_t) - \nabla F_S(x_t)\|^2$ increases as t increases. Combining inequalities (43), (44) and Theorem 1, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 &= \mathcal{O}\left(\frac{T^{\frac{4p-2}{3p-2}}}{n\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta}}\left(d+\ln\frac{1}{\delta}\right)\right) + \mathcal{O}(T^{\frac{2-2p}{3p-2}}\log^{\frac{2p-2}{2p-1}}\frac{T}{\delta}) \\ &\leq \mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{p-1}{3p-2}}\log^{\frac{2p-2}{2p-1}}\left(\frac{1}{\delta}\sqrt{\frac{n}{d}}\right)\right)\end{aligned} \tag{45}$$

holding with the probability at least $1-2\delta$, where we choose $T=\sqrt{\frac{n}{d}}$ in the last inequality. $\square$

C Federated Setting

Next, we begin to analyze the high-probability bounds for the federated setting. Compared with the Clipped SGD, there are more difficulties to be addressed.

C.1 High Probability Optimization Bound

Lemma 7 (Batch gradient variance bound). *If Assumptions 3 holds and $\|\nabla f_i(x_t)\| \leq \frac{\lambda}{2}, \forall i \in [m]$, for batch stochastic gradient $\nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k) = \frac{1}{b}\sum_{\xi_{t,i}^k \in \boldsymbol{\xi}_{t,i}^k} \nabla f_i(x_{t,i}^k; \xi_{t,i}^k)$, we have*

$$\mathbb{E}_t[\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k; \boldsymbol{\xi}_{t,i}^k)\|^2] \leq \frac{3\sigma^p\lambda^{2-p}}{b} \tag{46}$$

holding under the heavy-tailed noise condition, where $\boldsymbol{\xi}_{t,i}^k = \{(\xi_{t,i}^k)_j\}_{j=1}^b$.

Proof. According to the i.i.d. sampling, we have

$$\begin{aligned}&\mathbb{E}_t[\|\nabla f_i(x_{t,i}^k) - \frac{1}{b}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\nabla f(x_{t,i}^k;\xi_{t,i}^k)\|^2] = \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\mathbb{E}_t\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^2 \\ &= \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\mathbb{E}_t\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^p\Big[\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^{2-p}\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|\leq\lambda} \\ &\quad +\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^{2-p}\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|\geq\lambda}\Big] \\ &\overset{(a)}{\leq} \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\mathbb{E}_t\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^p\,(2\lambda)^{2-p} + \mathbb{E}_t[\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|\geq\lambda}] \\ &\overset{(b)}{\leq} \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\sigma^p\,(2\lambda)^{2-p} + [\mathbb{E}_t\|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|^p]^{\frac{2}{p}}\Big[\mathbb{E}_t[\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|\geq\lambda}]\Big]^{1-\frac{2}{p}} \\ &\overset{(c)}{\leq} \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\sigma^p\,(2\lambda)^{2-p} + \sigma^2\Big[\mathbb{E}_t[\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)-\nabla f_i(x_t)\|\geq\frac{\lambda}{2}}]\Big]^{1-\frac{2}{p}} \\ &\overset{(d)}{\leq} \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\sigma^p\,(2\lambda)^{2-p} + \sigma^2\Pr\left[\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k) - \nabla f_i(x_t)\|^p \geq \left(\frac{\lambda}{2}\right)^p\right]^{1-\frac{2}{p}} \\ &\overset{(e)}{\leq} \frac{1}{b^2}\sum_{\xi_{t,i}^k\in\boldsymbol{\xi}_{t,i}^k}\sigma^p\,(2\lambda)^{2-p} + 2^{p-2}\sigma^p\lambda^{2-p} \leq \frac{1}{b}[\sigma^p\,(2\lambda)^{2-p} + 2^{p-2}\sigma^p\lambda^{2-p}] \leq \frac{3\sigma^p\lambda^{2-p}}{b},\end{aligned} \tag{47}$$

where the inequality (a) holds due to $\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\|<\lambda} \leq \lambda + \|\nabla f_i(x_{t,i}^k)\| \leq \lambda + \|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_t)\| + \|\nabla f_i(x_t)\| \leq \frac{3\lambda}{2} + (L_0 + \frac{L_1\lambda}{2})\|x_{t,i}^k - x_t\| \leq \frac{3\lambda}{2} + (L_0 + \frac{L_1\lambda}{2})\eta k\lambda \leq 2\lambda$ when $\eta \leq \frac{1}{k(2L_0+L_1\lambda)}$, the inequality (b) holds due to Assumption 3 and Hölder Inequality, the inequality (c) holds due to Assumption 3 and $\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)\| \geq \lambda \Rightarrow$

$\|\nabla f_i(x_{t,i}^k;\xi_{t,i}^k)-\nabla f_i(x_t)\|\geq\frac{\lambda}{2}$ when $\|\nabla f_i(x_t)\|\leq\frac{\lambda}{2}$, the inequality (d) holds due to the definition of the indicator function $\mathbb{I}(\cdot)$, and the inequality (e) holds due to Markov's Inequaliy. □

Lemma 8. *We define*

$$\epsilon_t^a=\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K(\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]),$$

$$\epsilon_t^b=\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla F_S(x_t).$$

For Algorithm 2, $\forall t\in[T]$, we have $\|\epsilon_t^a\|\leq 2\lambda$. Besides, if $\|\nabla f_i(x_t)\|\leq\frac{\lambda}{2}$, there is

$$\|\epsilon_t^b\|\leq\frac{12\sigma^p\lambda^{1-p}}{b},\quad \mathbb{E}_t[\|\epsilon_t^a\|^2]\leq\frac{100\sigma^p\lambda^{2-p}}{mKb}\tag{48}$$

Proof. To simplify, we have the following notations: $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]:=\mathbb{E}_t[\cdot]$. We define $\epsilon_t=\epsilon_t^a+\epsilon_t^b$. We bound $\|\epsilon_t^b\|$ as follows:

$$\begin{aligned}
\|\epsilon_t^b\|\ &=\|\frac{1}{mK}\mathbb{E}_t[\widetilde{\Delta}_t^i]-\nabla F_S(x_t)\|=\|\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K(\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_t))\|\\
&\leq\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_t)\|\\
&\leq\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\left(\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_{t,i}^k)\|+\|\nabla f_i(x_{t,i}^k)-\nabla f_i(x_t)\|\right)\\
&\leq\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\left(\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_{t,i}^k)\|+(L_0+L_1\|\nabla f_i(x_t)\|)\|x_{t,i}^k-x_t\|\right)\\
&\leq\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\left(\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_{t,i}^k)\|+(L_0+\frac{L_1\lambda}{2})\eta\|\sum_{q=[k]}\widetilde{\nabla} f_i(x_{t,i}^q;\boldsymbol{\xi}_{t,i}^q)\|\right)\\
&\leq\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\left(\underbrace{\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]-\nabla f_i(x_{t,i}^k)\|}_{J_1}+(L_0+\frac{L_1\lambda}{2})\eta k\max\{1,\lambda\}\right),
\end{aligned}\tag{49}$$

where the last equality holds due to the definition of F_S, the first and second inequalities holds due to the Triangle Inequality, the four inequality holds due to the update rule of local update and $\|\nabla f_i(x_t)\|\leq\frac{\lambda}{2}$, and the last inequality holds due to the clipping operator.

Then, we bound the term J_1 as follows:

$$\begin{aligned}
J_1\ &=\|\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]\|\leq\mathbb{E}_t[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|]\\
&=\mathbb{E}_t\left[\left\|\left(\frac{\lambda}{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|}-1\right)\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\geq\lambda}\right\|\right]\\
&=\mathbb{E}_t\left[\left(\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|-\lambda\right)\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\geq\lambda}\right]\\
&\leq\mathbb{E}_t\left[\left(\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_t)\|+\|\nabla f_i(x_t)\|-\lambda\right)\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|}\geq\lambda\right]\\
&\leq\mathbb{E}_t\left[\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_t)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\geq\lambda}\right],
\end{aligned}\tag{50}$$

where $\mathbb{I}(\cdot)$ is an indicator function, the last inequality holds due to $\|\nabla f_i(x_t)\| \le \frac{\lambda}{2} < \lambda$. Then,

$$\begin{aligned}
&\mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\|\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda}\right] \\
&\le \mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} + \|\nabla f_i(x_{t,i}^k) - \nabla f_i(x_t)\|\right] \\
&\le \mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} + (L_0 + L_1\|\nabla f_i(x_t)\|)\|x_{t,i}^k - x_t\|\right] \\
&\le \mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} + (L_0 + \frac{L_1\lambda}{2})\eta\|\sum_{q=1}^{k}\nabla\widetilde{f}(x_{t,i}^q;\boldsymbol{\xi}_{t,i}^q)\|\right] \\
&\le \mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k\right].
\end{aligned} \tag{51}$$

According to $\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\| \ge \lambda \Rightarrow \|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\| \ge \frac{\lambda}{2}$ when $\|\nabla f_i(x_t)\| \le \frac{\lambda}{2}$, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k\right] \\
&\le \mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|\mathbb{I}_{\|\nabla f(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\|\ge\frac{\lambda}{2}} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k\right] \\
&\overset{(a)}{\le} [\mathbb{E}_t\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2]^{1/2}\left[\mathbb{E}_t[\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\|\ge\frac{\lambda}{2}}]\right]^{1-1/2} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k \\
&\le \frac{\sqrt{3\sigma^p\lambda^{2-p}}}{\sqrt{b}}\Pr\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_t)\|^2 \ge \left(\frac{\lambda}{2}\right)^2\right]^{1-1/2} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k \\
&\overset{(b)}{\le} \frac{\sqrt{3\sigma^p\lambda^{2-p}}}{\sqrt{b}}\left(\frac{28\sigma^p\lambda^{2-p}}{(\lambda)^2 b}\right)^{1-1/2} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k \\
&= \frac{10\sigma^p\lambda^{1-p}}{b} + (L_0 + \frac{L_1\lambda}{2})\eta\lambda k \overset{\eta\le\frac{1}{k\lambda(2L_0+L_1\lambda)}}{\le} \frac{11\sigma^p\lambda^{1-p}}{b},
\end{aligned} \tag{52}$$

where (a) is due to Hölder Inequality and (b) is due to the same reason as the inequality (f) in (55). Combining (49), (50) and (52), we have $\|\epsilon_t^b\| \le \frac{12\sigma^p\lambda^{1-p}}{b}$.

We bound the $\mathbb{E}_t[\|\epsilon_t^a\|^2]$ as follows:

$$\begin{aligned}
\mathbb{E}_t[\|\epsilon_t^a\|^2] &\le \frac{1}{m^2K^2}\mathbb{E}_t\|\sum_{i=1}^{m}\sum_{k=1}^{K}\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]\|^2 \\
&\overset{(b)}{\le} \frac{1}{m^2K^2}\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{E}_t[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]\|^2] \\
&\overset{(c)}{\le} \frac{1}{m^2K^2}\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{E}_t[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2] \\
&\le \frac{1}{m^2K^2}\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{E}_t\Big[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} \\
&\quad + \|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|<\lambda}\Big] \\
&= \frac{1}{m^2K^2}\sum_{i=1}^{m}\sum_{k=1}^{K}\mathbb{E}_t\Big[\|\frac{\lambda}{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|}\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda} \\
&\quad + \|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k) - \nabla f_i(x_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|<\lambda}\Big]
\end{aligned} \tag{53}$$

where the inequality (b) holds due to i.i.d. sampling, the inequality (c) holds due to $\mathbb{E}_t[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\mathbb{E}_t[\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)]\|^2]\le\mathbb{E}_t[\|\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-P\|^2]$ for any deterministic vector P [15]. According to

$$
\begin{aligned}
&\left\|\frac{\lambda}{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|}\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\right\|\le\lambda+\|\nabla f_i(x_{t,i}^k)\|\overset{\eta\le\frac{1}{k(2L_0+L_1\lambda)}}{\le}2\lambda,\\
&\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|<\lambda}\le\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2,
\end{aligned}
\tag{54}
$$

we have

$$
\begin{aligned}
&\mathbb{E}_t[\|\epsilon_t^a\|^2]\\
\le&\frac{1}{m^2K^2}\sum_{i=1}^m\sum_{k=1}^K\mathbb{E}_t\left[4\lambda^2\mathbb{I}_{\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|\ge\lambda}+\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2\right]\\
\overset{(d)}{\le}&\frac{1}{m^2K^2}\sum_{i=1}^m\sum_{k=1}^K4\lambda^2\Pr\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_t)\|^2\ge\left(\frac{\lambda}{2}\right)^2\right]+\mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2\right]\\
\overset{(e)}{\le}&\frac{1}{m^2K^2}\sum_{i=1}^m\sum_{k=1}^K4\lambda^2\Pr\left[2\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2+2\|\nabla f_i(x_{t,i}^k)-\nabla f_i(x_t)\|^2\ge\left(\frac{\lambda}{2}\right)^2\right]\\
&+\mathbb{E}_t\left[\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2\right]\\
\overset{(f)}{\le}&\frac{1}{m^2K^2}\sum_{i=1}^m\sum_{k=1}^K\mathbb{E}_t\left[4\lambda^2\frac{4[6\sigma^p\lambda^{2-p}/b+2(L_0+L_1\lambda/2)^2\eta^2k^2\lambda^2]}{\lambda^2}+\|\nabla f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)-\nabla f_i(x_{t,i}^k)\|^2\right]\\
\overset{(g)}{\le}&\frac{1}{m^2K^2}\sum_{i=1}^m\sum_{k=1}^K\left[\frac{97\sigma^p\lambda^{2-p}}{b}+\frac{3\sigma^p\lambda^{2-p}}{b}\right]=\frac{100\sigma^p\lambda^{2-p}}{mKb},
\end{aligned}
\tag{55}
$$

the inequality (d) holds due to the definition of indicator function, (e) holds due to the Young's inequality, and the inequality (f) holds due to the Markov's Inequaliy, Assumption 3 and $\|\nabla f_i(x_{t,i}^k)-\nabla f_i(x_t)\|\le(L_0+L_1\|\nabla f_i(x_t)\|)\|x_{t,i}^k-x_t\|\le(L_0+\frac{L_1\lambda}{2})\eta\|\sum_{q=0}^{k-1}\widetilde{\nabla} f_i(x_{t,i}^q;\boldsymbol{\xi}_{t,i}^q)\|\le(L_0+\frac{L_1\lambda}{2})\eta k\lambda$, and the inequality (g) holds due to $\eta\le\frac{\sigma^{\frac{p}{2}}}{k\sqrt{b}\lambda(2L_0+L_1\lambda)}$.

□

Lemma 9. *Assume that f satisfies Assumptions 1, 2 and 3, and $\gamma\le\frac{1}{mKL_0}$ then for all $t\ge1$, there is*

$$
\begin{aligned}
&\frac{\gamma_t mK}{2}\|\nabla F_S(x_t)\|^2\\
\le&\Delta_t'-\Delta_{t+1}'+\gamma mK(1+\frac{L_1}{L_0}\|\nabla f(x_t)\|)(\|\epsilon_t^a\|^2-\mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}])\\
&+\frac{L_1}{L_0}\|\nabla f(x_t)\|\gamma mK\langle\epsilon_t^a,\nabla f(x_t)\rangle+\gamma mK(1+\frac{L_1}{L_0}\|\nabla f(x_t)\|)\|\epsilon_t^b\|^2+\frac{L_1\|\nabla f(x_t)\|\gamma^2m^2K^2}{2}\|\nabla f(x_t)\|^2\\
&+\frac{L_1}{L_0}\|\nabla f(x_t)\|^2\gamma mK\|\epsilon_t^b\|+\gamma mK(1+\frac{L_1}{L_0}\|\nabla f(x_t)\|)\mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}]
\end{aligned}
\tag{56}
$$

holding, where $\Delta_t'=f(x_t)-f^*$.

Proof. We define $\Delta_{t+1}'=f(x_{t+1})-f^*$, $\widetilde{\Delta}_t=\frac{1}{mK}\sum_{i=1}^m\sum_{k=1}^K\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)$. According to our setting for local step size η and clipping parameter λ, we have the fact that $\|x_{t,i}^{k+1}-x_{t,i}^k\|\le\eta\lambda\le\frac{1}{L_1}$

for sufficiently large m, K, T. Due to the smoothness of F_S, we have

$$
\begin{aligned}
&F_S(x_{t+1}) - F_S(x_t)\\
&\leq \langle \nabla F_S(x_t), x_{t+1} - x_t\rangle + \frac{\ell}{2}\|x_{t+1} - x_t\|^2\\
&= -\gamma\langle \nabla F_S(x_t), \sum_{i=1}^{m}\sum_{k=1}^{K}\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\rangle + \frac{\ell\gamma^2}{2}\|\sum_{i=1}^{m}\sum_{k=1}^{K}\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)\|^2\\
&= \frac{\gamma mK}{2}\|\nabla F_S(x_t) - \widetilde{\Delta}_t\|^2 - \frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2 - \frac{\gamma}{2mK}\|mK\widetilde{\Delta}_t\|^2 + \frac{\ell\gamma^2}{2}\|mK\widetilde{\Delta}_t\|^2\\
&= \frac{\gamma mK}{2}\|\epsilon_t^a + \epsilon_t^b\|^2 - \frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2 - (\frac{\gamma}{2mK} - \frac{L_0\gamma^2}{2})\|mK\widetilde{\Delta}_t\|^2 + \frac{L_1\|\nabla F_S(x_t)\|\gamma^2}{2}\|mK\widetilde{\Delta}_t\|^2\\
&= \frac{\gamma mK}{2}\|\epsilon_t^a + \epsilon_t^b\|^2 - \frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2 + \frac{L_1\|\nabla F_S(x_t)\|\gamma^2 m^2K^2}{2}\|\epsilon_t + \nabla F_S(x_t)\|^2,
\end{aligned}
\tag{57}
$$

where, $\ell = L_0 + L_1\|\nabla f(x_t)\|$, and the last equality holds due to $\gamma \leq \frac{1}{L_0 mK}$. Then, by $\|\epsilon_t + \nabla F_S(x_t)\|^2 = \|\epsilon_t\|^2 + \|\nabla F_S(x_t)\|^2 + 2\langle \epsilon_t, \nabla F_S(x_t)\rangle$, we have

$$
\begin{aligned}
F_S(x_{t+1}) - F_S(x_t) \leq{}& \frac{\gamma mK}{2}(1 + \frac{L_1\|\nabla F_S(x_t)\|}{L_0})\|\epsilon_t^a + \epsilon_t^b\|^2 - \frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2\\
&+ \frac{L_1\|\nabla F_S(x_t)\|\gamma^2 m^2K^2}{2}\|\nabla F_S(x_t)\|^2 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|\gamma mK\langle \epsilon_t, \nabla F_S(x_t)\rangle.
\end{aligned}
\tag{58}
$$

We construct the martingale difference sequences as follows

$$
\begin{aligned}
\Delta'_{t+1} +{}& \frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2 \leq \Delta'_t + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^a\|^2 + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2\\
&+ \frac{L_1\|\nabla F_S(x_t)\|\gamma^2 m^2K^2}{2}\|\nabla F_S(x_t)\|^2 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|\gamma mK\langle \epsilon_t, \nabla F_S(x_t)\rangle\\
={}& \Delta'_t + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}])\\
&+ \frac{L_1}{L_0}\|\nabla F_S(x_t)\|\gamma mK\langle \epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \frac{L_1\|\nabla F_S(x_t)\|\gamma^2 m^2K^2}{2}\|\nabla F_S(x_t)\|^2\\
&+ \frac{L_1}{L_0}\|\nabla F_S(x_t)\|\gamma mK\langle \epsilon_t^b, \nabla F_S(x_t)\rangle + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}]\\
\leq{}& \Delta'_t + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}])\\
&+ \frac{L_1}{L_0}\|\nabla F_S(x_t)\|\gamma mK\langle \epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \frac{L_1\|\nabla F_S(x_t)\|\gamma^2 m^2K^2}{2}\|\nabla F_S(x_t)\|^2\\
&+ \frac{L_1}{L_0}\|\nabla F_S(x_t)\|^2\gamma mK\|\epsilon_t^b\| + \gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}[\|\epsilon_t^a\|^2|\mathcal{F}_{t-1}],
\end{aligned}
\tag{59}
$$

where the last inequality holds due to Cauchy-Schwarz inequality. □

Our convergence proof relies on the following inductive lemma.

Lemma 10. *For $1 \leq N \leq T+1$, let E'_N be the event that for all $l = 1, \cdots, N$,*

$$\Delta'_l + \frac{\gamma m K}{2}\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^2$$
$$\leq \Delta'_1 + \gamma m K \sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])$$
$$+ \frac{L_1}{L_0}\gamma m K\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|\langle \epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma m K\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \sum_{t=1}^{l-1}\frac{L_1\gamma^2 m^2 K^2}{2}\|\nabla F_S(x_t)\|^3$$
$$+ \frac{L_1}{L_0}\gamma m K\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\| + \gamma m K\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2] \leq 2\Delta'_1. \tag{60}$$

Then E'_N happens with probability at least $1 - \frac{(N-1)\delta}{T}$ for each $N \in [T]$.

Proof. We will show by induction that the probability of event E'_{N+1} occurring is at least $1 - \frac{N\delta}{T}$.

For $N = 1$, the event happens with probability 1. Suppose that for $N \leq T$, $\Pr[E'_N] \geq 1 - \frac{(N-1)\delta}{T}$. We will prove that $\Pr[E'_{N+1}] \geq 1 - \frac{N\delta}{T}$.

For $l \leq N$, under the event E'_N, the following inequality

$$\Delta'_l + \frac{\gamma m K}{2}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^2$$
$$\leq \Delta'_1 + \gamma m K \sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])$$
$$+ \frac{L_1}{L_0}\gamma m K\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|\langle \epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma m K\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \sum_{t=1}^{l-1}\frac{L_1\gamma^2 m^2 K^2}{2}\|\nabla F_S(x_t)\|^3$$
$$+ \frac{L_1}{L_0}\gamma m K\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\| + \gamma m K\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2] \leq 2\Delta'_1 \tag{61}$$

holds with the probability at least $1 - \frac{(N-1)\delta}{T}$. Thus, from the induction hypothesis, we have $\Delta'_l \leq 2\Delta'_1, \forall l \leq N$.

By the smoothness of F_S and F_S is bounded below, we have

$$\|\nabla F_S(x_t)\| \leq \sqrt{2\ell\Delta'_t} = \sqrt{2(L_0 + L_1\|\nabla F_S(x_t)\|)\Delta'_t} \tag{62}$$

Thus, we have

$$\|\nabla F_S(x_t)\| \leq \frac{2\Delta'_t L_1 + \sqrt{4(\Delta'_t)^2 L_1^2 + 8L_0\Delta'_t}}{2}, \quad \forall t \leq N. \tag{63}$$

Under the event E'_N, we have

$$\|\nabla F_S(x_t)\| \leq \frac{4\Delta'_1 L_1 + \sqrt{16(\Delta'_1)^2 L_1^2 + 16L_0\Delta'_1}}{2} \leq \frac{\lambda}{2}, \tag{64}$$

where we use the $\lambda \geq [4\Delta'_1 L_1 + 4\sqrt{(\Delta'_1)^2 L_1^2 + L_0\Delta'_1}]$. Similarly, we have

$$\|\nabla f_i(x_t)\| \leq 2R' \leq \frac{\lambda}{2}, \tag{65}$$

where $R' := \frac{1}{2}(L_0 + (4L_1+1)\Delta'_1)$.

Thus, according to Lemma 8 and we have

$$\|\epsilon_t^b\| \leq \frac{12\sigma^p\lambda^{1-p}}{b}, \quad \mathbb{E}_t[\|\epsilon_t^a\|^2] \leq \frac{100\sigma^p\lambda^{2-p}}{mKb}. \tag{66}$$

Now, we can analyze the event E'_{N+1}. By summing over t from 1 to N regarding inequality (59), we have

$$
\begin{aligned}
&\Delta'_{N+1} + \frac{\gamma mK}{2}\sum_{t=0}^{N}\|\nabla F_S(x_t)\|^2 \\
&\le \Delta'_1 + \underbrace{\gamma mK\sum_{t=1}^{N}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])}_{A} + \underbrace{\frac{L_1}{L_0}\gamma mK\sum_{t=1}^{N}\|\nabla F_S(x_t)\|\langle\epsilon_t^a, \nabla F_S(x_t)\rangle}_{B} \\
&\quad + \underbrace{\sum_{t=1}^{N}\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2}_{C_1} + \underbrace{\sum_{t=1}^{N}\frac{L_1\gamma^2 m^2 K^2}{2}\|\nabla F_S(x_t)\|^3}_{C_2} \\
&\quad + \underbrace{\frac{L_1}{L_0}\gamma mK\sum_{t=1}^{N}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\|}_{C_3} + \underbrace{\sum_{t=1}^{N}\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2]}_{C_4}.
\end{aligned}
\tag{67}
$$

Thus, we can compute $\Pr[E'_{N+1}]$ by the probability $\Pr[E_A \cap E_B \cap E_{C_1} \cap E_{C_2} \cap E_{C_3} \cap E_{C_4}]$.

Upper Bound for A

We check the conditions to apply Freedman's inequality. Firstly, $\mathbb{E}_t\left[\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])\right] = 0$. Further, with probability 1, $\|\epsilon_t^a\| \le 2\lambda$, thus $|\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])| \le 8(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 K$. Thus, the sequence $\{\gamma K(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])\}$ is a bounded martingale difference sequence. Applying Freedman's inequality for constants c and G to be chosen, we have

$$
\begin{aligned}
&\Pr\Bigg[\left|\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\sum_{t=1}^{N}\left(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]\right)\right| > c \text{ and} \\
&\sum_{t=1}^{N}\mathbb{E}_t\left[\left(\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])\right)\right] \le G\ln\frac{4T}{\delta}\Bigg] \\
&\le 2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta} + \frac{16}{3}(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 mKc}\right)
\end{aligned}
\tag{68}
$$

We choose c such that

$$
2\exp\left(-\frac{c^2}{2G\ln\frac{4T}{\delta} + \frac{16}{3}(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 mKc}\right) = \frac{\delta}{2T}
\tag{69}
$$

which gives

$$
c = \left(\frac{8}{3}(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 mK + \sqrt{\frac{64(1+\frac{2L_1R'}{L_0})^2\gamma^2\lambda^4 m^2K^2}{9} + 2G}\right)\ln\frac{4T}{\delta}
\tag{70}
$$

We use the same setting $K \le \mathcal{O}(\frac{T^{\frac{p}{2p-2}}}{m})$ as in [49]. We choose $\gamma \le \frac{\min\{\frac{32\mathcal{R}'\rho}{L_0}, \frac{3}{4}, \frac{3L_0}{8L_1R'}\}\Delta'_1}{32\mathcal{R}'\lambda^2\rho(mKT)^{\frac{1}{3}}}$ such that we have $\frac{16}{3}(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 mK \le \frac{\Delta'_1}{8}$. Besides, we choose $G = 800(1+\frac{2L_1R'}{L_0})^2\frac{\sigma^p\gamma^2\lambda^{4-p}m^2K^2T}{mKb}$, and we set $\lambda \ge \frac{(10\mathcal{R}')^{\frac{2}{p}}\sigma(mKT)^{\frac{1}{3(3p-2)}}}{(\mathcal{R}'^2)^{\frac{1}{p}}}$ and $b = \mathcal{O}((mKT)^{\frac{2p-2}{3(3p-2)}})$ such that we can obtain $c \le \frac{2\Delta'_1}{8}$.

Then, the following event happens

$$E_{C'} = \left\{ \text{either } C \leq |\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\sum_{t=1}^{N}(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])| \leq \frac{2\Delta_1'}{8} \text{ or} \right.$$
$$\left. \sum_{t=1}^{N}\mathbb{E}_t\left[(\gamma mK(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]))^2\right] \geq G\ln\frac{4T}{\delta} \right\}$$

Under the event E_N', we have

$$\begin{aligned}
&\sum_{t=1}^{N}\mathbb{E}_t\left[(\gamma K(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]))^2\right] \\
&\leq 8(1+\frac{2L_1R'}{L_0})\gamma\lambda^2 mK\sum_{t=1}^{N}\mathbb{E}_t[|\gamma K(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])|] \\
&\leq 8(1+\frac{2L_1R'}{L_0})^2\gamma^2\lambda^2 m^2K^2\sum_{t=1}^{N}\mathbb{E}_t[\|\epsilon_t^a\|^2] \\
&\leq 800(1+\frac{2L_1R'}{L_0})^2\frac{\sigma^p\gamma^2\lambda^{4-p}m^2K^2N}{mKb} \leq G \leq G\ln\frac{4T}{\delta}
\end{aligned} \tag{71}$$

Thus, when $E_N' \cap E_A$ happens (i.e., with the probability $(1-\frac{(N-1)\delta}{T})(1-\frac{\delta}{2T})$), we have $A \leq \frac{2\Delta_1'}{8}$.

Upper Bound for B

To bound B, we need the Freedman's inequality. To contruct the martingale difference sequence, we define the following random variables

$$X_t = \begin{cases} \nabla F_S(x) & \text{if } \Delta_t' \leq 2\Delta_1' \\ 0 & \text{otherwise} \end{cases} \tag{72}$$

Thus $\|X_t\| \leq \|\nabla F_S(x_t)\| \leq 2\Delta_1' L_1 + 2\sqrt{(\Delta_1')^2L_1^2 + L_0\Delta_1'}$ for all t.

Instead of bounding B, we bound $B' = \frac{L_1}{L_0}\gamma mK\sum_{t=1}^{N}\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle$. We check the conditions to apply Freedman's inequality. First $\mathbb{E}_t[\frac{L_1}{L_0}\gamma mK\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle] = 0$. Furthermore, with probability 1, $\|\epsilon_t^a\| \leq 2\lambda$, and $\|X_t\| \leq 2\Delta_1' L_1 + 2\sqrt{(\Delta_1')^2L_1^2 + L_0\Delta_1'}$, thus $|\frac{L_1}{L_0}\gamma mK\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle| \leq \frac{L_1}{L_0}\gamma mK\|\nabla F_S(x_t)\|\|X_t\|\|\epsilon_t^a\| \leq 8\frac{L_1}{L_0}\gamma mK(R')^2\lambda$. Hence, $\{\frac{L_1}{L_0}\gamma K\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle\}$ is a bounded martingale difference sequence. Therefore, for constant a and M to be chosen we have

$$\begin{aligned}
&\Pr\left[\left|\sum_{t=1}^{N}\frac{L_1}{L_0}\gamma mK\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle\right| > a \text{ and } \sum_{t=1}^{N}\mathbb{E}_t\left[(\frac{L_1}{L_0}\gamma mK\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle)^2\right] \leq M\ln\frac{4T}{\delta}\right] \\
&\leq 2\exp\left(-\frac{a^2}{2M\ln\frac{4T}{\delta} + \frac{16}{3}\frac{L_1}{L_0}(R')^2\gamma\lambda mKa}\right)
\end{aligned} \tag{73}$$

We choose a such that

$$2\exp\left(-\frac{a^2}{2M\ln\frac{4T}{\delta} + \frac{16}{3}\frac{L_1}{L_0}(R')^2\gamma\lambda mKa}\right) = \frac{\delta}{2T} \tag{74}$$

which gives

$$a = \left(\frac{8}{3}\frac{L_1}{L_0}(R')^2\lambda\gamma mK + \sqrt{\frac{64L_1^2(R')^4\lambda^2\gamma^2m^2K^2}{9L_0^2} + 2M}\right)\ln\frac{4T}{\delta} \tag{75}$$

If we choose $M = 800\frac{L_1^2}{L_0^2}\gamma^2 m^2 K^2 (R')^4 \left(\frac{\sigma}{\lambda}\right)^p \frac{\lambda^2 N}{mKb}$ and set $\lambda \geq \max\{\frac{(5L_1\Delta_1)^{\frac{2}{2+p}}\sigma^{\frac{p}{2+p}}(mKT)^{\frac{p}{3(3p-2)(2+p)}}}{(8)^{\frac{2}{2+p}}b^{\frac{1}{2+p}}}, \frac{\frac{8}{3}L_1 R'(mK)^{\frac{2}{3}}}{T^{\frac{1}{3}}}\}$, we can obtain $\sqrt{2M}\ln\frac{4T}{\delta} \leq \frac{\Delta_1'}{16}$ and $\frac{16}{3}L_1(R')^2\lambda\gamma mK \leq \frac{\Delta_1'}{16}$. In summary, we can obtain that $a \leq \frac{\Delta_1'}{8}$. Thus, we have

$$E_B = \left\{\text{either } |B'| \leq \frac{\Delta_1'}{8} \text{ or } \sum_{t=1}^{N}\mathbb{E}_t\left[(\frac{L_1}{L_0}\gamma K\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle)^2\right] \geq M\ln\frac{4T}{\delta}\right\} \tag{76}$$

holds with probability at least $1 - \frac{\delta}{2T}$.

Also notice that under the event E_N', we have

$$\begin{aligned}
&\sum_{t=1}^{N}\mathbb{E}_t\left[(\frac{L_1}{L_0}\gamma m^2 K\|\nabla F_S(x_t)\|\langle X_t, \epsilon_t^a\rangle)^2\right] \\
&\leq 2\frac{L_1^2}{L_0^2}\gamma^2 m^2 K^2 (R')^2 \sum_{t=1}^{N}\mathbb{E}_t\left[\|X_t\|^2\|\epsilon_t^a\|^2\right] \leq 8\frac{L_1^2}{L_0^2}\gamma^2 m^2 K^2 (R')^4 \sum_{t=1}^{N}\mathbb{E}_t[\|\epsilon_t^a\|^2] \\
&\leq 800\frac{L_1^2}{L_0^2}\gamma^2 m^2 K^2 (R')^4 \frac{\sigma^p\lambda^{2-p}N}{mKb} \leq 800\frac{L_1^2}{L_0^2}\gamma^2 m^2 K^2 (R')^4 \left(\frac{\sigma}{\lambda}\right)^p \frac{\lambda^2 N}{mKb} \leq M \leq M\ln\frac{4T}{\delta}.
\end{aligned} \tag{77}$$

Under E_N', we have that $X_t = \nabla f(x_t)$ for all $t \leq N$. Thus, when $E_N' \cap E_B$ happens (i.e., with the probability $(1 - \frac{(N-1)\delta}{T})(1 - \frac{\delta}{2T}) \geq 1 - \delta$), we have $B \leq \frac{\Delta_1'}{8}$.

Upper Bound for C_1

When the event E_N' happens,

$$\begin{aligned}
C_1 = &\sum_{t=1}^{N}\gamma mK(1 + \frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 \leq 144\gamma mK\sum_{t=1}^{N}(1 + \frac{2L_1R'}{L_0})\frac{\sigma^{2p}\lambda^{2-2p}}{b^2} \\
&\leq \frac{144\sigma^{2p}\lambda^{2-2p}}{b^2}(1 + \frac{2L_1R'}{L_0})\gamma mKT
\end{aligned} \tag{78}$$

We need to bound C_1 to the order $\mathcal{O}(\Delta_1)$. However, this requirement leads to the small γ, which leads to a slow convergence rate. Thus, we propose the `FedCBG` algorithm, where we can choose $b = \mathcal{O}((mKT)^{\frac{2p-2}{3(3p-2)}})$, which balances the trade-off between the convergence rate and the upper bound of C_1. We set $\gamma \leq \frac{\Delta_1'}{32(1+2(L_1+1)R')\lambda^2\max\{\ln\frac{4T}{\delta},1\}(mKT)^{\frac{1}{3}}}$ and $\lambda \geq \frac{36^{\frac{1}{2p}}\sigma^2(mKT)^{\frac{1}{3(3p-2)}}}{\rho^{\frac{1}{2p}}}$ and We have

$$C_1 \leq 144\left(\frac{\sigma}{\lambda}\right)^{2p}(1 + \frac{2L_1R'}{L_0})\frac{\gamma\lambda^2 mKT}{b^2} \leq \frac{\Delta_1'}{8}. \tag{79}$$

where $\mathcal{R}' = (L_0 + 2(L_1+1)R')$. Thus, when $E_N' \cap E_{C_1}$ happens, we have $C_1 \leq \frac{\Delta_1'}{8}$ with the probability $(1 - \frac{(N-1)\delta}{T})(1 - \frac{\delta}{2T}) \geq 1 - \delta$.

Upper Bound for C_2

Under the event E_N', we have

$$C_2 = \frac{L_1\gamma^2(mK)^2}{2L_0}\sum_{t=0}^{l-1}\|\nabla F_S(x_t)\|^3 \leq 4\frac{L_1}{L_0}\gamma^2(mK)^2(R')^3T \leq \frac{\Delta_1'}{8}, \tag{80}$$

This condition is naturally satisfied if we choose the step size γ as mentioned above.

Upper Bound for C_3

Under the event E_N', we have

$$\frac{L_1}{L_0}\gamma mK\sum_{t=1}^{N}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\| \leq 48\frac{L_1}{L_0}mK\gamma(R')^2\frac{\sigma^p\lambda^{1-p}}{b}T \leq \frac{\Delta_1'}{8}, \tag{81}$$

We set $\lambda \geq \frac{(12L_1R')^{\frac{1}{p+1}}\sigma^{\frac{p}{p+1}}(mKT)^{\frac{1}{3(p+1)}}}{(L_0\rho b)^{\frac{1}{p+1}}}$. According to our settings of b, λ and γ, we can ensure C_3 is upper bounded.

Upper bound for C_4

When the event E'_N happens,

$$C_4 = \gamma mK \sum_{t=1}^{N}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2] \leq \frac{100\sigma^p\lambda^{2-p}\mathcal{R}'\gamma mKN}{mKb} \leq 100\mathcal{R}'\left(\frac{\sigma}{\lambda}\right)^p\frac{\lambda^2\gamma T}{b} \leq \frac{\Delta'_1}{8}, \tag{82}$$

where according to our settings for γ and $\lambda \geq \frac{(25)^{\frac{1}{p}}\sigma}{(\rho)^{\frac{1}{p}}(mK)^{\frac{5p-4}{3p(3p-2)}}T^{\frac{10-4p}{3p(3p-2)}}}$, we have $100\mathcal{R}'\sigma^p\frac{\lambda^{2-p}\gamma T}{b} \leq \frac{\Delta'_1}{8}$.

Theorem 6. *In summary, we choose*

$$\begin{aligned}\lambda = & \max\{(10)^{\frac{2}{p}}\sigma(mKT)^{\frac{1}{3(3p-2)}}, \frac{(25)^{\frac{1}{p}}\sigma}{(\rho b)^{\frac{1}{p}}(mK)^{\frac{5p-4}{3p(3p-2)}}T^{\frac{10-4p}{3p(3p-2)}}}, \frac{(5L_1\Delta'_1)^{\frac{2}{p+2}}\sigma^{\frac{p}{p+2}}(mKT)^{\frac{p}{3(p+2)(3p-2)}}}{(8b)^{\frac{2}{p+2}}},\\ & \frac{36^{\frac{1}{2p}}\sigma^2(mKT)^{\frac{1}{3(3p-2)}}}{\rho^{\frac{1}{2p}}}, 4R', \frac{(12L_1R')^{\frac{1}{p+1}}\sigma^{\frac{p}{p+1}}(mKT)^{\frac{1}{3(p+1)}}}{(L_0\rho b)^{\frac{1}{p+1}}}\}\\ = & \mathcal{O}((mKT)^{\frac{1}{3(3p-2)}}/\rho^{\frac{1}{2p}}),\end{aligned} \tag{83}$$

$$b = \mathcal{O}((mKT)^{\frac{2p-2}{3(3p-2)}}), \tag{84}$$

$$\begin{aligned}\gamma &= \frac{\beta\Delta'_1}{32\mathcal{R}'\lambda^2(mKT)^{\frac{1}{3}}\rho} = \mathcal{O}((mKT)^{\frac{-p}{3p-2}}/\rho^{\frac{p-1}{p}}),\\ \eta &= \frac{1}{2L_0+L_1\lambda}\min\{\frac{1}{K}, \frac{1}{K\lambda}, \frac{\sigma^{\frac{p}{2}}}{K\sqrt{b}\lambda}\} \leq \frac{1}{2L_0+L_1\lambda}\frac{\rho^{\frac{1}{2p}}}{K(mKT)^{\frac{3p}{4(3p-2)}}} \leq \mathcal{O}(\frac{\log^{\frac{1}{p}}\frac{T}{\delta}}{K^{\frac{17p-8}{4(3p-2)}}(mT)^{\frac{5p}{4(3p-2)}}}),\end{aligned} \tag{85}$$

Summing the descent lemma from 1 to T, we have

$$\begin{aligned}&\sum_{t=1}^{T}\frac{\gamma mK}{2}\|\nabla F_S(x_t)\|^2 + \Delta_{T+1} \leq \Delta_1 + \gamma mK\sum_{t=1}^{T}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2])\\ &+ \frac{L_1}{L_0}\gamma mK\sum_{t=1}^{T}\|\nabla F_S(x_t)\|\langle\epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma mK\sum_{t=1}^{T}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \sum_{t=1}^{T}\frac{L_1\gamma^2m^2K^2}{2}\|\nabla F_S(x_t)\|^3\\ &+ \frac{L_1}{L_0}\gamma mK\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\| + \gamma mK\sum_{t=1}^{T}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2] \leq 2\Delta_1\end{aligned} \tag{86}$$

holding with the probability at least $\Pr(E_N \cap E_A \cap E_B \cap E_{C_1} \cap E_{C_2} \cap E_{C_3} \cap E_{C_4}) = (1 - \frac{(N-1)\delta}{T})(1-\frac{\delta}{2T})(1-\frac{\delta}{2T})\cdot 1 \cdot 1 \geq 1 - \frac{N\delta}{T}$. *Thus, the* `FedCBG` *algorithm (i.e., Algorithm 2) can achieve the convergence rate*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2 \leq \mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{T}{\delta}) \tag{87}$$

with the probability at least $1-\delta$ and find an ϵ-stationary point.

$\square$

C.2 High Probability Generalization Bound

Theorem 7. *Suppose Assumptions 1, 2 and 3 hold. Let $\{x_t\}$ be the sequence produced by Algorithm 2 with the local step size η, clipping parameter λ, global learning rate γ and batch size b set to be the same as in Theorem 1. If we choose $T = \mathcal{O}(\sqrt{\frac{n}{d}})$, then with probability $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 \le \mathcal{O}\left(m^{\frac{8p-3}{3p-2}}\left(\frac{d}{n}\right)^{\frac{2p-2}{10p-5}}\log\left(\frac{1}{\delta}\left(\frac{n}{d}\right)^{\frac{3p-2}{2(11p-6)}}\right)\right) \tag{88}$$

for any $\delta \in (0,1)$.

Proof. Since $x_{t+1} - x_t = \gamma\sum_{i=1}^{m}\sum_{k=0}^{K-1}\widetilde{\nabla} f_i(x_{t,i}^k;\boldsymbol{\xi}_{t,i}^k)$, we have

$$\begin{aligned}
&\|x_{t+1}\| \\
&= \gamma\|\sum_{l=1}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k)\| = \gamma\|\sum_{l=1}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \mathbb{E}_{j_l}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) \\
&\quad + \mathbb{E}_{j_l}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \nabla F_S(x_l) + \nabla F_S(x_l)\| \\
&\le \underbrace{\gamma\|\sum_{l=1}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \mathbb{E}_{j_l}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k)\|}_{A_1} + \underbrace{\gamma\|\sum_{l=0}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_{j_l}\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \nabla F_S(x_l)\|}_{A_2} \\
&\quad + \underbrace{\gamma\|\sum_{l=0}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_S(x_l)\|}_{A_3}.
\end{aligned} \tag{89}$$

Upper Bound for A_1

For A_1, it is clear that the sequence $\{\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \mathbb{E}_t\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k)\}$ is a martingale difference sequence. With probability 1, there is $|\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \mathbb{E}_t\widetilde{\nabla} f_i(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k)| \le 2\lambda$. Thus, $A_1 \le \gamma\sum_{l=0}^{t}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\|\widetilde{\nabla} f(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \mathbb{E}_t\widetilde{\nabla} f(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k)\| \le 2\gamma\lambda tmK$.

Upper Bound for A_2

According to Lemma 10, we know that

$$\begin{aligned}
&\gamma mK\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)(\|\epsilon_t^a\|^2 - \mathbb{E}_t[\|\epsilon_t^a\|^2]) \\
&+ \frac{L_1}{L_0}\gamma mK\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|\langle\epsilon_t^a, \nabla F_S(x_t)\rangle + \gamma mK\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\|\epsilon_t^b\|^2 + \sum_{t=1}^{l-1}\frac{L_1\gamma^2 m^2K^2}{2}\|\nabla F_S(x_t)\|^3 \\
&+ \frac{L_1}{L_0}\gamma mK\sum_{t=1}^{l-1}\|\nabla F_S(x_t)\|^2\|\epsilon_t^b\| + \gamma mK\sum_{t=1}^{l-1}(1+\frac{L_1}{L_0}\|\nabla F_S(x_t)\|)\mathbb{E}_t[\|\epsilon_t^a\|^2] \le \Delta_1'
\end{aligned} \tag{90}$$

holds with the probability of $1-\frac{(t-1)\delta}{T}$. Thus, we know that $\Delta_t' \le 2\Delta_1'$ holds for $0 \le t \le T$ with the probability $1-\frac{(t-1)\delta}{T}$. According to our setting for λ in Theorem 3, we know that $\|\nabla F_S(x_t)\| \le \frac{\lambda}{2}$ and $\|\nabla f_i(x_t)\| \le \frac{\lambda}{2}$ also holds with the probability $1-\frac{(t-1)\delta}{T}$. Thus, according to Lemma 8, $A_2 \le \gamma\sum_{l=0}^{t}\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}_t\widetilde{\nabla} f(x_{l,i}^k;\boldsymbol{\xi}_{l,i}^k) - \nabla F_S(x_l)\| \le \gamma tmK\frac{12\sigma^p\lambda^{1-p}}{b}$.

Upper Bound for A_3

Besides, we have the following inequality with probability at least $1-\delta$

$$\begin{aligned}\|\sum_{l=0}^{t}\nabla F_S(x_l)\|^2 \le & \left(\sum_{l=0}^{t}\|\nabla F_S(x_l)\|\right)^2 \le \left(\sum_{l=1}^{t}\right)\left(\sum_{l=0}^{t}\|\nabla F_S(x_l)\|^2\right)\\ & \le t^2\mathcal{O}\left((mK)^{\frac{2-2p}{3p-2}}t^{\frac{p}{3p-2}}\log^{\frac{p-1}{p}}\frac{t}{\delta}\right) \le \mathcal{O}\left((mK)^{\frac{2-2p}{3p-2}}t^{\frac{7p-4}{3p-2}}(\log\frac{t}{\delta})^{\frac{p-1}{p}}\right),\end{aligned} \tag{91}$$

where the second inequality holds due to the Cauchy-Schwarz inequality and the third inequality holds due to the Theorem 3. Thus, with the probability $1-\delta$,

$$\|\sum_{l=0}^{t}\nabla F_S(x_l)\| = \mathcal{O}\left((mK)^{\frac{1-p}{3p-2}}t^{\frac{3.5p-2}{3p-2}}\log^{\frac{p-1}{2p}}\frac{t}{\delta}\right). \tag{92}$$

Combining these bounds, we have the following inequality

$$\|x_{t+1}\| \le A_1 + A_2 + A_3 \le 2\gamma\lambda tmK + 12\gamma tmK\frac{\sigma^p\lambda^{1-p}}{b} + \gamma\mathcal{O}\left((mK)^{\frac{1-p}{3p-2}}t^{\frac{3.5p-2}{3p-2}}\log^{\frac{p-1}{2p}}\frac{t}{\delta}\right) \tag{93}$$

holding with the probability of $(1-\frac{(t-1)\delta}{T})(1-\delta) \ge 1-2\delta$. Since we choose the step size $\gamma = \mathcal{O}((mKT)^{\frac{-p}{3p-2}}/\rho^{\frac{p-1}{p}})$, $\lambda = \mathcal{O}((mKT)^{\frac{1}{2(3p-2)}}/\rho^{\frac{1}{2p}})$, we have

$$\begin{aligned}\|x_{t+1}\| &\le \mathcal{O}((mKT)^{\frac{4p-3}{2(3p-2)}}/\rho^{\frac{2p-1}{2p}}) + \mathcal{O}((mKT)^{\frac{2p-3}{2(3p-2)}}/\rho^{\frac{p-1}{2p}}) + \mathcal{O}((mK)^{\frac{1-2p}{3p-2}}T^{\frac{2.5p-2}{3p-2}}/\rho^{\frac{p-1}{2p}})\\ &\le \mathcal{O}((mK)^{\frac{2p-1.5}{3p-2}}T^{\frac{2.5p-2}{3p-2}}/\rho^{\frac{p-1}{2p}}).\end{aligned} \tag{94}$$

Thus, we know that x_t traverses a sphere, and as t increases, the radius of the sphere increases. We need to replace R in Lemma 5 with $\|x_{t+1}\|$. Plugging the bound of $\|x_{t+1}\|$ into Lemma 5, we have the following inequality

$$\begin{aligned}\|\nabla F(x_{t+1}) - \nabla F_S(x_{t+1})\| &\le (L_0\|x_{t+1}\| + (L_1\|x_{t+1}\|+1)\widetilde{b})n^{-\frac{1}{2}}\mu\\ &\le \mu\frac{L_0\mathcal{O}(\frac{(mK)^{\frac{2p-1.5}{3p-2}}T^{\frac{2.5p-2}{3p-2}}}{\log^{\frac{p-1}{2p}}\frac{T}{\delta}}) + (L_1\mathcal{O}(\frac{(mK)^{\frac{2p-1.5}{3p-2}}T^{\frac{2.5p-2}{3p-2}}}{\log^{\frac{p-1}{2p}}\frac{T}{\delta}})+1)\widetilde{b}}{\sqrt{n}}\end{aligned} \tag{95}$$

holding with the probability of $1-2\delta$, where $\mu = \left(2+2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log\frac{1}{\delta}}\right)$. This bound also means that the following inequality

$$\|\nabla F(x_{t+1}) - \nabla F_S(x_{t+1})\|^2 = \mathcal{O}\left(\frac{(mK)^{\frac{4p-3}{3p-2}}T^{\frac{6p-4}{3p-2}}}{n\log^{\frac{p-1}{p}}\frac{T}{\delta}}\left(d+\log\frac{1}{\delta}\right)\right) \tag{96}$$

holding with the probability $1-2\delta$. Finally, we begin to bound $\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2$. Firstly,

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 &\le 2\sum_{t=1}^{T}\|\nabla F(x_t) - \nabla F_S(x_t)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2\\ &\le 2T\max_{1\le t\le T}\|\nabla F(x_t) - \nabla F_S(x_t)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2\\ &= 2T\|\nabla F(x_T) - \nabla F_S(x_T)\|^2 + 2\sum_{t=1}^{T}\|\nabla F_S(x_t)\|^2,\end{aligned} \tag{97}$$

where the last inequality holds due to the fact that $\|\nabla F(x_t) - \nabla F_S(x_t)\|^2$ increases as t increases. Combining inequalities (96), (97) and Theorem 1, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 &= \mathcal{O}\left(\frac{(mK)^{\frac{4p-3}{3p-2}}T^{\frac{5p-4}{3p-2}}}{n\log^{\frac{p-1}{p}}\frac{T}{\delta}}\left(d+\log\frac{1}{\delta}\right)\right) + \mathcal{O}((mKT)^{\frac{2-2p}{3p-2}}\log^{\frac{p-1}{p}}\frac{T}{\delta})\\ &\le \mathcal{O}\left(\left(\frac{d}{n}\right)^{\frac{p-1}{7p-6}}\log\left(\frac{1}{\delta}\left(\frac{n}{d}\right)^{\frac{3p-2}{2(7p-6)}}\right)\right)\end{aligned} \tag{98}$$

holding with the probability $(1 - 2\delta)$, where we choose $T = \mathcal{O}\left(\left(\frac{n}{d}\right)^{\frac{3p-2}{2(7p-6)}} / m^{\frac{6p-5}{7p-6}}\right)$, $K = \mathcal{O}\left(\left(\frac{n}{d}\right)^{\frac{3p-2}{2(6p-5)}}\right)$ in the last inequality. $\square$

D More Experimental Results

Firstly, we provide a brief description of the dataset and the model structure. Secondly, we verified the heavy-tailed condition. Thirdly, we conducted ablation experiments on several hyperparameters. Finally, we verify the performance of our `FedCBG` method on unseen datasets.

D.1 Dataset and model

We use the Shakespeare dataset with the same train/test splits as in previous work [43]. For the text character prediction, we use a stacked LSTM, similar to [28] as the classification model.

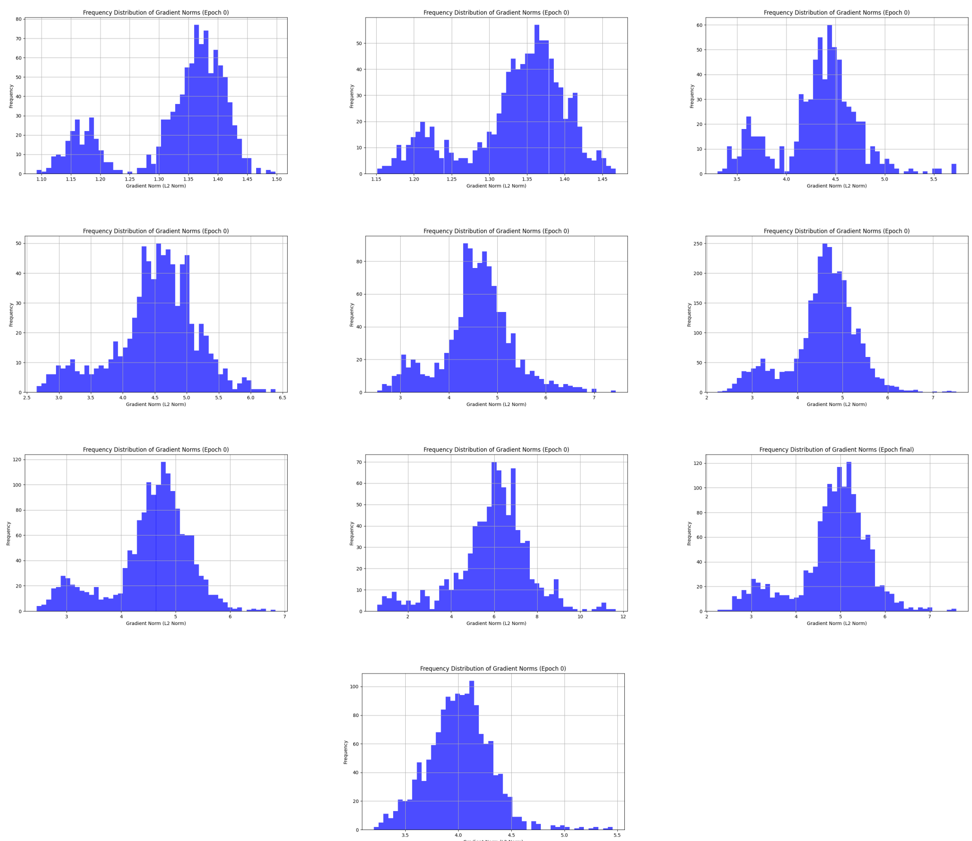

Figure 4: Distributions of the gradient noises during training for LSTM on the Shakespeare dataset at $T = 0$ with $m = 139$ clients participate in the training.

D.2 Heavy-tailed noise condition

We verify the heavy-tailed noise assumption 3 by training the text model on the Shakespeare dataset. We extracted the gradient noise vectors of the first epoch and the final epoch from 10 random clients and visualized their distributions as shown in Figs. 4 and 5, respectively. It can be seen that the noise follows a heavy-tailed distribution rather than a light-tailed Gaussian distribution. Combined with Table 3, it shows that our `FedCBG` algorithm does generalize better than existing algorithms under heavy-tailed noise.

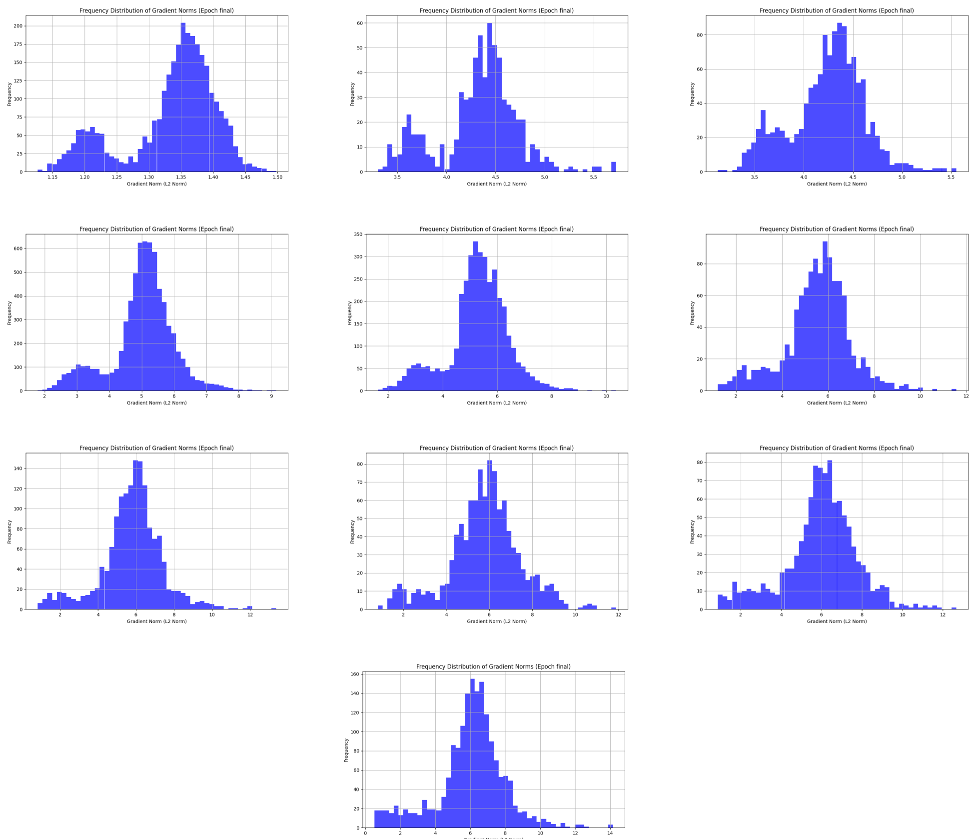

Figure 5: Distributions of the gradient noises during training for LSTM on the Shakespeare dataset at $T = 100$ with $m = 139$ clients participate in the training.

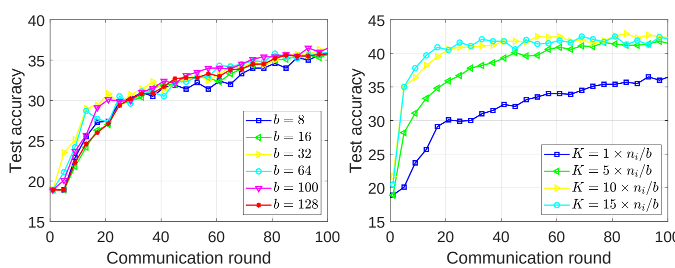

Figure 6: Selection of hyperparameters: batch size b (left) and the number of local updates K (right) on the Shakespeare dataset for our `FedCBG` algorithm.

D.3 Hyperparameter Selection

We conducted ablation experiments on hyperparameters γ, λ, batch size b and the number of local updates K for our `FedCBG` algorithm. During training, the test accuracy curve is shown in Figs. 3 and 7. When $\gamma = 0.2$ or 0.3, our `FedCBG` algorithm performs significantly better than other choices. The performance of $\lambda = 3.0$, $b = 100$ and $K = 10 \times n_i/b$ exceeds that of other values. For a fair comparison, we also conducted ablation experiments for FAT-Clipping-PR and FAT-Clipping-PI algorithms and the results are shown in Figs. 8 and 9, respectively.

For a fair comparison, we conduct comparative experiments again based on the hyperparameters in Table 6. The experimental results are shown in Table 7.

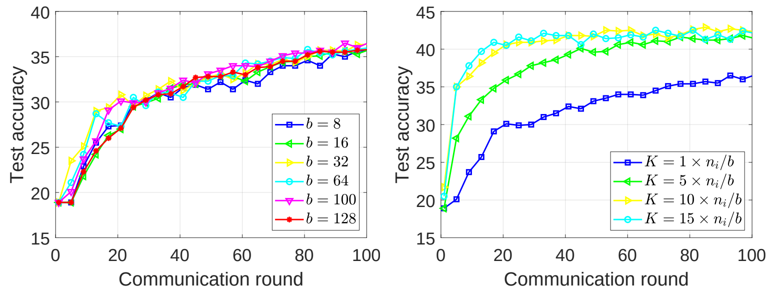

Figure 7: Selection of hyperparameters: mini-batch b (left) and local iterations K (right) on the Shakespeare dataset for our `FedCBG` algorithm.

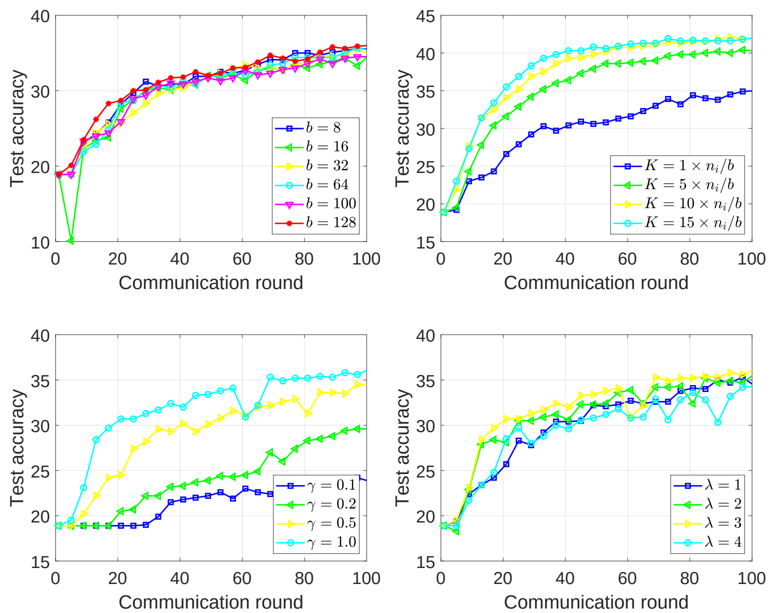

Figure 8: Selection of hyperparameters: mini-batch b, local iterations K, global learning rate γ and clipping parameter λ on the Shakespeare dataset for the FAT-clipping-PR algorithm.

Table 6: The hyperparameter settings for each algorithm are based on the fine-tuning experiments in Figs. 7, 9 and 8.

hyper-parameters / Algorithms	b	K	γ	λ
FAT-Clipping-PR	100	$10 \times n_i/b$	1.0	3.0
FAT-Clipping-PI	100	$10 \times n_i/b$	1.0	4.0
FedCBG (Ours)	100	$10 \times n_i/b$	0.3	3.0

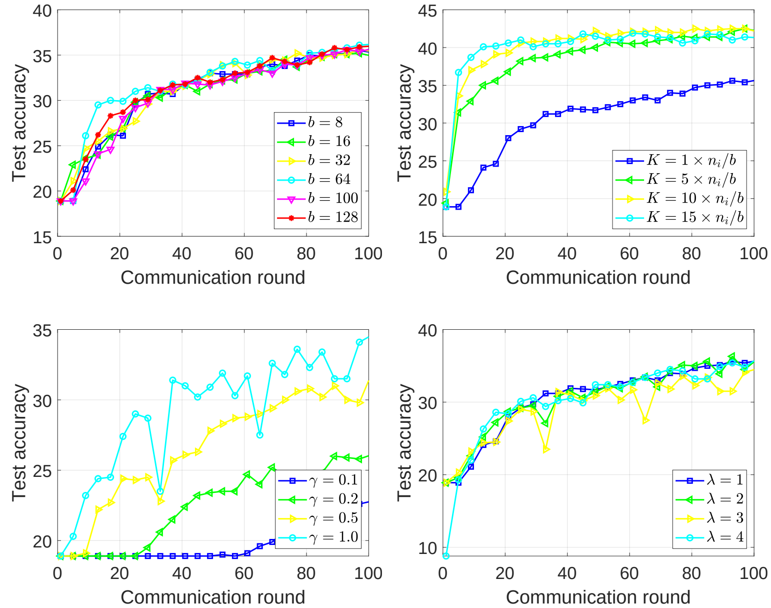

Figure 9: Selection of hyperparameters: mini-batch b, local iterations K, global learning rate γ and clipping parameter λ on the Shakespeare dataset for the FAT-clipping-PI algorithm.

Table 7: Comparison of the training loss (TLoss.), testing classification accuracy (TAcc.) and communication round to reach target test accuracy in FL with heavy-tailed noise under hyper-parameter in Table 6. "Round" refers to the number of communication rounds to achieve test accuracy 84.5% for CIFAR-10, 46.0% for CIFAR-100 and 42.0% for Shakespeare datasets.

Datasets	Evaluation	CIFAR-10	CIFAR-100	Shakespeare
FAT-Clipping-PR	TLoss	0.15	3.05	2.08
	TAcc. (%)	83.6	43.8	42.2
	Round	245 (2.9×)	296 (1.8×)	64 (1.4×)
FAT-Clipping-PI	TLoss	0.08	2.54	2.06
	TAcc. (%)	85.1	44.2	42.1
	Round	171 (2.0×)	235 (1.4×)	58 (1.3×)
`FedCBG` (Ours)	TLoss	**0.06**	**1.70**	**2.03**
	TAcc. (%)	**85.9**	**45.8**	**42.7**
	Round	**85**	**166**	**46**

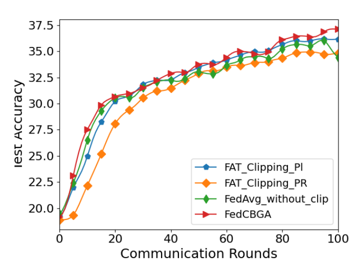

Figure 10: Comparison of test accuracy during training by FAT-clipping-PR, FAT-clipping-PI, FedAvg without clipping and our `FedCBG` algorithms optimizing text model on the Shakespeare dataset.

D.4 Generalization Performance on Unseen Datasets.

In the federated setting, we also conducted experiments to verify the generalization ability of our `FedCBG` algorithm on unseen datasets. We set the local learning rate $\eta = 1$, the clipping parameter $\lambda = 3.0$, global learning rate $\gamma = 0.3$ and $m = 139$ to train the model. Then, we test the well-trained model on testing datasets. For a fair comparison, we set $b = 100$ for all the methods.

We compare the testing accuracy during training and the results are shown in Fig. 10. It can be seen that our `FedCBG` algorithm still generalize better than the existing heavy-tailed algorithm on the unseen testing datasets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. Please see the sections Abstract and Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations of the work. Please see Section Conclusions and Future Work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides the full set of assumptions and a complete (and correct) proof in Sections 3, 4 and 5.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Please see the section Experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please see the section Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please see the section Experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see the section Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is of a theoretical nature. The main contribution is to weaken the conditions. There are no positive and negative societal impacts from this work, and as a result we did not discuss it.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used public datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We propose a new analytical method.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.