ON PROVABLE BENEFITS OF MUON IN FEDERATED LEARNING: IMPROVED COMMUNICATION COMPLEXITY AND BEYOND

Anonymous authors
Paper under double-blind review

ABSTRACT

The recently introduced optimizer, Muon, has gained increasing attention due to its superior performance across a wide range of applications. However, its effectiveness in federated learning remains unexplored. To address this gap, this paper investigates the performance of Muon in the federated learning setting. Specifically, we propose a new algorithm, FedMuon, and establish its convergence rate for nonconvex problems. Our theoretical analysis reveals multiple favorable properties of FedMuon. In particular, due to its orthonormalized update direction, FedMuon achieves significantly improved communication complexity compared to existing momentum-based federated learning methods. Furthermore, it does not rely on any heterogeneity assumptions or specialized operations to guarantee convergence, its learning rate is independent of problem-specific parameters, and, importantly, it can naturally accommodate heavy-tailed noise. Finally, extensive experiments on a variety of neural network architectures validate the effectiveness of the proposed algorithm.

1 INTRODUCTION

Recently, several new optimizers have been developed based on various inductive biases regarding machine learning models. Among them, Muon (Jordan et al., 2024) has gained more attention due to its superior performance across a wide range of applications. Muon is essentially a stochastic gradient descent with momentum (SGDM) algorithm. Its key difference from traditional SGDM lies in directly optimizing a two-dimensional matrix, rather than flattening it into a vector, using an orthonormalized momentum matrix. Specifically, assuming the momentum of the stochastic gradient in the t-th iteration is denoted by $M_t \in \mathbb{R}^{m\times n}$, Muon orthonormalizes it as follows:

$$O_t = \arg\min_{O} \|O - M_t\|_F^2, \quad s.t. \quad O^T O = I_n \ , \tag{1}$$

where $I_n \in \mathbb{R}^{n\times n}$ denotes the identity matrix. The optimal solution to this problem is $O_t = U_t V_t^T$ where $U_t \in \mathbb{R}^{m\times r}$ and $V_t \in \mathbb{R}^{n\times r}$ are obtained from the singular value decomposition (SVD) of M_t, i.e., $M_t = U_t S_t V_t^T$. Here, $S_t \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose diagonal entries are the singular values of M_t, and r denotes the rank of M_t. Muon proposes using Newton–Schulz approach (Bernstein & Newhouse, 2024) to approximately solve this problem, instead of SVD, in order to accelerate computation. Due to its orthonormalization step, Muon has demonstrated strong performance across a wide range of applications, such as the pretraining of large language models (Liu et al., 2025).

The theoretical convergence rate of Muon has been well studied this year (Li & Hong, 2025; Shen et al., 2025; An et al., 2025; Kovalev, 2025; Zhang et al., 2025; Sato et al., 2025; Sfyraki & Wang, 2025; Chen et al., 2025). For example, Li & Hong (2025) provided the first convergence analysis for Muon when the loss function is nonconvex. Shen et al. (2025) established the convergence rate of Muon when the loss function is nonconvex and star-convex. However, all these existing works focus solely on the single-machine setting, making it unclear how well Muon performs in the federated learning context. Federated learning (McMahan et al., 2017) is an important distributed machine learning framework that enables model training on decentralized data without sharing raw data. On the other hand, the orthonormalization step in Muon introduces new properties to the search direction, such as bounded magnitude. This naturally leads to the question: **how does Muon perform in**

the federated learning setting? Specifically, what convergence rate and communication complexity can Muon achieve in this context?

To answer this question, we first develop a new federated optimization algorithm, FedMuon, which employs Muon to update variables on each worker and periodically communicates these updates to the central server. Then, we established the convergence rate of FedMuon for nonconvex problems under mild assumptions. Specifically, our theoretical analyses show that FedMuon enjoys the following favorable properties:

- FedMuon achieves significantly better communication complexity than existing momentum-based federated optimization algorithms. More importantly, its communication complexity can even match that of variance-reduction-based federated optimization algorithms (see Remark 5.6).
- The convergence analysis of FedMuon naturally accommodates the heterogeneous setting, as it does not rely on any assumptions about heterogeneity (see Remark 5.2). Moreover, the learning rate of FedMuon is inherently independent of problem-specific parameters, such as the Lipschitz constant (see Remark 5.4).
- FedMuon can naturally accommodate heavy-tailed noise, as it does not require gradient clipping to guarantee convergence (see Section 6).

The detailed comparison between FedMuon and existing state-of-the-art methods can be found in Table 2.1. All these favorable properties are due to the orthonormalization operation in Muon. To the best of our knowledge, this is the first work revealing these favorable properties of Muon in federated learning. Finally, we performed extensive experiments to validate the performance of our new algorithm and the experimental results confirm the efficacy of FedMuon.

2 Related Work

2.1 Federated Optimization

To solve federated learning models, numerous federated optimization algorithms (McMahan et al., 2017; Stich, 2018; Yu et al., 2019b;a; Yang et al., 2021; Khanduri et al., 2021; Wu et al., 2023) have been proposed and analyzed in the past few years. For example, Yu et al. (2019b) established a convergence rate of $O(1/\epsilon^4)$ and a communication complexity of $O(1/\epsilon^3)$ for LocalSGD in nonconvex optimization problems by relying on a bounded gradient norm, where $\epsilon > 0$ denotes the solution accuracy. Yu et al. (2019a) established the same convergence rate and communication complexity for LocalSGD with momentum (LocalSGDM) in nonconvex optimization problems. Unlike LocalSGD, it does not require a bounded gradient norm but instead relies on a bounded heterogeneity assumption. Under the same heterogeneity assumption as in Yu et al. (2019a), Khanduri et al. (2021) proposed STEM, which uses the stochastic variance-reduced gradient to improve the convergence rate to $O(1/\epsilon^3)$ and communication complexity to $O(1/\epsilon^2)$ for nonconvex problems.

To mitigate the influence of heterogeneous data distributions, a couple of federated optimization algorithms (Karimireddy et al., 2020; Cheng et al., 2024; Yan et al., 2025) have been developed to establish convergence rates without making any assumptions about heterogeneity. Essentially, these methods introduce a global correction term to mitigate heterogeneity. For example, Karimireddy et al. (2020) proposed SCAFFOLD, which uses a global correction term to adjust the local stochastic gradient, and established established its convergence rate for both strongly convex and nonconvex problems. In particular, this algorithm achieves the same convergence rate and communication complexity as LocalSGD, LocalSGDM, and STEM for nonconvex problems. Later, Cheng et al. (2024) leveraged variance reduced techniques to improve the convergence rate to $O(1/\epsilon^3)$ and the communication complexity to $O(1/\epsilon^2)$. Building on this strategy, Yan et al. (2025) developed a problem-parameter-free algorithm, whose learning rate does not rely on problem-specific parameters, and established the same convergence rate and communication complexity as Cheng et al. (2024). However, all these federated optimization algorithms that do not rely on the heterogeneity assumption require communication of a global correction term, which introduces additional communication overhead in each round. An advantage of our FedMuon algorithm over these existing heterogeneous federated optimization algorithms is that it does not require a correction term or any heterogeneity assumption, as will be shown in the following sections. On the other hand, to handle the heavy-tailed noise, Lee et al. (2025) proposed using the gradient clipping technique on each worker to mitigate the influence of heavy-tailed noise. However, the gradient clipping approach requires a threshold to clip gradients, which is difficult to tune in practical applications.

Table 1: The comparison of convergence rate and communication complexity of different federated optimization algorithms for nonconvex problems. Note that all these algorithms can achieve linear speedup, so we omit this for simplicity. In the first column, **M** denotes momentum, **V** denotes variance reduction. In the fifth column, **Communicate control variate** indicates that there is no heterogeneity assumption, but this comes at the cost of communicating additional control variates.

	Algorithms	Convergence Rate	Communication Complexity	Heterogeneity	Parameter Free	Heavy-tailed Noise
-	**FedAvg/LocalSGD** (Yu et al., 2019b)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Bounded gradient	✗	✗
	SCAFFOLD (Karimireddy et al., 2020)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Communicate control variate	✗	✗
M	**LocalSGDM** (Yu et al., 2019a)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Bounded heterogeneity	✗	✗
	FedAvg-M (Cheng et al., 2024)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Communicate control variate	✗	✗
	SCAFFOLD-M (Cheng et al., 2024)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Communicate control variate	✗	✗
	PAdaMFed (Yan et al., 2025)	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^3}\right)$	Communicate control variate	✓	✗
V	**STEM** (Khanduri et al., 2021)	$O\left(\frac{1}{\epsilon^3}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	Bounded heterogeneity	✗	✗
	FedAvg-VR (Cheng et al., 2024)	$O\left(\frac{1}{\epsilon^3}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	Communicate control variate	✗	✗
	SCAFFOLD-VR (Cheng et al., 2024)	$O\left(\frac{1}{\epsilon^3}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	Communicate control variate	✗	✗
	PAdaMFed-VR (Yan et al., 2025)	$O\left(\frac{1}{\epsilon^3}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	Communicate control variate	✓	✗
M	**FedMuon** Corollary 5.5	$O\left(\frac{1}{\epsilon^4}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$	-	✓	✗
	FedMuon Corollary 6.5	$O\left(\frac{1}{\epsilon^{\frac{2p}{p-1}}}\right)$	$O\left(\frac{1}{\epsilon^{\frac{p}{p-1}}}\right)$	-	✓	✓

2.2 MUON

Muon was first proposed in Jordan et al. (2024) to optimize the hidden layer of deep neural networks, which showed great performance for various applications. Several recent works (Li & Hong, 2025; An et al., 2025; Kovalev, 2025; Shen et al., 2025; Riabinin et al., 2025; Zhang et al., 2025; Sato et al., 2025; Sfyraki & Wang, 2025; Chen et al., 2025; Pethick et al., 2025) have attempted to establish its convergence rate in the single-machine setting. In particular, Li & Hong (2025) established the convergence rate of Muon for nonconvex problems under the assumption of Frobenius-norm Lipschitz smoothness. An et al. (2025) provided its convergence rate under a generalized-norm Lipschitz smoothness assumption. Kovalev (2025) further analyzed Muon's convergence rate given the spectral-norm Lipschitz smoothness assumption. The recent work (Shen et al., 2025) provided convergence analysis for Muon under all these smoothness assumptions when the loss function is nonconvex and star-convex. In addition, Chen et al. (2025) established the convergence rate of Muon from the perspective of spectral norm constraints. Zhang et al. (2025) combined Muon with Adagrad to introduce the adaptive learning rate and then established its convergence rate for nonconvex problems. Moreover, Sfyraki & Wang (2025) established the convergence rate of Muon from the perspective of Frank-Wolfe method for nonconvex problems. It then introduced the gradient clipping technique to Muon to handle the heavy-tailed noise. In this paper, we will show that FedMuon can still guarantee convergence without relying on the clipping technique.

3 PROBLEM SETUP

3.1 PROBLEM DEFINITION

In this paper, K (where $K > 0$) workers collaboratively optimize the following problem:

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{K} \sum_{k=1}^{K} f^{(k)}(X)\,, \tag{2}$$

where $f^{(k)}(X) = \mathbb{E}[f^{(k)}(X;\xi)]$, $X \in \mathbb{R}^{m\times n}$ denotes the optimization variable, and the superscript $k \in \{1,\cdots,K\}$ denotes the index of workers. In the federated learning setting, all workers communicate with a central server to exchange updated variables or gradients.

In this paper, for a matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|_F$ denotes the Frobenius norm, $\|X\|_*$ denotes the nuclear norm, and $\|X\|_2$ denotes the spectral norm. In addition, for $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times n}$, we have $\langle X, Y\rangle = \text{Tr}(X^T Y)$, where $\text{Tr}(\cdot)$ denotes the trace of a matrix. Moreover, $\bar{X} = \frac{1}{K}\sum_{k=1}^{K} X^{(k)}$.

3.2 Assumption

In this paper, we introduce the following assumptions, which has been commonly used in Li & Hong (2025); Shen et al. (2025); Zhang et al. (2025); Sato et al. (2025); Sfyraki & Wang (2025).

Assumption 3.1. *For any $k \in \{1,\cdots,K\}$, the loss function $f^{(k)}(\cdot)$ is L-smooth, i.e., for any $X_1 \in \mathbb{R}^{m\times n}$ and $X_2 \in \mathbb{R}^{m\times n}$, it satisfies: $\|\nabla f^{(k)}(X_1) - \nabla f^{(k)}(X_2)\|_F \le L\|X_1 - X_2\|_F$, where $L > 0$ is a constant.*

Assumption 3.2. *For any $k \in \{1,\cdots,K\}$, the stochastic gradient $\nabla f^{(k)}(X;\xi)$ is an unbiased estimator of the full gradient and satisfies: $\mathbb{E}[\|\nabla f^{(k)}(X;\xi) - \nabla f^{(k)}(X)\|_F^2] \le \sigma^2$, where $\sigma > 0$ is a constant.*

Assumption 3.3. *For any $k \in \{1,\cdots,K\}$, the stochastic gradient $\nabla f^{(k)}(X;\xi)$ is an unbiased estimator of the full gradient and satisfies: $\mathbb{E}[\|\nabla f^{(k)}(X;\xi) - \nabla f^{(k)}(X)\|_F^p] \le \sigma^p$, where $\sigma > 0$ is a constant and $p \in (1,2]$.*

Note that Assumption 3.3 characterizes heavy-tailed noise. When $p = 2$, it reduces to the standard bounded noise assumption in Assumption 3.2.

It is worth noting that our algorithm does not require any assumptions regarding data heterogeneity, unlike many existing federated learning works (Yu et al., 2019a; Karimireddy et al., 2020) that rely on inequalities such as the following:

$$\mathbb{E}[\|\nabla f^{(k)}(X) - \nabla f(X)\|_F^2] \le \delta^2\,, \qquad \mathbb{E}[\|\nabla f^{(k)}(X)\|_F^2] \le B\mathbb{E}[\|\nabla f(X)\|_F^2] + \delta^2\,, \tag{3}$$

where $\delta > 0$ and $B > 0$ are constants.

Algorithm 1 FedMuon

Input: $\eta > 0, \beta > 0, \tau > 1$.
1: **for** $t = 0,\cdots,T-1$, the k-th worker **do**
2: **if** $t == 0$ **then**
3: $M_t^{(k)} = \nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)})$
4: **else**
5: $M_t^{(k)} = (1-\beta)M_{t-1}^{(k)} + \beta\nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)})$ // Update gradient momentum $M_t^{(k)}$
6: **end if**
7: $(U_t^{(k)}, S_t^{(k)}, V_t^{(k)}) = \text{SVD}(M_t^{(k)})$ // Orthonormalize $M_t^{(k)}$ with Newton–Schulz approach
8: $X_{t+1}^{(k)} = X_t^{(k)} - \eta U_t^{(k)}(V_t^{(k)})^T$ // Update variable $X_t^{(k)}$
9: **if** $\text{mod}(t+1,\tau) == 0$ **then**
10: $X_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^{K} X_{t+1}^{(k')}$ // Perform communication
11: **end if**
12: **end for**

4 Algorithm

In Algorithm 1, we developed a novel federated optimization algorithm based on Muon, i.e., FedMuon. In the t-th iteration, as shown in Step 5 of Algorithm 1 each worker k uses its local training samples to update the momentum $M_t^{(k)} \in \mathbb{R}^{m\times n}$ as follows:

$$M_t^{(k)} = (1-\beta)M_{t-1}^{(k)} + \beta\nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)})\,, \tag{4}$$

In Eq. (4), $\beta \in (0,1)$ denotes the hyperparameter. $\nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)})$ denotes the stochastic gradient, where $X_t^{(k)}$ denotes the variable and $\xi_t^{(k)}$ represents the randomly selected training samples on the k-th worker in the t-th iteration. In Step 7, each worker uses Newton–Schulz approach to perform SVD for $M_t^{(k)}$, where $U_t^{(k)} \in \mathbb{R}^{m\times r}$ is composed of left singular vectors, $V_t^{(k)} \in \mathbb{R}^{n\times r}$

consists of the right singular vectors, and $S_t^{(k)}$ is a diagonal matrix of singular values. With such a decomposition, FedMuon updates variable as follows:

$$X_{t+1}^{(k)} = X_t^{(k)} - \eta U_t^{(k)} (V_t^{(k)})^T \,, \tag{5}$$

where $\eta > 0$ denotes the learning rate. Every $\tau > 1$ iterations, as shown in Step 10 of Algorithm 1, each worker k uploads its local variable $X_{t+1}^{(k)}$ to the central server. The server then averages all received variables and broadcasts the result to all workers.

Remark 4.1. *In Algorithm 1, only the optimization variable is communicated between the workers and the central server. In contrast, vector-based methods that use momentum, including LocalSGDM (Yu et al., 2019a) and STEM (Khanduri et al., 2021), require communication of both the optimization variable and the momentum to guarantee convergence. Moreover, vector-based methods that do not rely on heterogeneity assumptions, such as SCAFFOLD Karimireddy et al. (2020), SCAFFOLD-M (Cheng et al., 2024), and PAdaMFed (Yan et al., 2025), require communication of a control variate in addition to the optimization variable.*

5 Theoretical Results Under Bounded Variance

In this section, we establish the convergence rate of FedMuon given Assumptions 3.1, 3.2.

5.1 Convergence Rate

Theorem 5.1. *Given Assumptions 3.1, 3.2, when $0 < \beta < 1$, FedMuon in Algorithm 1 can achieve the following convergence upper bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq \frac{f(X_0) - f(X_*)}{\eta T} + \frac{\eta n L}{2} + 4\eta\tau n L + \frac{2\sqrt{n}\sigma}{\beta T} + \frac{2\eta n L}{\beta} + \frac{2\sqrt{\beta}\sqrt{n}\sigma}{\sqrt{K}} \,. \tag{6}$$

Remark 5.2. ***(Benefit 1: No Heterogeneity Assumptions and No Need to Communicate Control Variates)*** *The proof of Theorem 5.1 does not require any assumptions regarding heterogeneity, such as Eq. (3), which is consistent with existing methods such as SCAFFOLD (Karimireddy et al., 2020), FedAvg-M/FedAvg-VR Cheng et al. (2024), SCAFFOLD-M/SCAFFOLD-VR Cheng et al. (2024), and PAdaMFed/PAdaMFed-VR (Yan et al., 2025). However, unlike these methods, FedMuon only needs to communicate the optimization variable, whereas the others require communication of an additional control variate to remove the heterogeneity assumption (see Table 2.1).*

Corollary 5.3. *In Theorem 5.1, for a sufficiently large T, by setting $\eta = \frac{K^{1/4}}{T^{3/4}}$, $\beta = \frac{K^{1/2}}{T^{1/2}}$, and $\tau = \frac{T^{1/2}}{K^{1/2}}$, FedMuon in Algorithm 1 can achieve the following convergence upper bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq O\left(\frac{f(X_0) - f(X_*) + nL + \sqrt{n}\sigma}{(KT)^{1/4}} + \frac{\sqrt{n}\sigma}{(KT)^{1/2}} + \frac{K^{1/4} n L}{T^{3/4}}\right) \,. \tag{7}$$

Since the first term in the convergence upper bound in Corollary 5.3 dominates the other terms, FedMuon achieves a convergence rate of $O\left(\frac{1}{(KT)^{1/4}}\right)$, which indicates a linear speedup with respect to the number of workers K. This convergence rate matches that of the vector-based counterparts that also use momentum, such as FedAvg-M Cheng et al. (2024), SCAFFOLD-M Cheng et al. (2024), and PAdaMFed (Yan et al., 2025).

Remark 5.4. ***(Benefit 2: Problem-Parameter-Free Hyperparameters)*** *In Corollary 5.3, the learning rate η, the momentum coefficient β, and the communication period τ do not depend on problem-specific parameters, such as the Lipschitz constant L. They depend only on the number of workers and the number of iterations, making them easy to tune. This is consistent with the vector-based method PAdaMFed (Yan et al., 2025). However, PAdaMFed has a higher communication complexity than ours, as depicted in Remark 5.6 and Table 2.1.*

Corollary 5.5. *In Theorem 5.1, by setting $T = O(\frac{1}{K\epsilon^4})$, $\eta = O(K\epsilon^3)$, $\beta = O(K\epsilon^2)$, $\tau = O(\frac{1}{K\epsilon^2})$, FedMuon achieves the ϵ-accuracy solution: $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq O(\epsilon)$, where $\epsilon > 0$ is a constant.*

Remark 5.6. ***(Benefit 3: Improved Communication Complexity)*** *From Corollary 5.5, it is easy to know that the communication complexity of FedMuon is $T/\tau = O(\frac{1}{\epsilon^2})$. This communication*

complexity $O(\frac{1}{\epsilon^2})$ is better than $O(\frac{1}{\epsilon^3})$ of the vector-based counterparts that also use momentum, such as LocalSGDM (Yu et al., 2019a), FedAvg-M Cheng et al. (2024), SCAFFOLD-M (Cheng et al., 2024), and PAdaMFed (Yan et al., 2025), and it can match the vector-based methods using the variance-reduced gradient, including SCAFFOLD-M-VR Cheng et al. (2024) and PAdaMFed-VR (Yan et al., 2025) (see Table 2.1). To the best of our knowledge, this is the first algorithm achieving such a small communication complexity without using the variance reduction technique. The improvement lies in the fact that ***a smaller learning rate*** *$\eta = O(K\epsilon^3)$, which results from the orthonormalization operation (explained in the next subsection),* ***allows a larger communication period*** *$\tau = O(\frac{1}{K\epsilon^2})$. Specifically, as shown in the third term, $4\eta\tau nL$, of the convergence upper bound in Theorem 5.1, a small learning rate η permits a large communication period τ. As a result, the communication complexity T/τ becomes smaller. On the contrary, the learning rate of existing momentum-based methods, such as LocalSGDM (Yu et al., 2019a), is in the order of $O(K\epsilon^2)$. As a result, the communication period τ must be as small as $O(\frac{1}{K\epsilon})$, leading to a worse communication complexity $T/\tau = O(\frac{1}{\epsilon^3})$.*

5.2 Sketch of the Proof and Reasons for the Benefits

The proof of Theorem 5.1 relies on the following three key lemmas.

Lemma 5.7. ***(Consensus Error)*** *Given Assumptions 3.1, 3.2, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F \leq 2\eta\tau\sqrt{n}\,. \tag{8}$$

Reason for Benefit 1. Bounding the consensus error regarding variables can be converted to bounding that regarding gradients, which can be seen in the proof of Lemma B.2 in Appendix. When bounding the consensus error regarding gradients, it typically requires an assumption about heterogeneity. To remove such assumptions, there are two commonly used approaches. The first one is to assume the second moment of the gradient is upper bounded, e.g., $\|\nabla f^{(k)}(X)\| \leq G$ in (Yu et al., 2019b), where $G > 0$ is a constant. With such an assumption, it is easy to bound the gradient consensus error as follows:

$$\|\nabla f^{(k)}(X) - \nabla f(X)\| \leq \|\nabla f^{(k)}(X)\| + \|\nabla f(X)\| \leq 2G. \tag{9}$$

However, this assumption is too strong for practical applications. The second approach is to introduce the control variate, which can convert the consensus error to bound some local updates. For example, when the local gradient estimator is composed of a local stochastic gradient $g_t^{(k)}$ and a control variate $c_t^{(k)}$, the consensus error regarding them can be converted in the following manner:

$$\|g_t^{(k)} - c_t^{(k)} - \frac{1}{K}\sum_{k'=1}^{K}(g_t^{(k')} - c_t^{(k')})\| \leq \|g_t^{(k)} - c_t^{(k)}\|. \tag{10}$$

The right-hand side does not involve the difference between the local variate and the global variate. It only involves the local variates. As a result, it avoids the influence from heterogeneity, eliminating the need for heterogeneity-related assumptions. Representative methods in this category include SCAFFOLD (Karimireddy et al., 2020), SCAFFOLD-M/SCAFFOLD-VR Cheng et al. (2024), and PAdaMFed/PAdaMFed-VR (Yan et al., 2025). **However, this approach requires the communication of an additional control variate, incurring higher communication costs.** In contrast, bounding the consensus error in FedMuon is naturally independent of heterogeneity. Specifically, due to the orthonormalization operation, the updating direction $\|U_t^{(k)}(V_t^{(k)})^T\|_F$ is naturally upper bounded. As a result, bounding the consensus error with respect to this direction does not require any assumptions on heterogeneity, since we have

$$\frac{1}{K}\sum_{k=1}^{K}\|U_t^{(k)}(V_t^{(k)})^T - \frac{1}{K}\sum_{k'=1}^{K}U_t^{(k')}(V_t^{(k')})^T\|_F \leq 2\frac{1}{K}\sum_{k=1}^{K}\|U_t^{(k)}(V_t^{(k)})^T\|_F \leq 2\sqrt{n}. \tag{11}$$

In summary, due to the orthonormalization operation, FedMuon does not require any assumptions regarding heterogeneity for convergence analysis.

Lemma 5.8. ***(Loss Function Update)*** *Given Assumptions 3.1, 3.2, the following inequality holds:*

$$f(\bar{X}_{t+1}) \leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}L\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F$$

$$+ 2\eta\sqrt{n}\|\frac{1}{K}\sum_{k=1}^{K} f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K} M_t^{(k)}\|_F + \frac{\eta^2 nL}{2} . \tag{12}$$

Reason for Benefit 2. This lemma characterizes how the loss function value is updated in each iteration. It is worth noting that there is no any requirement on the learning rate η. On the contrary, most existing methods require $\eta \leq O(\frac{1}{L})$, where the Lipschitz constant L is NOT easy to know. For example, considering the vector scenario and assuming the global updating direction is $\bar{q}_t$, then most momentum-based methods, such as Lemma A.5 in Khanduri et al. (2021), have the following inequality in their proof:

$$f(\bar{x}_{t+1}) \leq f(\bar{x}_t) - \frac{\eta}{2}\|\nabla f(\bar{x}_t)\|^2 + \left(\frac{\eta^2 L}{2} - \frac{\eta}{2}\right)\|\bar{q}_t\|^2 + \eta\|\bar{q}_t - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(x_t^{(k)})\| . \tag{13}$$

Here, a typical operation is to let $\frac{\eta^2 L}{2} - \frac{\eta}{2} \leq -\frac{\eta}{4}$ by setting $\eta \leq \frac{1}{2L}$. Since the updating direction $\|U_t^{(k)}(V_t^{(k)})^T\|_F$ in FedMuon is upper bounded, the constraint regarding the learning rate η can be avoided. The details can be found in the proof of Lemma B.1 in Appendix.

In summary, due to the orthonormalization operation, the hyperparameter of FedMuon does not rely on the problem-specific parameters such as the Lipschitz constant.

Lemma 5.9. ***(Gradient Error)*** *Given Assumptions 3.1, 3.2, by setting $\beta < 1$, the following inequality holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K} f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K} M_t^{(k)}\|_F] \leq \frac{1}{T}\frac{\sigma}{\beta} + \frac{\eta\sqrt{n}L}{\beta} + \frac{\sqrt{\beta}\sigma}{\sqrt{K}} . \tag{14}$$

Reason for Benefit 3. This lemma characterizes the gradient error. A key feature of the upper bound lies in its second term, which is in the order of $O(\frac{\eta}{\beta})$, where η is the learning rate and β is the momentum coefficient. Due to this term, η must have a higher order dependence on ϵ, i.e., $O(\epsilon^3)$, compared to β, which is in the order of $O(\epsilon^2)$, in order to ensure convergence, i.e, $O(\frac{\eta}{\beta}) = O(\epsilon)$. This is also due to the orthonormalization operation. The details can be found in the proof of T_1 in Lemma B.3 in the Appendix. In contrast, existing moving-average-based momentum methods include a term $O(\frac{\eta^2}{\beta})$ in the upper bound of their gradient error. For example, as shown in Eq. (34) in (Qiu et al., 2020), the coefficient of the second term is in the order of $O(\frac{v_t^2}{\alpha v_t})$. Then, its learning rate v_t can be set to a larger value $O(\epsilon^2)$, as shown in its Theorem 4.9.

In summary, the orthonormalization operation results in a smaller learning rate, which in turn allows a larger communication period and ultimately leads to a smaller communication complexity.

6 Theoretical Results Under Heavy-Tailed Noise

In this section, we establish the convergence rate of FedMuon given Assumptions 3.1, 3.3. To the best of our knowledge, this is the first work establishing the convergence rate for Muon in federated learning given heavy-tailed noise.

Theorem 6.1. *Given Assumptions 3.1, 3.3, when $0 < \beta < 1$, FedMuon in Algorithm 1 can achieve the following convergence upper bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq \frac{f(X_0) - f(X_*)}{\eta T} + \frac{\eta nL}{2} + 4\eta\tau nL + \frac{4\sqrt{2n}\sigma}{\beta T} + \frac{2\eta nL}{\beta} + \frac{4\sqrt{2n}\beta^{1-\frac{1}{p}}}{K^{1-\frac{1}{p}}}\sigma . \tag{15}$$

Remark 6.2. *The proof of Theorem 6.1 also does not rely on any assumptions regarding heterogeneity. In addition, the last term demonstrates how the tail index p affects the convergence upper bound. Note that Sfyraki & Wang (2025) requires the gradient clipping operation to establish the convergence rate of Muon in the single-machine setting. In contrast, our algorithm and proof do NOT require such a clipping operation.*

Corollary 6.3. *In Theorem 5.1, for a sufficiently large T, by setting $\eta = \frac{K^{1/4}}{T^{3/4}}$, $\beta = \frac{K^{1/2}}{T^{1/2}}$, and $\tau = \frac{T^{1/2}}{K^{1/2}}$, FedMuon in Algorithm 1 can achieve the following convergence upper bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq O\left(\frac{f(X_0) - f(X_*) + nL}{(KT)^{1/4}} + \frac{\sqrt{n}\sigma}{(KT)^{1/2}} + \frac{K^{1/4}nL}{T^{3/4}} + \frac{\sqrt{n}\sigma}{(KT)^{\frac{p-1}{2p}}}\right) . \tag{16}$$

Remark 6.4. *Since $p \in (1, 2]$, the last term in the convergence upper bound of Corollary 6.3 dominates the other terms. Therefore, the convergence rate of FedMuon under heavy-tailed noise is $O\left(\frac{1}{(KT)^{\frac{p-1}{2p}}}\right)$. This convergence rate also indicates a linear speedup with respect to the number of workers, and the hyperparameter does not rely on problem-specific parameters like Liptschitz constant and gradient noise. Moreover, when $K = 1$, it matches the convergence rate of the single-machine method (Liu & Zhou, 2024). When $p = 2$, it reduces to the convergence rate in Corollary 5.3. Furthermore, like the regular noise setting, the learning rate η, the momentum coefficient β, and the communication period τ do not depend on problem-specific parameters, such as the Lipschitz constant L. They depend only on the number of workers and the number of iterations.*

Corollary 6.5. *In Theorem 5.1, by setting $T = O\left(\frac{1}{K\epsilon^{\frac{2p}{p-1}}}\right)$, $\eta = O\left(K\epsilon^{\frac{3p}{2(p-1)}}\right)$, $\beta = O\left(K\epsilon^{\frac{p}{p-1}}\right)$, $\tau = O\left(\frac{1}{K\epsilon^{\frac{p}{p-1}}}\right)$, FedMuon achieves the ϵ-accuracy solution: $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq O(\epsilon)$, where $\epsilon > 0$ is a constant.*

Remark 6.6. *From Corollary 6.5, we can know that the communication complexity is $T/\tau = O\left(\frac{1}{\epsilon^{\frac{p}{p-1}}}\right)$. When $p = 2$, it reduces to the communication complexity in Corollary 5.5.*

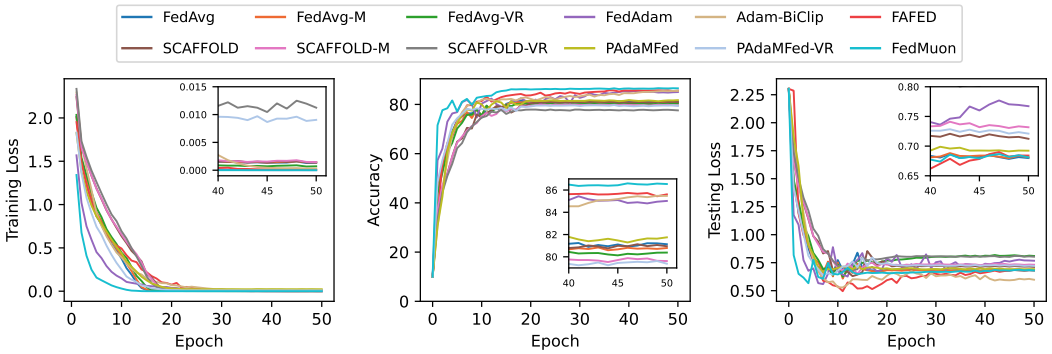

Figure 1: CIFAR-10 on ResNet-18 (period = 4).

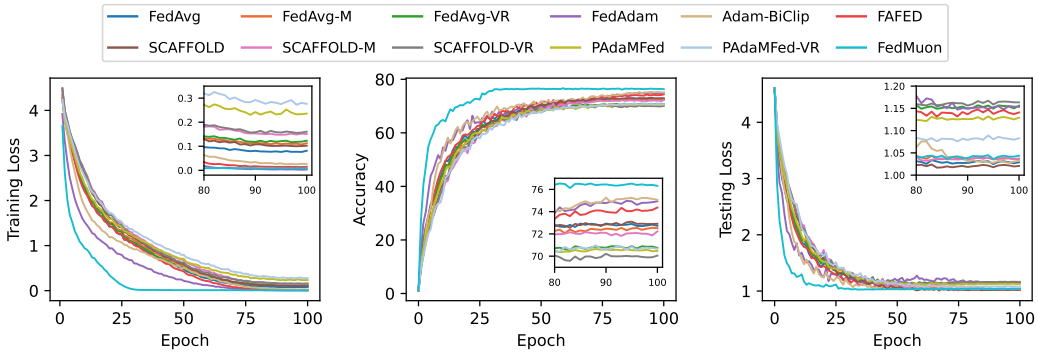

Figure 2: CIFAR-100 on ResNet-18 (period = 4).

7 Experiments

In our experiments, we evaluate the performance of FedMuon on three types of deep neural networks: convolutional neural networks, recurrent neural networks, and transformers, using both image and text datasets.

Experiment Settings. In our experiments, we include four categories of baselines to provide a comprehensive comparison. Specifically, we consider 1) the classical method, FedAvg (Yu et al., 2019b); 2) the control-variate-based method, including SCAFFOLD (Karimireddy et al., 2020), FedAvg-M/FedAvg-VR (Cheng et al., 2024), SCAFFOLD-M/SCAFFOLD-VR (Cheng et al., 2024); 3) the problem-parameter-free method, including PAdaMFed/PAdaMFed-VR (Yan et al., 2025);

and 4) adaptive methods, including FedAdam (Reddi et al., 2020), FAFED (Wu et al., 2023), and Adam-BiClip (Lee et al., 2025), where the last one uses the gradient clipping method to address heavy-tailed noise. Our federated environment is implemented on four NVIDIA RTX 6000 GPUs, where two workers are assigned to each GPU to simulate distributed clients, resulting in a total of eight workers ($K = 8$) participating in the federated training.

7.1 Image Classification with ResNet and Transformer

We conduct experiments on two widely used image classification benchmarks, CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009).

First, we adopt ResNet-18 (He et al., 2016) model for image classification. For fair comparisons, the hyperparameters of all baseline algorithms are carefully tuned through grid search to ensure their best performance. In particular, for FedMuon, the learning rate η is selected from $\{0.001, 0.002, 0.005, 0.01, 0.05\}$, and the weight decay from $\{0.0001, 0.001, 0.01, 0.05, 0.1, 0.2\}$. All methods are trained with a cosine decaying learning rate schedule. The momentum hyperparameters β for all methods are fixed at 0.9. The batch size of all datasets on each worker is 64.

The training loss, test accuracy, and testing loss are presented in Figure 1 and Figure 2 with communication period set to 4. The results demonstrate taht FedMuon exhibits a substantially faster decline in training loss, indicating fast convergence and improved learning efficiency compared with the baselines. Moreover, it consistently outperforms all competing baselines and achieves the highest testing accuracy over epochs. The testing loss curves further show that our approach attains a generalization ability comparable to or better than existing methods.

To further validate the generality of our approach on modern architectures, we also consider a Vision Transformer (ViT) model (Dosovitskiy et al., 2020) without pre-training, with the results shown in Figure 3. It can be observed that the improvement of FedMuon over the baselines is more significant on ViT compared to ResNet-18. In particular, adaptive baselines such as FedAdam, FAFED and Adam-BiClip, can achieve better performance than other methods, which is consistent with the analysis in (Zhang et al., 2020; Kunstner et al., 2024; Zhang et al., 2024). Furthermore, the block heterogeneity phenomenon in Transformers identified by Zhang et al. (2024) can also be effectively mitigated by FedMuon, contributing to its superior performance.

Additional results, including those with a communication period of 16, CIFAR10 under the heterogeneous settings, CIFAR100 with the ViT model, and the text classification task, are provided in Appendix A.

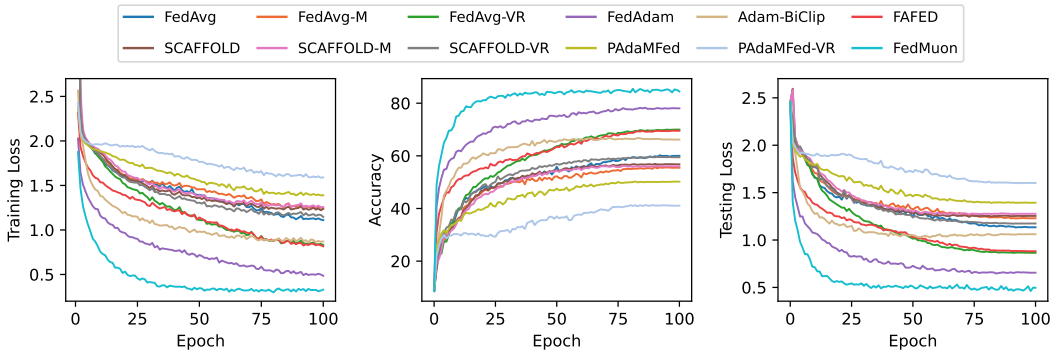

Figure 3: CIFAR-10 on ViT (period = 4).

8 Conclusion

In this paper, we developed a novel federated learning algorithm based on Muon optimizer. Our theoretical analysis identifies multiple favorable properties of Muon in the federated learning setting. In particular, its theoretical analysis does not require heterogeneity assumptions, its learning rate does not require the prior knowledge regarding the problem-specific parameter, learning rate, and it can naturally accommodate heavy-tailed noise. More importantly, it can achieve a much better communication complexity than its vector counterpart methods. The extensive experiments confirm the efficacy of our algorithm.

Ethics Statement This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, sensitive personal data, or applications that may directly cause societal harm. All datasets used are publicly available benchmark datasets and we follow the corresponding licenses. No additional risks regarding privacy, fairness, discrimination, or security are introduced. The proposed methodology focuses on algorithmic optimization and is intended solely for research purposes in machine learning.

Reproducibility Statement We have made efforts to ensure reproducibility of our results. All theoretical assumptions are explicitly stated in Section 3, and complete proofs are provided in Appendix B and Appendix C. All experimental details are described in Section 7 and Appendix A. The source code for reproducing all experiments will be made publicly available upon acceptance of the paper.

The Use of Large Language Models (LLMs) The use of LLMs in this work was limited to editing and polishing the text for readability. All research contributions were conducted entirely by the authors.

REFERENCES

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TdhkAcXkRi.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061, 2021.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

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

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *Advances in Neural Information Processing Systems*, 37:30106–30148, 2024.

Su Hyeong Lee, Manzil Zaheer, and Tian Li. Efficient distributed optimization under heavy-tailed noise. *arXiv preprint arXiv:2502.04164*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon and further. *arXiv e-prints*, pp. arXiv–2502, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

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

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

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *arXiv preprint arXiv:2008.10103*, 2020.

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Analysis of muon's convergence and critical batch size. *arXiv preprint arXiv:2507.01598*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster adaptive federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 10379–10387, 2023.

Wenjing Yan, Kai Zhang, Xiaolu Wang, and Xuanyu Cao. Problem-parameter-free federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZuazHmXTns.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Kentaro Yoshioka. vision-transformers-cifar10: Training vision transformers (vit) and related models on cifar-10. https://github.com/kentaroy47/vision-transformers-cifar10, 2024.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pp. 7184–7193. PMLR, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5693–5700, 2019b.

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

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adagrad meets muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024.

A Additional Experiments

A.1 More Experiments about Image Classification with ResNet and Transformer

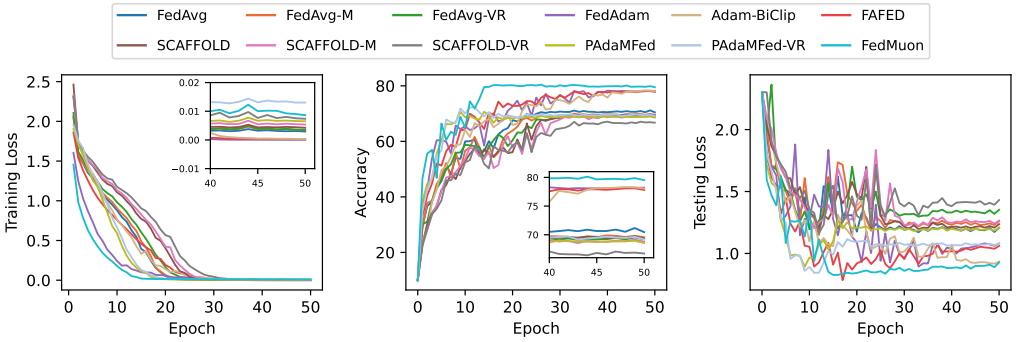

Figure 4: CIFAR-10 on ResNet-18 (period = 4, $Dir(0.5)$).

To evaluate the performance on heterogeneous setting, we further conduct experiments on CIFAR10 with ResNet-18. Specifically, the data are partitioned across clients using a Dirichlet distribution (Hsu et al., 2019) with $Dir(0.5)$. The results are shown in Figure 4. It can be observed that FedMuon consistently outperforms all baselines. While methods such as SCAFFOLD, SCAFFOLD-M/SCAFFOLD-VR, and PAdaMFed/PAdaMFed-VR mitigate date heterogeneity by introducing additional control variate, FedMuon stilll achieves superior performance without employing any such auxiliary mechanism, demonstrating its simplicity and effectiveness in heterogeneous environments.

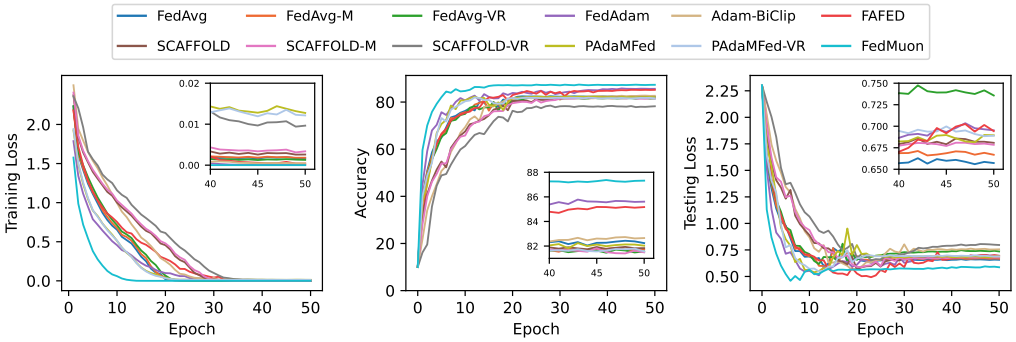

Figure 5: CIFAR-10 on ResNet-18 (period = 16).

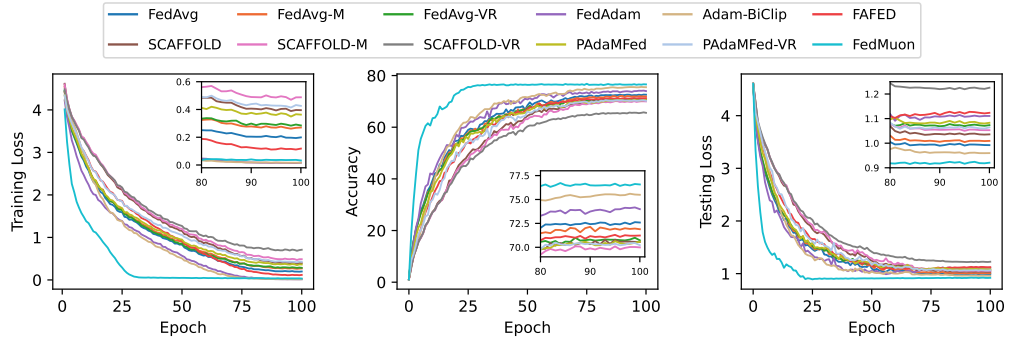

Figure 6: CIFAR-100 on ResNet-18 (period = 16).

Moreover, for the homogeneous setting, we also report the results with a communication period of 16, as shown in Figure 5 and Figure 6. The results show that FedMuon continues to outperform the

baselines even when the communication period is high, and the performance gap becomes larger at higher communication periods, particularly in terms of testing loss.

For the ViT model, we follow the implementation of Yoshioka (2024) and summarize the detail of architecture settings in Table 2.

Table 2: Architecture of the Vision Transformer (ViT) model.

Component	Configuration
Image patches (batches)	4
Attention heads per layer	8
Dimension per head	64
Transformer encoder depth	6
Dropout rate of encoder	0.1
MLP dimension	512
Dropout rate of MLP	0.1

More experimental results on CIFAR-10 with the ViT model are presented in Figure 7, and the results on CIFAR-100 are shown in Figure 8 and Figure 9. These results further confirm the effectiveness of FedMuon on Transformer-based architectures.

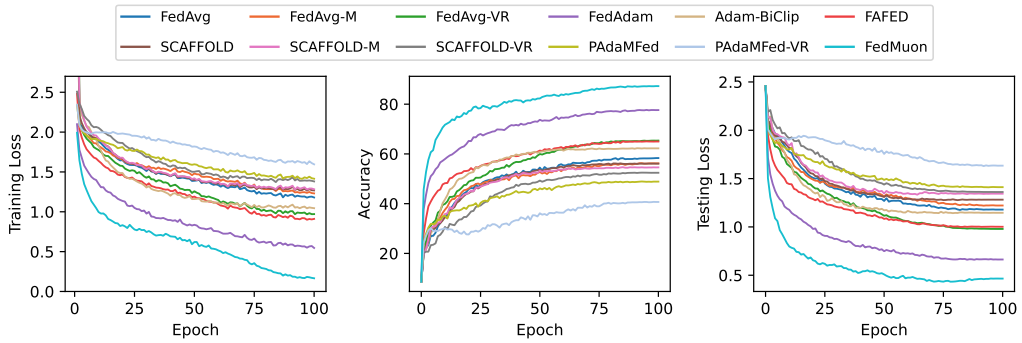

Figure 7: CIFAR-10 on ViT (period = 16).

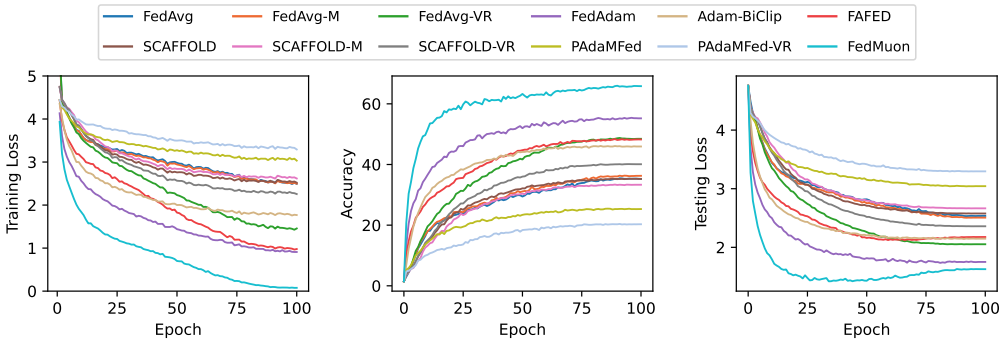

Figure 8: CIFAR-100 on ViT (period = 4).

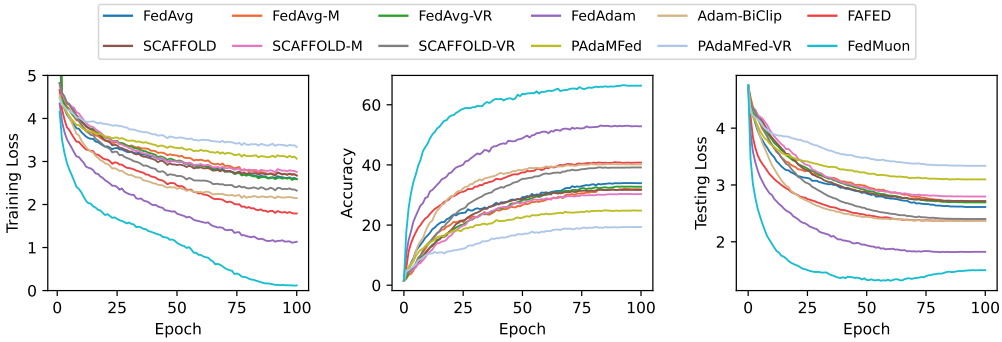

Figure 9: CIFAR-100 on ViT (period = 16).

A.2 Text Classification with RNN

Next, we evaluate our approach on a text classification task, where the data naturally exhibit heavy-tailed noise characteristics. We use the Sentiment140 dataset (Go et al., 2009) and adopt a recurrent neural network (RNN) (Elman, 1990). For the Sentiment140 dataset, the original corpus contains 1.6 million training samples and a testing set of merely 498 samples. To avoid overly fast convergence and better observe the training dynamics, we randomly subsample the training set and retain only 0.01% of the original training data for model training. The batch size of Sentiment140 dataset on each worker is 64. For the RNN model used in text classification task, we summarize the detail of architecture settings in Table 3.

Table 3: Architecture of the RNN model.

Component	Dimension
Input dimension	300
Hidden dimension	4096
Output dimension	2

The results are presented in Figure 10. FedMuon consistently outperforms the baselines across all metrics. Moreover, it can be observed that adaptive methods, such as FedAdam, FAFED, and Adam-BiClip, demonstrate greater robustness to heavy-tailed noise compared with other approaches, which is consistent with prior findings (Zhang et al., 2020; Kunstner et al., 2024). This observation aligns with explanations that in language tasks, the heavy-tailed class imbalance causes infrequent words to converge more slowly under gradient descent, whereas adaptive methods are less sensitive to this issue (Kunstner et al., 2024). FedMuon similarly benefits from this robustness, which explains its strong performance under heavy-tailed settings.

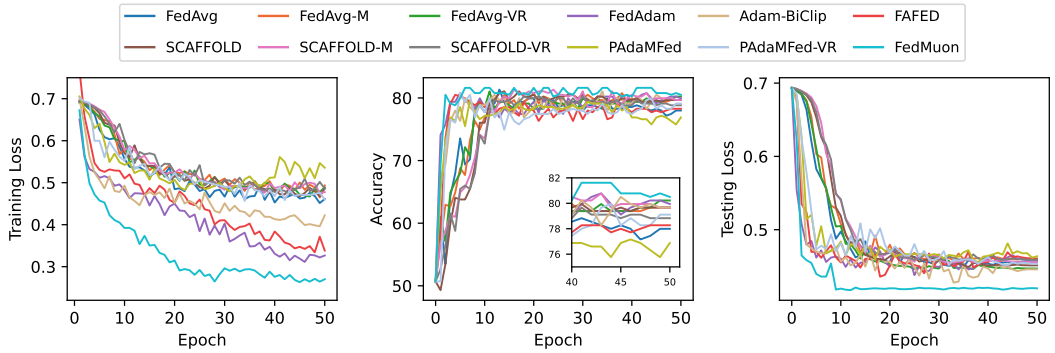

Figure 10: Sentiment140 on RNN (period = 4).

More experimental results with communication period of 16 is shown in Figure 11. It can be observed that when the communication period is increased to 16, FedMuon achieves even larger per-

formance gains over the baselines compared with the case of period 4. This demonstrates that FedMuon is particularly effective under infrequent communication, as it benefits from the reduced communication complexity while still maintaining superior performance.

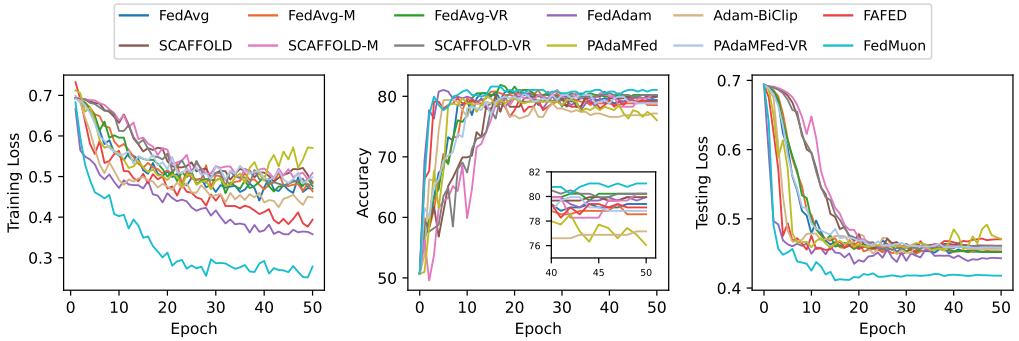

Figure 11: Sentiment140 on RNN (period = 16).

B Convergence Analysis Under Regular Noise

Lemma B.1. *Given Assumptions 3.1, 3.2, the following inequality holds:*

$$f(\bar{X}_{t+1}) \leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}L\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F + 2\eta\sqrt{n}\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F + \frac{\eta^2 nL}{2}\,. \tag{17}$$

Proof.

$$\begin{aligned}
f(\bar{X}_{t+1}) &\leq f(\bar{X}_t) + \langle\nabla f(\bar{X}_t), \bar{X}_{t+1} - \bar{X}_t\rangle + \frac{L}{2}\|\bar{X}_{t+1} - \bar{X}_t\|_F^2 \\
&\leq f(\bar{X}_t) - \eta\langle\nabla f(\bar{X}_t), \frac{1}{K}\sum_{k=1}^{K}U_t^{(k)}(V_t^{(k)})^T\rangle + \frac{\eta^2 L}{2}\|\frac{1}{K}\sum_{k=1}^{K}U_t^{(k)}(V_t^{(k)})^T\|_F^2 \\
&= f(\bar{X}_t) - \eta\frac{1}{K}\sum_{k=1}^{K}\langle\nabla f(\bar{X}_t) - \bar{M}_t, U_t^{(k)}(V_t^{(k)})^T\rangle - \eta\frac{1}{K}\sum_{k=1}^{K}\langle\bar{M}_t, U_t^{(k)}(V_t^{(k)})^T\rangle \\
&\quad + \frac{\eta^2 L}{2}\|\frac{1}{K}\sum_{k=1}^{K}U_t^{(k)}(V_t^{(k)})^T\|_F^2 \\
&\leq f(\bar{X}_t) - \eta\|\bar{M}_t\|_* + \eta\sqrt{n}\|\nabla f(\bar{X}_t) - \bar{M}_t\|_F + \frac{\eta^2 nL}{2} \\
&\leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}\|\nabla f(\bar{X}_t) - \bar{M}_t\|_F + \frac{\eta^2 nL}{2} \\
&\leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}\|\nabla f(\bar{X}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla f^{(k)}(X_t^{(k)})\|_F \\
&\quad + 2\eta\sqrt{n}\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F + \frac{\eta^2 nL}{2} \\
&\leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}L\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F \\
&\quad + 2\eta\sqrt{n}\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F + \frac{\eta^2 nL}{2}\,,
\end{aligned} \tag{18}$$

where the fourth step holds due to $\langle\bar{M}_t, U_t^{(k)}(V_t^{(k)})^T\rangle = \|\bar{M}_t\|_*$, $\langle\nabla f(\bar{X}_t) - \bar{M}_t, U_t^{(k)}(V_t^{(k)})^T\rangle \leq \|\nabla f(\bar{X}_t) - \bar{M}_t\|_F\|U_t^{(k)}(V_t^{(k)})^T\|_F \leq \sqrt{n}\|\nabla f(\bar{X}_t) - \bar{M}_t\|_F$, and $\|U_t^{(k)}(V_t^{(k)})^T\|_F^2 \leq n$, the fifth step holds due to $\|\nabla f(\bar{X}_t)\|_F \leq \|\nabla f(\bar{X}_t)\|_* = \|\nabla f(\bar{X}_t) - \bar{M}_t + \bar{M}_t\|_* \leq \|\nabla f(\bar{X}_t) - \bar{M}_t\|_* + \|\bar{M}_t\|_* \leq \sqrt{n}\|\nabla f(\bar{X}_t) - \bar{M}_t\|_F + \|\bar{M}_t\|_*$. □

Lemma B.2. *Given Assumptions 3.1, 3.2, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F \leq 2\eta\tau\sqrt{n}\,. \tag{19}$$

Proof.

$$\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F$$

$$
\begin{aligned}
&\leq \frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_{s_t\tau} - \eta\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k'=1}^{K}U_t^{(k')}(V_t^{(k')})^T - X_{s_t\tau}^{(k)} + \eta\sum_{t'=s_t\tau}^{t-1}U_t^{(k)}(V_t^{(k)})^T\|_F \\
&\leq \eta\frac{1}{K}\sum_{k=1}^{K}\|\sum_{t'=s_t\tau}^{t-1}U_t^{(k)}(V_t^{(k)})^T - \sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k'=1}^{K}U_t^{(k')}(V_t^{(k')})^T\|_F \\
&\leq \eta\frac{1}{K}\sum_{k=1}^{K}\|\sum_{t'=s_t\tau}^{t-1}U_t^{(k)}(V_t^{(k)})^T\|_F + \eta\frac{1}{K}\sum_{k=1}^{K}\|\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k'=1}^{K}U_t^{(k')}(V_t^{(k')})^T\|_F \\
&\leq 2\eta\tau\sqrt{n}\,, \qquad (20)
\end{aligned}
$$

where the last step holds due to $\|U_t^{(k)}(V_t^{(k)})^T\|_F \leq \sqrt{n}$. □

Lemma B.3. *Given Assumptions 3.1, 3.2, by setting $\beta < 1$, the following inequality holds:*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F] \leq \frac{1}{T}\frac{\sigma}{\beta} + \frac{\eta\sqrt{n}L}{\beta} + \frac{\sqrt{\beta}\sigma}{\sqrt{K}}\,. \qquad (21)
$$

Proof. According to the update of $M_t^{(k)}$, we can obtain

$$
\begin{aligned}
&f^{(k)}(X_t^{(k)}) - M_t^{(k)} \\
&= f^{(k)}(X_t^{(k)}) - (1-\beta)M_{t-1}^{(k)} - \beta\nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)}) \\
&= (1-\beta)f^{(k)}(X_{t-1}^{(k)}) - (1-\beta)M_{t-1}^{(k)} + (1-\beta)f^{(k)}(X_t^{(k)}) - (1-\beta)f^{(k)}(X_{t-1}^{(k)}) \\
&\quad + \beta f^{(k)}(X_t^{(k)}) - \beta\nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)}) \\
&= (1-\beta)(f^{(k)}(X_{t-1}^{(k)}) - M_{t-1}^{(k)}) + (1-\beta)(f^{(k)}(X_t^{(k)}) - f^{(k)}(X_{t-1}^{(k)})) \\
&\quad + \beta(f^{(k)}(X_t^{(k)}) - \nabla f^{(k)}(X_t^{(k)};\xi_t^{(k)})) \\
&= (1-\beta)^t(f^{(k)}(X_0^{(k)}) - M_0^{(k)}) + \sum_{i=1}^{t}(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - f^{(k)}(X_{i-1}^{(k)})) \\
&\quad + \sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - \nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\,. \qquad (22)
\end{aligned}
$$

Then, we can obtain

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F] \\
&\leq (1-\beta)^t\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_0^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_0^{(k)}\|_F] \\
&\quad \underbrace{+\,\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - f^{(k)}(X_{i-1}^{(k)}))\|_F]}_{T_1} \\
&\quad \underbrace{+\,\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - \nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F]}_{T_2}\,. \qquad (23)
\end{aligned}
$$

Regarding T_1, we have

$$
T_1 = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - f^{(k)}(X_{i-1}^{(k)}))\|_F]
$$

$$
\begin{aligned}
&\leq \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}(1-\beta)^{t-i+1}\|f^{(k)}(X_i^{(k)}) - f^{(k)}(X_{i-1}^{(k)})\|_F \\
&\leq L\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}(1-\beta)^{t-i+1}\|X_i^{(k)} - X_{i-1}^{(k)}\|_F \\
&\leq \eta\sqrt{n}L\sum_{i=1}^{t}(1-\beta)^{t-i+1} \\
&\leq \frac{\eta\sqrt{n}L}{\beta}\,, \qquad (24)
\end{aligned}
$$

where the third step holds due to Assumption 3.1, and the fourth step holds due to $\|X_i^{(k)} - X_{i-1}^{(k)}\|_F = \eta\|U_t^{(k)}(V_t^{(k)})^T\|_F \leq \eta\sqrt{n}$.

Regarding T_2, we have

$$
\begin{aligned}
T_2^2 &= \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)}) - \nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F^2] \\
&= \sum_{i=1}^{t}\beta^2(1-\beta)^{2(t-i+1)}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}(f^{(k)}(X_i^{(k)}) - \nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F^2] \\
&\leq \frac{\beta^2\sigma^2}{K}\sum_{i=1}^{t}(1-\beta)^{2(t-i+1)} \\
&\leq \frac{\beta^2\sigma^2}{K(1-(1-\beta)^2)} \\
&= \frac{\beta^2\sigma^2}{K(2\beta-\beta^2)} \\
&= \frac{\beta\sigma^2}{K(2-\beta)} \\
&\leq \frac{\beta\sigma^2}{K}\,, \qquad (25)
\end{aligned}
$$

where the second step holds due to $\mathbb{E}[\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)})] = f^{(k)}(X_i^{(k)})$, the third step holds due to Assumption 3.2, and the last step holds due to $\beta \in (0,1)$.

As a result, we can obtain

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F] \\
&\leq (1-\beta)^t\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_0^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_0^{(k)}\|_F] + \frac{\eta\sqrt{n}L}{\beta} + \frac{\sqrt{\beta}\sigma}{\sqrt{K}}\,. \qquad (26)
\end{aligned}
$$

By summing over t from 0 to $T-1$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}(1-\beta)^t\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_0^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_0^{(k)}\|_F] + \frac{\eta\sqrt{n}L}{\beta} + \frac{\sqrt{\beta}\sigma}{\sqrt{K}} \\
&\leq \frac{1}{T}\frac{\sigma}{\beta} + \frac{\eta\sqrt{n}L}{\beta} + \frac{\sqrt{\beta}\sigma}{\sqrt{K}}\,, \qquad (27)
\end{aligned}
$$

where the last step holds due to $M_0^{(k)} = \nabla f^{(k)}(X_0^{(k)};\xi_0^{(k)})$ and Assumption 3.2. □

Proof of Theorem 5.1.

Proof. Based on Lemma B.1, by summing over t from 0 to $T-1$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq \frac{f(\bar{X}_0)-f(\bar{X}_T)}{\eta T} + \frac{\eta n L}{2} + 2\sqrt{n}L\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{X}_t - X_t^{(k)}\|_F] + 2\sqrt{n}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\,. \tag{28}$$

According to Lemma B.2 and Lemma B.3, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \leq \frac{f(X_0)-f(X_*)}{\eta T} + \frac{\eta n L}{2} + 4\eta\tau n L + \frac{2\sqrt{n}\sigma}{\beta T} + \frac{2\eta n L}{\beta} + \frac{2\sqrt{\beta}\sqrt{n}\sigma}{\sqrt{K}}\,. \tag{29}$$

By setting $\eta = \frac{K^{1/4}}{T^{3/4}}$, $\beta = \frac{K^{1/2}}{T^{1/2}}$, and $\tau = \frac{T^{1/2}}{K^{1/2}}$, we can obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] &\leq \frac{f(X_0)-f(X_*)}{(KT)^{1/4}} + \frac{K^{1/4}nL}{2T^{3/4}} + \frac{4nL}{(KT)^{1/4}} + \frac{2\sqrt{n}\sigma}{(KT)^{1/2}} + \frac{2nL}{(KT)^{1/4}} + \frac{2\sqrt{n}\sigma}{(KT)^{1/4}} \\ &\leq O\left(\frac{f(X_0)-f(X_*)+nL+\sqrt{n}\sigma}{(KT)^{1/4}} + \frac{\sqrt{n}\sigma}{(KT)^{1/2}} + \frac{K^{1/4}nL}{T^{3/4}}\right)\,. \end{aligned} \tag{30}$$

□

C Convergence Analysis Under Heavy-Tailed Noise

To prove Theorem 6.1, we first introduce an important lemma, originally proved for vectors in Liu & Zhou (2024) (see Lemma 4.3), which can be trivially extended to matrices in the following.

Lemma C.1. *Given random matrices V_t and natural filtration $\mathcal{F}_{t-1}$ for $t \in \mathbb{N}$, assume that $\mathbb{E}[V_t|\mathcal{F}_{t-1}] = 0$. Then, the following inequality holds:*

$$\mathbb{E}[\|\sum_{t=1}^{T}V_t\|_F] \leq 2\sqrt{2}\mathbb{E}[(\sum_{t=1}^{T}\|V_t\|_F^p)^{\frac{1}{p}}]\,, \tag{31}$$

where $T \in \mathbb{N}$ and $p \in [1,2]$.

This is a matrix version of Lemma 4.3 in Liu & Zhou (2024). It can be trivially proved by following the proof in Liu & Zhou (2024).

Lemma C.2. *Given Assumptions 3.1, 3.3, the following inequality holds:*

$$f(\bar{X}_{t+1}) \leq f(\bar{X}_t) - \eta\|\nabla f(\bar{X}_t)\|_F + 2\eta\sqrt{n}L\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F + 2\eta\sqrt{n}\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F + \frac{\eta^2 nL}{2}\,. \tag{32}$$

This lemma is same as Lemma B.1.

Lemma C.3. *Given Assumptions 3.1, 3.3, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\|\bar{X}_t - X_t^{(k)}\|_F \leq 2\eta\tau\sqrt{n}\,. \tag{33}$$

This lemma is same as Lemma B.2.

Lemma C.4. *Given Assumptions 3.1, 3.2, by setting $\beta < 1$, the following inequality holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\leq\frac{1}{T}\frac{\sigma}{\beta}+\frac{\eta\sqrt{n}L}{\beta}+\frac{\sqrt{\beta}\sigma}{\sqrt{K}}\,. \tag{34}$$

Proof. Same as the proof of Lemma B.3, we can obtain

$$\begin{aligned}
&\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\\
&\leq(1-\beta)^t\underbrace{\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_0^{(k)}\|_F]}_{T_0}\\
&\quad+\underbrace{\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)})-f^{(k)}(X_{i-1}^{(k)}))\|_F]}_{T_1}\\
&\quad+\underbrace{\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F]}_{T_2}\,.
\end{aligned}\tag{35}$$

T_1 has the same bound as Lemma B.3:

$$T_1\leq\frac{\eta\sqrt{n}L}{\beta}\,. \tag{36}$$

Regarding T_0, for $p\in(1,2]$, we have

$$\begin{aligned}
T_0&=\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_0^{(k)}\|_F]\\
&=\frac{1}{K}\mathbb{E}[\|\sum_{k=1}^{K}(f^{(k)}(X_0^{(k)})-\nabla f^{(k)}(X_0^{(k)};\xi_0^{(k)}))\|_F]\\
&\leq\frac{2\sqrt{2}}{K}\mathbb{E}\left[\left(\sum_{k=1}^{K}\|(f^{(k)}(X_0^{(k)})-\nabla f^{(k)}(X_0^{(k)};\xi_0^{(k)}))\|_F^p\right)^{\frac{1}{p}}\right]\\
&\leq\frac{2\sqrt{2}}{K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|(f^{(k)}(X_0^{(k)})-\nabla f^{(k)}(X_0^{(k)};\xi_0^{(k)}))\|_F^p\right]\right)^{\frac{1}{p}}\\
&\leq\frac{2\sqrt{2}}{K^{1-\frac{1}{p}}}\sigma\\
&\leq 2\sqrt{2}\sigma\,,
\end{aligned}\tag{37}$$

where the third step holds due to Lemma C.1, the fourth step holds due to Hölder's inequality, the fifth step holds due to Assumption 3.3, and the last step holds due to $K>1$.

Regarding T_2, we have

$$\begin{aligned}
T_2&=\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F]\\
&=\frac{1}{K}\mathbb{E}[\|\sum_{k=1}^{K}\sum_{i=1}^{t}\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\sqrt{2}\frac{1}{K}\mathbb{E}[(\sum_{k=1}^{K}\sum_{i=1}^{t}\|\beta(1-\beta)^{t-i+1}(f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)}))\|_F^p)^{\frac{1}{p}}]\\
&\leq 2\sqrt{2}\frac{1}{K}\mathbb{E}[(\sum_{k=1}^{K}\sum_{i=1}^{t}\beta^p(1-\beta)^{p(t-i+1)}\|f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)})\|_F^p)^{\frac{1}{p}}]\\
&\leq 2\sqrt{2}\frac{1}{K}(\sum_{k=1}^{K}\sum_{i=1}^{t}\beta^p(1-\beta)^{p(t-i+1)}\mathbb{E}[\|f^{(k)}(X_i^{(k)})-\nabla f^{(k)}(X_i^{(k)};\xi_i^{(k)})\|_F^p])^{\frac{1}{p}}\\
&\leq 2\sqrt{2}\frac{1}{K}(\sum_{k=1}^{K}\sum_{i=1}^{t}\beta^p(1-\beta)^{p(t-i+1)}\sigma^p)^{\frac{1}{p}}\\
&=\frac{2\sqrt{2}\beta\sigma}{K^{1-\frac{1}{p}}}(\sum_{i=1}^{t}(1-\beta)^{p(t-i+1)})^{\frac{1}{p}}\\
&\leq\frac{2\sqrt{2}\beta\sigma}{K^{1-\frac{1}{p}}}(\frac{1}{1-(1-\beta)^p})^{\frac{1}{p}}\\
&\leq\frac{2\sqrt{2}\beta\sigma}{K^{1-\frac{1}{p}}}(\frac{1}{1-(1-\beta)})^{\frac{1}{p}}\\
&\leq\frac{2\sqrt{2}\beta^{1-\frac{1}{\beta}}}{K^{1-\frac{1}{p}}}\sigma\,,
\end{aligned}
\tag{38}
$$

where the third step holds due to Lemma C.1, the fifth step holds due to Holder's inequality, the sixth step holds due to Assumption 3.3.

As a result, we can obtain

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\\
&\leq(1-\beta)^t 2\sqrt{2}\sigma+\frac{\eta\sqrt{n}L}{\beta}+\frac{2\sqrt{2}\beta^{1-\frac{1}{p}}}{K^{1-\frac{1}{p}}}\sigma\,.
\end{aligned}
\tag{39}
$$

By summing over t from 0 to $T-1$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\\
&\leq\frac{1}{T}\sum_{t=0}^{T-1}(1-\beta)^t 2\sqrt{2}\sigma+\frac{\eta\sqrt{n}L}{\beta}+\frac{\sqrt{\beta}\sigma}{\sqrt{K}}\\
&\leq\frac{1}{T}\frac{2\sqrt{2}\sigma}{\beta}+\frac{\eta\sqrt{n}L}{\beta}+\frac{2\sqrt{2}\beta^{1-\frac{1}{p}}}{K^{1-\frac{1}{p}}}\sigma\,.
\end{aligned}
\tag{40}
$$

□

Proof of Theorem 6.1.

Proof. Based on Lemma C.2, by summing over t from 0 to $T-1$, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F]\leq\frac{f(\bar{X}_0)-f(\bar{X}_T)}{\eta T}+\frac{\eta nL}{2}+2\sqrt{n}L\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{X}_t-X_t^{(k)}\|_F]\\
+2\sqrt{n}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}f^{(k)}(X_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}M_t^{(k)}\|_F]\,.
\end{aligned}
\tag{41}
$$

According to Lemma C.3 and Lemma C.4, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] \le \frac{f(X_0)-f(X_*)}{\eta T} + \frac{\eta n L}{2} + 4\eta\tau n L + \frac{4\sqrt{2n}\sigma}{\beta T} + \frac{2\eta n L}{\beta} + \frac{4\sqrt{2n}\beta^{1-\frac{1}{p}}}{K^{1-\frac{1}{p}}}\sigma\,. \tag{42}$$

By setting $\eta = \frac{K^{1/4}}{T^{3/4}}$, $\beta = \frac{K^{1/2}}{T^{1/2}}$, and $\tau = \frac{T^{1/2}}{K^{1/2}}$, we can obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{X}_t)\|_F] &\le \frac{f(X_0)-f(X_*)}{(KT)^{1/4}} + \frac{K^{1/4}nL}{2T^{3/4}} + \frac{4nL}{(KT)^{1/4}} \\ &\quad + \frac{4\sqrt{2n}\sigma}{(KT)^{1/2}} + \frac{2nL}{(KT)^{1/4}} + \frac{4\sqrt{2n}\sigma}{(KT)^{\frac{p-1}{2p}}} \\ &\le O\left(\frac{f(X_0)-f(X_*)+nL}{(KT)^{1/4}} + \frac{\sqrt{n}\sigma}{(KT)^{1/2}} + \frac{K^{1/4}nL}{T^{3/4}} + \frac{\sqrt{n}\sigma}{(KT)^{\frac{p-1}{2p}}}\right).\end{aligned} \tag{43}$$

□