Taming Fat-Tailed ("Heavier-Tailed" with Potentially Infinite Variance) Noise in Federated Learning

Haibo Yang
Dept. of ECE
The Ohio State University
Columbus, OH 43210
yang.5952@osu.edu

Peiwen Qiu
Dept. of ECE
The Ohio State University
Columbus, OH 43210
qiu.617@osu.edu

Jia Liu
Dept. of ECE
The Ohio State University
Columbus, OH 43210
liu@ece.osu.edu

Abstract

In recent years, federated learning (FL) has emerged as an important distributed machine learning paradigm to collaboratively learn a global model with multiple clients, while keeping data local and private. However, a key assumption in most existing works on FL algorithms' convergence analysis is that the noise in stochastic first-order information has a finite variance. Although this assumption covers all light-tailed (i.e., sub-exponential) and some heavy-tailed noise distributions (e.g., log-normal, Weibull, and some Pareto distributions), it fails for many fat-tailed noise distributions (i.e., "heavier-tailed" with potentially infinite variance) that have been empirically observed in the FL literature. To date, it remains unclear whether one can design convergent algorithms for FL systems that experience fat-tailed noise. This motivates us to fill this gap in this paper by proposing an algorithmic framework called FAT-Clipping (federated averaging with two-sided learning rates and clipping), which contains two variants: FAT-Clipping per-round (FAT-Clipping-PR) and FAT-Clipping per-iteration (FAT-Clipping-PI). Specifically, for the largest tail-index $\alpha \in (1, 2]$ such that the fat-tailed noise in FL still has a bounded α-moment, we show that both variants achieve $\mathcal{O}((mT)^{\frac{2-\alpha}{\alpha}})$ and $\mathcal{O}((mT)^{\frac{1-\alpha}{3\alpha-2}})$ convergence rates in the strongly-convex and general non-convex settings, respectively, where m and T are the numbers of clients and communication rounds. Moreover, with more clipping operations compared to FAT-Clipping-PR, FAT-Clipping-PI further enjoys a linear speedup effect with respect to the number of local updates at each client and being lower-bound-matching (i.e., order-optimal). Collectively, our results advance the understanding of designing efficient algorithms for FL systems that exhibit fat-tailed first-order oracle information.

1 Introduction

In recent years, federated learning (FL) has emerged as an important distributed machine learning paradigm, where, coordinated by a server, a set of clients collaboratively learn a global model, while keeping their training data local and private. With intensive research in recent years, researchers have developed many FL algorithms (e.g., FedAvg [1] and many follow-ups [2–12]) that have been theoretically shown to achieve fast convergence rates in the presence of various types of randomness and heterogeneity resulted from training data, network environments, computing resources at clients, etc. Moreover, many of these algorithms enjoy the so-called "linear speedup" effect, i.e., the convergence time to a first-order stationary point is inversely proportional to the number of workers and local update steps.

However, despite the recent advances in FL algorithm design and theoretical understanding, a "cloud that remains obscures the sky of FL" is a common assumption that can be found in almost all works

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

on performance analysis of FL algorithms, which states that the random noise in stochastic first-order oracles (e.g., stochastic gradients or associated estimators) has a *finite variance*. Although this assumption is not too restrictive and can cover all light-tailed (i.e., sub-exponential) and some heavy-tailed noise distributions (e.g., log-normal, Weibull, and some Pareto distributions), it fails for many "fat-tailed" distributions (i.e., "heavier-tailed" with potentially infinite variance[1]). In fact, fat-tailed distributions have already been empirically observed under centralized learning settings [15–19], let alone in the more heterogeneous FL environments. Later in Section 3, we will also provide empirical evidence that shows that fat-tailed noise distributions can be easily induced by FL systems with non-i.i.d. datasets and heterogeneous local updates across clients.

The presence of fat-tailed noise poses two major challenges in FL algorithm design and analysis: i) Experimentally, it has been shown in [20] that many existing FL algorithms suffer severely from fat-tailed noise and frequently exhibit the so-called "catastrophic failure of model performance" (i.e., sudden and dramatic drops of learning accuracy during the training phase); ii) Theoretically, the infinite variance of the random noise in the stochastic first-order oracles renders most of the proof techniques in existing FL algorithmic convergence analysis inapplicable, which necessitates new algorithmic ideas and proof strategies. In light of these empirical and theoretical challenges, two foundational questions naturally emerge in FL algorithm design and analysis: *1) Can we develop FL algorithms with convergence guarantee under fat-tailed noise? 2) If the answer to 1) is "yes," could we characterize their finite-time convergence rates?* In this paper, we provide affirmative answer to the above questions. Our major contributions in this paper are highlighted as follows:

- To address the challenges of the fat-tailed noise in FL algorithm design, we propose an algorithmic framework called FAT-Clipping (federated averaging with two-sided learning rates and clipping), which leverages a clipping technique to mitigate the impact of fat-tailed noise and uses a two-sided learning rate mechanism to lower communication complexity. Our FAT-Clipping framework contains two variants: FAT-Clipping per-round (FAT-Clipping-PR) and FAT-Clipping per-iteration (FAT-Clipping-PI). We show that, for the largest tail-index $\alpha \in (1, 2]$ such that the fat-tailed noise in FL still has a bounded α-moment, both FAT-Clipping variants achieve $\mathcal{O}((mT)^{\frac{2-\alpha}{\alpha}})$ and $\mathcal{O}((mT)^{\frac{1-\alpha}{3\alpha-2}})$ convergence rates in the strongly-convex and general non-convex settings, respectively, where m and T are the numbers of clients and communication rounds.

- Between the proposed FAT-Clipping variants, FAT-Clipping-PR only performs one clipping operation in each communication round before client communicates to the server, while FAT-Clipping-PI performs clipping in each iteration of local model update. We show that, at the expense of more clipping operations compared to FAT-Clipping-PR, FAT-Clipping-PI further achieves a linear speedup effect with respect to the number local model updates at each client and is lower-bound matching in terms of convergence rate.

- In addition to theoretical analysis, we also conduct extensive numerical experiments to study the fat-tailed phenomenon in FL systems and verify the efficacy of our proposed FAT-Clipping algorithms for FL systems with fat-tailed noise. We first provide concrete empirical evidence that fail-tailed noise distributions are not uncommon in FL systems with non-i.i.d. datasets and heterogeneous local updates. We show that our FAT-Clipping algorithms render a much smoother FL training process, which effectively prevents the "catastrophic failure" in various FL settings.

For quick reference and easy comparisons, we summarize all convergence rate results in Table 1. The rest of the paper is organized as follows. In Section 2, we review the literature to put our work in comparative perspectives. In Section 3, we provide empirical fat-tailed evidence for FL to further motivate this work. Section 4 presents our FAT-Clipping algorithms and their convergence analyses. Section 5 presents numerical results and Section 6 concludes this paper. Due to space limitation, all proof details and some experiments are provided in the supplementary material.

[1] In the literature, the terminologies "heavy-tailed" and "fat-tailed" are not universally defined and could be interchangeable sometimes. In this paper, we follow the convention of those authors who reserve the term "fat-tailed" to mean the subclass of heavy-tailed distributions that exhibit power law decay behavior as well as infinite variance (see, e.g., [13, 14]). Thus, every fat-tailed distribution is heavy-tailed, but the reverse is not true.

Table 1: Convergence rate comparisons under fat-tailed noise distributions (shaded parts are our results; metrics: $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon$ and $\|\nabla f(\mathbf{x})\| \leq \epsilon$ for strongly-convex and non-convex functions, respectively): $\alpha = 2$ and $\alpha \in (1, 2)$ correspond to non-fat-tailed and fat-tailed noises, respectively. Here, R is the total number of iterations for centralized algorithms (SGD and GClip); K and T are local update steps and communication rounds in the FL setting, respectively; m is the number of clients. N/A means no theoretical guarantee for convergence. Note that the total number of iterations R in FL can be computed as $R = KT$, which relates to that in the centralized setting.

Methods	Strongly Convex Objective Functions		Nonconvex Objective Functions	
	Fat-Tailed	Non-Fat-Tailed	Fat-Tailed	Non-Fat-Tailed
SGD [21]	N/A	$\mathcal{O}(R^{-1})$	N/A	$\mathcal{O}(R^{-\frac{1}{4}})$
GClip [22]	$\mathcal{O}(R^{\frac{2-2\alpha}{\alpha}})$	$\mathcal{O}(R^{-1})$	$\mathcal{O}(R^{\frac{1-\alpha}{3\alpha-2}})$	$\mathcal{O}(R^{-\frac{1}{4}})$
FedAvg [3, 7]	N/A	$\tilde{\mathcal{O}}((mKT)^{-1})$	N/A	$\mathcal{O}((mKT)^{-\frac{1}{4}})$
FAT-Clipping-PR	$\mathcal{O}((mT)^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}})$	$\tilde{\mathcal{O}}((mKT)^{-1})$	$\mathcal{O}((mT)^{\frac{1-\alpha}{3\alpha-2}} K^{\frac{2-\alpha}{3\alpha-2}})$	$\mathcal{O}((mKT)^{-\frac{1}{4}})$
FAT-Clipping-PI	$\tilde{\mathcal{O}}((mKT)^{\frac{2-2\alpha}{\alpha}})$	$\tilde{\mathcal{O}}((mKT)^{-1})$	$\mathcal{O}((mKT)^{\frac{1-\alpha}{3\alpha-2}})$	$\mathcal{O}((mKT)^{-\frac{1}{4}})$
Lower Bound	$\Omega((mKT)^{\frac{2-2\alpha}{\alpha}})$	$\Omega((mKT)^{-1})$	$\Omega((mKT)^{\frac{1-\alpha}{3\alpha-2}})$	$\Omega((mKT)^{-\frac{1}{4}})$

2 Related work

In this section, we will provide a quick overview on three related topics in the literature: i) federated learning, ii) heavy-tailed noise in learning, and iii) the clipping techniques, thus putting our work into comparative perspective to highlight our novelty and differences.

1) Federated Learning: As mentioned earlier, FL has recently emerged as an important distributed learning paradigm. The first and perhaps the most popular FL method, the federated averaging (FedAvg) algorithm [1], was initially proposed as a heuristic to improve communication efficiency and data privacy. Since then, FedAvg has sparked many follow-ups to further address the challenges of data/system heterogeneity and further reduce iteration and communication complexities. Notable approaches include adding regularization for the local loss function [2,5,6], using variance reduction techniques [3], taking adaptive learning rate strategy [8] or adaptive communication strategy [23], and many momentum variants [4,9,10]. Empirically, these algorithms are shown to be communication-efficient [1] and enjoy better generalization performance [24]. Moreover, many state-of-the-art algorithms enjoy the "linear speed" effect in terms of the numbers of clients and local update steps in different FL settings [3,7,25,26]. We note, however, that all these theoretical results are built upon the finite variance assumption of stochastic gradient noise. Unfortunately, when the stochastic gradient noise is fat-tailed, the finite variance assumption no longer holds, and hence the associated theoretical analysis is also invalid. This motivates us to fill this gap in this paper and conduct the first theoretical analysis for FL systems that experience fat-tailed noise.

2) Heavy-Tailed Noise in Learning: Recently, heavy-tailed noise has been empirically observed in modern machine learning systems and theoretically analyzed [15,16,18,22,27–30]. Heavy-tailed noise significantly affects the learning dynamics and computational complexity, such as the first exit time escaping from saddle point [27] and iteration complexity [22]. This is dramatically different from classic dynamic analysis often based on sub-Gaussian noise assumption [31,32] and algorithmic convergence analysis with bounded variance assumption [21,33]. However, for FL, there exist few investigations about heavy-tailed behaviors. In this paper, we first demonstrate through extensive experiments that fat-tailed (i.e., heavier-tailed) noise in FL can be easily induced by data heterogeneity and local update steps. We then propose efficient algorithms to mitigate the impacts of fat-tails.

3) The Clipping Technique: Since our FAT-Clipping algorithms are based on the idea of clipping, here we provide an overview on this technique. As far as we know, dating back to at least 1985 [34], gradient clipping has been an effective technique to ensure convergence for optimization problems with fast-growing objective functions. In deep learning, clipping is a widely adopted technique to address the exploding gradient problem. Recently, gradient clipping was theoretically shown to be able to accelerate the training of centralized learning [17,35–37]. Also, clipping is an effective approach to mitigate heavy-tailed noise [17,18] in centralized learning. In FL, clipping has been used as the preconditioning step for preserving differential privacy (DP) [38–40]. Unlike these works, in this paper, we utilize clipping to address algorithmic divergence caused by fat-tailed noise in FL.

Algorithm 1 Generalized FedAvg Algorithm (GFedAvg).

Initialize $\mathbf{x}_1$.
for $t = 1, \cdots, T$ (communication round) **do**
  for each client $i \in [m]$ in parallel **do**
    Update local model: $\mathbf{x}_{t,i}^1 = \mathbf{x}_t$.
    for $k = 1, \cdots, K$ (local update step) **do**
      Compute an unbiased estimate $\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ of $\nabla f_i(\mathbf{x}_{t,i}^k)$.
      Local update: $\mathbf{x}_{t,i}^{k+1} = \mathbf{x}_{t,i}^k - \eta_L \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
    end for
    Send $\Delta_t^i = \sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ to the server.
  end for
  Global Aggregation At Server:
    Receive $\Delta_t^i, i \in [m]$.
    Server Update: $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta\eta_L}{m} \sum_{i\in[m]} \Delta_t^i$.
    Broadcasting $\mathbf{x}_{t+1}$ to clients.
end for

3 Fat-tailed noise phenomenon in federated learning

In this section, we first introduce the basic FL problem statement and the standard FedAvg algorithm for FL. Then, we provide some necessary background of fat-tailed distributions and provide empirical evidence to show that fat-tailed noise can be easily induced by heterogeneity of data and local updates in FL, which further motivates this work. Lastly, we demonstrate the algorithmic divergence and frequently catastrophic model failure under fat-tailed noise.

1) Problem Statement of Federated Learning and the FedAvg Algorithm: The goal of FL is to solve the following optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{1}$$

where m is the number of clients and $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i\sim D_i}[f_i(\mathbf{x}, \xi_i)]$ is the local loss function associated with a local data distribution D_i. A key challenge in FL stems from data heterogeneity, i.e., $D_i \neq D_j, \forall i \neq j$. In FL, the standard and perhaps the most popular algorithm is the federated averaging (FedAvg) method. Here in Algorithm 1, we illustrate a more generalized version of the original FedAvg (GFedAvg) with separate learning rates on the client and server sides [3,7,8]. Note that when $\eta = 1$, GFedAvg reduces to the original FedAvg [1]. In each communication round of GFedAvg, each client performs local update steps and returns the update difference Δ_t^i. The server then aggregates these results and update the global model [2] and the updated model parameters will then be retrieved by the clients to start the next round of local updates.

2) Empirical Evidence of Fat-Tailed Noise Phenomenon in Federated Learning: With the basics of FL and the FedAvg algorithm, we are now in a position to demonstrate the empirical evidence of the existence of fat-tailed noise in FL systems. As mentioned earlier, in most performance analyses of FL algorithms, a common assumption is the bounded variance assumption of the local stochastic gradients: $\mathbb{E}[\|\nabla f_i(\mathbf{x}, \xi) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$. This assumption holds for all light-tailed noise distributions (i.e., the sub-exponential family) and some heavy-tailed distributions (e.g., log-normal, Weibull, and some Pareto distributions).

However, the finite-variance assumption fails to hold for many fat-tailed noise distributions. For instance, for a random variable X, if its density $p(x)$ has a power-law tail decreasing as $1/|x|^{\alpha+1}$ with $\alpha \in (0, 2)$, then only the α-moment of this noise exists with $\alpha < 2$. To more precisely characterize fat-tailed distributions, in this paper, we adopt the notion of tail-index α [15] to parameterize fat-tailed and heavy-tailed distributions. More specifically, if the density of a random variable X's distribution decays with a power law tail as $1/|x|^{\alpha+1}$ where $\alpha \in (0, 2]$, then α is called the *tail-infex*. This α-parameter determines the behavior of the distribution: the smaller the α-value, the heavier the tail

[2] We assume all clients participate in the training at each communication round, but the results can be extended to that with (uniformly random sampled) subset of clients in each communication round [3,7].

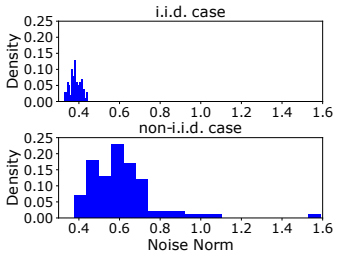


Figure 1: Distributions of the norms of the pseudo-gradient noises computed with CNN on CIFAR-10 dataset in i.i.d. case (top) and non-i.i.d. case (bottom). $m = 100$ clients participate in the training.

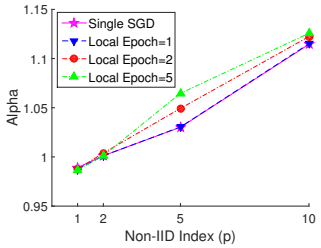


Figure 2: Estimation of α for CIFAR-10 dataset. The non-IID index p represents the data heterogeneity level, and $p = 10$ is the IID case. The smaller the p, the more heterogeneous the data across clients.

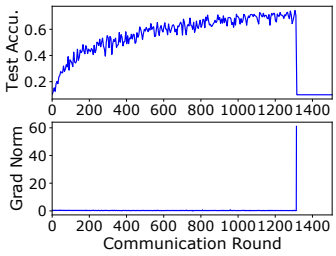


Figure 3: Catastrophic training failures happen when applying GFedAvg on CIFAR-10 dataset, where the test accuracy experiences a sudden and dramatic drop and the pseudo-gradient norm increases substantially.

of the distribution. Also, the α-parameter also determines the moments: $\mathbb{E}[|X|^r] < \infty$ if and only if $r < \alpha$, which implies that X has infinite variance when $\alpha < 2$, i.e., being *fat-tailed*.

Next, we investigate the tail property of model updates returned by clients in the GFedAvg algorithm. Due to multiple local steps in the GFedAvg algorithm, we view the whole update vector Δ_t^i returned by each client, which we called "pseudo-gradient," as a random vector and then analyze its statistical properties. Note that in the special case with the number of local update $K = 1$, Δ_t^i coincides with a single stochastic gradient of a random sample, (i.e., $\Delta_t^i = \nabla f_i(\mathbf{x}_t, \xi_t)$).

We study the mismatch between the "non-fat-tailed" condition ($\alpha = 2$) and the empirical behavior of the stochastic psudo-gradient noise. In Fig. 1, we illustrate the distributions of the norms of the stochastic pseudo-gradient noises computed with convolutional neural network (CNN) on the CIFAR-10 dataset in both i.i.d. and non-i.i.d. client dataset settings. We can clearly observe that the non-i.i.d. case exhibits a rather fat-tailed behavior, where the pseudo-gradient norm could be as large as 1.6. Although the i.i.d. case appears to have a much lighter tail, our detailed analysis shows that it still exhibits a fat-tailed behavior. To see this, in Fig. 2, we estimate α-value for the CIFAR-10 dataset in different scenarios: 1) different local update steps, and 2) different data heterogeneity. We use a parameter p to characterize the data heterogeneity level, with $p = 10$ corresponding to the i.i.d. case. The smaller the p, the more heterogeneous the data among clients. Fig. 2 shows that the α-value is smaller than 1.15 in all scenarios, and α increases as the non-i.i.d. index p increases (i.e., closer to the i.i.d. case). This implies that the stochastic pseudo gradient noise is fat-tailed and the "fatness" increases as the clients' data become more heterogeneous.

3) The Impacts of Fat-Tailed Noise on Federated Learning: Next, we show that the fat-tailed noise could lead to a "catastrophic model failure" (i.e., a sudden and dramatic drop of learning accuracy), consistent with previous observations in the FL literature [20]. To demonstrate this, we apply GFedAvg on the CIFAR-10 dataset and randomly sample five clients among $m = 10$ clients in each communication round. In Fig. 3, we illustrate a trial where a catastrophic training failure occurred. Correspondingly, we can observe in Fig. 3 a spike in the norm of the pseudo-gradient. This exceedingly large pseudo-gradient norm motivates us to apply the clipping technique to curtail the gradient updates. It is also worth noting that even if the squared norm of stochastic gradient may not be infinitely large in practice (i.e., having a bounded support empirically), it could still be too large to cause catastrophic model failures. In fact, under fat-tailed noise, the FedAvg algorithm could *diverge*, which follows from the fact that there exists one function that SGD diverges under heavy-tailed noise (see Remark 1 in [22]). As a result, the returned value by one client might be exceedingly large, leading to divergence of the FedAvg-type algorithms.

It is worth pointing out that, although we have empirically shown heavy/fat-tailed noise in FL for the first time in this paper, we are by no means the only one to have observed heavy-tailed or fat-tailed noise phenomenon property in learning. Previous works have also found heavy/fat-tailed noise phenomenon in centralized training with SGD-type algorithms. For example, the work in [15] showed the heavy-tailed noise phenomenon while (centralized) training the AlexNet on CIFAR-10. Here, we adopt a procedure similar to that in [15] to evaluate the tail index α of the noise norm distribution in

Algorithm 2 The FAT-Clipping-PR Algorithm.

Initialize $\mathbf{x}_1$.
for $t = 1, \cdots, T$ (communication round) **do**
  for each client $i \in [m]$ in parallel **do**
    Update local model: $\mathbf{x}_{t,i}^1 = \mathbf{x}_t$.
    for $k = 1, \cdots, K$ (local update step) **do**
      Compute an unbiased estimate $\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ of $\nabla f_i(\mathbf{x}_{t,i}^k)$.
      Local update: $\mathbf{x}_{t,i}^{k+1} = \mathbf{x}_{t,i}^k - \eta_L \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
    end for
    Let $\Delta_t^i = \sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
    Clipping: $\tilde{\Delta}_t^i = \min\{1, \frac{\lambda}{\|\Delta_t^i\|}\}\Delta_t^i$, where $\Delta_t^i = \sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
    Send $\tilde{\Delta}_t^i$ to the server.
  end for
  Global Aggregation At Server:
    Receive $\tilde{\Delta}_t^i, i \in [m]$.
    Server Update: $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta\eta_L}{m} \sum_{i\in[m]} \tilde{\Delta}_t^i$.
    Broadcasting $\mathbf{x}_{t+1}$ to clients.
end for

FL. As indicated above, we also observe that the (pseudo-)stochastic gradient noise is heavy/fat-tailed rather than Gaussian.

It is also worth noting that it remains controversial whether the heavy/fat-tailed noise phenomenon exists in all models and datasets. For example, the work in [19] showed that the stochastic gradient noise is Gaussian at least in the early phases of training, while [44] showed that the stationary distribution of stochastic gradient noise is heavy-tailed and state-dependent. Also, the evaluation methodologies of α could be different in different works with different statistical errors, thus leading to different observations [19, 22]. We believe that the phenomenon of heavy/fat-tailed noise in training with SGD-type methods is an under-explored area that deserves more efforts from the community.

To conclude this section, we would also like to leave a caveat regarding catastrophic training failures. In this section, we have shown that, under heavy/fat-tailed noises, catastrophic training failures happen in FL training, which is consistent with the observations in large-cohort FL training [20]. However, this does not necessarily mean that all FL trainings will suffer from catastrophic failures. Sometimes, such catastrophic failures may not happen at all (see the appendix for such empirical evidence). Here, we hypothesize that the heavy/fat-tailed noise phenomenon in FL is highly correlated with catastrophic failures in FL. This is based on our subsequent observations that such catastrophic failures in FL can be effectively mitigated by employing clipping methods. However, whether or not the heavy/fat-tailed noise phenomenon is truly the culprit for catastrophic failures still needs further investigations. Nonetheless, the mere existence of such a correlation between heavy/fat-tailed noise and catastrophic failures in FL warrants our study on mitigating heavy/fat-tailed noise in this paper.

4 The FAT-Clipping algorithmic framework for fat-tailed federated learning

Given the evidence of fat-tailed noise in FL and its potential catastrophic training failure as shown in Section 3, there is a compelling need to design an efficient FL algorithm with provable convergence guarantee under fat-tailed noise in FL. Interestingly, the observation of an exceedingly large pseudo-gradient norm in Fig. 3 suggests a natural idea to mitigate fat-tailed noise: *clipping*. Toward this end, in Section 4.1 we first propose a clipping-based algorithmic framework called FAT-Clipping, which contains two variants: FAT-Clipping per-round (FAT-Clipping-PR) and FAT-Clipping per-iteration (FAT-Clipping-PI). Then in Section 4.2, we analyze their convergence rate performance.

4.1 The FAT-Clipping-PR and FAT-Clipping-PI algorithms

We illustrate the FAT-Clipping-PR and FAT-Clipping-PI algorithms in Algorithms 2 and 3, respectively. It can be seen that both FAT-Clipping-PR and FAT-Clipping-PI share a similar algorithmic

Algorithm 3 The FAT-Clipping-PI Algorithm.

1: Initialize $\mathbf{x}_1$.
2: **for** $t = 1, \cdots, T$ (communication round) **do**
3: **for** each client $i \in [m]$ in parallel **do**
4: Update local model: $\mathbf{x}_{t,i}^1 = \mathbf{x}_t$.
5: **for** $k = 1, \cdots, K$ (local update step) **do**
6: Compute an unbiased estimate $\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ of $\nabla f_i(\mathbf{x}_{t,i}^k)$.
7: **Clipping:** $\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) = \min\{1, \frac{\lambda}{\|\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|}\} \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
8: Local update: $\mathbf{x}_{t,i}^{k+1} = \mathbf{x}_{t,i}^k - \eta_L \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$.
9: **end for**
10: Send $\tilde{\Delta}_t^i = \sum_{k\in[K]} \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ to the server.
11: **end for**
12: Global Aggregation At Server:
13: Receive $\tilde{\Delta}_t^i, i \in [m]$.
14: Server Update: $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta\eta_L}{m} \sum_{i\in[m]} \tilde{\Delta}_t^i$.
15: Broadcasting $\mathbf{x}_{t+1}$ to clients.
16: **end for**

structure with GFedAvg, with the key differences lying in the additional clipping operations. In FAT-Clipping-PR, each client performs a clipping in each communication round on the returned Δ_t^i:

$$\tilde{\Delta}_t^i = \min\left\{1, \frac{\lambda}{\|\Delta_t^i\|}\right\} \Delta_t^i, \tag{2}$$

and then sends $\tilde{\Delta}_t^i$ instead of Δ_t^i to the server (Line 10 in Algorithm 2). By contrast, in FAT-Clipping-PI, each client clips the stochastic gradient before each local update step (Line 7 in Algorithm 3):

$$\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) = \min\left\{1, \frac{\lambda}{\|\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|}\right\} \nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k), \tag{3}$$

$$\mathbf{x}_{t,k+1}^i = \mathbf{x}_{t,i}^k - \eta_L \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k). \tag{4}$$

Then, $\tilde{\Delta}_t^i = \sum_{k\in[K]} \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)$ is sent to the server for aggregation (Line 10 in Algorithm 3).

4.2 Convergence analysis of the FAT-Clipping algorithms

Before conducting the convergence analysis for the FAT-Clipping algorithms, we first state two standard assumptions that are commonly used in the literature of first-order stochastic methods.

Assumption 1 (L-Lipschitz Continuous Gradient)**.** *There exists a constant $L > 0$, such that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and $i \in [m]$.*

Assumption 2 (Unbiased Local Gradient Estimator)**.** *The local gradient estimator is unbiased, i.e., $\mathbb{E}[\nabla f_i(\mathbf{x}, \xi)] = \nabla f_i(\mathbf{x})$, $\forall i \in [m]$, where ξ is a random local data sample at the i-th worker.*

Next, we state the key bounded α-moment assumption for *fat-tailed* the stochastic first-order oracle, which leverages the notion of tail-index introduced in Section 3:

Assumption 3 (Bounded α-Moment)**.** *There exists a real number $\alpha \in (1, 2]$ and a constant $G \geq 0$, such that $\mathbb{E}[\|\nabla f_i(\mathbf{x}, \xi)\|^\alpha] \leq G^\alpha$, $\forall i \in [m], \mathbf{x} \in \mathbb{R}^d$.*

1) Convergence Rates of the FAT-Clipping-PR Algorithm: We first state the convergence rates of FAT-Clipping-PR for μ-strongly convex and non-convex objective functions.

Theorem 1. (Convergence Rate of FAT-Clipping-PR in the Strongly Convex Case) *Suppose that $f(\cdot)$ is a μ-strongly convex function. Under Assumptions 1–3, if $\eta\eta_L K \geq \frac{2}{\mu T}$, then the output $\bar{\mathbf{x}}_T$ of* FAT-Clipping-PR *being chosen in such a way that $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$, where $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$, satisfies:*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{\mu}{2} \exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} + \frac{4}{\mu}\left[2G^{2\alpha}\lambda^{2-2\alpha} + 2L^2\eta_L^2 K^2 G^\alpha \lambda^{2-\alpha}\right],$$

where $\mathbf{x}^*$ *denotes the global optimal solution. Further, let* $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$*, where* $c \geq 1$ *is a constant satisfying* $m^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}} T^{c+\frac{2-2\alpha}{\alpha}} \geq 1$*, and let* $\eta_L \leq (mKT)^{\frac{1-\alpha}{\alpha}}$*. It then follows that*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = \mathcal{O}((mT)^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}}).$$

Theorem 2. (Convergence Rate of FAT-Clipping-PR in the Nonconvex Case) *Suppose that* $f(\cdot)$ *is a nonconvex function. Under Assumptions 1–3, if* $\eta\eta_L KL \leq 1$*, then the sequence of outputs* $\{\mathbf{x}_k\}$ *generated by* FAT-Clipping-PR *satisfies:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1) - f(x_T)\right)}{\eta\eta_L KT} + \left(L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}\right) + \frac{L\eta\eta_L}{m}\left(KG^\alpha\lambda^{2-\alpha}\right).$$

Further, choosing learning rates and clipping parameter in such a way that $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}} K^{\frac{-\alpha-2}{3\alpha-2}} T^{\frac{-\alpha}{3\alpha-2}}, \eta_L \leq (mT)^{\frac{1-\alpha}{3\alpha-2}} K^{\frac{4-4\alpha}{3\alpha-2}}$*, and* $\lambda = (mK^4T)^{\frac{1}{3\alpha-2}}$*, we have*

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}((mT)^{\frac{2-2\alpha}{3\alpha-2}} K^{\frac{4-2\alpha}{3\alpha-2}}).$$

Remark 1. We note that the above convergence rates for FAT-Clipping-PR does not generalize the results of FedAvg when $\alpha = 2$ (non-fat-tailed noise). Specifically, FedAvg is able to achieve $\tilde{\mathcal{O}}((mKT)^{-1})$ and $\mathcal{O}((mKT)^{-\frac{1}{4}})$ convergence rates for strongly convex and non-convex function, respectively [3, 41]. In contrast, FAT-Clipping-PR achieves $\mathcal{O}((mT)^{-1}K)$ and $\mathcal{O}((mT)^{-\frac{1}{4}})$ for strongly-convex and non-convex functions, respectively. These two rates are consistent with those of FedAvg in terms of m and T, but not in terms of K.

Interestingly, with a separate proof for non-fat-tailed noise ($\alpha = 2$), we can show that clipping does not affect the dependence on K in the convergence rates. Thus, FAT-Clipping-PR has the *same* convergence rates as those of FedAvg. Due to space limitation, we state an informal version of these theorems here. The full versions of Theorem 5, 6 and their proofs are formally stated in Appendix.

Theorem 6 & **7 (informal)** (Convergence Rates of FAT-Clipping-PR for Non-Fat-Tailed Noise): *For* $\alpha = 2$*, CPR-FedAvg achieves convergence rate* $\tilde{\mathcal{O}}((mKT)^{-1})$ *for strongly-convex and* $\mathcal{O}((mKT)^{-\frac{1}{4}})$ *for non-convex functions, respectively.*

2) Convergence Rate of the FAT-Clipping-PI **Algorithm:** Next, we provide the convergence rates of FAT-Clipping-PI for μ-strongly convex and non-convex objective functions.

Theorem 3. (Convergence Rate of FAT-Clipping-PI in the Strongly Convex Case) *Suppose that* $f(\cdot)$ *is a* μ*-strongly convex function. Under Assumptions 1–3, if* $\eta\eta_L K \geq \frac{2}{\mu T}$*, then the output* $\bar{\mathbf{x}}_T$ *of* FAT-Clipping-PI *being chosen in such a way that* $\bar{\mathbf{x}}_T = \mathbf{x}_t$ *with probability* $\frac{w_t}{\sum_{j\in[T]} w_j}$*, where* $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$*, satisfies:*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2} G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2 K^2 G^\alpha\lambda^{2-\alpha}],$$

where $\mathbf{x}^*$ *denotes the global optimal solution. Further, let* $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$*, where* $c \geq 1$ *is a constant satisfying* $(mK)^{\frac{2-2\alpha}{\alpha}} T^{c+\frac{2-2\alpha}{\alpha}} \geq 1$*, and let* $\lambda = (mKT)^{\frac{1}{\alpha}}$*, and* $\eta_L \leq (mT)^{-\frac{1}{2}} K^{-\frac{3}{2}}$*). It then follows that*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = \tilde{\mathcal{O}}((mKT)^{\frac{2-2\alpha}{\alpha}}).$$

Theorem 4. (Convergence Rate of FAT-Clipping-PI in the Nonconvex Case) *Suppose that* $f(\cdot)$ *is a non-convex function. Under Assumptions 1–3, if* $\eta\eta_L KL \leq 1$*, then the sequence of outputs* $\{\mathbf{x}_k\}$ *generated by* FAT-Clipping-PI *satisfies:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1) - f(x_T)\right)}{\eta\eta_L KT} + \left(2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2 K^2 G^\alpha\lambda^{2-\alpha}\right) + \frac{L\eta\eta_L}{m}\left(G^\alpha\lambda^{2-\alpha}\right).$$

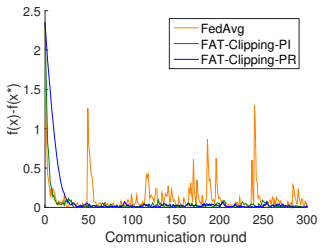


Figure 4: Convergence comparisons of FedAvg, FAT-Clipping-PI, and FAT-Clipping-PR for solving strongly convex models: synthetic data with ξ having Cauchy tails (fat).

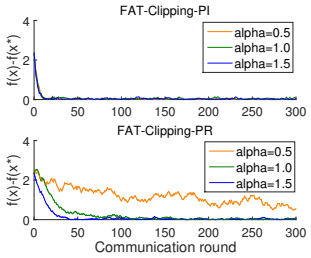


Figure 5: Convergence comparisons of FAT-Clipping-PI, and FAT-Clipping-PR for solving strongly convex models: synthetic data with ξ having different fat tails represented by α.

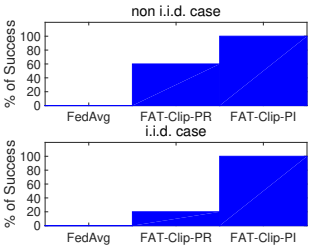


Figure 6: Percentage of successful training over 5 trials when applying FedAvg, FAT-Clipping-PR and FAT-Clipping-PI to CIFAR-10 dataset in non-i.i.d. case and i.i.d. case.

Further, choosing learning rates and clipping parameter in such a way that $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}}(KT)^{\frac{-\alpha}{3\alpha-2}}, \eta_L \leq (mKT)^{\frac{-\alpha}{6\alpha-4}}$, *and* $\lambda = (mKT)^{\frac{1}{3\alpha-2}}$, *we have*

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mKT)^{\frac{2-2\alpha}{3\alpha-2}}).$$

Remark 2. In comparison to FAT-Clipping-PR, convergence rates of FAT-Clipping-PI generalize the results of FedAvg for the non-fat-tailed noise case (i.e., $\alpha = 2$). Specifically, when $\alpha = 2$, FAT-Clipping-PI achieves $\tilde{\mathcal{O}}((mKT)^{-1})$ and $\mathcal{O}((mKT)^{-\frac{1}{4}})$ convergence rates for strongly convex and nonconvex objective functions, respectively. These two convergence rates are consistent with those of FedAvg in terms of m, K and T (ignoring logarithmic factors in the strongly-convex case).

Next, we show that the convergence rates for FAT-Clipping-PI is *order-optimal* for $\alpha \in (1, 2]$ by proving the following lower bounds.

Corollary 1 (Convergence Rate Lower Bound)**.** *Given any* $\alpha \in (1, 2]$, *for any potentially randomized algorithm, there exists a stochastic strongly-convex function satisfying Assumption 3 with* $G \leq 1$, *such that the output of* $\mathbf{x}_T$ *after* T *communication rounds has an expected error lower bounded by*

$$\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}_*) = \Omega((mKT)^{\frac{2-2\alpha}{\alpha}}).$$

Also, there exists a non-convex function satisfying Assumption 3, such that the output of $\mathbf{x}_T$ *after* T *communication rounds has an expected error lower bounded by*

$$\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|]^2 = \Omega((mKT)^{\frac{2-2\alpha}{3\alpha-2}}).$$

With T communication rounds, the total number of stochastic gradients is mKT. Thus, the lower bounds above can be obtained from the centralized SGD with fat-tailed noise [22, Theorems 5 and 6]). Clearly, the above lower bounds imply the optimality of the convergence rates of FAT-Clipping-PI .

5 Numerical results

In this section, we conduct numerical experiments to verify the theoretical findings in Section 4 using 1) a synthetic function, 2) a convolutional neural network (CNN) with two convolutional layers on CIFAR-10 dataset [42], and 3) RNN on Shakespeare dataset. Due to space limitation, we relegate experiment details and extra experimental results to the supplementary material.

1) Strongly Convex Model with Synthetic Data: We consider a strongly convex model for Problem (1) as follows: $f_i(x) = \mathbb{E}_\xi[f_i(x, \xi)]$ and $f_i(x, \xi) = \frac{1}{2}\|x\|^2 + \langle \xi, x\rangle$, where ξ is a random vector. We compare FedAvg, FAT-Clipping-PI, and FAT-Clipping-PR, where the noise ξ is Cauchy distributed ($\alpha < 2$, fat-tailed). Also, we compare FAT-Clipping-PI and FAT-Clipping-PR with ξ having different tail-indexes ($\alpha = 0.5, 1.0,$ and 1.5). For each distribution, we use the same experimental setup, and $m = 5$ clients participate in the training. We show the trajectories of FedAvg, FAT-Clipping-PI, and FAT-Clipping-PR for solving Problem (1) with ξ having Cauchy tails in Fig. 4

and with ξ having different α-values in Fig. 5. We can clearly observe from Fig. 4 that FAT-Clipping-PI and FAT-Clipping-PR converge rapidly in the Cauchy case, and FAT-Clipping-PI converges faster than FAT-Clipping-PR as our theoretical results predict. In contrast, FedAvg is not convergent in the Cauchy case. In Fig. 5, we can see that the convergence processes of FAT-Clipping-PR and FAT-Clipping-PI become slower as the α-value increases as our theoretical results predict, but the differences in FAT-Clipping-PI is much less obvious compared to those of FAT-Clipping-PR.

2) CNN (Non-convex Model) on the CIFAR-10: This setting has $m = 10$ clients in total, and five clients are randomly selected to participate in each round of the training. We compare FAT-Clipping algorithms with FedAvg under different data heterogeneity. To simulate data heterogeneity across clients, we distribute the data to each client in a label-based partition following the same procedure as in existing works (e.g., [1, 7, 43]): we use a parameter p to represent the number of labeled classes in each client, with $p = 10$ corresponding to the i.i.d. case and the rest corresponding to non-i.i.d. cases. The smaller the p-value, the more heterogeneous the data across clients. In Fig. 6, we present the percentage of successful training over 5 trials when applying FedAvg, FAT-Clipping-PR and FAT-Clipping-PI on CIFAR-10 in non-i.i.d. case ($p = 2$) and i.i.d. case ($p = 10$). FAT-Clipping-PI has 100% successful rates (i.e., no catastrophic model failures) in both non-i.i.d. and i.i.d cases, and FAT-Clipping-PR has 60% and 20% successful rates in non-i.i.d. and i.i.d. cases, respectively. However, FedAvg fails in all 5 trials. Thus, compared to FedAvg, FAT-Clipping methods (FAT-Clipping-PI in particular) significantly reduce catastrophic training failures.

6 Conclusions and future work

In this paper, we investigated the problem of designing efficient federated learning algorithms with convergence performance guarantee in the presence of fat-tailed noise in the stochastic first-order oracles. We first showed empirical evidence that fat-tailed noise in federated learning can be induced by data heterogeneity and heterogeneous local update steps. To address the fat-tailed noise challenge in FL algorithm design, we proposed a clipping-based algorithmic framework called FAT-Clipping . The FAT-Clipping framework contains two variants FAT-Clipping-PR and FAT-Clipping-PI , which perform clipping operations in each communication round and in each local update step, respectively. Then, we derived the convergence rate bounds of FAT-Clipping-PR and FAT-Clipping-PI for strongly convex and non-convex loss functions under fat-tailed noise. Not only does our work shed light on theoretical understanding of FL under fat-tailed noise, it also opens the doors to many new interesting questions in FL systems that experience fat-tailed noise.

Acknowledgements

This work has been supported in part by NSF grants CAREER CNS-2110259, CNS-2112471, ECCS-2140277, and CCF-2110252.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, pp. 429–450.

[3] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 5132–5143.

[4] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[5] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data," *arXiv preprint arXiv:2005.11418*, 2020.

[6] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *International Conference on Learning Representations*, 2021.

[7] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-IID federated learning," in *International Conference on Learning Representations*, 2021.

[8] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konecný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.

[9] P. Khanduri, P. SHARMA, H. Yang, M. Hong, J. Liu, K. Rajawat, and P. Varshney, "STEM: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[10] X. Gu, K. Huang, J. Zhang, and L. Huang, "Fast federated learning in the presence of arbitrary device unavailability," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[11] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No fear of heterogeneity: Classifier calibration for federated learning with non-IID data," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[12] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Breaking the centralized barrier for cross-device federated learning," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[13] J. Nair, A. Wierman, and B. Zwart, *The Fundamentals of Heavy Tails*. Cambridge University Press, 2022, vol. 53.

[14] J. E. Taylor. Heavy-tailed distributions. [Online]. Available: https://math.la.asu.edu/~jtaylor/teaching/Spring2016/STP421/lectures/stable.pdf

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

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

[17] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJgnXpVYwS

[18] E. Gorbunov, M. Danilova, and A. Gasnikov, "Stochastic optimization with heavy-tailed noise via accelerated gradient clipping," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 042–15 053, 2020.

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

[20] Z. Charles, Z. Garrett, Z. Huo, S. Shmulyian, and V. Smith, "On large-cohort training for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[21] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[22] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra, "Why are adaptive methods good for attention models?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 383–15 393, 2020.

[23] J. Wang and G. Joshi, "Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd," in *Proceedings of Machine Learning and Systems*, A. Talwalkar, V. Smith, and M. Zaharia, Eds., vol. 1, 2019, pp. 212–229. [Online]. Available: https://proceedings.mlsys.org/paper/2019/file/c8ffe9a587b126f152ed3d89a146b445-Paper.pdf

[24] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi, "Don't use large mini-batches, use local sgd," *arXiv preprint arXiv:1808.07217*, 2018.

[25] H. Yang, X. Zhang, P. Khanduri, and J. Liu, "Anarchic federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 331–25 363.

[26] X. Zhang, M. Fang, Z. Liu, H. Yang, J. Liu, and Z. Zhu, "Net-fleet: achieving linear convergence speedup for fully decentralized federated learning with heterogeneous data," *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2022.

[27] T. H. Nguyen, U. Simsekli, M. Gurbuzbalaban, and G. Richard, "First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise," *Advances in neural information processing systems*, vol. 32, 2019.

[28] U. Simsekli, L. Zhu, Y. W. Teh, and M. Gurbuzbalaban, "Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8970–8980.

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

[30] H. Wang, M. Gurbuzbalaban, L. Zhu, U. Simsekli, and M. A. Erdogdu, "Convergence rates of stochastic gradient descent under infinite noise variance," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[31] S. Yaida, "Fluctuation-dissipation relations for stochastic gradient descent," *arXiv preprint arXiv:1810.00004*, 2018.

[32] W. Hu, C. J. Li, L. Li, and J.-G. Liu, "On the diffusion approximation of nonconvex stochastic gradient descent," *Annals of Mathematical Sciences and Applications*, vol. 4, no. 1, 2019.

[33] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[34] N. Z. Shor, "Minimization methods for non-differentiable functions," in *Springer Series in Computational Mathematics*, 1985.

[35] B. Zhang, J. Jin, C. Fang, and L. Wang, "Improved analysis of clipping algorithms for non-convex optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 511–15 521, 2020.

[36] X. Chen, S. Z. Wu, and M. Hong, "Understanding gradient clipping in private sgd: A geometric perspective," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 773–13 782, 2020.

[37] J. Qian, Y. Wu, B. Zhuang, S. Wang, and J. Xiao, "Understanding gradient clipping in incremental gradient methods," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1504–1512.

[38] X. Zhang, X. Chen, M.-F. Hong, Z. S. Wu, and J. Yi, "Understanding clipping for federated learning: Convergence and client-level differential privacy," *ArXiv*, vol. abs/2106.13673, 2021.

[39] R. Das, A. Hashemi, S. Sanghavi, and I. S. Dhillon, "Privacy-preserving federated learning via normalized (instead of clipped) updates," *arXiv preprint arXiv:2106.07094*, 2021.

[40] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy, "Differentially private learning with adaptive clipping," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[41] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth, "Lower bounds for non-convex stochastic optimization," *arXiv preprint arXiv:1912.02365*, 2019.

[42] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[43] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020.

[44] Q. Meng, S. Gong, W. Chen, Z.-M. Ma, and T.-Y. Liu, "Dynamic of stochastic gradient descent with state-dependent noise," *arXiv preprint arXiv:2006.13719*, 2020.

[45] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] As a theoretical paper towards further understanding of federated optimization, we do not see a direct path to any negative applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The dataset used in this paper is widely-used public datasets and we have provided detailed instruction for the partition about datasets.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We run the experiments multiple times and report the failure rates instead.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs for Fat-Tailed Federated Learning

A.1 Proof of FAT-Clipping-PR

For notional clarity, we have the following update:

$$\begin{aligned}
\text{Local update: } \mathbf{x}_{t,i}^{k+1} &= \mathbf{x}_{t,i}^{k} - \eta_L \nabla f_i(\mathbf{x}_{t,i}^{k}, \xi_{t,i}^{k}), k \in [K], \\
\text{Clipping: } \mathbf{x}_{t,i}^{K+1} &= \mathbf{x}_{t,i}^{k} - \eta_L \text{clipping}(\sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^{k}, \xi_{t,i}^{k})), \\
\Delta_{t,i} &= \sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^{k}, \xi_{t,i}^{k}), \tilde{\Delta}_{t,i} = \text{clipping}(\sum_{k\in[K]} \nabla f_i(\mathbf{x}_{t,i}^{k}, \xi_{t,i}^{k}), \lambda), \\
\Delta_t &= \frac{1}{m}\sum_{i\in[m]} \Delta_{t,i}, \tilde{\Delta}_t = \frac{1}{m}\sum_{i\in[m]} \tilde{\Delta}_{t,i} \\
\mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\eta_L \tilde{\Delta}_t.
\end{aligned}$$

Lemma 1 (Bounded Variance of Stochastic Local Updates for FAT-Clipping-PR)**.** *Assume $f_i(\mathbf{x}, \xi)$ satisfies the Bounded $\alpha-$Moment assumption 3, then for* FAT-Clipping-PR *we have:*

$$\begin{aligned}
\mathbb{E}[\|\tilde{\Delta}_t\|^2] &\leq K^2 G^\alpha \lambda^{2-\alpha}, \\
\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 &\leq \frac{K^2}{m} G^\alpha \lambda^{2-\alpha}, \\
\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(\mathbf{x}_t)\|^2 &\leq L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}.
\end{aligned}$$

Note here the expectation is on the random samples $\xi_{t,i}^k$.

Proof.

$$\begin{aligned}
\mathbb{E}[\|\tilde{\Delta}_t\|^2] &= \max_{i\in[m]} \mathbb{E}[\|\tilde{\Delta}_{t,i}\|^2] \\
&\leq \mathbb{E}[\|\tilde{\Delta}_{t,j}\|^\alpha]\lambda^{2-\alpha} \\
&\leq K \sum_{k\in K} \mathbb{E}[\|\nabla f(\mathbf{x}_{t,j}^k, \xi_{t,j}^k)\|^\alpha]\lambda^{2-\alpha} \\
&\leq K^2 G^\alpha \lambda^{2-\alpha},
\end{aligned}$$

where $j = argmax_{i\in[m]}\mathbb{E}[\|\tilde{\Delta}_{t,i}\|^2]$, and the first inequality is due to the clipping, i.e., $\|\tilde{\Delta}_{t,i}\| \leq \lambda$.

$$\begin{aligned}
\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 &= \mathbb{E}\left\|\frac{1}{m}\sum_{i\in[m]}\left(\tilde{\Delta}_{t,i} - \mathbb{E}[\tilde{\Delta}_{t,i}]\right)\right\|^2 \\
&\leq \frac{1}{m^2}\sum_{i\in[m]} \mathbb{E}\|\tilde{\Delta}_{t,i} - \mathbb{E}[\tilde{\Delta}_{t,i}]\|^2 \\
&\leq \frac{1}{m^2}\sum_{i\in[m]} \mathbb{E}\|\tilde{\Delta}_{t,i}\|^2 \\
&\leq \frac{K^2}{m} G^\alpha \lambda^{2-\alpha}.
\end{aligned}$$

$$\begin{aligned}
\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(\mathbf{x}_t)\| &\leq \left\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\Delta_t]\right\| + \frac{1}{K}\left\|\mathbb{E}[\Delta_t] - \mathbb{E}[\tilde{\Delta}_t]\right\| \\
&\leq \frac{1}{mK}\sum_{i\in[m]}\sum_{k\in[K]} \mathbb{E}\left\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_{t,i}^k)\right\| + \frac{1}{mK}\sum_{i\in[m]}\left\|\mathbb{E}[\Delta_{t,i} - \tilde{\Delta}_{t,i}]\right\|
\end{aligned}$$

$$\leq \frac{L\eta_L}{mK}\sum_{i\in[m]}\sum_{k\in[K]}\mathbb{E}\left\|\sum_{j\in[k]}\nabla f_i(\mathbf{x}_{t,i}^j,\xi_{t,i}^j)\right\| + \frac{1}{mK}\sum_{i\in[m]}\mathbb{E}[\|\Delta_{t,i}\|\mathbf{1}_{\{\|\Delta_{t,i}\|\geq\lambda\}}]$$
$$\leq L\eta_L KG + KG^\alpha\lambda^{1-\alpha}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, the last inequality follows from the fact that $\Delta_{t,i} = \tilde{\Delta}_{t,i}$ if $\|\Delta_{t,i}\| \leq \lambda$ and $\mathbb{E}[\|\Delta_{t,i}\|\mathbf{1}_{\{\|\Delta_{t,i}\|\geq\lambda\}}] \leq \mathbb{E}[\|\Delta_{t,i}\|^\alpha]\lambda^{1-\alpha} \leq K^2G^\alpha\lambda^{1-\alpha}$; the second last inequality is due to L-smoothness, Jenson's inequality (i.e.,$\mathbb{E}[\Delta_{t,i} - \tilde{\Delta}_{t,i}]\| \leq \mathbb{E}\|[\Delta_{t,i} - \tilde{\Delta}_{t,i}]\|$) and the clipping step. Then, we have

$$\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(x)\|^2 \leq L^2\eta_L^2K^2G^2 + K^2G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_LK^2G^{1+\alpha}\lambda^{1-\alpha}$$

□

Theorem 1. (Convergence Rate of FAT-Clipping-PR in the Strongly Convex Case) *Suppose that $f(\cdot)$ is a μ-strongly convex function. Under Assumptions 1–3, if $\eta\eta_LK \geq \frac{2}{\mu T}$, then the output $\bar{\mathbf{x}}_T$ of* FAT-Clipping-PR *being chosen in such a way that $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]}w_j}$, where $w_t = (1-\frac{1}{2}\mu\eta\eta_LK)^{1-t}$, satisfies:*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_LKT\right) + \frac{\eta\eta_LK}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}\left[2G^{2\alpha}\lambda^{2-2\alpha} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}\right],$$

where $\mathbf{x}^*$ denotes the global optimal solution. Further, let $\eta\eta_LK = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$, where $c \geq 1$ is a constant satisfying $m^{\frac{2-2\alpha}{\alpha}}K^{\frac{2}{\alpha}}T^{c+\frac{2-2\alpha}{\alpha}} \geq 1$, and let $\eta_L \leq (mKT)^{\frac{1-\alpha}{\alpha}}$. It then follows that

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = \mathcal{O}((mT)^{\frac{2-2\alpha}{\alpha}}K^{\frac{2}{\alpha}}).$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] = \mathbb{E}[\|\mathbf{x}_t - \eta\eta_L\tilde{\Delta}_t - x_*\|^2] \\
&= \|\mathbf{x}_t - x_*\|^2 + \eta^2\eta_L^2\mathbb{E}[\|\tilde{\Delta}_t\|^2] - 2\left\langle \mathbf{x}_t - x_*, \eta\eta_L\left(\mathbb{E}[\tilde{\Delta}_t] - K\nabla f(\mathbf{x}_t) + K\nabla f(\mathbf{x}_t)\right)\right\rangle \\
&\leq (1-\mu\eta\eta_LK)\|\mathbf{x}_t - x_*\|^2 + \eta^2\eta_L^2\mathbb{E}[\|\tilde{\Delta}_t\|^2] - 2\eta\eta_LK\left\langle \mathbf{x}_t - x_*, \left(\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(\mathbf{x}_t)\right)\right\rangle \\
&\quad - 2\eta\eta_LK\left(f(\mathbf{x}_t) - f(\mathbf{x}_*)\right) \\
&\leq (1-\frac{1}{2}\mu\eta\eta_LK)\|\mathbf{x}_t - x_*\|^2 + \eta^2\eta_L^2\mathbb{E}[\|\tilde{\Delta}_t\|^2] + \frac{8\eta\eta_LK}{\mu}\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(\mathbf{x}_t)\|^2 \\
&\quad - 2\eta\eta_LK\left(f(\mathbf{x}_t) - f(\mathbf{x}_*)\right).
\end{aligned}$$

The first inequality follows from the strongly-convex property, i.e., $-\langle \mathbf{x}_t - x_*, \nabla f(\mathbf{x}_t)\rangle \leq -(f(\mathbf{x}_t) - f(\mathbf{x}_*) + \frac{\mu}{2}\|x_t - \mathbf{x}_*\|^2)$, and the last inequality is due to Young's inequality. Then we have

$$\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}_*) &\leq \frac{1}{2\eta\eta_LK}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1-\frac{1}{2}\mu\eta\eta_LK)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L}{2K}\mathbb{E}[\|\tilde{\Delta}_t\|^2] + \frac{4}{\mu}\|\mathbb{E}[\frac{1}{K}\tilde{\Delta}_t - \nabla f(\mathbf{x}_t)\|^2 \\
&\leq \frac{1}{2\eta\eta_LK}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1-\frac{1}{2}\mu\eta\eta_LK)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L}{2K}K^2G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[L^2\eta_L^2K^2G^2 + K^2G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_LK^2G^{1+\alpha}\lambda^{1-\alpha}],
\end{aligned}$$

where the last inequality is due to Lemma 1.

Let $w_t = (1-\frac{1}{2}\mu\eta\eta_L K)^{1-t}$, $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$.

$$
\begin{aligned}
f(\bar{x}_T) - f(\mathbf{x}_*) &\leq \frac{1}{\sum_{j\in[T]} w_j} \sum_{t\in[T]} \left(\frac{w_t}{2\eta\eta_L K} \left[-\|\mathbf{x}_{t+1} - x_*\|^2 + (1-\frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2 \right] \right) \\
&\quad + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} + \frac{4}{\mu}[L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}] \\
&\leq \frac{1}{\sum_{j\in[T]} w_j} \sum_{t\in[T]} \left(\frac{1}{2\eta\eta_L K} \left[-w_t\|\mathbf{x}_{t+1} - x_*\|^2 + w_{t-1}\|\mathbf{x}_t - x_*\|^2 \right] \right) \\
&\quad + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} + \frac{4}{\mu}[L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}] \\
&\leq \frac{1}{\sum_{j\in[T]} w_j} \frac{1}{2\eta\eta_L K} \|\mathbf{x}_1 - x_*\|^2 \\
&\quad + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} + \frac{4}{\mu}[L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}],
\end{aligned}
$$

where the second inequality follows from $w_t \leq w_{t-1}$.

$$
\begin{aligned}
2\eta\eta_L K \sum_{t\in[T]} w_t &= 2\eta\eta_L K \left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T} \sum_{t\in[T]} \left(1-\frac{1}{2}\mu\eta\eta_L K\right)^t \\
&= \frac{4}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T} \left[1 - \left(1-\frac{1}{2}\mu\eta\eta_L K\right)^T\right] \\
&\geq \frac{4}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T} \left[1 - \exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right)\right] \\
&\geq \frac{2}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T},
\end{aligned}
$$

where the last inequality follows from that $\eta\eta_L K \geq \frac{2}{\mu T}$, the second last inequality is due to $\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^T \leq \exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right)$.

$$
\begin{aligned}
f(\bar{x}_T) - f(\mathbf{x}_*) &\leq \frac{\mu}{2}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^T + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} \\
&\quad + \frac{4}{\mu}[L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}] \\
&\leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2} G^\alpha \lambda^{2-\alpha} \\
&\quad + \frac{4}{\mu}[L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}].
\end{aligned}
$$

Let $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$ ($c \geq 1$ is a constant and $T^{-c-\frac{2-2\alpha}{\alpha}} \leq m^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}}$), $\lambda = (mKT)^{\frac{1}{\alpha}}$, and $\eta_L \leq (mKT)^{\frac{1-\alpha}{\alpha}}$,

$$
f(\bar{x}_T) - f(\mathbf{x}_*) \leq \frac{1}{T^c} + (mKT)^{\frac{2-2\alpha}{\alpha}} \ln(T) + (mT)^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}} = \mathcal{O}((mT)^{\frac{2-2\alpha}{\alpha}} K^{\frac{2}{\alpha}}).
$$

□

Theorem 2. (Convergence Rate of FAT-Clipping-PR in the Nonconvex Case) *Suppose that $f(\cdot)$ is a nonconvex function. Under Assumptions 1-3, if $\eta\eta_L KL \leq 1$, then the sequence of outputs $\{\mathbf{x}_k\}$ generated by* FAT-Clipping-PR *satisfies:*

$$
\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1) - f(x_T)\right)}{\eta\eta_L KT} + \left(L^2\eta_L^2 K^2 G^2 + K^2 G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2 G^{1+\alpha}\lambda^{1-\alpha}\right)
$$

$$+\frac{L\eta\eta_L}{m}\left(KG^{\alpha}\lambda^{2-\alpha}\right).$$

Further, choosing learning rates and clipping parameter in such a way that $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}}K^{\frac{-\alpha-2}{3\alpha-2}}T^{\frac{-\alpha}{3\alpha-2}}, \eta_L \leq (mT)^{\frac{1-\alpha}{3\alpha-2}}K^{\frac{4-4\alpha}{3\alpha-2}}$, *and* $\lambda = (mK^4T)^{\frac{1}{3\alpha-2}}$, *we have*

$$\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}((mT)^{\frac{2-2\alpha}{3\alpha-2}}K^{\frac{4-2\alpha}{3\alpha-2}}).$$

Proof. Due to the smoothness in Assumption 1, taking expectation of $f(\mathbf{x}_{t+1})$ over the randomness at communication round t, we have:

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_{t+1})] - f(\mathbf{x}_t) \leq \left\langle \nabla f(\mathbf{x}_t), \mathbb{E}[\mathbf{x}_{t+1}-\mathbf{x}_t]\right\rangle + \frac{L}{2}\mathbb{E}[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2] \\
&= -\eta\eta_L\left\langle \nabla f(\mathbf{x}_t), \mathbb{E}[\tilde{\Delta}_t]\right\rangle + \frac{L}{2}\eta^2\eta_L^2\mathbb{E}[\|\tilde{\Delta}_t\|^2] \\
&= -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\eta\eta_L}{2K}\|\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{L\eta^2\eta_L^2}{2}\mathbb{E}[\|\tilde{\Delta}_t\|^2] \\
&= -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \left(-\frac{\eta\eta_L}{2K} + \frac{L\eta^2\eta_L^2}{2}\right)\|\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2 \\
&\quad + \frac{L\eta^2\eta_L^2}{2}\mathbb{E}[\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2] \\
&\leq -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta\eta_L K}{2}\underbrace{\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2}_{A_1} + \frac{L\eta^2\eta_L^2}{2}\underbrace{\mathbb{E}[\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2]}_{A_2},
\end{aligned} \tag{5}$$

where the last inequality follows from $\left(-\frac{\eta\eta_L}{2K} + \frac{L\eta^2\eta_L^2}{2}\right) \leq 0$ if $\eta\eta_L KL \leq 1$.

From Lemma 1, we have the bound of A_1 and A_2 in (5). By rearranging and telescoping, we have:

$$\begin{aligned}
\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1) - f(x_T)\right)}{\eta\eta_L KT} + \left(L^2\eta_L^2K^2G^2 + K^2G^{2\alpha}\lambda^{-2(\alpha-1)} + L\eta_L K^2G^{1+\alpha}\lambda^{1-\alpha}\right) \\
+ \frac{L\eta\eta_L}{m}\left(KG^{\alpha}\lambda^{2-\alpha}\right).
\end{aligned}$$

Suppose $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}}K^{\frac{-\alpha-2}{3\alpha-2}}T^{\frac{-\alpha}{3\alpha-2}}, \eta_L \leq (mT)^{\frac{1-\alpha}{3\alpha-2}}K^{\frac{4-4\alpha}{3\alpha-2}}$, and $\lambda = (mK^4T)^{\frac{1}{3\alpha-2}}$,

$$\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mT)^{\frac{2-2\alpha}{3\alpha-2}}K^{\frac{4-2\alpha}{3\alpha-2}}).$$

□

A.2 Proof of FAT-Clipping-PI

For FAT-Clipping-PI, we have the following notions:

$$\begin{aligned}
\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) &= \min\{1, \frac{\lambda_t}{\|\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|}\}\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k), \tilde{\nabla} f(\mathbf{x}_{t,i}^k) = \mathbb{E}[\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)]; \\
\text{Local steps: } \mathbf{x}_{t,i}^{k+1} &= \mathbf{x}_{t,i}^k - \eta_L\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k), k \in [K]; \\
\Delta_{t,i} &= \sum_{k\in[K]}\nabla f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k), \tilde{\Delta}_{t,i} = \sum_{k\in[K]}\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) \\
\Delta_t &= \frac{1}{m}\sum_{i\in[m]}\Delta_{t,i}, \tilde{\Delta}_t = \frac{1}{m}\sum_{i\in[m]}\tilde{\Delta}_{t,i} \\
\mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\eta_L\frac{1}{m}\sum_{i\in[m]}\sum_{k\in[K]}\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) = \mathbf{x}_t - \eta\eta_L\tilde{\Delta}_t.
\end{aligned}$$

Lemma 2 (Bounded Variance of Stochastic Local Updates for FAT-Clipping-PI)**.** *Assume $f_i(\mathbf{x}, \xi)$ satisfies the Bounded $\alpha-$Moment assumption 3, then we have:*

$$\mathbb{E}[\|\tilde{\Delta}_t\|^2] \leq K^2 G^\alpha \lambda^{2-\alpha},$$
$$\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 \leq \frac{K}{m} G^\alpha \lambda^{2-\alpha},$$
$$\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(x)\|^2 \leq 2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2 K^2 G^\alpha \lambda^{2-\alpha}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[\|\tilde{\Delta}_t\|^2] &= \frac{1}{m}\sum_{i\in[m]} \mathbb{E}[\|\tilde{\Delta}_{t,i}\|^2] \\
&\leq \frac{1}{m}\sum_{i\in[m]} \mathbb{E}[\|\sum_{j\in[K]} \tilde{\nabla} f(\mathbf{x}_{t,i}^j, \xi_{t,i}^j)\|^2] \\
&\leq \frac{K}{m}\sum_{i\in[m]}\sum_{j\in[K]} \mathbb{E}[\|\tilde{\nabla} f(\mathbf{x}_{t,i}^j, \xi_{t,i}^j)\|^2] \\
&\leq K^2 G^\alpha \lambda^{2-\alpha},
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^2 \leq \mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^\alpha \lambda^{2-\alpha} \leq G^\alpha \lambda^{2-\alpha}$ (see Lemma 9 in [22]).

$$\begin{aligned}
\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 &= \mathbb{E}\left\|\frac{1}{m}\sum_{i\in[m]}\sum_{k\in[K]} \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) - \frac{1}{m}\sum_{i\in[m]}\sum_{k\in[K]} \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k)\right\|^2 \\
&\leq \frac{1}{m^2}\sum_{i\in[m]}\sum_{k\in[K]} \mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) - \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k)\|^2 \\
&\leq \frac{1}{m^2}\sum_{i\in[m]}\sum_{k\in[K]} \mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^2 \\
&\leq \frac{K}{m} G^\alpha \lambda^{2-\alpha},
\end{aligned}$$

where the first inequality follows from the fact that $\{\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k) - \tilde{\nabla} f_i(\mathbf{x}_{t,i}^k)\}$ form a martingale difference sequence (Lemma 4 in [3]), the second inequalities is due to $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$, and the third inequality follows from the fact that $\mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^2 \leq \mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^\alpha \lambda^{2-\alpha} \leq G^\alpha \lambda^{2-\alpha}$ (see Lemma 9 in [22]).

$$\begin{aligned}
\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(x)\|^2 &\leq \left\|\frac{1}{mK}\sum_{i\in[m]}\sum k \in [K]\left(\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k) - f_i(\mathbf{x}_t)\right)\right\|^2 \\
&\leq \frac{1}{mK}\sum_{i\in[m]}\sum_{k\in[K]} \left\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k) - f_i(\mathbf{x}_t)\right\|^2 \\
&\leq \frac{1}{mK}\sum_{i\in[m]}\sum_{k\in[K]} \left(2\left\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k) - \nabla f_i(\mathbf{x}_{t,i}^k)\right\|^2 + 2\left\|\nabla f_i(\mathbf{x}_{t,i}^k) - f_i(\mathbf{x}_t)\right\|^2\right) \\
&\leq 2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\frac{1}{mK}\sum_{i\in[m]}\sum_{k\in[K]} \|\mathbf{x}_{t,i}^k - \mathbf{x}_t\|^2 \\
&\leq 2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2\frac{1}{mK}\sum_{i\in[m]}\sum_{k\in[K]} \left\|\sum_{j\in[K]} \nabla f(x_{t,i}^j, \xi_{t,i}^j)\right\|^2
\end{aligned}$$

$$\leq 2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha},$$

where the forth inequality is due to $\|\tilde{\nabla} f_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2 \leq G^{2\alpha}\lambda^{-2(\alpha-1)}$ (see Lemma 9 in [22]), and the last inequality follows from the fact that $\mathbb{E}\|\tilde{\nabla} f_i(\mathbf{x}_{t,i}^k, \xi_{t,i}^k)\|^2 \leq G^\alpha\lambda^{2-\alpha}$. □

Theorem 3. (Convergence Rate of FAT-Clipping-PI in the Strongly Convex Case) *Suppose that $f(\cdot)$ is a μ-strongly convex function. Under Assumptions 1–3, if $\eta\eta_L K \geq \frac{2}{\mu T}$, then the output $\bar{\mathbf{x}}_T$ of* FAT-Clipping-PI *being chosen in such a way that $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$, where $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$, satisfies:*

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}],$$

where $\mathbf{x}^*$ denotes the global optimal solution. Further, let $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$, where $c \geq 1$ is a constant satisfying $(mK)^{\frac{2-2\alpha}{\alpha}}T^{c+\frac{2-2\alpha}{\alpha}} \geq 1$, and let $\lambda = (mKT)^{\frac{1}{\alpha}}$, and $\eta_L \leq (mT)^{-\frac{1}{2}}K^{-\frac{3}{2}}$). It then follows that

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = \tilde{\mathcal{O}}((mKT)^{\frac{2-2\alpha}{\alpha}}).$$

Proof. Similarly, we have the following one step iteration:

$$\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}_*) &\leq \frac{1}{2\eta\eta_L K}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L}{2K}\mathbb{E}[\|\tilde{\Delta}_t\|^2] + \frac{4}{\mu}\|\mathbb{E}[\frac{1}{K}\tilde{\Delta}_t - \nabla f(\mathbf{x}_t)\|^2 \\
&\leq \frac{1}{2\eta\eta_L K}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L}{2K}K^2G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}],
\end{aligned}$$

where the last inequality is due to Lemma 2.

Let $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$, $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$.

$$\begin{aligned}
f(\bar{x}_T) - f(\mathbf{x}_*) &\leq \frac{1}{\sum_{j\in[T]} w_j}\sum_{t\in[T]}\left(\frac{w_t}{2\eta\eta_L K}\left[-\|\mathbf{x}_{t+1} - x_*\|^2 + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right]\right) \\
&\quad + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}] \\
&\leq \frac{1}{\sum_{j\in[T]} w_j}\sum_{t\in[T]}\left(\frac{1}{2\eta\eta_L K}\left[-w_t\|\mathbf{x}_{t+1} - x_*\|^2 + w_{t-1}\|\mathbf{x}_t - x_*\|^2\right]\right) \\
&\quad + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}] \\
&\leq \frac{1}{\sum_{j\in[T]} w_j}\frac{1}{2\eta\eta_L K}\|\mathbf{x}_1 - x_*\|^2 \\
&\quad + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}].
\end{aligned}$$

$$\begin{aligned}
2\eta\eta_L K\sum_{t\in[T]} w_t &= 2\eta\eta_L K\left(1 - \frac{1}{2}\mu\eta\eta_L K\right)^{-T}\sum_{t\in[T]}\left(1 - \frac{1}{2}\mu\eta\eta_L K\right)^t \\
&= \frac{4}{\mu}\left(1 - \frac{1}{2}\mu\eta\eta_L K\right)^{-T}\left[1 - \left(1 - \frac{1}{2}\mu\eta\eta_L K\right)^T\right]
\end{aligned}$$

$$\geq \frac{4}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T}\left[1-\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right)\right]$$
$$\geq \frac{2}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T},$$

where the last inequality follows from that $\eta\eta_L K \geq \frac{2}{\mu T}$, teh second last inequality is due to $\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^T \leq \exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right)$.

$$f(\bar{x}_T)-f(\mathbf{x}_*) \leq \frac{\mu}{2}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^T + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}]$$
$$\leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2}G^\alpha\lambda^{2-\alpha} + \frac{4}{\mu}[2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}].$$

Let $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$ ($c \geq 1$ is a constant and $T^{-c-\frac{2-2\alpha}{\alpha}} \leq (mK)^{\frac{2-2\alpha}{\alpha}}$), $\lambda = (mKT)^{\frac{1}{\alpha}}$, and $\eta_L \leq (mT)^{-\frac{1}{2}}K^{-\frac{3}{2}}$,

$$f(\bar{x}_T)-f(\mathbf{x}_*) \leq \frac{1}{T^c} + (mKT)^{\frac{2-2\alpha}{\alpha}}\ln(T) = \tilde{\mathcal{O}}((mKT)^{\frac{2-2\alpha}{\alpha}}).$$

□

Theorem 4. (Convergence Rate of FAT-Clipping-PI in the Nonconvex Case) *Suppose that $f(\cdot)$ is a non-convex function. Under Assumptions 1–3, if $\eta\eta_L KL \leq 1$, then the sequence of outputs $\{\mathbf{x}_k\}$ generated by* FAT-Clipping-PI *satisfies:*

$$\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1)-f(x_T)\right)}{\eta\eta_L KT} + \left(2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}\right)$$
$$+ \frac{L\eta\eta_L}{m}\left(G^\alpha\lambda^{2-\alpha}\right).$$

Further, choosing learning rates and clipping parameter in such a way that $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}}(KT)^{\frac{-\alpha}{3\alpha-2}}, \eta_L \leq (mKT)^{\frac{-\alpha}{6\alpha-4}}$, and $\lambda = (mKT)^{\frac{1}{3\alpha-2}}$, we have

$$\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mKT)^{\frac{2-2\alpha}{3\alpha-2}}).$$

Proof. Due to the smoothness in Assumption 1, taking expectation of $f(\mathbf{x}_{t+1})$ over the randomness at communication round t, we have:

$$\mathbb{E}[f(\mathbf{x}_{t+1})] - f(\mathbf{x}_t) \leq \left\langle\nabla f(\mathbf{x}_t), \mathbb{E}[\mathbf{x}_{t+1}-\mathbf{x}_t]\right\rangle + \frac{L}{2}\mathbb{E}[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2]$$
$$= -\eta\eta_L\left\langle\nabla f(\mathbf{x}_t), \mathbb{E}[\tilde{\Delta}_t]\right\rangle + \frac{L}{2}\eta^2\eta_L^2\mathbb{E}[\|\tilde{\Delta}_t\|^2]$$
$$= -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\eta\eta_L}{2K}\|\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{L\eta^2\eta_L^2}{2}\mathbb{E}[\|\tilde{\Delta}_t\|^2]$$
$$= -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \left(-\frac{\eta\eta_L}{2K} + \frac{L\eta^2\eta_L^2}{2}\right)\|\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2 + \frac{L\eta^2\eta_L^2}{2}\mathbb{E}[\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2]$$
$$\leq -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta\eta_L K}{2}\underbrace{\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2}_{A_1} + \frac{L\eta^2\eta_L^2}{2}\underbrace{\mathbb{E}[\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2]}_{A_2}, \tag{6}$$

where the last inequality follows from $\left(-\frac{\eta\eta_L}{2K} + \frac{L\eta^2\eta_L^2}{2}\right) \leq 0$ if $\eta\eta_L KL \leq 1$.

From Lemma 2, we have the bound of A_1 and A_2 in (6). By rearranging and telescoping, we have:

$$\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_1)-f(x_T)\right)}{\eta\eta_L KT} + \left(2G^{2\alpha}\lambda^{-2(\alpha-1)} + 2L^2\eta_L^2K^2G^\alpha\lambda^{2-\alpha}\right) + \frac{L\eta\eta_L}{m}\left(G^\alpha\lambda^{2-\alpha}\right).$$

Suppose $\eta\eta_L = m^{\frac{2\alpha-2}{3\alpha-2}}(KT)^{\frac{-\alpha}{3\alpha-2}}$, $\eta_L \leq (mKT)^{\frac{-\alpha}{6\alpha-4}}$, and $\lambda = (mKT)^{\frac{1}{3\alpha-2}}$,

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mKT)^{\frac{2-2\alpha}{3\alpha-2}}).$$

□

A.3 Proof of FAT-Clipping-PR in Gaussian Noise

In this subsection, we utilize the classic bounded variance and bounded gradient assumption.

Assumption 4. *(Bounded Stochastic Gradient Variance) There exists a constant $\sigma > 0$, such that the variance of each local gradient estimator is bounded by $\mathbb{E}[\|\nabla f_i(\mathbf{x},\xi) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$, $\forall i \in [m]$.*

Assumption 5. *(Bounded Gradient) There exists a constant $G \geq 0$, such that gradient is bounded by $\|\nabla f_i(\mathbf{x})\|^2 \leq G^2$, $\forall i \in [m]$.*

Lemma 3 (Lemma F.5 [18])**.** *Suppose there exists a constant σ such that the variance of the stochastic gradient of F has bounded variance, i.e., $\mathbb{E}[\|\nabla F(\mathbf{x},\xi) - \nabla F(\mathbf{x})\|^2] \leq \sigma^2$, and $\|\nabla F(\mathbf{x})\|^2 \leq \frac{\lambda}{2}$, then we have the following inequalities for the clipping $\tilde{\nabla} F(\mathbf{x}_t) = \mathbb{E}[\tilde{\nabla} F(\mathbf{x},\xi)] = \mathbb{E}[\min\{1, \frac{\lambda}{\|\nabla F(\mathbf{x},\xi)\|}\}\nabla F(\mathbf{x},\xi)]$:*

$$\|\mathbb{E}[\tilde{\nabla} F(\mathbf{x},\xi)] - \nabla F(\mathbf{x})\|^2 \leq \frac{16\sigma^4}{\lambda^2},$$
$$\mathbb{E}\|\tilde{\nabla} F(\mathbf{x},\xi) - \nabla F(\mathbf{x})\|^2 \leq 18\sigma^2,$$
$$\mathbb{E}\|\tilde{\nabla} F(\mathbf{x},\xi) - \mathbb{E}[\tilde{\nabla} F(\mathbf{x},\xi)]\|^2 \leq 18\sigma^2.$$

We remark that for any stochastic estimator satisfies the above conditions, the above inequalities hold. The proof is the exactly same as that in original proof [18].

Lemma 4 (Bounded Variance of Clipping Stochastic Local Updates in FAT-Clipping-PR)**.** *Assume f_i satisfies the bounded variance assumption, then we have:*

$$\mathbb{E}[\|\Delta_{t,i} - \mathbb{E}[\Delta_{t,i}]\|^2] \leq K\sigma^2.$$

In addition, assume there exists a constant G such that gradient is bounded $\|\nabla f_i(\mathbf{x})\|^2 \leq G^2$, if we set clipping parameter as $\lambda^2 \geq 2K^2G^2$, i.e., $\|\nabla f_i(\mathbf{x})\| \leq \frac{\lambda}{2}$, then we have:

$$\left\|\mathbb{E}[\Delta_{t,i}] - \mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2 \leq \frac{16K\sigma^4}{\lambda^2},$$
$$\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 \leq \frac{18K}{m}\sigma^4.$$

.

Proof.

$$\mathbb{E}[\|\Delta_{t,i} - \mathbb{E}[\Delta_{t,i}]\|^2] = \mathbb{E}[\|\nabla f(\mathbf{x}_{t,i}^j, \xi_{t,i}^j) - \mathbb{E}[\nabla f(\mathbf{x}_{t,i}^j)]\|^2]$$
$$\leq K\sigma^2,$$

where $\{\nabla f(\mathbf{x}_{t,i}^j, \xi_{t,i}^j) - \mathbb{E}[\nabla f(\mathbf{x}_{t,i}^j)\}$ forms martingale difference sequence (Lemma 4 in [3]).

Then by applying Lemma 3, we have the bound of $\left\|\mathbb{E}[\Delta_{t,i}] - \mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2$.

$$\mathbb{E}\|\tilde{\Delta}_t - \mathbb{E}[\tilde{\Delta}_t]\|^2 = \mathbb{E}\left\|\frac{1}{m}\sum_{i\in[m]}\tilde{\Delta}_{t,i} - \frac{1}{m}\sum_{i\in[m]}\mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2$$
$$\leq \frac{18K}{m}\sigma^4.$$

where the last inequality follows from the fact that $\mathbb{E}[\|\Delta_{t,i} - \mathbb{E}[\Delta_{t,i}]\|^2] \leq K\sigma^2$, $\{\Delta_{t,i} - \mathbb{E}[\Delta_{t,i}]\}$ forms martingale difference sequence and Lemma 3. □

Theorem 5. *Suppose f is non-convex function, under Assumptions 1, 2, 4, and 5, if $\eta\eta_L KL \leq 1$, then the sequence of outputs $\{\mathbf{x}_k\}$ generated by Algorithm* FAT-Clipping-PR *satisfies:*

$$\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_0)-f(\mathbf{x}_T)\right)}{\eta\eta_L KT} + \frac{1}{T}\sum_{t\in[T]}\left(2L^2\eta_L^2K^2(\sigma^2+G^2)+\frac{32\sigma^4}{K^2\lambda_t^2}\right)+\left(\frac{18L\eta\eta_L}{m}\sigma^2\right).$$

Choosing learning rates and clipping parameter as $\eta\eta_L = \frac{m^{1/2}}{(KT)^{1/2}}$, $\eta_L \leq \frac{1}{(mT)^{1/2}K^{5/2}}$, and $\lambda_t \geq (mT)^{1/4}K^{-3/4}$,

$$\min_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mKT)^{-\frac{1}{2}}).$$

Proof. Due to the smoothness in Assumption 1, taking expectation of $f(\mathbf{x}_{t+1})$ over the randomness at communication round t, we have the same inequality:

$$\mathbb{E}[f(\mathbf{x}_{t+1})]-f(\mathbf{x}_t) \leq -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta\eta_L K}{2}\underbrace{\|\nabla f(\mathbf{x}_t)-\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2}_{A_1} + \frac{L\eta^2\eta_L^2}{2}\underbrace{\mathbb{E}[\|\tilde{\Delta}_t-\mathbb{E}[\tilde{\Delta}_t]\|^2]}_{A_2}, \tag{7}$$

where it requires $\eta\eta_L KL \leq 1$.

Note that the term A_1 in (7) can be bounded as follows:

$$\begin{aligned}
A_1 &= \|\nabla f(\mathbf{x}_t)-\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2 \\
&= 2\left\|\nabla f(\mathbf{x}_t)-\frac{1}{K}\mathbb{E}[\Delta_t]\right\|^2 + \frac{2}{K^2}\left\|\mathbb{E}[\Delta_t]-\mathbb{E}[\tilde{\Delta}_t]\right\|^2 \\
&\leq \frac{2}{mK}\sum_{i\in[m]}\sum_{k\in[K]}\left\|\nabla f_i(\mathbf{x}_t)-\nabla f_i(\mathbf{x}_{t,i}^k)\right\|^2 + \frac{2}{mK^2}\sum_{i\in[m]}\left\|\mathbb{E}[\Delta_{t,i}]-\mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2 \\
&\leq \frac{2L^2\eta_L^2}{mK}\sum_{i\in[m]}\sum_{k\in[K]}\mathbb{E}\left\|\sum_{j\in[k]}\nabla f_i(\mathbf{x}_{t,i}^j,\xi_{t,i}^j)\right\|^2 + \frac{2}{mK^2}\sum_{i\in[m]}\left\|\mathbb{E}[\Delta_{t,i}]-\mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2 \\
&\leq 2L^2\eta_L^2K^2\left(\mathbb{E}\left\|\nabla f_i(\mathbf{x}_{t,i}^k,\xi_{t,i}^k)-\nabla f_i(\mathbf{x}_{t,i}^k)\right\|^2+\left\|\nabla f_i(\mathbf{x}_{t,i}^k)\right\|^2\right)+\frac{32\sigma^4}{K^2\lambda_t^2} \\
&\leq 2L^2\eta_L^2K^2(\sigma^2+G^2)+\frac{32\sigma^4}{K^2\lambda_t^2}
\end{aligned}$$

where the second inequality is due to smoothness assumption 1, the third inequality is due to Lemma 4, and the last inequality follows from bounded variance assumption 4 and bounded gradient assumption 5.

From Lemma 4, the term A_2 in (7) can be bounded as follows:

$$A_2 \leq \frac{18K\sigma^2}{m}.$$

Putting pieces together, we can have the one communication round descent in expectation:

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_{t+1})]-f(\mathbf{x}_t) \leq -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta\eta_L K}{2}\underbrace{\|\nabla f(\mathbf{x}_t)-\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t]\|^2}_{A_1} + \frac{L\eta^2\eta_L^2}{2}\underbrace{\mathbb{E}[\|\tilde{\Delta}_t-\mathbb{E}[\tilde{\Delta}_t]\|^2]}_{A_2} \\
&\leq -\frac{\eta\eta_L K}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta\eta_L K}{2}\left(2L^2\eta_L^2K^2(\sigma^2+G^2)+\frac{32\sigma^4}{K^2\lambda_t^2}\right)+\frac{18LK\eta^2\eta_L^2}{2m}\sigma^2.
\end{aligned}$$

Rearranging and telescoping, we have the final convergence result:

$$\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\left(f(\mathbf{x}_0)-f(\mathbf{x}_T)\right)}{\eta\eta_L KT} + \frac{1}{T}\sum_{t\in[T]}\left(2L^2\eta_L^2K^2(\sigma^2+G^2)+\frac{32\sigma^2}{K^2\lambda_t^2}\right)+\left(\frac{18L\eta\eta_L}{m}\sigma^2\right).$$

Suppose $\eta\eta_L = \frac{m^{1/2}}{(KT)^{1/2}}, \eta_L \leq \frac{1}{(mT)^{1/2}K^{5/2}}$, and $\lambda_t \geq (mT)^{1/4}K^{-3/4}$,

$$\min_{t\in[T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}((mKT)^{-\frac{1}{2}}).$$

□

Theorem 6. *Suppose f is μ-strongly convex function, under Assumptions 1–3, if $\eta\eta_L K \geq \frac{2}{\mu T}$, then the outputs $\bar{\mathbf{x}}_T$ in Algorithm 2 (*FAT-Clipping-PR*) by $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$ where $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$ satisfies:*

$$f(\bar{x}_T) - f(\mathbf{x}_*) \leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right) + \frac{\eta\eta_L K}{2}G^2 + \frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2) + \frac{32\sigma^4}{\lambda^2}].$$

Suppose $\eta\eta_L K = \frac{2c}{\mu}\frac{\ln(T)}{mKT}$ ($c > 0$ is a constant and $T^{-c+1} \leq (mK)^{-1}$), $\lambda \geq (mKT)^{\frac{1}{2}}$, and $\eta_L \leq (mT)^{-\frac{1}{2}}K^{-\frac{3}{2}}$,

$$f(\bar{x}_T) - f(\mathbf{x}_*) = \tilde{\mathcal{O}}((mKT)^{-1}).$$

Proof.

$$\begin{aligned}
\|\frac{1}{K}\mathbb{E}[\tilde{\Delta}_t] - \nabla f(x)\|^2 &\leq 2\left\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\mathbb{E}[\Delta_t]\right\|^2 + \frac{2}{K}\left\|\mathbb{E}[\Delta_t] - \mathbb{E}[\tilde{\Delta}_t]\right\|^2 \\
&\leq \frac{2}{mK}\sum_{i\in[m]}\sum_{k\in[K]}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_{t,i}^k)\right\|^2 + \frac{2}{mK}\sum_{i\in[m]}\left\|\mathbb{E}[\Delta_{t,i}] - \mathbb{E}[\tilde{\Delta}_{t,i}]\right\|^2 \\
&\leq \frac{2L^2\eta_L^2}{mK}\sum_{i\in[m]}\sum_{k\in[K]}\mathbb{E}\left\|\sum_{j\in[k]}\nabla f_i(\mathbf{x}_{t,i}^j, \xi_{t,i}^j)\right\|^2 + \frac{32\sigma^4}{\lambda^2} \\
&\leq 2L^2\eta_L^2K^2(G^2+\sigma^2) + \frac{32\sigma^4}{\lambda^2}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}_*) &\leq \frac{1}{2\eta\eta_L K}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L}{2K}\mathbb{E}[\|\tilde{\Delta}_t\|^2] + \frac{4}{\mu}\|\mathbb{E}[\frac{1}{K}\tilde{\Delta}_t - \nabla f(\mathbf{x}_t)\|^2 \\
&\leq \frac{1}{2\eta\eta_L K}\left[-\mathbb{E}[\|\mathbf{x}_{t+1} - x_*\|^2] + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right] \\
&\quad + \frac{\eta\eta_L K}{2}G^2 + \frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2) + \frac{32\sigma^4}{\lambda^2}],
\end{aligned}$$

Let $w_t = (1 - \frac{1}{2}\mu\eta\eta_L K)^{1-t}$, $\bar{\mathbf{x}}_T = \mathbf{x}_t$ with probability $\frac{w_t}{\sum_{j\in[T]} w_j}$.

$$\begin{aligned}
f(\bar{x}_T) - f(\mathbf{x}_*) &\leq \frac{1}{\sum_{j\in[T]} w_j}\sum_{t\in[T]}\left(\frac{w_t}{2\eta\eta_L K}\left[-\|\mathbf{x}_{t+1} - x_*\|^2 + (1 - \frac{1}{2}\mu\eta\eta_L K)\|\mathbf{x}_t - x_*\|^2\right]\right) \\
&\quad + \frac{\eta\eta_L K}{2}G^2 + \frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2) + \frac{32\sigma^4}{\lambda^2}] \\
&\leq \frac{1}{\sum_{j\in[T]} w_j}\frac{1}{2\eta\eta_L K}\|\mathbf{x}_1 - x_*\|^2 + \frac{\eta\eta_L K}{2}G^2 + \frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2) + \frac{32\sigma^4}{\lambda^2}].
\end{aligned}$$

Same as that in heavy-tailed noise case, we have the same bound for $2\eta\eta_L K\sum_{t\in[T]} w_t$:

$$2\eta\eta_L K\sum_{t\in[T]} w_t \geq \frac{2}{\mu}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{-T},$$

where it requires $\eta\eta_L K \geq \frac{2}{\mu T}$.

$$\begin{aligned} f(\bar{x}_T)-f(\mathbf{x}_*) &\leq \frac{\mu}{2}\left(1-\frac{1}{2}\mu\eta\eta_L K\right)^{T}+\frac{\eta\eta_L K}{2}G^2+\frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2)+\frac{32\sigma^4}{\lambda^2}] \\ &\leq \frac{\mu}{2}\exp\left(-\frac{1}{2}\mu\eta\eta_L KT\right)+\frac{\eta\eta_L K}{2}G^2+\frac{4}{\mu}[2L^2\eta_L^2K^2(G^2+\sigma^2)+\frac{32\sigma^4}{\lambda^2}]. \end{aligned}$$

Let $\eta\eta_L K=\frac{2c}{\mu}\frac{\ln(T)}{mKT}$ ($c>0$ is a constant and $T^{-c+1}\leq (mK)^{-1}$), $\lambda\geq (mKT)^{\frac{1}{2}}$, and $\eta_L\leq (mT)^{-\frac{1}{2}}K^{-\frac{3}{2}}$,

$$f(\bar{x}_T)-f(\mathbf{x}_*)=\tilde{\mathcal{O}}((mKT)^{-1}).$$

□

B Experiments in Section 3

In this section, we provide experimental details to demonstrate the fat-tailed noise phenomenon in federated learning. We conduct experiments with CNN on CIFAR-10 dataset as shown in Section 3, and provide additional results of RNN model on Shakespeare dataset. Furthermore, we verify the accuracy of α estimation with logistic regression on MNIST dataset.

B.1 CNN on CIFAR-10 Dataset

B.1.1 Experiment details

We run a convolutional neural network (CNN) model on CIFAR-10 dataset using FedAvg. The CNN architecture is shown in Table 2. To simulate data heterogeneity across clients, we manually distribute the the data to each client in a label-based partition. Specifically, we split the data according to the classes (p) of images that each client has. Then, we randomly distribute these partitioned data to $m=100$ clients such that each client has only p classes of images in both training and test data, which causes the heterogeneity of data among different clients. For example, for $p=10$, each client contains training/test data samples with ten classes. Since CIFAR-10 has 10 classes of images, $p=10$ is the nearly i.i.d case. For the remaining p, each client contains data samples with class p. Therefore, the classes (p) of images in each client's local dataset can be used to represent the non-i.i.d. degree. The smaller the p-value, the more heterogeneous the data between clients.

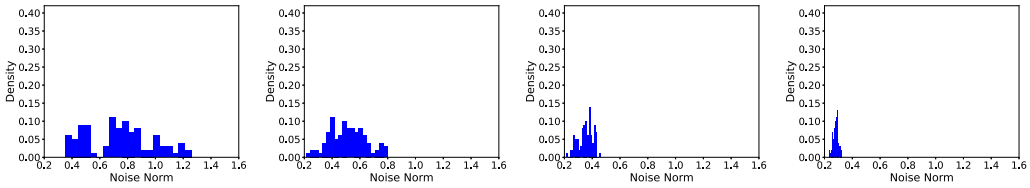
In this experimental setting, we use the global learning rate $\frac{\eta\eta_L}{m}=1.0$ and the local learning rate $\eta_L=0.1$. The batch size is set to 500, and the communication round is $T=4000$. We run this experiment in different cases, including singleSGD and different local epochs $\{1,2,5\}$ and non-iid index $p\in\{1,2,5,10\}$. Single SGD means one local update step, which is equivalent to mini-batch SGD.

Table 2: CNN architecture for CIFAR-10.

LAYER TYPE	SIZE
Convolution + ReLu	$3 \times 32 \times 5$
Max Pooling	2×2
Convolution + ReLu	$32 \times 64 \times 5$
Max Pooling	2×2
Fully Connected + ReLU	1600×512
Fully Connected + ReLU	512×128
Fully Connected	128×10

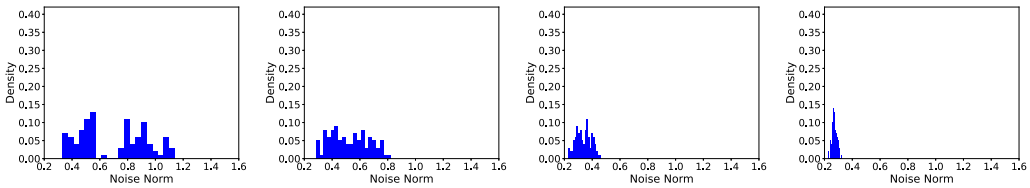
B.1.2 Additional experimental results

We provide additional distributions of the norms of the pseudo-gradient noises in different cases as follows. From Fig. 7- 10, the observation is that the gradient norm statistics are contracted together for more iid cases while dispersed uniformly for more non-iid cases. This is



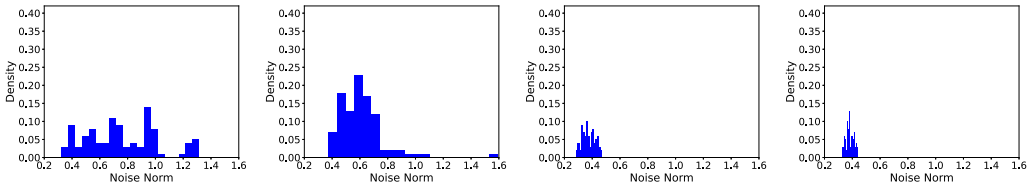
(a) $p = 1$ (b) $p = 2$ (c) $p = 5$ (d) $p = 10$

Figure 7: Distributions of the norms of the pseudo-gradient noises for CIFAR-10 dataset in the case of *Single SGD*.



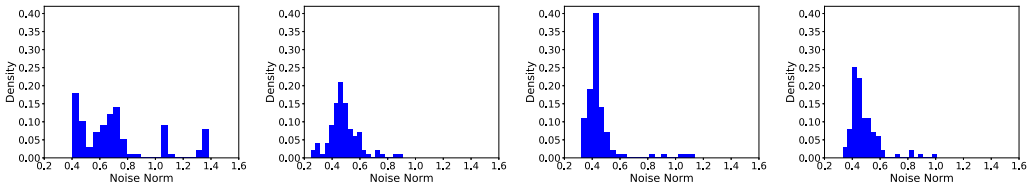
(a) $p = 1$ (b) $p = 2$ (c) $p = 5$ (d) $p = 10$

Figure 8: Distributions of the norms of the pseudo-gradient noises for CIFAR-10 dataset in the case of *Local Epoch=1*.



(a) $p = 1$ (b) $p = 2$ (c) $p = 5$ (d) $p = 10$

Figure 9: Distributions of the norms of the pseudo-gradient noises for CIFAR-10 dataset in the case of *Local Epoch=2*.



(a) $p = 1$ (b) $p = 2$ (c) $p = 5$ (d) $p = 10$

Figure 10: Distributions of the norms of the pseudo-gradient noises for CIFAR-10 dataset in the case of *Local Epoch=5*.

B.2 RNN on Shakespeare Dataset

B.2.1 Experiment details

To provide more evidences of the fat-tailed noise phenomenon, we further run a recurrent neural network (RNN) model on Shakespeare dataset.

Shakespeare dataset is a natural non-iid dataset, and it is built from *The Complete Works of William Shakespeare* [1]. The learning task is to predict next character, and there are 80 classes of characters in total. We use a two-layer LSTM classifier containing 100 hidden units with an 8-dimensional (8D) embedding layer. The model inputs a sequence of 80 characters, embeds each of the characters into a learned 8D space, and then outputs one character per training sample after two LSTM layers and a densely-connected layer. The dataset and model are taken from [45].

There are $m = 143$ clients participating in this experiment. The global learning rate is chosen as 1.0, and the local learning rate is chosen as 0.8. The batch size is set to 10, and the communication round is $T = 150$.

B.2.2 Experimental results

We show the results when local step is set to be one (Single SGD), and multiple local epochs $\{1, 2, 5\}$. In Fig. 11, we observe that the α-value is smaller than 2, and it increases when the number of local epoch increases. This implies that the gradient noise is fat-tailed. Fig. 12 shows that the distributions of the norms of the pseudo-gradient noises are fat-tailed.

B.3 Accuracy of Alpha Estimation (Logistic Regression on MNIST Dataset)

Accurate α-value computation requires the full-gradient calculation, and we have to compute both full-gradient and stochastic gradient in each local step. This is computationally expensive. Instead, we use an estimation to approximate the exact α-value. The full-gradient is replaced by the mean value of the stochastic gradients. We verify the accuracy of this estimation method by running logistic regression on MNIST dataset [**?**]. The details and the results are described as follows.

B.3.1 Experiment details

MNIST dataset contains ten classes of images, and it is manually partitioned using the same method as to partition CIFAR-10 dataset (see details in Appendix B.1.1). The number of classes (p) that each client has can be used to represent the non-iid level.

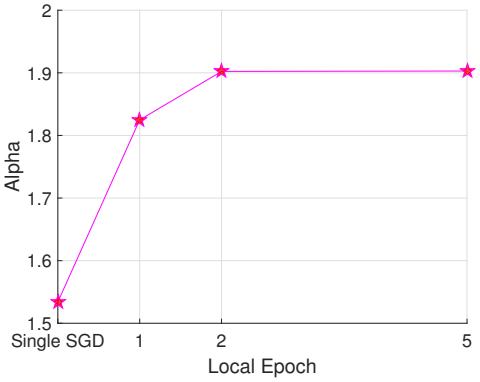
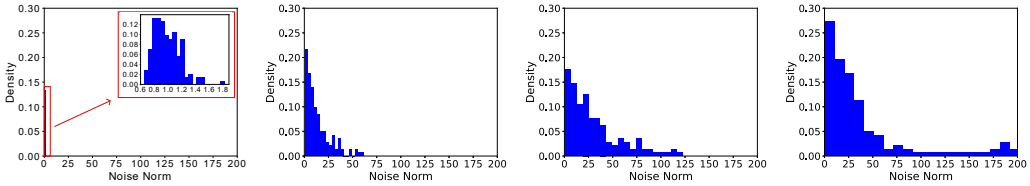


Figure 11: Estimation of α for Shakespeare dataset.



(a) Single SGD (b) Local Epoch=1 (c) Local Epoch=2 (d) Local Epoch=5

Figure 12: Distributions of the norms of the pseudo-gradient noises for Shakespeare dataset.

$m = 100$ clients participate in the experiment. The communication round is $T = 150$. The global learning rate is set to 1.0, and the local learning rate is set to 0.1. The batch size is chosen to be 64.

B.3.2 Experimental results

Table 3 shows the error rate of α-value estimation in different cases, and this implies that the estimation of α-value is within an acceptable margin of error.

C Experiments in Section 5

In this section, we describe the details of the numerical experiments from Section 5 and provide some extra experimental results.

C.1 Experiment details

C.1.1 Strongly Convex Model with Synthetic Data

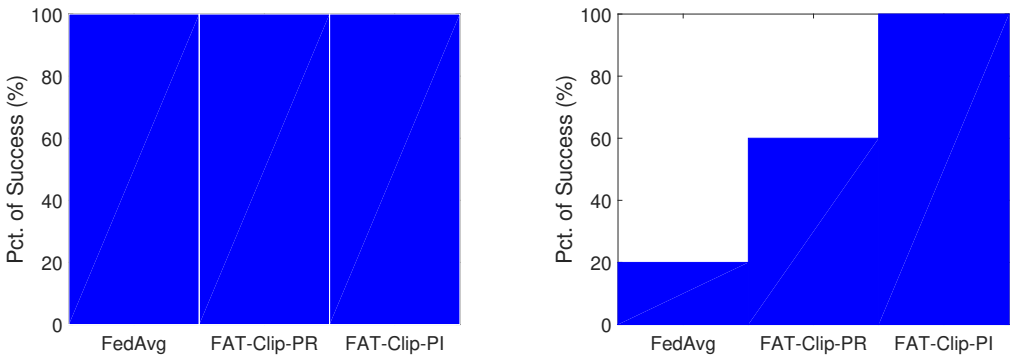
In these experiments, we consider a strongly convex model for Problem (1) as follows:

$$f_i(x) = \mathbb{E}_\xi \left[f_i(x, \xi) \right]$$
$$f_i(x, \xi) = \frac{1}{2} \|x\|^2 + \langle \xi, x \rangle ,$$

where $x \in \mathbb{R}^{3\times 1}$ and ξ is a random vector. The optimal solution is $f(x^*) = 0$ with $x^* = [0; 0; 0]$.

Table 3: Error rate (%) of α-value estimation.

	NonIID Index (p)			
	1	2	5	10
Single SGD	-2.82	-1.09	-0.12	3.12
Local Epoch=1	1.19	0.37	1.4	2.08
Local Epoch=2	1.8	1.4	1.43	1.74
Local Epoch=5	1.86	0.23	0.56	0.25



(a) non-i.i.d. index $p = 1$

(b) non-i.i.d. index $p = 5$

Figure 13: Percentage of successful training over 5 trials when applying FedAvg, FAT-Clipping-PR and FAT-Clipping-PI to CIFAR-10 dataset in non-i.i.d. cases.

To compare the performance of FedAvg, FAT-Clipping-PI and FAT-Clipping-PR , we consider the noise ξ to be a Cauchy distribution($\alpha < 2$, fat-tailed) with a location parameter of 0 and a scale parameter of 2.1.

To compare the performance of FAT-Clipping-PI and FAT-Clipping-PR under different scenarios, we consider the noise ξ having different tail-indexes ($\alpha = 0.5, 1.0,$ and 1.5) with the same location parameters of 0 and the same scale parameters of 1.

For all the distributions of ξ mentioned above, we use the same experimental setup. There are $m = 5$ clients participating in the training. We choose the starting point $x_0 = [2; 1; 1.5]$. We set the global learning rate $\frac{\eta\eta_L}{m} = 0.1$ and the local learning rate $\eta_L = 0.1$. The local steps we use is $K = 2$, and the communication round is $T = 300$. The clipping parameter in FAT-Clipping-PI we select is $\lambda = 3$, and the clipping parameter in FAT-Clipping-PR is $\lambda = 5$.

C.1.2 CNN (Non-convex Model) on the CIFAR-10

To test the performance of FAT-Clipping-PI and FAT-Clipping-PR for non-convex function, we run a convolutional neural network (CNN) on CIFAR-10 dataset. We compare FAT-Clipping-PI and FAT-Clipping-PR with FedAvg under different data heterogeneity.

In this experimental setting, we randomly select five clients from $m = 10$ clients to participate in each round of the training. The local epoch we use is two. The clipping parameter in FAT-Clipping-PI we select is $\lambda = 50$, and the clipping parameter in FAT-Clipping-PR is $\lambda = 2$. All the remaining settings are the same as described in Appendix B.1.1.

C.2 Additional experimental Results

We provide two additional results when applying FedAvg, FAT-Clipping-PI and FAT-Clipping-PR to the CNN model on CIFAR-10 dataset. In Fig. 13, we show the percentage of successful training over 5 trials in non-i.i.d. cases when the non-i.i.d. index $p = 1$ and $p = 5$. These results further support our finding that FAT-Clipping methods and especially FAT-Clipping-PI reduce catastrophic training failures compared to FedAvg.