Journal of Machine Learning Research 26 (2025) 1-55 Submitted 5/24; Revised 11/24; Published 3/25

Sharp Bounds for Sequential Federated Learning on Heterogeneous Data

Yipeng Li LIYIPENG@BUPT.EDU.CN
National Engineering Research Center for Mobile Network Technologies
Beijing University of Posts and Telecommunications
Beijing, 100876, China

Xinchen Lyu* LVXINCHEN@BUPT.EDU.CN
National Engineering Research Center for Mobile Network Technologies
Beijing University of Posts and Telecommunications
Beijing, 100876, China

Editor: Peter Richtárik

Abstract

There are two paradigms in Federated Learning (FL): parallel FL (PFL), where models are trained in a parallel manner across clients, and sequential FL (SFL), where models are trained in a sequential manner across clients. Specifically, in PFL, clients perform local updates independently and send the updated model parameters to a global server for aggregation; in SFL, one client starts its local updates only after receiving the model parameters from the previous client in the sequence. In contrast to that of PFL, the convergence theory of SFL on heterogeneous data is still lacking. To resolve the theoretical dilemma of SFL, we establish sharp convergence guarantees for SFL on heterogeneous data with both upper and lower bounds. Specifically, we derive the upper bounds for the strongly convex, general convex and non-convex objective functions, and construct the matching lower bounds for the strongly convex and general convex objective functions. Then, we compare the upper bounds of SFL with those of PFL, showing that SFL outperforms PFL on heterogeneous data (at least, when the level of heterogeneity is relatively high). Experimental results validate the counterintuitive theoretical finding.

Keywords: stochastic gradient descent, random reshuffling, parallel federated learning, sequential federated learning, convergence analysis

1. Introduction

Federated Learning (FL) (McMahan et al., 2017; Chang et al., 2018) is a popular distributed machine learning paradigm, where multiple clients collaborate to train a global model, while preserving data privacy and security. Commonly, FL can be categorized into two types: (i) parallel FL (PFL), where models are trained in a parallel manner across clients, with periodic aggregation, such as Federated Averaging (FedAvg) (McMahan et al., 2017) and Local SGD (Stich, 2019), and (ii) sequential FL (SFL), where models are trained in a sequential manner across clients, such as Cyclic Weight Transfer (CWT) (Chang et al.,

*. Corresponding author.

2018) and peer-to-peer FL (Yuan et al., 2023). A simple illustration of SFL and PFL is presented in Figure 1.

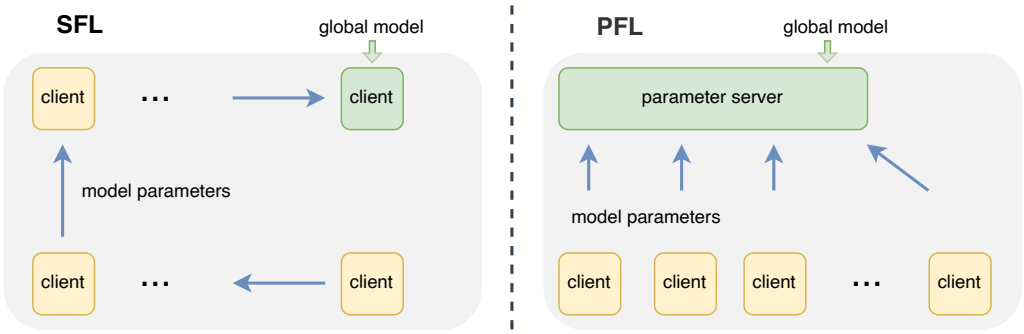

Figure 1: Illustration of SFL and PFL.

SFL has recently attracted much attention in the FL community (Lee et al., 2020; Yuan et al., 2024) with various applications in medicine (Chang et al., 2018; Huang et al., 2024), automated driving (Yuan et al., 2023) and so on. Compared with PFL, SFL shows the advantages in the following aspects: First, as one of the most popular decentralized FL paradigms (Yuan et al., 2024), *SFL operates in a peer-to-peer manner, avoiding the reliance on a central parameter server.* This is a more practical option for medical applications, as establishing a central server is a costly endeavor (Huang et al., 2024). In addition, it avoids single points of failure and bottlenecks in communication, computation, and storage resources found in central servers (Yuan et al., 2024). Second, *SFL is regarded as a network-resilient, communication- and computation-efficient alternative to PFL* (Yuan et al., 2024; Yan et al., 2024). Recently, SFL has been increasingly combined with PFL to complement each other. For example, in Clustered Federated Learning, where clients are grouped into multiple clusters, Zaccone et al. (2022); Chen et al. (2023) adopts SFL for clients within each cluster to speed up the training process and reduce communication overhead; Chen et al. (2020); Yan et al. (2024) adopts SFL for inter-cluster training to make frequent updates and overcome the shortcomings of network dynamics and instability. Third, *SFL has played a great role in Split Learning (SL) (Gupta and Raskar, 2018; Thapa et al., 2022)*, an emerging distributed learning technology at the network edge, where the full model is split into client-side and server-side portions to alleviate the excessive computation overhead for resource-constrained devices. In SL, one popular way to enable multi-client training is the sequential model training manner, where clients collaborate with the server to perform the updates sequentially (one by one), as the full model is divided between the client side and the server side. Our theory on the convergence of SFL is also applicable to such sequential SL (Li and Lyu, 2023).

Both PFL and SFL suffer from "data heterogeneity", one of the most persistent problems in FL. Up to now, there have been numerous works to study the convergence of PFL on heterogeneous data (Li et al., 2020; Khaled et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020b). These theoretical works not only helped understand the effect of heterogeneity, but also spawned new algorithms like SCAFFOLD (Karimireddy et al., 2020; Mishchenko et al., 2022b). In contrast, the convergence of SFL on heterogeneous data

has not been well studied. Recent works (Cho et al., 2023; Malinovsky et al., 2023) studied the convergence of FL with cyclic client participation, which can be seen as an extension of SFL. However, its convergence analysis is still in an infancy stage, and existing works do not cover the SFL setups in this paper (see Section 2). Our earlier conference paper (Li and Lyu, 2023) proved the upper bounds of SFL (which is shown to be applicable to SL). However, notably, the lower bounds for SFL are still missing in existing works. The lack of theoretical study can hinder further development of SFL and even SL.

To resolve the theoretical dilemma of SFL, this paper, extending our conference paper (Li and Lyu, 2023),[1] aims to establish sharp convergence guarantees for SFL with both upper and lower bounds. In the case of homogeneous data, this task is trivial, as SFL is reduced to SGD (Stochastic Gradient Descent). However, in the case of heterogeneous data, it is more challenging than existing works, including PFL and SGD-RR (Random Reshuffling), primarily due to the following reasons:

(i) Sequential and shuffling training manner across clients (vs. PFL). In PFL, the local updates at each client only depend on the randomness of the current client within each training round. However, in SFL, the local updates additionally depend on the randomness of all previous clients.

(ii) Multiple local update steps at each client (vs. SGD-RR). In contrast to its with-replacement sibling SGD, SGD-RR samples data samples "without replacement" and then performs one step of GD (Gradient Descent) on each data sample. Similarly, SFL samples clients without replacement and then performs multiple steps of SGD at each client. In fact, SGD-RR can be regarded as a special case of SFL.

In this paper, we establish the convergence guarantees for SFL (Algorithm 1), and then compare them with those of PFL (Algorithm 2). The main contributions are as follows:

- We derive the upper bounds of SFL for the strongly convex, general convex and non-convex cases on heterogeneous data with the standard assumptions in FL in Subsection 4.2 (Theorem 3 and Corollary 4).

- We construct the lower bounds of SFL for the strongly convex and general convex cases in Subsection 4.3 (Theorem 5 and Theorem 6). They match the derived upper bounds for the large number of training rounds.

- We compare the upper bounds of SFL with those of PFL in Subsections 5.1 and 5.2. In the convex cases,[2] the comparison results show a subtle difference under different heterogeneity assumptions. That is, under Assumption 3, the upper bounds of SFL are better than those of PFL strictly, while under Assumption 5, the upper bounds of SFL are still better unless the level of heterogeneity is very low. In the non-convex case under Assumption 4, the upper bounds of SFL are better without exception.

1. The conference paper is available at `https://arxiv.org/abs/2311.03154`. Notably, by default, we use the latest arXiv version for all the references.
2. For clarity, we use the term "the convex cases" to collectively refer to both the strongly convex case and the general convex case in this paper.

- The comparison results imply that SFL outperforms PFL in heterogeneous settings (at least, when the level of heterogeneity is relatively high). We then validate this counterintuitive result with experiments on quadratic functions (Subsection 6.1), logistic regression (Subsection 6.2) and deep neural networks (Subsection 6.3).

2. Related Work

The most relevant research topics are the convergence analyses of PFL and SGD-RR.

So far, there have been a wealth of works to study the upper bounds of PFL on data heterogeneity (Li et al., 2020; Khaled et al., 2020; Karimireddy et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020b), system heterogeneity (Wang et al., 2020), partial client participation (Li et al., 2020; Yang et al., 2021; Wang and Ji, 2022) and other variants (Karimireddy et al., 2020; Wang et al., 2020; Reddi et al., 2021). The lower bounds of PFL have also been studied in Woodworth et al. (2020a,b); Yun et al. (2022); Glasgow et al. (2022). In this work, we make a comparison between the upper bounds of PFL and those of SFL on heterogeneous data (see Subsections 5.1 and 5.2).

SGD-RR has been gaining significant attention as a more practical alternative to SGD. Nagaraj et al. (2019); Ahn et al. (2020); Mishchenko et al. (2020); Nguyen et al. (2021); Lu et al. (2022); Koloskova et al. (2024) have proved the upper bounds and Safran and Shamir (2020, 2021); Rajput et al. (2020); Cha et al. (2023) have proved the lower bounds of SGD-RR. In particular, to the best of our knowledge, Mishchenko et al. (2020) provided the tightest upper bounds and Cha et al. (2023) provided the tightest lower bounds of SGD-RR. In this work, we adopt them to exam the tightness of the convergence bounds of SFL (see Subsection 4.2).

Recently, the shuffling-based method, SGD-RR, has been applied to FL. One line of these works is Local RR (or FedRR) (Mishchenko et al., 2022a; Yun et al., 2022; Horváth et al., 2022; Sadiev et al., 2023; Malinovsky et al., 2023), which adopts SGD-RR (instead of SGD) as the local solver. Another line is FL with cyclic client participation (Eichner et al., 2019; Wang and Ji, 2022; Cho et al., 2023; Malinovsky et al., 2023), which can be seen as an extension of SFL. However, its convergence analysis is still in an infancy stage, and existing works do not cover the SFL setups in this paper. Eichner et al. (2019) considered that clients can form different blocks due to diurnal variation, and propose training seperate models for each of these blocks. This differs from our setting, which aims to train a single global model. In Wang and Ji (2022); Cho et al. (2023), when it reduces to SFL, the client training order is deterministic (not random), and thus the analyses cannot be directly extended to our setting. In Malinovsky et al. (2023), although their bounds are slightly tighter on the optimization term with SGD-RR as the local solver, their analysis is limited to the case where the number of local steps equals the size of the local data set. Most importantly, Cho et al. (2023) considered upper bounds for PL objective functions and Malinovsky et al. (2023) considered upper bounds for strongly convex objective functions,[3] while we consider both upper bounds (for both convex and non-convex cases) and lower bounds. Detailed comparisons are in Appendix G and Li and Lyu (2023).

3. PL condition can be thought as a non-convex generalization of strong convexity.

3. Setup

Notation. We let $[n] \coloneqq \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}^+$ and $\{x_i\}_{i\in\mathcal{S}} \coloneqq \{x_i : i \in \mathcal{S}\}$ for any set $\mathcal{S}$. We use $|\mathcal{S}|$ to denote the size of any set $\mathcal{S}$. We use $\lesssim$ to denote "less than" up to some absolute constants and polylogarithmic factors, and $\gtrsim$ and $\asymp$ are defined likewise. We also use the big O notations, $\tilde{O}$, $\mathcal{O}$, Ω, where $\mathcal{O}$, Ω hide numerical constants, $\tilde{\mathcal{O}}$ hides numerical constants and polylogarithmic factors. We use $\|\cdot\|$ to denote the L_2-norm for both vectors and matrices. More notations are in Table 4.

Problem formulation. The basic FL problem is to minimize a global objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ F(\mathbf{x}) \coloneqq \frac{1}{M}\sum_{m=1}^{M} \left(F_m(\mathbf{x}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_m}\left[f_m(\mathbf{x};\xi)\right]\right) \right\},$$

where F_m and f_m denote the local objective function and the local component function of Client m ($m \in [M]$), respectively. The local objective function is the average of the local component functions, $F_m(\mathbf{x}) = \frac{1}{|\mathcal{D}_m|}\sum_{i\in\mathcal{D}_m} f_m(\mathbf{x};\xi_m^i)$, when the local data set $\mathcal{D}_m$ contains a finite number of data samples.

Update rule of SFL (Algorithm 1). At the beginning of each training round, the indices $\pi_1, \pi_2, \ldots, \pi_M$ are sampled without replacement from $[M]$ randomly as the clients' training order. Within a round, each client (i) initializes its model with the latest parameters from its previous client, (ii) performs K steps of local updates over its local data set, and (iii) passes the updated parameters to the next client. This process continues until all clients finish their local training. Let $\mathbf{x}_{m,k}^{(r)}$ denote the local parameters of the m-th client (that is, Client π_m) after k local steps in the r-th round, and $\mathbf{x}^{(r)}$ denote the global parameters in the r-th round. Then, if SGD is chosen as the local solver (with a constant learning rate η), the update rule of SFL is as follows:

$$\begin{aligned}
&\text{Local update}: \ \mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta\mathbf{g}_{\pi_m,k}^{(r)}, \quad \text{initializing } \mathbf{x}_{m,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)}, & m = 1 \\ \mathbf{x}_{m-1,K}^{(r)}, & m > 1 \end{cases}, \\
&\text{Global model}: \ \mathbf{x}^{(r+1)} = \mathbf{x}_{M,K}^{(r)}.
\end{aligned}$$

Here we use $\mathbf{g}_{\pi_m,k}^{(r)} \coloneqq \nabla f_{\pi_m}(\mathbf{x}_{m,k}^{(r)};\xi_{m,k}^{(r)})$ to denote the stochastic gradient generated at the m-th client for its $k+1$-th local update in the r-th round.

Update rule of PFL (Algorithm 2). Within a round, each client (i) initializes its model with the global parameters, (ii) performs K steps of local updates, and (iii) sends the updated parameters to the central server. The server will aggregate the local parameters to generate the global parameters. With the the same notations as those of SFL, the update rule of PFL is as follows:

$$\begin{aligned}
&\text{Local update}: \ \mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta\mathbf{g}_{m,k}^{(r)}, \quad \text{initializing } \mathbf{x}_{m,0}^{(r)} = \mathbf{x}^{(r)}, \forall m \in [M] \\
&\text{Global model}: \ \mathbf{x}^{(r+1)} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_{m,K}^{(r)}.
\end{aligned}$$

The mechanism of "two learning rates". Karimireddy et al. (2020) has proven that the mechanism of "two learning rates" can improve the convergence rate of PFL. This

mechanism involves using a client-specific learning rate for local updates and a different server-specific learning rate for global updates on the server. In fact, *This mechanism can also be applied to SFL. Theoretically, it can achieve the same improvement as that in PFL (Karimireddy et al., 2020).* We show how to implement this mechanism in SFL, and compare the upper bounds with those of PFL (Karimireddy et al., 2020) in Appendix E.

Algorithm 1: Sequential FL

Output: $\{\mathbf{x}^{(r)}\}$

for $r = 0, \ldots, R-1$ **do**
  Sample a permutation $\pi_1, \pi_2, \ldots, \pi_M$ of $\{1, 2, \ldots, M\}$
  for $m = 1, \ldots, M$ **in sequence do**
    $\mathbf{x}_{m,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)}, & m = 1 \\ \mathbf{x}_{m-1,K}^{(r)}, & m > 1 \end{cases}$
    for $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{\pi_m,k}^{(r)}$
  Global model: $\mathbf{x}^{(r+1)} = \mathbf{x}_{M,K}^{(r)}$

Algorithm 2: Parallel FL

Output: $\{\mathbf{x}^{(r)}\}$

for $r = 0, \ldots, R-1$ **do**
  for $m = 1, \ldots, M$ **in parallel do**
    $\mathbf{x}_{m,0}^{(r)} = \mathbf{x}^{(r)}$
    for $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{m,k}^{(r)}$
  Global model: $\mathbf{x}^{(r+1)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_{m,K}^{(r)}$

4. Convergence Analysis of SFL

We consider three typical cases: the strongly convex case, the general convex case and the non-convex case, where all the local objective functions $F_1, F_2, \ldots, F_M$ are μ-strongly convex, general convex (see Definition 1) and non-convex, respectively.

4.1 Assumptions

We assume that (i) F is lower bounded by F^* for all cases and there exists a global minimizer $\mathbf{x}^*$ such that $F(\mathbf{x}^*) = F^*$ for the convex cases; (ii) all local objective functions are differentiable and smooth (see Definition 2). Furthermore, we need to make assumptions on the diversities: (iii) the assumptions on the stochasticity bounding the diversity of local component functions $\{f_m(\cdot; \xi_m^i)\}_i^{|\mathcal{D}_m|}$ with respect to i inside each client (Assumptions 1 and 2); (iv) the assumptions on the heterogeneity bounding the diversity of local objective functions $\{F_m\}_m^M$ with respect to m across clients (Assumptions 3, 4 and 5).

Definition 1 *A differentiable function F is μ-strongly convex if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

If $\mu = 0$, we say that F is general convex.

Definition 2 *A differentiable function F is L-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

Assumptions on the stochasticity. Since both Algorithms 1 and 2 use SGD (data samples are chosen with replacement) as the local solver, the stochastic gradient generated at each client is an (conditionally) unbiased estimate of the gradient of the local objective function, $\mathbb{E}_{\xi\sim\mathcal{D}_m}[\nabla f_m(\mathbf{x};\xi) \mid \mathbf{x}] = \nabla F_m(\mathbf{x})$. In the FL literature, there are two common assumptions, Assumptions 1 and 2, to bound the stochasticity, where σ_*, σ measure the level of stochasticity. Assumptions 1 only assumes the bounded stochasticity at the optimum, and therefore it is weaker than Assumption 2. However, if using Assumption 1, we need to assume that each local component function $f_m(\mathbf{x};\xi)$ is smooth, rather than merely assuming that each local objective function $F_m(\mathbf{x})$ is smooth (Khaled et al., 2020; Koloskova et al., 2020). Besides, we prioritize studying the effects of heterogeneity. For these two reasons, we use Assumption 2 for all cases in this paper.

Assumption 1 *There exists a constant σ_* such that for the global minimizer $\mathbf{x}^* \in \mathbb{R}^d$,*

$$\mathbb{E}_{\xi\sim\mathcal{D}_m} \left\|\nabla f_m(\mathbf{x}^*;\xi) - \nabla F_m(\mathbf{x}^*)\right\|^2 \leq \sigma_*^2.$$

Assumption 2 *There exists a constant σ such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\xi\sim\mathcal{D}_m} \left\|\nabla f_m(\mathbf{x};\xi) - \nabla F_m(\mathbf{x})\right\|^2 \leq \sigma^2.$$

Assumptions on the heterogeneity. Now we make assumptions on the diversity of the local objective functions in Assumption 3, 4 and 5, also known as the heterogeneity in FL. For the convex cases, we use Assumption 3 as Koloskova et al. (2020) did, which assumes the bounded diversity only at the optimum. Assumption 4 is made for the non-convex case, where the constants β and ζ measure the heterogeneity of the local objective functions. Assumption 5, the strongest assumption, is only made in Subsection 5.2. Notably, that all the local objective functions are identical (that is, no heterogeneity) means that $\zeta_*, \beta, \zeta, \hat{\zeta}$ equal zero in these assumptions. Yet the reverse may not be true, as they only assume the first-order relationships (Karimireddy et al., 2020).

Assumption 3 *Let $\mathbf{x}^* \in \mathbb{R}^d$ be a minimizer of the global objective function F. Define*

$$\zeta_*^2 \coloneqq \tfrac{1}{M}\textstyle\sum_{m=1}^M \left\|\nabla F_m(\mathbf{x}^*)\right\|^2,$$

where ζ_*^2 is assumed to be bounded.

Assumption 4 *There exist bounded constants β^2 and ζ^2 such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\tfrac{1}{M}\textstyle\sum_{m=1}^M \left\|\nabla F_m(\mathbf{x}) - \nabla F(\mathbf{x})\right\|^2 \leq \beta^2 \left\|\nabla F(\mathbf{x})\right\|^2 + \zeta^2.$$

Assumption 5 *There exists a bounded constant $\hat{\zeta}^2$ such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\max_m \left\|\nabla F_m(\mathbf{x}) - \nabla F(\mathbf{x})\right\|^2 \leq \hat{\zeta}^2.$$

4.2 Upper Bounds of SFL

Theorem 3 *Let all the local objectives be L-smooth (Definition 2). For SFL (Algorithm 1), there exist a constant effective learning rate $\tilde{\eta} \coloneqq \eta MK$ and weights $\{w_r\}_{r\geq 0}$, such that the weighted average of the global model parameters $\bar{\mathbf{x}}^{(R)} \coloneqq \frac{\sum_{r=0}^{R} w_r \mathbf{x}^{(r)}}{\sum_{r=0}^{R} w_r}$ satisfies the following upper bounds:*

***Strongly convex**: Under Assumptions 2, 3, there exist $\tilde{\eta} \leq \frac{1}{6L}$ and $w_r = (1-\frac{\mu\tilde{\eta}}{2})^{-(r+1)}$, such that for $R \geq 6\kappa$ ($\kappa \coloneqq L/\mu$),*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] \leq \frac{9}{2}\mu D^2 \exp\left(-\frac{\mu\tilde{\eta}R}{2}\right) + \frac{12\tilde{\eta}\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{M}.$$

***General convex**: Under Assumptions 2, 3, there exist $\tilde{\eta} \leq \frac{1}{6L}$ and $w_r = 1$, such that*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] \leq \frac{3D^2}{\tilde{\eta}R} + \frac{12\tilde{\eta}\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{M}.$$

***Non-convex**: Under Assumptions 2, 4, there exist $\tilde{\eta} \leq \frac{1}{6L(1+\beta^2/M)}$ and $w_r = 1$, such that*

$$\min_{0\leq r\leq R} \mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{10A}{\tilde{\eta}R} + \frac{20L\tilde{\eta}\sigma^2}{MK} + \frac{75L^2\tilde{\eta}^2\sigma^2}{4MK} + \frac{75L^2\tilde{\eta}^2\zeta^2}{4M}.$$

Here $D \coloneqq \left\|x^{(0)} - x^*\right\|$ for the convex cases and $A \coloneqq F(\mathbf{x}^{(0)}) - F^*$ for the non-convex case.

Proof We provide intuitive proof sketches of Theorem 3 as done in Karimireddy et al. (2020). Ideally, we want to update the model with the gradients of the global objective function. For any local gradient in some training round of SFL, it can be decomposed into two vectors (or estimated by Taylor formula),

$$\nabla F_m(\mathbf{x}_{m,k}) \approx \left(\nabla F_m(\mathbf{x}) + \nabla^2 F_m(\mathbf{x})(\mathbf{x}_{m,k} - \mathbf{x})\right).$$

Then, the global update of SFL can be written as

$$\begin{aligned}\Delta_{\text{SFL}} &= -\eta \sum_{m=1}^{M}\sum_{k=0}^{K-1} \left\{\nabla F_m(\mathbf{x}_{m,k}) \approx \left(\nabla F_m(\mathbf{x}) + \nabla^2 F_m(\mathbf{x})(\mathbf{x}_{m,k} - \mathbf{x})\right)\right\} \\ &= \underbrace{-\eta MK \nabla F(\mathbf{x})}_{\text{optimization vector}} \underbrace{-\eta \sum_{m=1}^{M}\sum_{k=0}^{K-1} \nabla^2 F_m(\mathbf{x})(\mathbf{x}_{m,k} - \mathbf{x})}_{\text{error vector}}.\end{aligned}$$

The optimization vector is beneficial while the error vector is detrimental. Thus, our goal is to suppress the error vector. Theorem 3 is aimed to prove that $\sum_{m=1}^{M}\sum_{k=0}^{K-1} \|\mathbf{x}_{m,k} - \mathbf{x}\|^2$ is bounded ($\nabla^2 F_m(\mathbf{x}) \approx L$). Intuitively, for there are about mK update steps between $\mathbf{x}$ and $\mathbf{x}_{m,k}$, it is estimated to be $\mathcal{O}\left(\sum_{m=1}^{M}\sum_{k=0}^{K-1}(\eta\sqrt{m}K\zeta)^2\right) = \mathcal{O}\left(\eta^2 M^2 K^3 \zeta^2\right)$, where $\sqrt{m}$ is due to the shuffling-based manner. The formal proofs are in Li and Lyu (2023). ■

The effective learning rate $\tilde{\eta} \coloneqq \eta MK$ is used in the upper bounds as done in Karimireddy et al. (2020); Wang et al. (2020). All these upper bounds consist of two parts: the optimization part (the first term) and the error part (the last three terms). Setting $\tilde{\eta}$ larger makes the optimization part vanishes at a higher rate, yet causes the error part to be larger. This implies that we need to choose an appropriate $\tilde{\eta}$ to achieve a balance between these two parts, which is actually done in Corollary 4. Here we choose the appropriate learning rate with a prior knowledge of the total training rounds R (Karimireddy et al., 2020).

Corollary 4 *By choosing an appropriate learning rate for the results of Theorem 3, we can obtain the upper bounds of SFL:*

***Strongly convex**: If $\tilde{\eta} = \eta MK \asymp \min\{\frac{1}{L}, \frac{1}{\mu R}\}$ for Theorem 3, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta_*^2}{\mu^2 MR^2} + \mu D^2 \exp\left(\frac{-\mu R}{L}\right)\right).$$

***General convex**: If $\tilde{\eta} = \eta MK \asymp \min\{\frac{1}{L}, \frac{D}{c_1^{1/2} R^{1/2}}, \frac{D^{2/3}}{c_2^{1/3} R^{2/3}}\}$ with $c_1 \asymp \frac{\sigma^2}{MK}$ and $c_2 \asymp \frac{L\sigma^2}{MK} + \frac{L\zeta^2}{M}$ for Theorem 3, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{(MK)^{1/3} R^{2/3}} + \frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{M^{1/3} R^{2/3}} + \frac{LD^2}{R}\right).$$

***Non-convex**: If $\tilde{\eta} = \eta MK \asymp \min\{\frac{1}{L(1+\beta^2/M)}, \frac{A^{1/2}}{c_1^{1/2} R^{1/2}}, \frac{A^{1/3}}{c_2^{1/3} R^{2/3}}\}$ with $c_1 \asymp \frac{L\sigma^2}{MK}$ and $c_2 \asymp \frac{L^2\sigma^2}{MK} + \frac{L^2\zeta^2}{M}$ for Theorem 3, then*

$$\min_{0 \le r \le R} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^{(r)})\right\|^2\right] = \mathcal{O}\left(\frac{\left(L\sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2\sigma^2 A^2\right)^{1/3}}{(MK)^{1/3} R^{2/3}} + \frac{\left(L^2\zeta^2 A^2\right)^{1/3}}{M^{1/3} R^{2/3}} + \frac{LA(1+\frac{\beta^2}{M})}{R}\right).$$

Here $D \coloneqq \left\|x^{(0)} - x^*\right\|$ for the convex cases and $A \coloneqq F(\mathbf{x}^{(0)}) - F^*$ for the non-convex case.

Similar to Theorem 3, all these upper bounds consist of two parts, the optimization part (the last term), and the error part (the first three terms). Specifically, the first two terms (containing σ) is called stochasticity terms, the third term (containing ζ_*, ζ) is called heterogeneity terms, the last term is called optimization terms.

Generally, for a sufficiently large number of training rounds R, the convergence rate is determined by the first term for all cases, resulting in rates of $\tilde{\mathcal{O}}(1/MKR)$, $\mathcal{O}(1/\sqrt{MKR})$, $\mathcal{O}(1/\sqrt{MKR})$ for the strongly convex, general convex and non-convex cases, respectively.

Recall that SGD-RR can be seen as one special case of SFL, where one step of GD is performed on each local objective F_m, which implies $K = 1$ and $\sigma = 0$. We now compare the upper bounds of SFL with those of SGD-RR to exam the tightness. As shown in Mishchenko et al. (2020)'s Corollaries 1, 2, 3, the upper bounds of SGD-RR are $\tilde{\mathcal{O}}\left(\frac{L}{\mu}\left(\frac{L\zeta_*^2}{\mu^2 MR^2} + \mu D^2 \exp\left(\frac{-\mu MR}{L}\right)\right)\right)$, $\mathcal{O}\left(\frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{M^{1/3} R^{2/3}} + \frac{LD^2}{R}\right)$, $\mathcal{O}\left(\frac{\left(L^2\zeta^2 A^2\right)^{1/3}}{M^{1/3} R^{2/3}} + \frac{LA}{R}\right)$ for the

strongly convex, general convex and non-convex cases, respectively. We see that our bounds match those of SGD-RR in the general convex, and non-convex cases. For the strongly convex case, the bound of SGD-RR shows an advantage on the optimization term (marked in red). This advantage is due to the advanced technique of Shuffling Variance introduced by Mishchenko et al. (2020, Definition 2). We leave the investigation on introducing this advanced technique to SFL for future work.

4.3 Lower Bounds of SFL

The lower bounds of SFL are stated in Theorem 5 and Theorem 6. Theorem 5 is for arbitrary learning rates $\eta > 0$ and Theorem 6 is for small learning rates $0 < \eta \lesssim \frac{1}{LMK}$.

Theorem 5 *There exist a multi-dimensional global objective function, whose local objective functions are μ-strongly convex (Definition 1) and L-smooth (Definition 2), and satisfy Assumptions 2 and 4, and an initialization point $\mathbf{x}^{(0)}$ such that for any $\eta > 0$ and $R \geq 1$, $M \geq 4$, $K \geq 1$, the last-round global parameters $\mathbf{x}^{(R)}$ satisfy*

$$\mathbb{E}\left[F(\mathbf{x}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{\sigma^2}{\mu MKR} + \frac{\zeta^2}{\mu MR^2}\right).$$

Proof The proofs are in Appendices B, C and D. We construct the lower bounds for stochasticity terms and heterogeneity terms, separately. For stochasticity, we let all the local objective functions be the same, that is $F = F_1 = \cdots = F_M$, and then aim to derive the lower bound for SGD (see Appendix C). For heterogeneity, we let the local component functions be the same inside each client $F_m = f_m(\cdot; \xi_m^1) = \cdots = f_m(\cdot; \xi_m^{|\mathcal{D}_m|})$ for each m, and then aim to extend the works of SGD-RR to SFL, that is, from performing one update step to performing multiple update steps on each local objective function (see Appendix D). We use the techniques in Woodworth et al. (2020b)'s Theorem 2, Yun et al. (2022)'s Theorem 4 and Proposition 5 and Cha et al. (2023)'s Theorem 3.1. ■

This lower bound, which holds for arbitrary $\eta > 0$, matches the error terms in the strongly convex case in Corollary 4, up to a factor of κ and some polylogarithmic factors. In the next theorem, for small learning rates $\eta \lesssim \frac{1}{LMK}$, we can remove the gap of κ.

Theorem 6 *Under the same conditions of Theorem 5 (unless explicitly stated), there exist a multi-dimensional global objective function and an initialization point, such that for $\eta \leq \frac{1}{101LMK}$, the arbitrary weighted average global parameters $\bar{\mathbf{x}}^{(R)}$ satisfy the lower bounds:*

***Strongly convex**: If $R \geq \frac{1}{1010}\kappa$ and $\kappa \geq 1010$, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta^2}{\mu^2 MR^2}\right).$$

***General convex**: If $R \geq 51^3 \max\left\{\frac{\sigma}{LM^{1/2}K^{1/2}D}, \frac{L^2MKD^2}{\sigma^2}, \frac{\zeta}{LM^{1/2}D}, \frac{L^2MD^2}{\zeta^2}\right\}$, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L\zeta^2 D^4\right)^{1/3}}{M^{1/3}R^{2/3}}\right).$$

Proof The proofs are in Appendix B.2. We use similar techniques in Woodworth et al. (2020b); Cha et al. (2023). ■

Theorem 6 provides the lower bounds in the convex cases with any arbitrary weighted average parameter $\bar{\mathbf{x}}^{(R)}$ for small learning rates $\eta \lesssim \frac{1}{LMK}$. Although the constraint on small learning rates seems stringent, Theorem 3 indicates that such small learning rates $\tilde{\eta} = \eta MK \lesssim \frac{1}{L}$ also exist in our upper bounds. Therefore, it is justified to use the lower bounds to assess the tightness of our upper bounds.

In the strongly convex case in Corollary 4, if $R \gtrsim \kappa$, the learning rate choice becomes

$$\tilde{\eta} \asymp \min\left\{\frac{1}{L}, \frac{1}{\mu R}\right\} \asymp \frac{1}{\mu R} \not\asymp \frac{1}{L},$$

which yields the upper bound of $\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta^2}{\mu^2 MR^2}\right)$ for SFL. This upper bound exactly matches the lower bound in Theorem 6 (ignoring polylogarithmic factors). Notably, the upper bound differs from the original bound in Corollary 4 due to the optimization term (the last term) in Corollary 4 being present only when $\tilde{\eta} \asymp \frac{1}{L}$ in the convex cases (see the proofs of Li and Lyu 2023's Lemmas 7 and 8). Moreover, it is appropriate to compare the upper bounds (with ζ_*) and the lower bounds (with ζ), since Assumption 4 is stronger than Assumption 3 (Cha et al., 2023).

In the general convex case, with the same logic as in the strongly convex case, if $R \gtrsim \max\left\{\frac{\sigma}{LM^{1/2}K^{1/2}D}, \frac{L^2MKD^2}{\sigma^2}, \frac{\zeta}{LM^{1/2}D}, \frac{L^2MD^2}{\zeta^2}\right\}$, then the upper bound of SFL becomes $\mathcal{O}\left(\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{M^{1/3}R^{2/3}}\right)$. It matches the lower bound in Theorem 6.

Key points and limitations of Subsection 4.3. The matching lower bounds in Theorem 5 and Theorem 6 verify that our upper bounds are tight in the convex cases for the sufficiently large number of training rounds R. However, the lower bounds for small R are loose and the lower bounds for the non-convex case are still lacking, for both SGD-RR and SFL.

5. Comparison Between PFL and SFL

Unless otherwise stated, our comparisons are in terms of training rounds, which is also adopted in Gao et al. (2021). This comparison (running for the same number of total training rounds R) is fair when considering the same total computation cost for both methods. We summarize the existing convergence results of PFL in Table 1.

5.1 Comparison under Assumption 3

Theorem 7 *Under the same conditions as those of the strongly convex case in Theorem 3, there exist* $\tilde{\eta} = \eta K \asymp \min\left\{\frac{1}{L}, \frac{1}{\mu R}\right\}$ *and* $w_r = (1 - \frac{\mu\tilde{\eta}}{2})^{-(r+1)}$*, such that for* $R \gtrsim \kappa$*,*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + \mu D^2 \exp\left(\frac{-\mu R}{L}\right)\right).$$

Proof Applying Karimireddy et al. (2020)'s Lemma 1 instead of Koloskova et al. (2020)'s Lemma 15 to the final recursion in Koloskova et al. (2020) yields this theorem. The detailed

Method	Upper Bound
Strongly Convex	
PFL (Karimireddy et al., 2020)	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta^2}{\mu^2 R^2} + \mu D^2 \exp\left(\frac{-\mu R}{L}\right)$ (1)
PFL (Koloskova et al., 2020)	$\frac{\sigma_*^2}{\mu MKR} + \frac{L\sigma_*^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + LKD^2 \exp\left(\frac{-\mu R}{L}\right)$ (2)
PFL (Theorem 7)	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + \mu D^2 \exp\left(\frac{-\mu R}{L}\right)$
PFL (Woodworth et al., 2020b)	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\hat{\zeta}^2}{\mu^2 R^2} + \mu D^2 \exp\left(\frac{-\mu \boldsymbol{K} R}{L}\right)$ (3)
SFL (Theorem 3)	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 \boldsymbol{M} KR^2} + \frac{L\zeta_*^2}{\mu^2 \boldsymbol{M} R^2} + \mu D^2 \exp\left(\frac{-\mu R}{L}\right)$
Convex	
PFL (Karimireddy et al., 2020)	$\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\left(L\zeta^2 D^4\right)^{1/3}}{R^{2/3}} + \frac{LD^2}{R}$
PFL (Koloskova et al., 2020)	$\frac{\sigma_* D}{\sqrt{MKR}} + \frac{\left(L\sigma_*^2 D^4\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{R^{2/3}} + \frac{LD^2}{R}$
PFL (Woodworth et al., 2020b)	$\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\left(L\hat{\zeta}^2 D^4\right)^{1/3}}{R^{2/3}} + \frac{LD^2}{\boldsymbol{K} R}$
SFL (Theorem 3)	$\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{\boldsymbol{M^{1/3}} K^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{\boldsymbol{M^{1/3}} R^{2/3}} + \frac{LD^2}{R}$
Non-convex	
PFL (Karimireddy et al., 2020; Koloskova et al., 2020)	$\frac{\left(L\sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2\sigma^2 A^2\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2 A^2\right)^{1/3}}{R^{2/3}} + \frac{LA}{R}$ (4)
SFL (Theorem 3)	$\frac{\left(L\sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2\sigma^2 A^2\right)^{1/3}}{\boldsymbol{M^{1/3}} K^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2 A^2\right)^{1/3}}{\boldsymbol{M^{1/3}} R^{2/3}} + \frac{LA}{R}$ (5)

(1) (i) We use $\frac{3L\eta^3 K^3 \sigma^2}{K}$ (see the last inequality of the proof of their Lemma 8) while Karimireddy et al. (2020) use $\frac{\eta^2 K^2 \sigma^2}{2K}$ with $\eta \leq 8LK$ (their Lemma 8), which causes the difference between their original bounds and our recovered bounds. (ii) This difference also exists in the other two cases. (iii) Their Assumption A1 is essentially equivalent to Assumption 4. For simplicity, we let $B = 1$ in their Assumption A1 for all three cases.

(2) Even the weaker Assumption 1 is used in Koloskova et al. (2020), we do not consider it is a improvement over ours in this paper, given the discussions in Subsection 4.1.

(3) Applying Karimireddy et al. (2020)'s Lemma 1 instead of their Theorem 3 yields this bound. Notably, Woodworth et al. (2020b) assume the average of the local parameters for all iterations can be obtained, which is in fact impractical in FL. Similar assumptions are made in Khaled et al. (2020); Koloskova et al. (2020). In this paper, we omit this difference.

(4) We let $P = 1, M = 0$ in Koloskova et al. (2020)'s Assumption 3b.

(5) We let $\beta = 0$ in Assumption 4.

Table 1: Upper bounds of PFL and SFL with absolute constants and polylogarithmic factors omitted. We highlight the upper bounds of "PFL under Assumption 5"/"SFL" with a blue/green background. Main differences are marked in red fonts.

proofs (specialized for PFL) are in Li and Lyu (2023). ■

To the best of our knowledge, the existing tightest upper bounds that uses Assumption 3 to catch the heterogeneity for PFL are introduced in Koloskova et al. (2020). Many works (Woodworth et al., 2020b; Yun et al., 2022; Glasgow et al., 2022) have constructed lower bounds to show these bounds are almost the tightest for the convex cases. Glasgow et al. (2022) has shown that this upper bound for the general convex case is not improvable.

For the following comparisons in this subsection, we mainly focus on the strongly convex case. For fairness, we slightly improve the bound of PFL in the strongly convex case in Theorem 7 by combining the works of Karimireddy et al. (2020); Koloskova et al. (2020). Unless otherwise stated, the conclusions also hold for the other two cases.

- *The upper bounds of SFL are better than PFL on heterogeneous data.* As shown in Table 1 (Theorems 3 and 7), the upper bound of SFL is better than that of PFL, with an advantage of $1/M$ on the second and third terms (marked in red). This benefits from its sequential and shuffling-based training manner of SFL.

- *Partial client participation.* In the more challenging cross-device settings, only a small fraction of clients participate in each training round. Following the work in Karimireddy et al. (2020); Yang et al. (2021), we provide the upper bounds for PFL and SFL with partial client participation as follows:

$$\text{PFL:}\quad \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu SKR}+\frac{\zeta_*^2}{\mu R}\frac{M-S}{S(M-1)}+\frac{L\sigma^2}{\mu^2KR^2}+\frac{L\zeta_*^2}{\mu^2R^2}+\mu D^2\exp\left(\frac{-\mu R}{L}\right)\right),$$

$$\text{SFL:}\quad \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu SKR}+\frac{\zeta_*^2}{\mu R}\frac{(M-S)}{S(M-1)}+\frac{L\sigma^2}{\mu^2\boldsymbol{S}KR^2}+\frac{L\zeta_*^2}{\mu^2\boldsymbol{S}R^2}+\mu D^2\exp\left(\frac{-\mu R}{L}\right)\right),$$

  where a subset of clients $\mathcal{S}$ (its size is $|\mathcal{S}|=S$) are selected randomly without replacement in each training round. There are additional terms (the second terms) for both PFL and SFL, which is due to partial client participation and random sampling (Yang et al., 2021). *It can be seen that the advantage of $1/S$ (marked in red) of SFL still exists, similar to the full client participation setup.* The proofs are in Li and Lyu (2023).

  Notably, the proofs of SFL with partial client participation are nontrivial considering $\mathbb{E}_\pi\left[\frac{1}{SK}\sum_{m\in\mathcal{S},k}\nabla F_{\pi_m}(\mathbf{x}_{m,k})\right]\neq\frac{1}{MK}\sum_{m,k}\mathbb{E}\left[\nabla F_m(\mathbf{x}_{m,k})\right]$ (updates in different clients are not independent) and we cannot transform them into the full participation setup directly as done in PFL (Karimireddy et al., 2020; Yang et al., 2021).

Key points of Subsection 5.1. The discussions above show that the upper bounds of SFL are better than PFL with both full client participation and partial client participation under Assumption 3 in the convex cases and under Assumption 4 in the non-convex case.

5.2 Comparison under Assumption 5

Since it is hard to achieve an improvement for SFL even with the stronger Assumption 5, we next compare Corollary 4 with Woodworth et al. (2020b)'s Theorem 3 (under Assumption 5) to show that PFL can outperform SFL when the heterogeneity is very small. For comparison on bounds with different heterogeneity assumptions, we note that if Assumption 5 holds, then Assumption 3 holds, and $\zeta_*\leq\hat{\zeta}$.

As shown in Table 1, the results of PFL under Assumption 5 are highlighted with a blue background and the results of SFL (under Assumption 3) are highlighted with a green background. These bounds closely resembles each other, with three error terms (the first three terms containing σ, ζ) and one optimization term (the last one). To emphasize the role of heterogeneity, we let $\sigma = 0, \mu = L = D = 1$ as done in Woodworth et al. (2020b).

In the strongly convex case, it can be seen that the upper bound of PFL shows better on its optimization term, while worse in the error terms. Consequently, to make the upper bound of PFL smaller, one sufficient (not necessary) condition is $\frac{\hat{\zeta}^2}{R^2} \lesssim \exp(-KR)$, or equivalently $\hat{\zeta}^2 \lesssim R^2 \cdot \exp(-KR)$, which implies that $\hat{\zeta}$ should be very small, or the level of heterogeneity is very low. In this condition, the optimization terms become dominant for both PFL and SFL,

$$\frac{\zeta_*^2}{MR^2} \lesssim \frac{\hat{\zeta}^2}{R^2} \lesssim \exp(-KR) \lesssim \exp(-R),$$

and then the bound of PFL will be better than that of SFL. However, similarly, once $\zeta_*^2 \gtrsim MR^2 \exp(-R)$, the error terms will become dominant and SFL becomes better.

In the general convex case, with the same logic as the strongly convex case, the sufficient (not necessary) condition is $\hat{\zeta}^2 \lesssim 1/(K^3 R)$, which still implies that $\hat{\zeta}$ should be very small.

Key points of Subsection 5.2. The discussions above show that the upper bounds of PFL can be better than SFL only when the heterogeneity is very small under Assumption 5 in the convex cases. However, it is unclear whether this superiority still exists under Assumption 3 in the convex cases, and in the non-convex case.

6. Experiments

We conduct experiments on quadratic functions (Subsection 6.1), logistic regression (Subsection 6.2) and deep neural networks (Subsection 6.3) to validate our theoretical finding that SFL outperforms PFL in heterogeneous settings, at least when the level of heterogeneity is relatively high. The code is available at `https://github.com/liyipeng00/SFL`.

6.1 Experiments on Quadratic Functions

According to the analyses in Subsections 5.1 and 5.2, SFL outperforms PFL in heterogeneous settings (at least when the level of heterogeneity is relatively high). Here we show that the counterintuitive result (in contrast to Gao et al. 2021) can appear even for simple one-dimensional quadratic functions (Karimireddy et al., 2020).

To further catch the heterogeneity, in addition to Assumption 3, we also consider Hessian of objective functions (Karimireddy et al., 2020; Glasgow et al., 2022; Patel et al., 2024):

$$\max_m \left\| \nabla^2 F_m(\mathbf{x}) - \nabla^2 F(\mathbf{x}) \right\| \leq \delta.$$

Larger value of δ means higher heterogeneity on Hessian.

$$\left\| \nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y}) \right\| \leq H \left\| \mathbf{x} - \mathbf{y} \right\|.$$

Larger value of H means more drastic Hessian change.

As shown in Table 2, we use ten groups of objective functions with various degrees of heterogeneity. In fact, we construct the lower bounds in Theorem 5 with similar functions.

As suggested by our theory, we set the learning rate of SFL be half of that of PFL. The experimental results of Table 2 are shown in Figure 2.

Overall, SFL outperforms PFL in all settings except the settings $\delta = 0$ and $H = 0$ (Groups 1, 6), which coincides with our theoretical conclusion. We attribute the unexpected cases to the limitations of existing works under Assumptions 3, 4 and 5, which omit the function of the global aggregation and thus underestimate the capacity of PFL (Wang et al., 2024). More specifically, the second-order information (Hessian) is not fully studied in existing works (Patel et al., 2023, 2024).

Group 1	Group 2	Group 3	Group 4	Group 5
$\begin{cases} F_1 = \frac{1}{2}x^2 + x \\ F_2 = \frac{1}{2}x^2 - x \end{cases}$ $\zeta_* = 1, \delta = 0, H = 0$	$\begin{cases} F_1 = \frac{3}{4}x^2 + x \\ F_2 = \frac{1}{4}x^2 - x \end{cases}$ $\zeta_* = 1, \delta = \frac{1}{2}, H = 0$	$\begin{cases} F_1 = x^2 + x \\ F_2 = -x \end{cases}$ $\zeta_* = 1, \delta = 1, H = 0$	$\begin{cases} F_1 = (\frac{3}{4}\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 + x \\ F_2 = (\frac{3}{4}\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 - x \end{cases}$ $\zeta_* = 1, \delta = 0, H = \frac{1}{2}$	$\begin{cases} F_1 = (1\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 + x \\ F_2 = (1\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 - x \end{cases}$ $\zeta_* = 1, \delta = 0, H = 1$
Group 6	**Group 7**	**Group 8**	**Group 9**	**Group 10**
$\begin{cases} F_1 = \frac{1}{2}x^2 + 10x \\ F_2 = \frac{1}{2}x^2 - 10x \end{cases}$ $\zeta_* = 10, \delta = 0, H = 0$	$\begin{cases} F_1 = \frac{3}{4}x^2 + 10x \\ F_2 = \frac{1}{4}x^2 - 10x \end{cases}$ $\zeta_* = 10, \delta = \frac{1}{2}, H = 0$	$\begin{cases} F_1 = x^2 + 10x \\ F_2 = -10x \end{cases}$ $\zeta_* = 10, \delta = 1, H = 0$	$\begin{cases} F_1 = (\frac{3}{4}\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 + 10x \\ F_2 = (\frac{3}{4}\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 - 10x \end{cases}$ $\zeta_* = 10, \delta = 0, H = \frac{1}{2}$	$\begin{cases} F_1 = (1\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 + 10x \\ F_2 = (1\mathbb{1}_{x<0} + \frac{1}{2}\mathbb{1}_{x\geq 0})x^2 - 10x \end{cases}$ $\zeta_* = 10, \delta = 0, H = 1$

Table 2: Settings of the experiments on quadratic functions. Each group has two local objectives ($M = 2$). Strictly speaking, the functions in Groups 4, 5, 9, 10 are not quadratic functions.

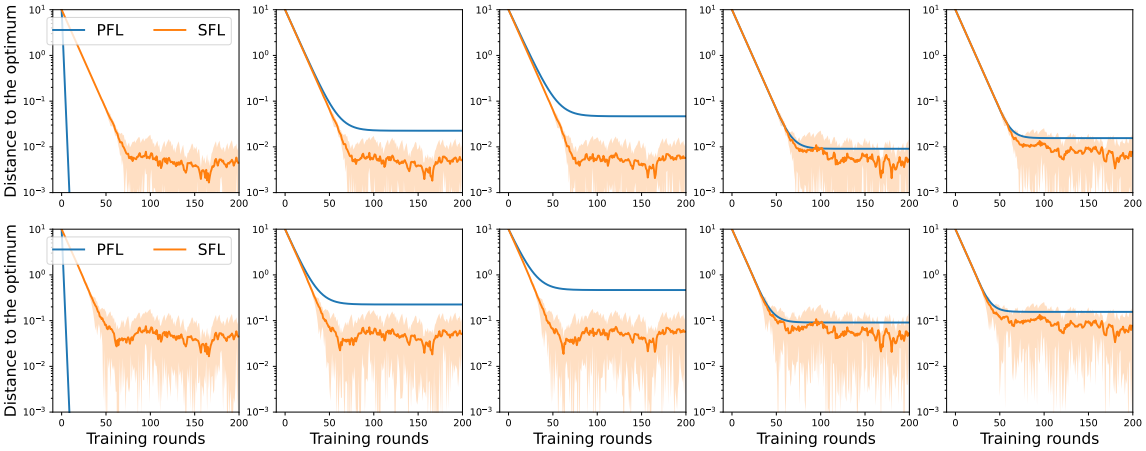

Figure 2: Results of the experiments on quadratic functions. It displays the experimental results of ten groups in Table 2. The top (bottom) row shows the first (last) five groups from left to right. We set $K = 10$. The shaded areas show the min-max values across 10 random seeds.

6.2 Experiments on Logistic Regression

We consider the classic logistic regression for the binary classification problem (Khaled et al., 2020; Mishchenko et al., 2020, 2022a,b; Malinovsky et al., 2023; Sadiev et al., 2023). Specifically, the local objective function F_m is defined as

$$F_m(\mathbf{x}) = -\left(b_m \log\left(h\left(a_m^{\mathrm{T}}\mathbf{x}\right)\right) + (1-b_m)\log\left(1-h\left(a_m^{\mathrm{T}}\mathbf{x}\right)\right)\right) + \frac{1}{2}\omega\left\|\mathbf{x}\right\|^2,$$

where $\mathbf{x} \in \mathbb{R}^d$ is the model parameters, $a_n \in \mathbb{R}^d$, $b_n \in \{0,1\}$ are the data samples, $h : x \to 1/(1+e^{-x})$ is the sigmoid function and ω is the L2 regularization parameter.

We use two data sets "a9a" and "w8a" from LIBSVM library (Chang and Lin, 2011). We partition them into $M = 1000$ clients by Extended Dirichlet strategy (Li and Lyu, 2023), with each client containing data samples from $C = 1, 2$ labels. Larger value of C means higher data heterogeneity. We set the number of local steps to $K = 5$, the number of participating clients to $S = 10$, and the mini-batch size to 8. The local solver is SGD with learning rate being constant, momentum being 0 and weight decay being 0. We tune the learning rate by the grid search. We run each experiment with 10 different random seeds.

The experimental results of PFL and SFL are in Figure 3. It can be observed that when the level of heterogeneity is relatively high ($C = 1$), the performance of SFL is better than that of PFL, and when the level of heterogeneity is low ($C = 2$), the performances are close. This is consistent with our theoretical finding.

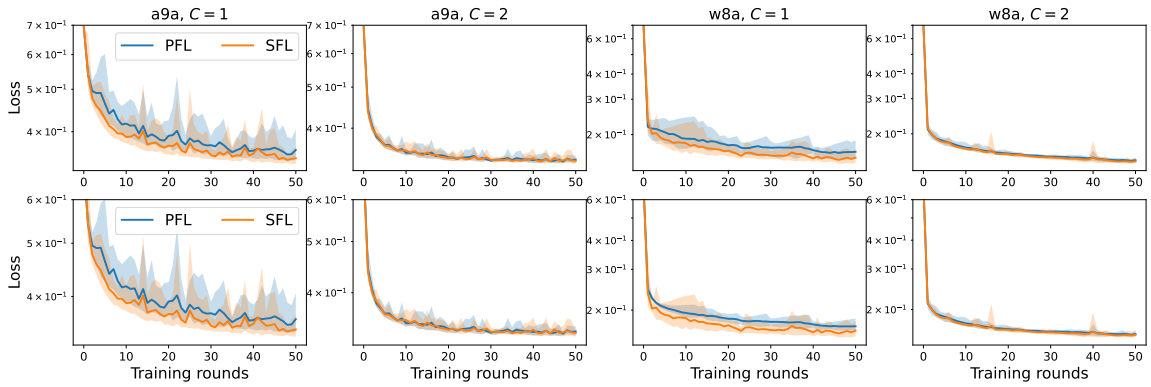

Figure 3: Training loss results of PFL and SFL. The top row shows the results when $\omega = 0.0$ and the bottom row shows the results when $\omega = 0.0001$. The shaded areas show the min-max values across 10 random seeds.

6.3 Experiments on Deep Neural Networks

Setup. We consider the common CV tasks, with data sets including Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), CINIC-10 (Darlow et al., 2018). Specifically, we train a CNN model from Wang and Ji (2022) on Fashion-MNIST and a VGG-9 model (Simonyan and Zisserman, 2014) from Lin et al. (2020) on CIFAR-10 and CINIC-10. We partition the training sets of Fashion-MNIST/CIFAR-10/CINIC-10 into 500/500/1000

clients by Extended Dirichlet strategy (Li and Lyu, 2023), with each client containing data samples from $C = 1, 2, 5$ labels. Larger value of C means higher data heterogeneity. We spare the original test sets for computing test accuracy. We fix the number of participating clients per round to $S = 10$. We fix the number of local update steps to $K = 5$ and the mini-batch size to 20 (about one single pass over the local data for each client) (Reddi et al., 2021). The local solver is SGD with learning rate being constant, momentem being 0 and weight decay being 0. We apply gradient clipping to both algorithms and tune the learning rate by grid search with a grid of $\{10^{-2.5}, 10^{-2.0}, 10^{-1.5}, 10^{-1.0}, 10^{-0.5}\}$.

SFL outperforms PFL on heterogeneous data. The accuracy results on training data and test data for various tasks are collected in Table 3. In particular, the test accuracy curves on CIFAR-10 are shown in Figure 4. It can be observed (i) that when the level of heterogeneity is relatively high (for example, $C = 1, 2$) the performance of SFL is much better than that of PFL, and (ii) that when the level of heterogeneity is low, the performances of both are close to each other. This is consistent with our theoretical finding.

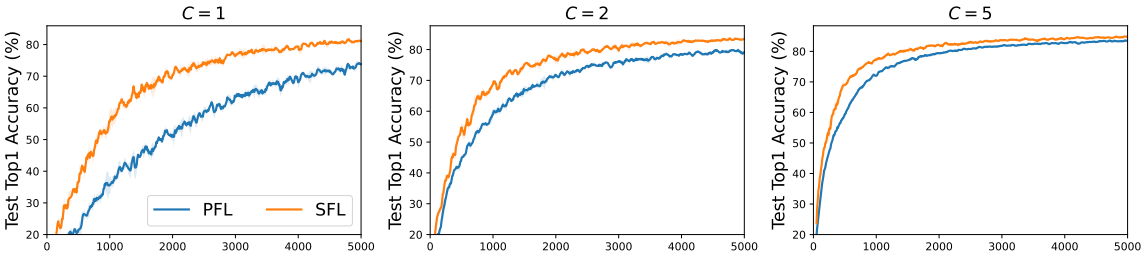

Figure 4: Test accuracy results of PFL and SFL on CIFAR-10. For visualization, we apply moving average over a window length of 5 data points. The shaded areas show the standard deviation across 3 random seeds.

Data set	Method	$C = 1$		$C = 2$		$C = 5$	
		Train	Test	Train	Test	Train	Test
Fashion-MNIST	PFL	$86.71_{\pm 1.87}$	$85.77_{\pm 1.96}$	$89.73_{\pm 1.23}$	$88.55_{\pm 1.19}$	$92.27_{\pm 0.57}$	$90.70_{\pm 0.50}$
	SFL	$88.86_{\pm 1.60}$	$87.60_{\pm 1.56}$	$91.33_{\pm 1.49}$	$89.66_{\pm 1.41}$	$92.83_{\pm 0.69}$	$90.92_{\pm 0.64}$
CIFAR-10	PFL	$76.48_{\pm 2.03}$	$73.84_{\pm 1.90}$	$85.92_{\pm 1.77}$	$78.99_{\pm 1.43}$	$94.55_{\pm 0.36}$	$83.47_{\pm 0.48}$
	SFL	$89.60_{\pm 2.29}$	$81.05_{\pm 1.78}$	$94.01_{\pm 0.97}$	$83.34_{\pm 0.68}$	$96.72_{\pm 0.50}$	$84.73_{\pm 0.44}$
CINIC-10	PFL	$53.36_{\pm 3.80}$	$52.27_{\pm 3.61}$	$65.38_{\pm 2.01}$	$61.96_{\pm 1.81}$	$74.97_{\pm 0.95}$	$68.45_{\pm 0.81}$
	SFL	$65.40_{\pm 3.57}$	$61.52_{\pm 3.14}$	$73.58_{\pm 2.32}$	$67.31_{\pm 1.87}$	$79.58_{\pm 1.42}$	$70.82_{\pm 1.05}$

Table 3: Training and test accuracy results of PFL and SFL on Fashion-MNIST, CIFAR-10 and CINIC-10. We run PFL and SFL for 2000/5000/5000 training rounds on Fashion-MNIST/CIFAR-10/CINIC-10. Results are computed across 3 random seeds and the last 100 training rounds.

7. Conclusion

In this paper, we have derived the upper bounds of SFL for the strongly convex, general convex and non-convex objective functions on heterogeneous data. We validate that the upper bounds of SFL are tight by constructing the corresponding lower bounds of SFL in the strongly convex and general convex cases. We also make comparisons between the upper bounds of SFL and those of PFL. In the convex cases, the comparison results show a subtle difference under different heterogeneity assumptions. That is, under Assumption 3, the upper bounds of SFL are better than those of PFL strictly, while under Assumption 5, the upper bounds of SFL are still better unless the level of heterogeneity is very low. In the non-convex case under Assumption 4, the upper bounds of SFL are better without exception. Experiments on quadratic functions, logistic regression and deep neural networks validate the theoretical finding that SFL outperforms PFL on heterogeneous data, at least, when the level of heterogeneity is relatively high.

Although this work has proved that SFL outperforms PFL on heterogeneous data with the standard assumptions, we believe the comparisons are still open. Are there any other conditions to overturn this conclusion? For example, new assumptions beyond the standard assumptions, new factors beyond data heterogeneity, new algorithms beyond vanilla PFL and SFL. One possible future direction is the convergence of PFL and SFL under Hessian assumptions to explain the unexpected results (Groups 1, 6) in Subsection 6.1. Another promising future direction is to investigate whether the server extrapolation can be applied to SFL to achieve faster convergence as proven in PFL (Jhunjhunwala et al., 2023; Li et al., 2024; Li and Richtárik, 2024).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62371059, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60006.

We thank the anonymous reviewers of NeurIPS 2023 and JMLR for their insightful suggestions.

Appendix

Appendix A. Notations

Table 4 summarizes the notations appearing in this paper.

Symbol	Description
R, r	number, index of training rounds
M, m	number, index of clients
K, k	number, index of local update steps
$\mathcal{S}, S$	the set of participating clients and its size
π	$\{\pi_1, \pi_2, \ldots, \pi_M\}$ is a permutation of $[M]$
η, $\tilde{\eta}$	learning rate, effective learning rate ($\eta_{\mathrm{SFL}} \coloneqq \eta MK$ and $\eta_{\mathrm{PFL}} \coloneqq \eta K$)
L, μ, κ	constants in Asm. 2 and Asm. 1; conditional number $\kappa \coloneqq L/\mu$
σ_*, σ	constants in Asm. 1 and Asm. 2 for stochasticity
$\zeta_*, \zeta(\beta), \hat{\zeta}$	constants in Asm. 3, Asm. 4 and Asm. 5 for heterogeneity
δ, H	constants in Subsection 6.1 for Hessian
F, F_m, f_m	global objective, local objective and local component function
$\mathbf{x}^{(r)}$	global model parameters in the r-th round
$\mathbf{x}_{m,k}^{(r)}$	local model parameters of the m-th client after k local steps in the r-th round
$\mathbf{g}_{\pi_m,k}^{(r)}$	$\mathbf{g}_{\pi_m,k}^{(r)} \coloneqq \nabla f_{\pi_m}(\mathbf{x}_{m,k}^{(r)}; \xi_{m,k}^{(r)})$ is the stochastic gradients of F_{π_m} regarding $\mathbf{x}_{m,k}^{(r)}$
C	each client containing data samples from C labels (Subsec. 6.3)

Table 4: Summary of key notations.

Appendix B. Proofs of Theorem 5 and Theorem 6

In this section, we use the results in Appendices C and D to compose the final lower bounds.

For clarity, we have summarized the lower bounds and the corresponding setups for these regimes in Tables 5 and 6. Since we use the typical objective functions in Appendices C and D, we omit the step to verify that these functions satisfy the assumptions in Theorem 5 (see the proofs of Cha et al. 2023's Theorem 3.1 about this step if needed).

B.1 Proof of Theorem 5

Proof For the stochasticity terms, we let $\lambda = \mu$ for $\eta \leq \frac{1}{102010\lambda NR}$; we let $\lambda_1 = \mu$ and $\lambda_0 = 1010\mu$ for $\frac{1}{102010\lambda_1 NR} \leq \eta \leq \frac{1}{101\lambda_0 N}$; we let $\lambda = 1010\mu$ for the other regimes in Table 5. There exist a 4-dimensional global objective function and an initialization point $\mathbf{x}^{(0)} = \left[\frac{\sigma}{\mu}, \frac{1}{8160800}\frac{\sigma}{\mu N^{\frac{1}{2}}R}, 0, \frac{\sigma}{\mu}\right]^{\mathrm{T}}$, such that for $\kappa = \frac{L}{\mu} \geq 2020$, $M \geq 4$ and $R \geq 1$, the lower bound for stochasticity terms satisfies

$$\mathbb{E}\left[F(\mathbf{x}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\min\left\{\frac{\sigma^2}{\mu}, \frac{\sigma^2}{\mu NR^2}, \frac{\sigma^2}{\mu N}, \frac{\sigma^2}{\mu}\right\}\right) = \Omega\left(\frac{\sigma^2}{\mu NR^2}\right),$$

where $\mathbf{x}^{(R)} = \left[x_1^{(R)}, x_2^{(R)}, x_3^{(R)}, x_4^{(R)}\right]^{\mathrm{T}}$ and $\mathbf{x}^* = [0,0,0,0]^{\mathrm{T}}$.

Notably, these dimensions are orthogonal. For one single round in SFL, with $N = MK$, the lower bound is $\Omega\left(\frac{\sigma^2}{\mu MKR^2}\right)$. It is well known that any first-order method which accesses at most MKR stochastic gradients with variance σ^2 for a μ-strongly convex objective will suffer error at least $\mathcal{O}\left(\frac{\sigma^2}{\mu MKR}\right)$ in the worst case (Nemirovskij and Yudin, 1983; Woodworth et al., 2020a,b). Therefore, we get the lower bound of $\Omega\left(\frac{\sigma^2}{\mu MKR}\right)$ for stochasticity terms.

For the heterogeneity terms, we let $\lambda = \mu$ for $\eta \leq \frac{1}{102010\lambda MKR}$; we let $\lambda_1 = \mu$ and $\lambda_0 = 1010\mu$ for $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$; we let $\lambda = 1010\mu$ for the other regimes. There exist a 5-dimensional global objective function and an initialization point $\mathbf{x}^{(0)} = \left[\frac{\zeta}{\mu}, \frac{1}{81608000}\frac{\zeta}{\mu M^{\frac{1}{2}}R}, 0, 0, \frac{\zeta}{\mu}\right]^{\mathrm{T}}$, such that for $\kappa = \frac{L}{\mu} \geq 1010$, $M \geq 4$ and $R \geq 1$, the lower bound for heterogeneity terms satisfies

$$\mathbb{E}\left[F(\mathbf{x}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\min\left\{\frac{\zeta^2}{\mu}, \frac{\zeta^2}{\mu MR^2}, \frac{\zeta^2}{\mu M}, \frac{\zeta^2}{\mu}, \frac{\zeta^2}{\mu}\right\}\right) = \Omega\left(\frac{\zeta^2}{\mu MR^2}\right).$$

Combining these cases, we get the final lower bound of $\Omega\left(\frac{\sigma^2}{\mu MKR} + \frac{\zeta^2}{\mu MR^2}\right)$. ■

B.2 Proof of Theorem 6

Proof In this theorem, by using the small $\eta = \mathcal{O}\left(\frac{1}{LMK}\right)$, we can extend the lower bound in Theorem 5 to arbitrary weighted average global parameters $\bar{\mathbf{x}}^{(R)} = \frac{\sum_{r=0}^{R} w_r \mathbf{x}^{(r)}}{\sum_{r=0}^{R} w_r}$, and even the general convex case.

When choosing the small $\eta = \mathcal{O}\left(\frac{1}{LMK}\right)$, we only need to consider the first two regimes $\eta \leq \frac{1}{102010\lambda MKR}$ and $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$ for both stochasticity terms and heterogeneity terms. Take the heterogeneity terms as an example. In these two regimes, we can lower bound $\mathbb{E}\left[x^{(r)}\right]$, that is, $\mathbb{E}\left[x^{(r)}\right] \geq \frac{51004}{51005} D_1$ ($D_1 = \left|x^{(0)} - x^*\right|$ is the initial distance for the first dimension) for $\eta \leq \frac{1}{102010\lambda MKR}$ and $\mathbb{E}\left[x^{(r)}\right] \geq \frac{\zeta}{81608000\lambda_1 M^{\frac{1}{2}}R}$ for $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$. With these lower bounds, the arbitrary weighted average parameters $\bar{x}^{(R)}$ can also be bounded

$$\bar{x}^{(R)} = \frac{\sum_{r=0}^{R} w_r x^{(r)}}{\sum_{r=0}^{R} w_r} \geq \frac{\sum_{r=0}^{R} w_r c}{\sum_{r=0}^{R} w_r} \geq c,$$

where $c = \frac{51004}{51005} D_1$ for $\eta \leq \frac{1}{102010\lambda MKR}$ and $c = \frac{\zeta}{81608000\lambda_1 M^{\frac{1}{2}}R}$ for $\frac{1}{102010\lambda_1 MKR} \leq \frac{1}{101\lambda_0 MK}$. The stochasticity terms are similar and thus be omitted here. Refer to the proofs of Cha et al. (2023)'s Theorem 3.3 for details if needed. In summary, we can get the same lower bound for $\bar{\mathbf{x}}^{(R)}$ and $\mathbf{x}^{(R)}$ in the first two regimes for stochasticity and heterogeneity.

B.2.1 Strongly Convex Case

For the stochasticity terms, we let $\lambda = \mu$ for $0 < \eta \leq \frac{1}{102010\lambda NR}$; we let $\lambda_1 = \mu, \lambda_0 = L$ for $\frac{1}{102010\lambda_1 NR} \leq \eta \leq \frac{1}{101\lambda_0 N}$ (with $N = MK$) in Table 5. There exist a 2-dimensional global

objective function and an initialization point $\mathbf{x}^{(0)} = \left[\frac{\sigma}{\mu}, \frac{1}{8160800}\frac{\sigma}{\mu M^{\frac{1}{2}} K^{\frac{1}{2}} R}\right]^{\mathrm{T}}$, such that for $\kappa \geq 1010$, $M \geq 4$ and $R \geq \frac{1}{1010}\kappa$, the lower bound for stochasticity terms satisfies

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\min\left\{\frac{\sigma^2}{\mu}, \frac{L\sigma^2}{\mu^2 MKR^2}\right\}\right) = \Omega\left(\frac{L\sigma^2}{\mu^2 MKR^2}\right).$$

Therefore, with the same logic in the proof of Theorem 5, we get the lower bound of $\Omega\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2}\right)$ for stochasticity terms.

For heterogeneity terms, we let $\lambda = \mu$ for $0 < \eta \leq \frac{1}{102010\lambda MKR}$; we let $\lambda_1 = \mu, \lambda_0 = L$ for $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$ in Table 6. There exist a 2-dimensional global objective function and an initialization point $\mathbf{x}^{(0)} = \left[\frac{\zeta}{\mu}, \frac{1}{81608000}\frac{\zeta}{\mu M^{\frac{1}{2}} R}\right]^{\mathrm{T}}$, such that for $\kappa \geq 1010$, $M \geq 4$ and $R \geq \frac{1}{1010}\kappa$, the lower bound for heterogeneity terms satisfies

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\min\left\{\frac{\zeta^2}{\mu}, \frac{L\zeta^2}{\mu^2 MR^2}\right\}\right) = \Omega\left(\frac{L\zeta^2}{\mu^2 MR^2}\right).$$

Therefore, we get the lower bound of $\Omega\left(\frac{L\zeta^2}{\mu^2 MR^2}\right)$ for heterogeneity terms.

Combining them, we get the lower bound of $\Omega\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta^2}{\mu^2 MR^2}\right)$.

B.2.2 General Convex Case

As done in Woodworth et al. (2020b)'s Theorem 2 and Cha et al. (2023)'s Corollary 3.5, we need to choose λ, λ_0 and λ_1 more carefully for the general convex case.

For the stochasticity terms, we let $\lambda = \frac{L^{1/3}\sigma^{2/3}}{M^{1/3}K^{1/3}R^{2/3}D^{2/3}}$ for the first regime; we let $\lambda_1 = \frac{L^{1/3}\sigma^{2/3}}{M^{1/3}K^{1/3}R^{2/3}D^{2/3}}$, $\lambda_0 = L$ for the second regime in Table 5. For the heterogeneity terms, we let $\lambda = \frac{L^{1/3}\zeta^{2/3}}{M^{1/3}R^{2/3}D^{2/3}}$ for the first regime; we let $\lambda_1 = \frac{L^{1/3}\zeta^{2/3}}{M^{1/3}R^{2/3}D^{2/3}}$, $\lambda_0 = L$ for the second regime in Table 6. Here D is the initial distance to the optimum $\mathbf{x}^* = [0,0,0,0]^{\mathrm{T}}$.

Considering that D is affected by both stochasticity and heterogeneity, we consider a 4-dimensional global objective function. We let D_1, D_2, D_3 and D_4 are the initial distance in the first, second, third and forth dimensions, respectively. Then, if

$$\mathbf{x}^{(0)} = \left[D_1, \frac{1}{8160800}\frac{\sigma^{1/3}D^{2/3}}{L^{1/3}M^{1/6}K^{1/6}R^{1/3}}, D_1, \frac{1}{81608000}\frac{\zeta^{1/3}D^{2/3}}{L^{1/3}M^{1/6}R^{1/3}}\right]^{\mathrm{T}},$$

then

$$\begin{aligned}\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\min\left\{\frac{L^{1/3}\sigma^{2/3}D_1^2}{M^{1/3}K^{1/3}R^{2/3}D^{2/3}}, \frac{L^{1/3}\sigma^{2/3}D^{4/3}}{M^{1/3}K^{1/3}R^{2/3}}\right\}\right) \\ + \Omega\left(\min\left\{\frac{L^{1/3}\zeta^{2/3}D_1^2}{M^{1/3}R^{2/3}D^{2/3}}, \frac{L^{1/3}\zeta^{2/3}D^{4/3}}{M^{1/3}R^{2/3}}\right\}\right).\end{aligned}$$

Then, since

$$\begin{aligned}2D_1^2 &= D^2 - D_2^2 - D_4^2 \\ &= D^2\left(1 - \frac{1}{8160800^2}\frac{\sigma^{2/3}}{L^{2/3}M^{1/3}K^{1/3}R^{2/3}D^{2/3}} - \frac{1}{816080000^2}\frac{\zeta^{2/3}}{L^{2/3}M^{1/3}R^{2/3}D^{2/3}}\right) \\ &\geq D^2\left(1 - \frac{1}{8160800^2} - \frac{1}{816080000^2}\right) \geq \frac{1}{2}D^2,\end{aligned}$$

where we use the conditions $R \geq \frac{\sigma}{LM^{1/2}K^{1/2}D}$ and $R \geq \frac{\zeta}{LM^{1/2}D}$. Thus, we get

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{L^{1/3}\sigma^{2/3}D^{4/3}}{M^{1/3}K^{1/3}R^{2/3}} + \frac{L^{1/3}\zeta^{2/3}D^{4/3}}{M^{1/3}R^{2/3}}\right).$$

Adding the classic bound $\Omega\left(\frac{\sigma D}{\sqrt{MKR}}\right)$ for the general convex case (Woodworth et al., 2020a,b) yields the final result.

To ensure that the objective functions satisfy the assumptions, we use the conditions

$$\begin{aligned}2\lambda \leq L \implies R &\geq 2^{\frac{3}{2}} \cdot \frac{\sigma}{LM^{1/2}K^{1/2}D}, \\ R \geq \frac{1}{1010}\frac{\lambda_0}{\lambda_1} \implies R &\geq \frac{1}{1010^3} \cdot \frac{L^2MKD^2}{\sigma^2}.\end{aligned}$$

for stochasticity terms and the conditions

$$\begin{aligned}2\lambda \leq L \implies R &\geq 2^{\frac{3}{2}} \cdot \frac{\zeta}{LM^{1/2}D}, \\ R \geq \frac{1}{1010}\frac{\lambda_0}{\lambda_1} \implies R &\geq \frac{1}{1010^3} \cdot \frac{L^2MD^2}{\zeta^2}.\end{aligned}$$

for heterogeneity terms. Note that the choices of λ, λ_0 and λ_1 are different for stochasticity terms and heterogeneity terms. For simplicity, we can use a stricter condition

$$R \geq 4\max\left\{\frac{\sigma}{LM^{1/2}K^{1/2}D}, \frac{L^2MKD^2}{\sigma^2}, \frac{\zeta}{LM^{1/2}D}, \frac{L^2MD^2}{\zeta^2}\right\}.$$

■

Appendix C. Proof of Stochasticity Terms in Theorem 5

Proof We assume $F_1 = F_2 = \cdots = F_M = F$, and then the task is to construct the lower bound of vanilla SGD, where one objective function is sampled with replacement for updates in each step: $x_{n+1} = x_n - \eta\nabla f_{\pi_n}(x_n)$. The results are summarized in Table 5. Notably, λ_0, λ_1 and λ in different regimes can be different.

Bound	Regime	Objective functions	Initialization	$\nabla^2 F_m \in$
$\Omega\left(\frac{\sigma^2}{\lambda}\right)$	$\eta \leq \frac{1}{102010\lambda NR}$	$f_{\pi_n}(x) = \lambda x^2$	$x^{(0)} = \frac{\sigma}{\lambda}$	$[2\lambda, 2\lambda]$
$\Omega\left(\frac{\lambda_0\sigma^2}{\lambda_1^2 NR^2}\right)$	$\frac{1}{102010\lambda_1 NR} \leq \eta$ and $\eta \leq \frac{1}{101\lambda_0 N}$	$f_{\pi_n}(x)$ $= (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} + \sigma\tau_n x$	$x^{(0)} =$ $\frac{1}{8160800}\frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R}$	$\left[\frac{\lambda_0}{1010}, \lambda_0\right]$
$\Omega\left(\frac{\sigma^2}{\lambda N}\right)$	$\frac{1}{101\lambda N} \leq \eta \leq \frac{1}{\lambda}$	$f_{\pi_n}(x) = \frac{\lambda}{2}x^2 + \sigma\tau_n x$	$x^{(0)} = 0$	$[\lambda, \lambda]$
$\Omega\left(\frac{\sigma^2}{\lambda}\right)$	$\eta \geq \frac{1}{\lambda}$	$f_{\pi_n}(x) = \lambda x^2$	$x^{(0)} = \frac{\sigma}{\lambda}$	$[2\lambda, 2\lambda]$

Table 5: Lower bounds of SFL for stochasticity terms. It requires that $\lambda = \frac{\lambda_0}{1010}$ and $R \geq \frac{1}{1010}\frac{\lambda_0}{\lambda_1}$ in the regime $\frac{1}{102010\lambda_1 NR} \leq \eta \leq \frac{1}{101\lambda_0 N}$. We set $N = MK$ in SFL.

C.1 Lower Bounds for $0 < \eta \leq \frac{1}{102010\lambda NR}$ and $\eta \geq \frac{1}{\lambda}$

In this regime, we consider the following objective functions

$$f_{\pi_n}(x) = \lambda x^2,$$
$$f(x) = \mathbb{E}\left[f_{\pi_n}(x)\right] = \lambda x^2.$$

We can soon build the relationship between $x^{(R)}$ and $x^{(0)}$: $x^{(R)} = (1 - 2\lambda\eta)^{NR} x^{(0)}$.

C.1.1 Lower Bound for $0 < \eta \leq \frac{1}{102010\lambda NR}$

Since $\eta \leq \frac{1}{\lambda NR}$ and $(1 - \frac{1}{51005} \cdot \frac{1}{x})^x$ is monotonically increasing when $x \geq 1$, we have

$$x^{(R)} = (1 - 2\lambda\eta)^{NR} x^{(0)} \geq (1 - \frac{1}{51005} \cdot \frac{1}{NR})^{NR} x^{(0)} \geq \frac{51004}{51005} x^{(0)} \qquad (\because N \geq 1)$$
$$F(x^{(R)}) = \lambda\left(x^{(R)}\right)^2 \geq \frac{51004^2}{51005^2}\lambda\left(x^{(0)}\right)^2.$$

If $x^{(0)} = \frac{\sigma}{\lambda}$, we can get

$$\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] \geq \frac{51004^2}{51005^2}\lambda\left(x^{(0)}\right)^2 = \Omega\left(\frac{\sigma^2}{\lambda}\right).$$

C.1.2 Lower Bound for $\eta \geq \frac{1}{\lambda}$

Since $\eta \geq \frac{1}{\lambda}$ implies that $(1 - 2\lambda\eta)^2 \geq 1$, we have

$$F(x^{(R)}) = \lambda\left(x^{(R)}\right)^2 = \lambda(1 - 2\lambda\eta)^{2NR}\left(x^{(0)}\right)^2 \geq \lambda\left(x^{(0)}\right)^2.$$

If $x^{(0)} = \frac{\sigma}{\lambda}$, we can get

$$\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] \geq \lambda\left(x^{(0)}\right)^2 = \frac{\sigma^2}{\lambda} = \Omega\left(\frac{\sigma^2}{\lambda}\right).$$

C.2 Lower Bound for $\frac{1}{102010\lambda_1 NR} \leq \eta \leq \frac{1}{101\lambda_0 N}$

In this regime, we consider the following functions

$$f_{\pi_n}(x) = (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} + \tau_n \sigma x,$$
$$f(x) = \mathbb{E}\left[f_{\pi_n}(x)\right] = (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} \quad (\lambda_0/\lambda \geq 1010),$$

where τ_n is a random variable with equal probabilities of being either "+1" or "−1". Next, we focus on a single round r, including N local steps in total, and thus we drop the superscripts r for a while, for example, replacing $x_n^{(r)}$ with x_n. Unless otherwise stated, the expectation is conditioned on x_0 when we focus on one single round.

The relationship between the current parameter x_n and the initial parameter x_0 satisfies

$$x_n = x_0 - \eta \sum_{i=0}^{n-1} (\lambda_0 \mathbb{1}_{x_i<0} + \lambda \mathbb{1}_{x_i\geq 0}) x_i - \eta\sigma\mathcal{E}_n. \tag{1}$$

C.2.1 Lower Bound of $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]$

We first bound $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]$. Since $(\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})x \leq \lambda_0 x$ and $(\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})x \leq \lambda x$, we have

$$\begin{aligned}
\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_n<0} + \lambda \mathbb{1}_{x_n\geq 0})x_n\right] &= \mathbb{P}(\mathcal{E}_n > 0)\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_n<0} + \lambda \mathbb{1}_{x_n\geq 0})x_n \mid \mathcal{E}_n > 0\right] \\
&\quad + \mathbb{P}(\mathcal{E}_n \leq 0)\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_n<0} + \lambda \mathbb{1}_{x_n\geq 0})x_n \mid \mathcal{E}_n \leq 0\right] \\
&\leq \lambda_0 \mathbb{P}(\mathcal{E}_n > 0)\mathbb{E}\left[x_n \mid \mathcal{E}_n > 0\right] + \lambda \mathbb{P}(\mathcal{E}_n \leq 0)\mathbb{E}\left[x_n \mid \mathcal{E}_n \leq 0\right]. \quad (2)
\end{aligned}$$

According to Eq. (1), we can get

$$\begin{aligned}
\mathbb{E}\left[x_n \mid \mathcal{E}_n > 0\right] &= \mathbb{E}\left[x_0 - \eta \sum_{i=0}^{n-1} (\lambda_0 \mathbb{1}_{x_i<0} + \lambda \mathbb{1}_{x_i\geq 0}) x_i - \eta\sigma\mathcal{E}_n \mid \mathcal{E}_n > 0\right] \\
&\leq x_0 + \lambda_0 \eta \sum_{i=0}^{n-1} \mathbb{E}\left[|x_i - x_0| \mid \mathcal{E}_n > 0\right] - \eta\sigma\mathbb{E}\left[\mathcal{E}_n \mid \mathcal{E}_n > 0\right].
\end{aligned}$$

Then using $\mathbb{E}\left[|x_i - x_0|\right] \geq \mathbb{E}\left[|x_i - x_0| \mid \mathcal{E}_n > 0\right]\mathbb{P}(\mathcal{E}_n > 0)$ with $\mathbb{P}(\mathcal{E}_n > 0) \geq \frac{1}{4}$ for the second term, and $\mathbb{E}\left[|\mathcal{E}_n|\right] = 2\mathbb{E}\left[\mathcal{E}_n > 0 \mid \mathcal{E}_n > 0\right]\mathbb{P}(\mathcal{E}_n > 0)$ with $\mathbb{P}(\mathcal{E}_n > 0) \leq \frac{1}{2}$, we can get

$$\begin{aligned}
\mathbb{E}\left[x_n \mid \mathcal{E}_n > 0\right] &\leq x_0 + 4\lambda_0 \eta \sum_{i=0}^{n-1} \mathbb{E}\left[|x_i - x_0|\right] - \frac{1}{2}\eta\sigma\mathbb{E}\left[|\mathcal{E}_n|\right] \\
&\leq (1 + \frac{1}{2525})x_0 - \frac{6}{100}\eta\sigma\sqrt{n},
\end{aligned}$$

where Lemmas 10, 8 and $\lambda_0 \eta N \leq \frac{1}{101}$ are applied in the last inequality. We bound $\mathbb{E}\left[x_n \mid \mathcal{E}_n \leq 0\right]$ with a looser bound as

$$\begin{aligned}
\mathbb{E}\left[x_n \mid \mathcal{E}_n \leq 0\right] &\leq x_0 + \mathbb{E}\left[|x_n - x_0| \mid \mathcal{E}_n \leq 0\right] \\
&\leq x_0 + 2\mathbb{E}\left[|x_n - x_0|\right] \qquad (\because \mathbb{P}(\mathcal{E}_n \leq 0) \geq \tfrac{1}{2}) \\
&\leq \frac{51}{50}\lambda_0 x_0 + \frac{101}{50}\eta\sigma\sqrt{n}.
\end{aligned}$$

Then, back to Ineq. (2), we have

$$\begin{aligned}\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_n<0} + \lambda \mathbb{1}_{x_n\geq 0})x_n\right] &\leq \frac{1}{2}\lambda_0\left((1+\frac{1}{2525})x_0 - \frac{6}{100}\eta\sigma\sqrt{n}\right) + \frac{3}{4}\lambda\left(\frac{51}{50}x_0 + \frac{101}{50}\eta\sigma\sqrt{n}\right)\\ &\leq \frac{253}{505}x_0 - \frac{1}{40}\eta\sigma\sqrt{n}.\end{aligned}$$

Now, we have

$$\begin{aligned}\mathbb{E}[x_N] &= x_0 - \eta\sum_{n=0}^{N-1}(\lambda_0\mathbb{1}_{x_n<0} + \lambda\mathbb{1}_{x_n\geq 0})x_n && (\because \mathbb{E}[\mathcal{E}_N] = 0)\\ &\geq x_0 - \eta\sum_{n=0}^{N-1}\left(\frac{253}{505}\lambda_0 x_0 - \frac{1}{40}\eta\sigma\sqrt{n}\right)\\ &\geq \left(1 - \frac{2}{3}\lambda_0\eta N\right)x_0 + \frac{1}{60}\eta N^{\frac{3}{2}}\sigma.\end{aligned}$$

That is,

$$\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right] \geq \left(1 - \frac{2}{3}\lambda_0 N\eta\right)\mathbb{E}\left[x^{(r)} \mid x^{(r)} \geq 0\right] + \frac{1}{60}\eta N^{\frac{3}{2}}\sigma.$$

C.2.2 Lower Bound of $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right]$

With similar analyses in Subsection D.2.2 and Cha et al. (2023)' Lemma B.4, we get

$$\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right] \geq \left(1 - \frac{2}{3}\lambda_0 N\eta\right)\mathbb{E}\left[x^{(r)} \mid x^{(r)} < 0\right].$$

C.2.3 Relationship Between $\mathbb{P}(x^{(r)}) \geq 0$ and $\mathbb{P}(x^{(r)}) < 0$.

With similar analyses in Subsection D.2.3 and Cha et al. (2023)' Lemma B.4, we get

$$\mathbb{P}\left(x^{(r)} \geq 0\right) \geq \frac{1}{2},$$

when $x^{(0)} \geq 0$.

C.2.4 Lower Bound for $\frac{1}{102010\lambda_1 NR} \leq \eta \leq \frac{1}{101\lambda_0 N}$

With the above bounds for $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]$ and $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right]$, we have

$$\begin{aligned}\mathbb{E}\left[x^{(r+1)}\right] &= \mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]\mathbb{P}\left(x^{(r)} \geq 0\right) + \mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right]\mathbb{P}\left(x^{(r)} < 0\right)\\ &\geq \left(1 - \frac{2}{3}\lambda_0 N\eta\right)x^{(r)} + \frac{1}{120}\lambda_0 N^{\frac{3}{2}}\eta^2\sigma. && (\because \mathbb{P}\left(x^{(r)} \geq 0\right) \geq \tfrac{1}{2})\end{aligned}$$

If $x^{(r)} \geq \frac{1}{8160800} \cdot \frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R}$, then using $\eta \geq \frac{1}{102010\lambda_1 NR}$, we have

$$\begin{aligned} x^{(r+1)} &\geq \left(1 - \frac{2}{3}\lambda_0 N\eta\right) x^{(r)} + \frac{1}{120}\lambda_0 N^{\frac{3}{2}}\eta^2\sigma \\ &\geq \left(1 - \frac{2}{3}\lambda_0 N\eta\right) \cdot \frac{1}{8160800} \cdot \frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R} + \frac{1}{120}\lambda_0 N^{\frac{3}{2}}\eta\sigma \cdot \frac{1}{102010\lambda_1 NR} \\ &\geq \frac{1}{8160800} \cdot \frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R}. \end{aligned}$$

Therefore, if we set $x^{(0)} \geq \frac{1}{8160800} \cdot \frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R}$, then the final parameters will also maintain $x^{(R)} \geq \frac{1}{8160800} \cdot \frac{\sigma}{\lambda_1 N^{\frac{1}{2}} R}$. Then, noting that $\frac{\lambda_0}{\lambda} \geq 1010$, we can choose $\frac{\lambda_0}{\lambda} = 1010$. Then,

$$\mathbb{E}\left[F(x^{(R)}) - F(x^*)\right] = \mathbb{E}\left[F(x^{(R)})\right] \geq \frac{1}{2} \cdot \frac{\lambda_0}{1010}\mathbb{E}\left[\left(x^{(R)}\right)^2\right] = \Omega\left(\frac{\lambda_0\sigma^2}{\lambda_1^2 NR^2}\right).$$

C.3 Lower Bound for $\frac{1}{101\lambda N} \leq \eta < \frac{1}{\lambda}$

In this regime, we consider the following functions

$$\begin{aligned} f_{\pi_n}(x) &= \frac{1}{2}\lambda x^2 + \tau_n \sigma x, \\ f(x) &= \mathbb{E}\left[f_{\pi_n}(x)\right] = \frac{1}{2}\lambda x^2, \end{aligned}$$

where τ_n is a random variable with equal probabilities of being either "+1" or "−1".

According to the result in Cha et al. (2023)'s Appendix B.3, we have

$$\begin{aligned} x_N &= (1 - \lambda\eta)^N x_0 - \eta\sigma \sum_{n=1}^{N} (1 - \lambda\eta)^{N-n}\tau_n \\ \mathbb{E}\left[x_N^2\right] &= (1 - \lambda\eta)^{2N} x_0^2 + \eta^2\sigma^2 \mathbb{E}\left(\sum_{n=1}^{N} (1 - \lambda\eta)^{N-n}\tau_n\right)^2. \end{aligned}$$

Similar to Safran and Shamir (2020)'s Lemma 1, we have

$$\begin{aligned} \mathbb{E}\left(\sum_{n=1}^{N} (1 - \lambda\eta)^{N-n}\tau_n\right)^2 &= \sum_{n=1}^{N} (1 - \lambda\eta)^{2(N-n)}\mathbb{E}\left[\tau_n^2\right] + \sum_{n=1}^{N}\sum_{i\neq n}^{N} (1 - \lambda\eta)^{2N-n-i}\mathbb{E}\left[\tau_n\tau_i\right] \\ &= \sum_{n=1}^{N} (1 - \lambda\eta)^{2(N-n)}. \qquad (\because \mathbb{E}[\tau_n\tau_i] = \mathbb{E}[\tau_n]\mathbb{E}[\tau_i] = 0) \end{aligned}$$

Thus, we get

$$\eta^2\sigma^2 \sum_{n=1}^{N} (1 - \lambda\eta)^{2(N-n)} = \frac{1 - (1 - \lambda\eta)^{2N}}{1 - (1 - \lambda\eta)^2}\eta^2\sigma^2 \geq \frac{1 - \exp(-2\lambda\eta N)}{\lambda\eta(2 - \lambda\eta)}\eta^2\sigma^2 \geq 0.009\frac{\eta\sigma^2}{\lambda},$$

where we use $\frac{1}{101\lambda N} \leq \eta \leq \frac{1}{\lambda}$. Then,

$$\mathbb{E}\left[\left(x^{(R)}\right)^2\right] \geq (1-\lambda\eta)^{2NR}\left(x^{(0)}\right)^2 + \sum_{r=0}^{R-1}(1-\lambda\eta)^{2Nr}0.009\frac{\eta\sigma^2}{\lambda} \geq 0.009\frac{\sigma^2}{\lambda^2 N},$$
$$\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] = \frac{\lambda}{2}\mathbb{E}\left[\left(x^{(R)}\right)^2\right] = \Omega\left(\frac{\sigma^2}{\lambda N}\right).$$

Now we complete the proofs for all regimes for stochasticity terms in Theorem 5. The setups and final results are summarized in Table 5. ■

C.4 Helpful Lemmas for Stochasticity Terms

Lemma 8 *Let $\tau_1, \tau_2, \ldots, \tau_n$ be independent random variables, each with equal probabilities of being either "+1" or "−1". Let $\mathcal{E}_n \coloneqq \sum_{i=1}^n \tau_i$ (with $\mathcal{E}_0 = 0$). Then for any $n \geq 0$,*

$$\frac{\sqrt{n}}{5} \leq \mathbb{E}\left|\mathcal{E}_n\right| \leq \sqrt{n}.$$

Proof For the upper bound, similar to Rajput et al. (2020)'s Lemma 12 and Cha et al. (2023)'s Lemma B.5, we have

$$\mathbb{E}\left|\mathcal{E}_n\right| = \mathbb{E}\left[\left|\sum_{i=1}^n \tau_i\right|\right] \leq \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n \tau_i\right)^2\right]} \leq \sqrt{\sum_{i=1}^n \mathbb{E}[(\tau_i)^2] + 2\sum_{i<j\leq m-1}\mathbb{E}[\tau_i\tau_j]} = \sqrt{n},$$

where $\mathbb{E}[\tau_i\tau_j] = 0$ for $i \neq j$, due to independence.

For the lower bound, similar to Rajput et al. (2020)'s Lemma 12, we have

$$\mathbb{E}\left|\mathcal{E}_n\right| = \mathbb{E}\left|\mathcal{E}_{n-1}\right| + \mathbb{P}(\mathcal{E}_{n-1}\tau_n = 0) + \mathbb{P}(\mathcal{E}_{n-1}\tau_n > 0) - \mathbb{P}(\mathcal{E}_{n-1}\tau_n < 0).$$

It can be seen that the last two terms can be canceled out,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{n-1}\tau_n > 0) &= \mathbb{P}(\mathcal{E}_{n-1} > 0, \tau_n > 0) + \mathbb{P}(\mathcal{E}_{n-1} < 0, \tau_n < 0)\\ &= \mathbb{P}(\mathcal{E}_{n-1} > 0)\cdot\mathbb{P}(\tau_n > 0) + \mathbb{P}(\mathcal{E}_{n-1} < 0)\mathbb{P}(\tau_n < 0)\\ &= \frac{1}{2}\mathbb{P}(\mathcal{E}_{n-1} > 0) + \frac{1}{2}\mathbb{P}(\mathcal{E}_{n-1} < 0).\end{aligned}$$

Then, since $\mathbb{P}(\mathcal{E}_{n-1}\tau_n = 0) = \mathbb{P}(\mathcal{E}_{n-1} = 0) = \mathbb{1}_{n-1 \text{ is even}}\frac{\binom{n-1}{n-1/2}}{2^{n-1}}$, we have

$$\mathbb{E}\left|\mathcal{E}_n\right| = \mathbb{E}\left|\mathcal{E}_{n-1}\right| + \mathbb{1}_{n-1 \text{ is even}}\frac{\binom{n-1}{n-1/2}}{2^{n-1}}.$$

According to Cha et al. (2023)'s Lemma B.8, $\binom{n}{n/2}$ can be estimated as $2^n \cdot \frac{\sqrt{2n+\alpha_n}}{\sqrt{\pi}(n+\alpha_{n/2})}$ with $0.333 \leq \alpha_{n/2}, \alpha_n \leq 0.354$ (Mortici, 2011), so we can get $\binom{n}{n/2}/2^n \geq \frac{1}{2\sqrt{n}}$.

When $n \geq 2$ is an even integer,

$$\begin{aligned}
\mathbb{E}\,|\mathcal{E}_n| &= \mathbb{E}\,|\mathcal{E}_{n-1}| + \mathbb{1}_{n-1 \text{ is even}} \frac{\binom{n-1}{n-1/2}}{2^{n-1}} \\
&= \mathbb{E}\,|\mathcal{E}_{n-2}| + \mathbb{1}_{n-2 \text{ is even}} \frac{\binom{n-1}{n-1/2}}{2^{n-1}} \\
&\vdots \\
&= \mathbb{E}\,|\mathcal{E}_2| + \sum_{p=1}^{n/2-1} \frac{1}{2\sqrt{n-2p}} \qquad (\because \mathbb{E}\,|\mathcal{E}_2| = 1) \\
&\geq \frac{n}{2} \cdot \frac{1}{2\sqrt{n}} \geq \frac{\sqrt{n}}{5}.
\end{aligned}$$

When $n \geq 1$ is an odd integer,

$$\mathbb{E}\,|\mathcal{E}_{n+1}| = \mathbb{E}\,|\mathcal{E}_n| + \mathbb{1}_{n \text{ is even}} \frac{\binom{n}{n/2}}{2^n} \implies \mathbb{E}\,|\mathcal{E}_n| = \mathbb{E}\,|\mathcal{E}_{n+1}| \geq \frac{\sqrt{n}}{5}.$$

Now we complete the proof of the bounds of $\mathbb{E}\,|\mathcal{E}_n|$ for any $n \geq 0$. ■

Lemma 9 *Let $\tau_1, \tau_2, \ldots, \tau_n$ be independent random variables, each with equal probabilities of being either "+1" or "−1". Let $\mathcal{E}_n \coloneqq \sum_{i=1}^n \tau_i$ (with $\mathcal{E}_0 = 0$). Then, for any $n \geq 0$, the probability distribution of $\mathcal{E}_n$ is symmetric with respect 0, and for any $n \geq 1$,*

$$\frac{1}{4} \leq \mathbb{P}(\mathcal{E}_n > 0) = \mathbb{P}(\mathcal{E}_n < 0) \leq \frac{1}{2}.$$

Proof When $n = 0$, $\mathbb{P}(\mathcal{E}_n > 0) = \mathbb{P}(\mathcal{E}_n < 0) = 0$. The distribution is symmetric trivially. When $n > 1$, $\mathbb{P}(\mathcal{E}_n = p) = \begin{cases} \frac{\binom{n}{(n+p)/2}}{2^n} = \frac{\binom{n}{(n-p)/2}}{2^n}, & n+p \mod 2 = 0 \\ 0, & n+p \mod 2 \neq 0 \end{cases}$ where the integer p satisfies $-n \leq p \leq n$. Thus, the distribution is symmetric with respect to 0.

When any integer $n > 1$, we have $\mathbb{P}(\mathcal{E}_n = 0) = \mathbb{1}_{n \text{ is even}} \frac{\binom{n}{n/2}}{2^n}$. Letting $g(n) = \frac{\binom{n}{n/2}}{2^n}$, it can be validated that $\frac{g(n+2)}{g(n)} = \frac{n+1}{n+2} < 1$, so $\mathbb{P}(\mathcal{E}_n = 0) = g(n) \leq g(n-2) \leq \cdots \leq g(2) = \frac{1}{2}$ when n is even. Therefore, $\frac{1}{2} \geq \mathbb{P}(\mathcal{E}_n > 0) = \mathbb{P}(\mathcal{E}_n < 0) = \frac{1}{2}(1 - \mathbb{P}(\mathcal{E}_n = 0)) \geq \frac{1}{4}$. ■

Lemma 10 *Suppose that $x_0 \geq 0$, $\frac{\lambda_0}{\lambda} \geq 1010$ and $\eta \leq \frac{1}{101\lambda N}$. Then for $0 \leq n \leq N$,*

$$\mathbb{E}\,|x_n - x_0| \leq \frac{1}{100} x_0 + \frac{101}{100} \eta\sigma\sqrt{n}$$

Proof With result of Lemma 8, the proof is identical to Cha et al. (2023)'s Lemma B.7, except the numerical factors, to be consistent with the constants used in Appendix D. ■

Appendix D. Proof of Heterogeneity Terms in Theorem 5

Proof We assume $F_m = f_m(\cdot;\xi_m^1) = \cdots = f_m(\cdot;\xi_m^{|\mathcal{D}_m|})$ for each m, and then extend the works of SGD-RR to SFL, from performing one update step to performing multiple update steps on each local objective function. The results are summarized in Table 6. Notably, λ_0, λ_1 and λ in different regimes can be different.

Bound	Regime	Objective functions	Initialization	$\nabla^2 F_m \in$
$\Omega\left(\frac{\zeta^2}{\lambda}\right)$	$\eta \le \frac{1}{102010\lambda MKR}$	$F_m(x) = \lambda x^2$	$x^{(0)} = \frac{\zeta}{\lambda}$	$[2\lambda, 2\lambda]$
$\Omega\left(\frac{\lambda_2\zeta^2}{\lambda_1^2 MR^2}\right)$	$\frac{1}{102010\lambda_1 MKR} \le \eta$ and $\eta \le \frac{1}{101\lambda_0 MK}$	$F_m(x) = \begin{cases}(\lambda_0\mathbb{1}_{x<0} + \lambda\mathbb{1}_{x\ge 0})\frac{x^2}{2} + \zeta x, & \text{if } m \le \frac{M}{2}\\ (\lambda_0\mathbb{1}_{x<0} + \lambda\mathbb{1}_{x\ge 0})\frac{x^2}{2} - \zeta x, & \text{otherwise}\end{cases}$	$x^{(0)} = \frac{1}{81608000}\frac{\zeta}{\lambda_1 M^{\frac{1}{2}} R}$	$\left[\frac{\lambda_0}{1010}, \lambda_0\right]$
$\Omega\left(\frac{\zeta^2}{\lambda MK}\right)$	$\frac{1}{101\lambda MK} \le \eta \le \frac{1}{\lambda K}$	$F_m(x) = \begin{cases}\frac{\lambda}{2}x^2 + \zeta x, & \text{if } m \le \frac{M}{2}\\ \frac{\lambda}{2}x^2 - \zeta x, & \text{otherwise}\end{cases}$	$x^{(0)} = 0$	$[\lambda, \lambda]$
$\Omega\left(\frac{\zeta^2}{\lambda}\right)$	$\frac{1}{\lambda K} \le \eta \le \frac{1}{\lambda}$	$F_m(x) = \begin{cases}\frac{\lambda}{2}x^2 + \zeta x, & \text{if } m \le \frac{M}{2}\\ \frac{\lambda}{2}x^2 - \zeta x, & \text{otherwise}\end{cases}$	$x^{(0)} = 0$	$[\lambda, \lambda]$
$\Omega\left(\frac{\zeta^2}{\lambda}\right)$	$\eta \ge \frac{1}{\lambda}$	$F_m(x) = \lambda x^2$	$x^{(0)} = \frac{\zeta}{\lambda}$	$[2\lambda, 2\lambda]$

Table 6: Lower bounds of SFL for heterogeneity terms. It requires that $\lambda_0 = 1010\lambda$ and $R \ge \frac{1}{1010}\frac{\lambda_0}{\lambda_1}$ in the regime $\frac{1}{102010\lambda_1 MKR} \le \eta \le \frac{1}{101\lambda_0 MK}$.

D.1 Lower Bounds for $0 < \eta \le \frac{1}{102010\lambda MKR}$ and $\eta \ge \frac{1}{\lambda}$

In this regime, we consider the following objective functions

$$F_m(x) = \lambda x^2,$$
$$F(x) = \frac{1}{M}\sum_{m=1}^{M} F_m(x) = \lambda x^2.$$

We can soon build the relationship between $x^{(R)}$ and $x^{(0)}$: $x^{(R)} = (1-2\lambda\eta)^{MKR}x^{(0)}$.

D.1.1 Lower Bound for $0 < \eta \le \frac{1}{102010\lambda MKR}$

Since $\eta \le \frac{1}{102010\lambda MKR}$ and $\left(1 - \frac{1}{51005}\cdot\frac{1}{x}\right)^x$ is monotonically increasing when $x \ge 1$, we have

$$x^{(R)} = (1-2\lambda\eta)^{MKR}x^{(0)} \ge \left(1 - \frac{1}{51005MKR}\right)^{MKR} x^{(0)} \ge \frac{51004}{51005}x^{(0)}, \qquad (\because M \ge 1)$$
$$F(x^{(R)}) = \lambda\left(x^{(R)}\right)^2 \ge \frac{51004^2}{51005^2}\lambda\left(x^{(0)}\right)^2.$$

If $x^{(0)} = \frac{\zeta}{\lambda}$, we can get

$$\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] \ge \frac{51004^2}{51005^2}\lambda\left(x^{(0)}\right)^2 = \frac{51004^2}{51005^2}\frac{\zeta^2}{\lambda} = \Omega\left(\frac{\zeta^2}{\lambda}\right).$$

D.1.2 Lower Bound for $\eta \geq \frac{1}{\lambda}$

Since $\eta \geq \frac{1}{\lambda}$ implies that $(1-2\lambda\eta)^2 \geq 1$, we have

$$F(x^{(R)}) = \lambda \left(x^{(R)}\right)^2 = \lambda(1-2\lambda\eta)^{2MKR} \left(x^{(0)}\right)^2 \geq \lambda \left(x^{(0)}\right)^2.$$

If $x^{(0)} = \frac{\zeta}{\lambda}$, we can get

$$\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] \geq \lambda \left(x^{(0)}\right)^2 = \frac{\zeta^2}{\lambda} = \Omega\left(\frac{\zeta^2}{\lambda}\right).$$

D.2 Lower Bound for $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$

In this regime, we consider the following functions

$$F_m(x) = \begin{cases} (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} + \zeta x, & \text{if } m \leq \frac{M}{2} \\ (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} - \zeta x, & \text{otherwise} \end{cases},$$

$$F(x) = \frac{1}{M}\sum_{m=1}^{M} F_m(x) = (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} \quad (\lambda_0/\lambda \geq 1010).$$

Next, we focus on a single round r, and thus we drop the superscripts r for a while. Unless otherwise stated, the expectation is conditioned on $x_{1,0}$ when considering one single round. The proofs in this regime have a similar structure as Cha et al. (2023)'s Theorem 3.1.

In each training round, we sample a random permutation $\pi = (\pi_1, \pi_2, \ldots, \pi_M)$ (π_m is its m-th element) from $\{1, 2, \ldots, M\}$ as the clients' training order. Then, we can denote F_{π_m} for $m \in \{1, 2, \ldots, M\}$ as $F_{\pi_m}(x) = (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})\frac{x^2}{2} + \zeta\tau_m x$, where $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$ (τ_m is its m-th element) is a random permutation of $\frac{M}{2}$ $+1$'s and $\frac{M}{2}$ -1's. For example, assuming that $\pi = (4, 2, 3, 1)$ (with $M = 4$), we can get the corresponding coefficients $\tau = (-1, +1, -1, +1)$.

Then, the relationship between $x_{m,k}$ and $x_{1,0}$ satisfies

$$\begin{aligned} x_{m,k} = x_{1,0} &- \eta \sum_{i=1}^{m}\sum_{j=0}^{K-1} \left((\lambda_0 \mathbb{1}_{x_{i,j}<0} + \lambda \mathbb{1}_{x_{i,j}\geq 0}) x_{i,j}\right) - K\zeta \sum_{i=1}^{m-1} \tau_i \\ &- \eta \sum_{j=0}^{k-1} \left((\lambda_0 \mathbb{1}_{x_{i,j}<0} + \lambda \mathbb{1}_{x_{i,j}\geq 0}) x_{m,j}\right) - k\zeta\tau_m. \end{aligned}$$

For convenience, we write it as

$$x_{m,k} = x_{1,0} - \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \left((\lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)}<0} + \lambda \mathbb{1}_{x_{b_1(i),b_2(i)}\geq 0}) x_{b_1(i),b_2(i)}\right) - K\eta\zeta\mathcal{A}_{m,k}, \tag{3}$$

where $b_1(i) \coloneqq \lfloor \frac{i}{K} \rfloor + 1$, $b_2(i) \coloneqq i - K\lfloor \frac{i}{K} \rfloor$, $\mathcal{A}_{m,k} \coloneqq \mathcal{E}_{m-1} + a_k\tau_m = \tau_1 + \tau_2 \cdots + a_k\tau_m$ ($\mathcal{E}_m \coloneqq \tau_1 + \tau_2 \cdots + \tau_m$ and $a_k \coloneqq k/K$) and $\mathcal{B}_{m,k} = \sum_{i=0}^{(m-1)K+k-1} 1 = (m-1)K + k$. In

particular, when $m = M+1$ and $k = 0$, it follows that

$$x_{M+1,0} = x_{1,0} - \eta \sum_{m=1}^{M} \sum_{k=0}^{K-1} \left((\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k} \right). \tag{4}$$

Notably, the following notations $x_{m+1,0}$ and $x_{m,K}$ are equivalent. In this paper, we use $x_{m+1,0}$ instead of $x_{m,K}$.

D.2.1 Lower Bound of $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]$.

We first give a stricter upper bound of $\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k}\right]$ for $1 \leq m \leq \frac{M}{2}+1$, and then give a general upper bound for $1 \leq m \leq M$. These two upper bounds are then plugged into Eq. (4), yielding the targeted lower bound of $\mathbb{E}\left[x_{M+1,0} \mid x_{1,0} \geq 0\right]$.

Since $(\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})x \leq \lambda_0 x$ and $(\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0})x \leq \lambda x$, we have

$$\begin{aligned}
&\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k}\right] \\
&= \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k} \mid \mathcal{A}_{m,k} > 0\right] \mathbb{P}(\mathcal{A}_{m,k} > 0) \\
&\quad + \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k} \mid \mathcal{A}_{m,k} \leq 0\right] \mathbb{P}(\mathcal{A}_{m,k} \leq 0) \\
&\leq \lambda_0 \mathbb{E}\left[x_{m,k} \mid \mathcal{A}_{m,k} > 0\right] \cdot \mathbb{P}(\mathcal{A}_{m,k} > 0) + \lambda \mathbb{E}\left[x_{m,k} \mid \mathcal{A}_{m,k} \leq 0\right] \cdot \mathbb{P}(\mathcal{A}_{m,k} \leq 0).
\end{aligned} \tag{5}$$

Intuitively, this is a trick, using λ_0 for the former term and λ for the latter term, so we can make the former term (which has a stricter bound) dominant by controlling the value of λ_0/λ. Next, we bound the terms on the right hand side in Ineq. (5). For the first term in Ineq. (5), according to Ineq. (3), we have

$$\begin{aligned}
&\mathbb{E}\left[x_{m,k} \mid \mathcal{A}_{m,k} > 0\right] \\
&= \mathbb{E}\left[x_{1,0} - \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \left((\lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)}<0} + \lambda \mathbb{1}_{x_{b_1(i),b_2(i)}\geq 0}) x_{b_1(i),b_2(i)} \right) - K\eta\zeta \mathcal{A}_{m,k} \mid \mathcal{A}_{m,k} > 0 \right] \\
&= x_{1,0} + \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)}<0} + \lambda \mathbb{1}_{x_{b_1(i),b_2(i)}\geq 0}) \left| x_{b_1(i),b_2(i)} - x_{1,0} \right| \mid \mathcal{A}_{m,k} > 0 \right] \\
&\quad - \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)}<0} + \lambda \mathbb{1}_{x_{b_1(i),b_2(i)}\geq 0}) x_{1,0} \mid \mathcal{A}_{m,k} > 0 \right] - K\eta\zeta \mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k} > 0\right] \\
&\leq x_{1,0} + \lambda_0 \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left| x_{b_1(i),b_2(i)} - x_{1,0} \right| \mid \mathcal{A}_{m,k} > 0 \right] - K\eta\zeta \mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k} > 0\right],
\end{aligned}$$

where we use $\lambda \leq (\lambda_0 \mathbb{1}_{x<0} + \lambda \mathbb{1}_{x\geq 0}) \leq \lambda_0$ and $x_{1,0} \geq 0$ in the last inequality. Then, we have

$$\begin{aligned}
\text{Term}_1 \text{ in } (5) \leq \lambda_0 x_{1,0} \mathbb{P}(\mathcal{A}_{m,k} > 0) &+ \lambda_0^2 \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left| x_{b_1(i),b_2(i)} - x_{1,0} \right| \mid \mathcal{A}_{m,k} > 0 \right] \mathbb{P}(\mathcal{A}_{m,k} > 0) \\
&- \lambda_0 K\eta\zeta \mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k} > 0\right] \mathbb{P}(\mathcal{A}_{m,k} > 0).
\end{aligned}$$

Since $\mathbb{E}\left[|x_{i,j} - x_{1,0}| \mid \mathcal{A}_{m,k} > 0\right] \mathbb{P}(\mathcal{A}_{m,k} > 0)$ and $\mathbb{E}\left[|x_{i,j} - x_{1,0}| \mid \mathcal{A}_{m,k} \leq 0\right] \mathbb{P}(\mathcal{A}_{m,k} \leq 0)$ are non-negative, the term $\mathbb{E}\left[|x_{i,j} - x_{1,0}| \mid \mathcal{A}_{m,k} > 0\right] \mathbb{P}(\mathcal{A}_{m,k} > 0)$ appearing in the second term

in the preceding inequality can be bounded by $\mathbb{E}\left[|x_{i,j}-x_{1,0}|\right]$ for any integers i,j. Since the probability distribution of $\mathcal{A}_{m,k}$ is symmetric with respect to 0 (Lemma 13), we get

$$\begin{aligned}\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] &= \mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k}>0\right]\mathbb{P}(\mathcal{A}_{m,k}>0)+\mathbb{E}\left[-\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k}<0\right]\mathbb{P}(\mathcal{A}_{m,k}<0)\\ &= \mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k}>0\right]\mathbb{P}(\mathcal{A}_{m,k}>0)+\mathbb{E}\left[-\mathcal{A}_{m,k} \mid -\mathcal{A}_{m,k}>0\right]\mathbb{P}(-\mathcal{A}_{m,k}>0)\\ &= 2\mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k}>0\right]\mathbb{P}(\mathcal{A}_{m,k}>0). \qquad \text{(by symmetry)}\end{aligned}$$

After using $\mathbb{P}(\mathcal{A}_{m,k}>0)\le\frac{1}{2}$ by symmetry, $\mathbb{E}\left[|x_{i,j}-x_{1,0}| \mid \mathcal{A}_{m,k}>0\right]\mathbb{P}(\mathcal{A}_{m,k}>0)\le\mathbb{E}\left[|x_{i,j}-x_{1,0}|\right]$ and $\mathbb{E}\left[|\mathcal{A}_{m,k}|\right]=2\mathbb{E}\left[\mathcal{A}_{m,k} \mid \mathcal{A}_{m,k}>0\right]\mathbb{P}(\mathcal{A}_{m,k}>0)$ for the first, the second and the last terms on the right hand side, respectively, we have

$$\text{Term}_1 \text{ in } (5) \le \frac{1}{2}\lambda_0 x_{1,0}+\lambda_0^2\eta\sum_{i=0}^{\mathcal{B}_{m,k}-1}\mathbb{E}\left[\left|x_{b_1(i),b_2(i)}-x_{1,0}\right|\right]-\frac{1}{2}\lambda_0 K\eta\zeta\mathbb{E}\left[|\mathcal{A}_{m,k}|\right]. \tag{6}$$

For the second term on the right hand side in Ineq. (5), we have

$$\begin{aligned}\text{Term}_2 \text{ in } (5) &\le \lambda\mathbb{E}\left[|x_{m,k}-x_{1,0}| \mid \mathcal{A}_{m,k}\le 0\right]\mathbb{P}(\mathcal{A}_{m,k}\le 0)+\lambda\mathbb{E}\left[x_{1,0} \mid \mathcal{A}_{m,k}\le 0\right]\mathbb{P}(\mathcal{A}_{m,k}\le 0)\\ &\le \lambda\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right]+\frac{5}{6}\lambda x_{1,0},\end{aligned} \tag{7}$$

where we use $\mathbb{P}(\mathcal{A}_{m,k}<0)=\mathbb{P}(\mathcal{A}_{m,k}>0)\ge\frac{1}{6}$ (Lemma 13) for the second term on the right hand side in the last inequality. Plugging Ineq. (6) and Ineq. (7) into Ineq. (5), we have

$$\begin{aligned}\mathbb{E}\left[(\lambda_0\mathbb{1}_{x_{m,k}<0}+\lambda\mathbb{1}_{x_{m,k}\ge 0})x_{m,k}\right] \le{}& \frac{1}{2}\lambda_0 x_{1,0}+\lambda_0^2\eta\sum_{i=0}^{\mathcal{B}_{m,k}-1}\mathbb{E}\left[\left|x_{b_1(i),b_2(i)}-x_{1,0}\right|\right]\\ &-\frac{1}{2}\lambda_0 K\eta\zeta\mathbb{E}\left[|\mathcal{A}_{m,k}|\right]+\lambda\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right]+\frac{5}{6}\lambda x_{1,0}.\end{aligned}$$

Using $\mathbb{E}\left[|x_{i,j}-x_{1,0}|\right]\le\frac{1}{100}x_{1,0}+\frac{101}{100}K\eta\zeta\sqrt{m-1+a_k^2}$ for $(i-1)K+j\le(m-1)K+k$ (Lemma 14), $\mathbb{E}\left[|\mathcal{A}_{m,k}|\right]\ge\frac{1}{20}\sqrt{m-1+a_k^2}$ (Lemma 12) and $\lambda_0/\lambda_0\ge 1010$ and $\lambda_0\eta\mathcal{B}_{m,k}\le\lambda_0 MK\eta\le\frac{1}{101}$, we can simplify it as

$$\mathbb{E}\left[(\lambda_0\mathbb{1}_{x_{m,k}<0}+\lambda\mathbb{1}_{x_{m,k}\ge 0})x_{m,k}\right]\le\frac{501}{1000}\lambda_0 x_{1,0}+\frac{14}{1000}\lambda_0 K\eta\zeta\sqrt{m-1+a_k^2}. \tag{8}$$

Ineq. (8) holds for $1\le m\le\frac{M}{2}+1$ ($M\ge 4$) and $0\le k\le K-1$, due to the constraints of Lemmas 12, 13 and 14. Even though the constraint of Lemma 13 excludes the case $m=1,k=0$, we can verify $\mathbb{E}\left[(\lambda_0\mathbb{1}_{x_{1,0}<0}+\lambda\mathbb{1}_{x_{1,0}\ge 0})x_{1,0}\right]=\lambda x_{1,0}\le\frac{1}{1010}\lambda_0 x_{1,0}$ ($x_{1,0}\ge 0$).

Next, we give a general upper bound of $\mathbb{E}\left[(\lambda_0\mathbb{1}_{x_{m,k}<0}+\lambda\mathbb{1}_{x_{m,k}\ge 0})x_{m,k}\right]$ for $1\le m\le M$.

$$\begin{aligned}\mathbb{E}\left[(\lambda_0\mathbb{1}_{x_{m,k}<0}+\lambda\mathbb{1}_{x_{m,k}\ge 0})x_{m,k}\right] &\le \mathbb{E}\left[\lambda x_{m,k}\right] \quad (\because (\lambda_0\mathbb{1}_{x<0}+\lambda\mathbb{1}_{x\ge 0})x\le\lambda x \text{ for all } x\in\mathbb{R})\\ &\le \lambda\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right]+\lambda x_{1,0}\\ &\le \frac{1}{1000}\lambda_0 x_{1,0}+\frac{1}{1000}\lambda_0 K\eta\zeta\sqrt{m-1+a_k^2},\end{aligned} \tag{9}$$

where we use Lemma 14 and $\lambda_0/\lambda \geq 1010$ in the last inequality. Notably, this inequality holds for $1 \leq m \leq M$ and $0 \leq k \leq K-1$ ($\because$ Lemma 14).

Recalling Eq. (4), we first separate the sum of $\mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k}\right]$ into two parts, and then use the tighter bound, Ineq. (8), for $1 \leq m \leq \frac{M}{2}+1$ (the first part) and the general bound, Ineq. (9), for $\frac{M}{2}+2 \leq m \leq M$ (the second part).

$$\begin{aligned}
\mathbb{E}\left[x_{M+1,0} - x_{1,0}\right] &= -\eta \sum_{m=1}^{\frac{M}{2}+1} \sum_{k=0}^{K-1} \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k}\right] \\
&\quad - \eta \sum_{m=\frac{M}{2}+2}^{M} \sum_{k=0}^{K-1} \mathbb{E}\left[(\lambda_0 \mathbb{1}_{x_{m,k}<0} + \lambda \mathbb{1}_{x_{m,k}\geq 0}) x_{m,k}\right] \\
&\geq -\eta \sum_{m=1}^{\frac{M}{2}+1} \sum_{k=0}^{K-1} \left(\frac{501}{1000} \lambda_0 x_{1,0} - \frac{14}{1000} \lambda_0 K \eta \zeta \sqrt{m-1+a_k^2} \right) \\
&\quad - \eta \sum_{m=\frac{M}{2}+2}^{M} \sum_{k=0}^{K-1} \left(\frac{1}{1000} \lambda_0 x_{1,0} + \frac{1}{1000} \lambda_0 K \eta \zeta \sqrt{m-1+a_k^2} \right).
\end{aligned}$$

Then, after simplifying the preceding inequality, we get

$$\mathbb{E}\left[x_{M+1,0}\right] \geq \left(1 - \frac{2}{3}\lambda_0 M K \eta\right) x_{1,0} + \frac{1}{600} \lambda_0 M^{\frac{3}{2}} K^2 \eta^2 \zeta.$$

Taking unconditional expectations and putting back the superscripts, we can get

$$\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right] \geq \left(1 - \frac{2}{3}\lambda_0 M K \eta\right) \mathbb{E}\left[x^{(r)} \mid x^{(r)} \geq 0\right] + \frac{1}{600} \lambda_0 M^{\frac{3}{2}} K^2 \eta^2 \zeta. \tag{10}$$

where we note that the following notations are interchangeable: $x_{1,0}^{(r)}$ and $x^{(r)}$, $x_{M+1,0}^{(r)}$ and $x^{(r+1)}$. Notably, Ineq. (10) holds for $M \geq 4$, because of the constraints of Ineq. (8).

D.2.2 Lower Bound of $\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right]$.

We introduce a new function $H(x)$ (see Section D.3 for details) as follows:

$$H_m(x) = \begin{cases} \frac{\lambda_0}{2} x^2 + \zeta x, & \text{if } m \leq \frac{M}{2} \\ \frac{\lambda_0}{2} x^2 - \zeta x, & \text{otherwise} \end{cases}$$

$$H(x) = \frac{1}{M} \sum_{m=1}^{M} H_m(x) = \frac{\lambda_0}{2} x^2.$$

Let Algorithm 1 run on the two functions $F(x)$ and $H(x)$, where both algorithms start from the same initial point and share the same random variables $\{\tau_{m,k}^{(r)}\}_{k,m,r}$ for all training rounds. Then, according to Part 1 of Cha et al. (2023)'s Lemma B.4, the model parameters generated on $F(x)$ and $H(x)$ satisfy $(x_{m,k})_F \geq (x_{m,k})_H$. Here, we let both cases share the same initial point $x_{1,0}$ and the same random variables π (accordingly, the same τ).

The relationship between $(x_{M+1,0})_H$ and $x_{1,0}$ satisfies:

$$\begin{aligned}\mathbb{E}\left[(x_{M+1,0})_H\right] &= \mathbb{E}\left[(1-\lambda_0\eta)^{MK}x_{1,0} - \eta\zeta\sum_{m=0}^{M-1}(1-\lambda_0\eta)^{mK}\tau_{M-m}\sum_{k=0}^{K-1}(1-\lambda_0\eta)^k\right] \\ &= (1-\lambda_0\eta)^{MK}x_{1,0} - \eta\zeta\sum_{m=0}^{M-1}(1-\lambda_0\eta)^{mK}\mathbb{E}\left[\tau_{M-m}\right]\sum_{k=0}^{K-1}(1-\lambda_0\eta)^k \\ &= (1-\lambda_0\eta)^{MK}x_{1,0}. \qquad (\because \mathbb{E}[\tau_{M-m}] = 0)\end{aligned}$$

Then since $(1-z)^K \le 1 - Kz + K^2z^2$, $\forall x \in [0,1]$, we can get

$$(1-\lambda_0\eta)^{MK} \le 1 - \lambda_0 MK\eta + \lambda_0^2M^2K^2\eta^2 \le 1 - \frac{2}{3}\lambda_0 MK\eta. \qquad (\because \lambda_0 MK\eta \le \tfrac{1}{101})$$

Then using $(x_{m,k})_F \ge (x_{m,k})_H$ and $x_{1,0} < 0$, we have

$$\mathbb{E}\left[(x_{M+1,0})_F\right] \ge \mathbb{E}\left[(x_{M+1,0})_H\right] = (1-\lambda_0\eta)^{MK}x_{1,0} \ge \left(1 - \frac{2}{3}\lambda_0 MK\eta\right)x_{1,0}.$$

Taking unconditional expectations and putting back the superscripts, we can get

$$\mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right] \ge \left(1 - \frac{2}{3}\lambda_0 MK\eta\right)\mathbb{E}\left[x^{(r)} \mid x^{(r)} < 0\right]. \tag{11}$$

D.2.3 Relationship Between $\mathbb{P}(x^{(r)} \ge 0)$ and $\mathbb{P}(x^{(r)}) < 0$.

We still use the function $H(x)$ for comparison. This time, we let both cases share the same initial point $x^{(0)}$ and the same $\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(r-1)}$ for the first r rounds. For any round r, we can build the relationship of $(x^{(r)})_H$ with x^0:

$$(x^{(r)})_H = (1-\lambda_0\eta)^{rMK}x^{(0)} - \eta\zeta\sum_{k=0}^{K-1}(1-\lambda_0\eta)^k\sum_{m=0}^{M-1}(1-\lambda_0\eta)^{mK}\sum_{s=0}^{r-1}(1-\lambda_0\eta)^{sMK}\tau_{M-m}^{(r-1-s)}.$$

For all possible permutations $\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(r-1)}$, we can find the corresponding permutations $(\tau^{(0)})', (\tau^{(1)})', \ldots, (\tau^{(r-1)})'$ satisfy $\tau_m^{(s)} = -(\tau_m^{(s)})'$ for all $m \in \{1, 2, \ldots, M\}$ and $s \in \{0, 1, \ldots, r-1\}$. Denoting the parameters obtained with $\tau^{(0)}, \tau^{(1)}, \ldots, \tau^{(r-1)}$ and $(\tau^{(0)})', (\tau^{(1)})', \ldots, (\tau^{(r-1)})'$ as $(x^{(r)})_H$ and $(x^{(r)})'_H$, respectively, we have

$$\frac{1}{2}\left((x^{(r)})_H + (x^{(r)})'_H\right) = (1-\lambda_0\eta)^{rMK}x^{(0)}.$$

where the second terms of $(x^{(r)})_H$ and $(x^{(r)})'_H$ are canceled out with $\frac{\tau_{M-m}^{(r-1-s)} + (\tau_{M-m}^{(r-1-s)})'}{2} = 0$. This means that for any possible parameter $(x^{(r)})_H$, there exits one corresponding parameter $(x^{(r)})'_H$ to make their average be $(1-\lambda_0\eta)^{rMK}x^{(0)}$, and further implies

$$\mathbb{P}\left((x^{(r)})_H \ge (1-\lambda_0\eta)^{rMK}x^{(0)}\right) \ge \frac{1}{2}.$$

Then, for the same initial point $x^{(0)} \geq 0$, we have

$$\mathbb{P}\left((x^{(r)})_F \geq 0\right) \geq \mathbb{P}\left((x^{(r)})_H \geq 0\right) \geq \mathbb{P}\left((x^{(r)})_H \geq (1-\lambda_0\eta)^{rMK}x^{(0)}\right) \geq \frac{1}{2}.$$

Intuitively, the total possible number of events (the permutations) are identical for both F and H. Since $(x^{(r)})_F \geq (x^{(r)})_H$ for the same permutations, the permutations that make $(x^{(r)})_H \geq 0$ always make $(x^{(r)})_F \geq 0$, causing $\mathbb{P}\left((x^{(r)})_F \geq 0\right) \geq \mathbb{P}\left((x^{(r)})_H \geq 0\right)$. The reasoning for the second inequality is similar. The permutations that make $(x^{(r)})_H \geq (1-\lambda_0\eta)^{rMK}x^{(0)}$ always make $(x^{(r)})_H \geq 0$ for $x^{(0)} \geq 0$.

D.2.4 Lower Bound for $\frac{1}{102010\lambda_1 MKR} \leq \eta \leq \frac{1}{101\lambda_0 MK}$

Using Ineq. (10), Ineq. (11) and $\mathbb{P}\left(x^{(r)} \geq 0\right) \geq \frac{1}{2}$ (when $x^{(0)} \geq 0$), we have

$$\begin{aligned}
\mathbb{E}\left[x^{(r+1)}\right] &= \mathbb{E}\left[x^{(r+1)} \mid x^{(r)} \geq 0\right]\mathbb{P}\left(x^{(r)} \geq 0\right) + \mathbb{E}\left[x^{(r+1)} \mid x^{(r)} < 0\right]\mathbb{P}\left(x^{(r)} < 0\right) \\
&\geq \left(\left(1 - \frac{2}{3}\lambda_0 MK\eta\right)x^{(r)} + \frac{1}{600}\lambda_0 M^{\frac{3}{2}}K^2\eta^2\zeta\right)\mathbb{P}\left(x^{(r)} \geq 0\right) \\
&\quad + \left(\left(1 - \frac{2}{3}\lambda_0 MK\eta\right)x^{(r)}\right)\mathbb{P}\left(x^{(r)} < 0\right) \\
&\geq \left(1 - \frac{2}{3}\lambda_0 MK\eta\right)x^{(r)} + \frac{1}{1200}\lambda_0 M^{\frac{3}{2}}K^2\eta^2\zeta. \qquad (\because \mathbb{P}\left(x^{(r)} \geq 0\right) \geq \tfrac{1}{2})
\end{aligned}$$

If $x^{(r)} \geq \frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R}$, then using $\eta \geq \frac{1}{102010\lambda_1 MKR}$, we have

$$\begin{aligned}
x^{(r+1)} &\geq \left(1 - \frac{2}{3}\lambda_0 MK\eta\right)x^{(r)} + \frac{1}{1200}\lambda_0 M^{\frac{3}{2}}K^2\eta^2\zeta \\
&\geq \left(1 - \frac{2}{3}\lambda_0 MK\eta\right) \cdot \frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R} + \frac{1}{1200}\lambda_0 M^{\frac{3}{2}}K^2\eta\zeta \cdot \frac{1}{102010\lambda_1 MKR} \\
&\geq \frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R}.
\end{aligned}$$

Therefore, if we set $x^{(0)} \geq \frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R}$, then the final parameters will also maintain $x^{(R)} \geq \frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R}$. Then, noting that $\frac{\lambda_0}{\lambda} \geq 1010$, we can choose $\frac{\lambda_0}{\lambda} = 1010$. Then,

$$\begin{aligned}
\mathbb{E}\left[F(x^R) - F(x^*)\right] = \mathbb{E}\left[F(x^R)\right] &\geq \frac{1}{2} \cdot \lambda\mathbb{E}\left[\left(x^R\right)^2\right] \\
&\geq \frac{1}{2} \cdot \frac{\lambda_0}{1010} \cdot \left(\frac{1}{81608000} \cdot \frac{\zeta}{\lambda_1 M^{\frac{1}{2}}R}\right)^2 \\
&= \Omega\left(\frac{\lambda_0\zeta^2}{\lambda_1^2 MR^2}\right).
\end{aligned}$$

Notably, this inequality holds for $M \geq 4$ (see Ineq. 10).

D.3 Lower Bound for $\frac{1}{101\lambda MK} \leq \eta \leq \frac{1}{\lambda K}$ and $\frac{1}{\lambda K} \leq \eta \leq \frac{1}{\lambda}$

In these two regimes, we consider the following functions

$$F_m(x) = \begin{cases} \frac{\lambda}{2}x^2 + \zeta x, & \text{if } m \leq \frac{M}{2} \\ \frac{\lambda}{2}x^2 - \zeta x, & \text{otherwise} \end{cases},$$

$$F(x) = \frac{1}{M}\sum_{m=1}^{M} F_m(x) = \frac{\lambda}{2}x^2.$$

In each training round, we sample a random permutation $\pi = (\pi_1, \pi_2, \ldots, \pi_M)$ (π_m is its m-th element) from $\{1, 2, \ldots, M\}$ as the clients' training order. Thus, we can denote F_{π_m} for $m \in \{1, 2, \ldots, M\}$ as $F_{\pi_m}(x) = \frac{\lambda}{2}x^2 + \zeta\tau_m x$, where $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$ is a random permutation of $\frac{M}{2}$ $+1$'s and $\frac{M}{2}$ -1's.

For a single training round, we can get

$$x^{(r)} = (1-\lambda\eta)^{MK}x^{(r-1)} - \eta\zeta\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\tau_{M-m}\sum_{k=0}^{K-1}(1-\lambda\eta)^k.$$

Taking expectation conditional on $x^{(r-1)}$, we can get

$$\begin{aligned}\mathbb{E}\left(x^{(r)}\right)^2 &= \mathbb{E}\left[\left((1-\lambda\eta)^{MK}x^{(r-1)} - \eta\zeta\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\tau_{M-m}\sum_{k=0}^{K-1}(1-\lambda\eta)^k\right)^2\right] \\ &= (1-\lambda\eta)^{2MK}\left(x^{(r-1)}\right)^2 + \eta^2\zeta^2\left(\sum_{k=0}^{K-1}(1-\lambda\eta)^k\right)^2\mathbb{E}\left[\left(\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\tau_m\right)^2\right],\end{aligned}$$

where we note that the cross terms on the right hand side equal zero since $\mathbb{E}[\tau_{M-m}] = 0$. Following Safran and Shamir (2020)'s Lemma 1, we first focus on the term

$$\mathbb{E}\left[\left(\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\tau_m\right)^2\right] = \sum_{m=0}^{M-1}(1-\lambda\eta)^{2mK}\mathbb{E}\left[\tau_m^2\right] + \sum_{i=0}^{M-1}\sum_{j\neq i}^{M-1}(1-\lambda\eta)^{(i+j)K}\mathbb{E}\left[\langle\tau_i, \tau_j\rangle\right].$$

Since $\tau_m^2 = 1$ and $\mathbb{E}\left[\langle\tau_i, \tau_j\rangle\right] = -\frac{1}{M-1}$ (see Safran and Shamir (2020)'s Lemma 2), we get

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\tau_m\right)^2\right] &= \sum_{m=0}^{M-1}(1-\lambda\eta)^{2mK} - \frac{1}{M-1}\sum_{i=0}^{M-1}\sum_{j\neq i}^{M-1}(1-\lambda\eta)^{(i+j)K} \\ &= \left(1+\frac{1}{M-1}\right)\sum_{m=0}^{M-1}(1-\lambda\eta)^{2mK} - \frac{1}{M-1}\left(\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK}\right)^2.\end{aligned}$$

Returning back to $\mathbb{E}\left[\left(x^{(r)}\right)^2\right]$ and using $\sum_{m=0}^{M-1}(1-\lambda\eta)^{mK} = \frac{1-(1-\lambda\eta)^{MK}}{1-(1-\lambda\eta)^K}$ and $\sum_{m=0}^{M-1}(1-\lambda\eta)^{2mk} = \frac{1-(1-\lambda\eta)^{2MK}}{1-(1-\lambda\eta)^{2K}}$, we can get

$$\begin{aligned}&\mathbb{E}\left(x^{(r)}\right)^2 \\ &= (1-d)^{2MK}\left(x^{(r-1)}\right)^2 + \eta^2\zeta^2\frac{M}{M-1}\cdot\frac{1}{d^2}\cdot\frac{1-(1-d)^K}{1+(1-d)^K}\cdot\left(1-(1-d)^{MK}\right)T(d), \quad (12)\end{aligned}$$

where we define $d = \lambda\eta$ and $T(d) = 1 + (1-d)^{MK} - \frac{1}{M} \cdot \frac{1+(1-d)^K}{1-(1-d)^K} \cdot \left(1 - (1-d)^{MK}\right)$. For convenience, we also define a intermediate variable $t = 1 - (1-d)^K$. Then

$$T(t) = \left(1 - \frac{1}{M} \cdot \frac{2-t}{t}\right) + \left(1 + \frac{1}{M} \cdot \frac{2-t}{t}\right)(1-t)^M. \tag{13}$$

According to Lemma 11, $T(t)$ is monotonically increasing on the interval $0 < t < 1$, and $T(d)$ is monotonically increasing on the interval $0 < d < 1$.

D.3.1 Lower Bound for $\frac{1}{101\lambda MK} \le \eta \le \frac{1}{\lambda K}$

In this regime, $T(d)$ is lower bounded by $T(\frac{1}{101MK})$ for $d \in [\frac{1}{101MK}, \frac{1}{K}]$. Here, we first lower bound $t(d)$, and then lower bound $T(d)$. According to the fact $(1-x)^n \le 1 - nx + \frac{1}{2}n^2x^2$ when $x \in (0,1)$ (it can be proved with Taylor expansion of $(1-x)^n$ at $x = 0.$), we get

$$t \ge 1 - (1-d)^K \ge 1 - (1 - dK + \frac{1}{2}d^2K^2) \ge dK - \frac{1}{2}d^2K^2 \ge \frac{1}{2}dK \ge \frac{1}{202M}$$

Then following the proofs of Safran and Shamir (2020)'s Lemma 1, we deal with the lower bound of $T(t)$ on $t \in [\frac{1}{202M}, 1]$,

$$T = 405\left(1 - \frac{1}{202M}\right)^M - 403 + \left(1 - \left(1 - \frac{1}{202M}\right)^M\right) \cdot \frac{1}{M}$$

Since the first two terms are increasing as M increases and the third term is positive, T is lower bounded by one numerical constant, as long as there exists M_0 such that $405\left(1 - \frac{1}{202M_0}\right)^{M_0} - 403 > 0$ and $T(t) > 0$ for all $2 \le t \le M_0$ (This can be done by a simple code). We get that $405\left(1 - \frac{1}{202\cdot 1212}\right)^{1212} - 403 > 1.3 \cdot 10^{-11}$ when $M_0 = 1212$. Hence, T is lower bounded by some numerical constant c. Returning to $\mathbb{E}\left[\left(x^{(r)}\right)^2\right]$, we get

$$\begin{aligned}
\mathbb{E}\left[\left(x^{(r)}\right)^2\right] &\ge (1-d)^{2MK}\left(x^{(r-1)}\right)^2 + c \cdot \eta^2\zeta^2 \frac{M}{M-1} \cdot \frac{1}{d^2} \cdot \frac{1-(1-d)^K}{1+(1-d)^K} \cdot \left(1 - (1-d)^{MK}\right) \\
&\ge (1-d)^{2MK}\left(x^{(r-1)}\right)^2 + \left(1 - \exp\left(-\frac{1}{101}\right)\right)\frac{c}{2} \cdot \frac{1}{202M} \cdot \eta^2\zeta^2 \frac{1}{d^2} \\
&\ge (1-\lambda\eta)^{2MK}\left(x^{(r-1)}\right)^2 + \frac{(1-\exp(-1/101))c}{404}\frac{\zeta^2}{\lambda^2 M} \\
&\ge (1-\lambda\eta)^{2MK}\left(x^{(r-1)}\right)^2 + c'\frac{\zeta^2}{\lambda^2 M}, \qquad (c' = \tfrac{(1-\exp(-1/101))}{404} \cdot c)
\end{aligned}$$

where we use $\frac{M}{M-1} \ge 1$, $1-(1-d)^{MK} \ge 1 - \exp(-dMK) \ge 1 - \exp(-1/101)$, $\frac{1}{1+(1-d)^K} \ge \frac{1}{2}$ and $1 - (1-d)^K \ge \frac{1}{202M}$ in the second inequality. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left(x^{(R)}\right)^2\right] \ge (1-\lambda\eta)^{2MKR}\left(x^{(0)}\right)^2 + \sum_{r=0}^{R-1}(1-\lambda\eta)^{2MKr}c'\frac{\zeta^2}{\lambda^2 M} \ge c'\frac{\zeta^2}{\lambda^2 M}, \\
&\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] = \frac{\lambda}{2}\mathbb{E}\left[\left(x^{(R)}\right)^2\right] \ge \frac{c'}{2}\frac{\zeta^2}{\lambda M} = \Omega\left(\frac{\zeta^2}{\lambda M}\right).
\end{aligned}$$

D.3.2 Lower Bound for $\frac{1}{\lambda K} \leq \eta \leq \frac{1}{\lambda}$

We still start from Eq. (12) and Eq. (13). For $d = \lambda\eta = 1$, we have $\mathbb{E}\left[\left(x^{(r)}\right)^2\right] = \eta^2\zeta^2 = \frac{\zeta^2}{\lambda^2}$. For $\frac{1}{K} \leq d < 1$, we have $t = 1-(1-d)^K \geq 1-\exp(-dK) \geq 1-\exp(-1) \approx 0.63 > 0.5$. Then we can get the lower bound of $T(t)$ on $\frac{1}{2} < t < 1$,

$$T \geq \left(1 - \frac{1}{M}\cdot\frac{2-\frac{1}{2}}{\frac{1}{2}}\right) + \left(1 + \frac{1}{M}\cdot\frac{2-\frac{1}{2}}{\frac{1}{2}}\right)(1-\frac{1}{2})^M \geq \left(1-\frac{3}{M}\right) + \left(1+\frac{3}{M}\right)\frac{1}{2^M}.$$

It can be seen that $T > 1 - \frac{3}{M} \geq \frac{1}{4}$ when $M \geq 4$, $T = \frac{1}{4}$ when $M = 3$, and $T = \frac{1}{8}$ when $M = 2$. Thus, we can obtain that $T \geq c$ for some numerical constant c. In fact, since $t \geq 1-\exp(-1) > \frac{1}{202M}$, we can also use the conclusion of the lower bound for $\frac{1}{101\lambda MK} \leq \eta \leq \frac{1}{\lambda K}$. Then,

$$\begin{aligned}
\mathbb{E}\left(x^{(r)}\right)^2 &\geq (1-d)^{2MK}\left(x^{(r-1)}\right)^2 + c\cdot\eta^2\zeta^2\frac{M}{M-1}\cdot\frac{1}{d^2}\cdot\frac{1-(1-d)^K}{1+(1-d)^K}\cdot\left(1-(1-d)^{MK}\right) \\
&\geq (1-\lambda\eta)^{2MK}\left(x^{(r-1)}\right)^2 + \frac{c}{8}\cdot\frac{\zeta^2}{\lambda^2},
\end{aligned}$$

where we use $\frac{M}{M-1} \geq 1$, $1-(1-d)^{MK} \geq 1-(1-d)^K \geq 1-\exp(-dK) \geq \frac{1}{2}$, $\frac{1}{1+(1-d)^K} \geq \frac{1}{2}$. Then, for $\frac{1}{K} \leq d = \lambda\eta < 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(x^{(R)}\right)^2\right] \geq (1-\lambda\eta)^{2MKR}\left(x^{(0)}\right)^2 + \sum_{r=0}^{R-1}(1-\lambda\eta)^{2MKr}\frac{c}{8}\frac{\zeta^2}{\lambda^2} \geq \frac{c}{8}\frac{\zeta^2}{\lambda^2}, \\
&\mathbb{E}\left[F(x^{(R)}) - F^*\right] = \mathbb{E}\left[F(x^{(R)})\right] = \frac{\lambda}{2}\mathbb{E}\left[\left(x^{(R)}\right)^2\right] = \Omega\left(\frac{\zeta^2}{\lambda}\right).
\end{aligned}$$

Now we complete the proofs for all regimes for heterogeneity terms in Theorem 5. The setups and final results are summarized in Table 6. ■

D.4 Helpful Lemmas for Heterogeneity Terms

Lemma 11 *The function $T(d)$ defined below is monotonically increasing on the interval $0 < d < 1$, for integers $M \geq 2$ and $K \geq 1$.*

$$T(d) = 1 + (1-d)^{MK} - \frac{1}{M}\cdot\frac{1+(1-d)^K}{1-(1-d)^K}\cdot\left(1-(1-d)^{MK}\right).$$

Proof Here we introduce an intermediate variable $t = 1-(1-d)^K$ (it implies that $(1-d)^K = 1-t$, $(1-d)^{MK} = (1-t)^M$) and analyze the function $T(t)$ on $0 < t < 1$ at first.

$$\begin{aligned}
T &= 1 + (1-t)^M - \frac{1}{M}\cdot\frac{2-t}{t}\cdot\left(1-(1-t)^M\right) \\
&= \left(1-\frac{1}{M}\cdot\frac{2-t}{t}\right) + \left(1+\frac{1}{M}\cdot\frac{2-t}{t}\right)(1-t)^M.
\end{aligned}$$

Then, we follow a similar way to th proof of Lemma 1 in Safran and Shamir (2020) to prove $T(t)$ is increasing. The derivative of the function $T(t)$ is

$$\begin{aligned} T(t)' &= \frac{2}{Mt^2} - \frac{2}{Mt^2}(1-t)^M - \left(1 + \frac{1}{M} \cdot \frac{2-t}{t}\right) \cdot M(1-t)^{M-1} \\ &= \frac{2}{Mt^2}\left(1 - (1-t)^{M-1} \cdot \left(1 + (M-1)t + \frac{1}{2}M(M-1)t^2\right)\right). \end{aligned}$$

The Taylor expansion of $\left((1-t)^{1-M}\right)$ at $t=0$ is

$$(1-t)^{1-M} = 1 + (M-1)t + \frac{1}{2}M(M-1)t^2 + \frac{1}{3!}(M+1)M(M-1)(1-\xi)^{-M-2}t^3,$$

where $\xi \in [0,t]$. When $0 < t < 1$, the remainder $\frac{1}{3!}(M+1)M(M-1)(1-\xi)^{-M-2}t^3 > 0$ for $M \geq 2$. So we can get $(1-t)^{1-M} > 1 + (M-1)t + \frac{1}{2}M(M-1)t^2$. It follows that

$$T' > \frac{2}{Mt^2}(1 - (1-t)^{M-1} \cdot (1-t)^{1-M}) = 0.$$

Thus, $T(t)$ is monotonically increasing on $0 < t < 1$. Since $t = 1-(1-d)^K$ is monotonically increasing on $0 < d < 1$, we can get that $T(d)$ is monotonically increasing on $0 < d < 1$. ■

Lemma 12 *Let $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$ be a random permutation of $\frac{M}{2}$ $+1$'s and $\frac{M}{2}$ -1's. Let $\mathcal{A}_{m,k} \coloneqq \mathcal{E}_{m-1} + a_k\tau_m$, where $\mathcal{E}_{m-1} = \sum_{i=1}^{m-1}\tau_i$ and $a_k = k/K$ $(0 \leq k \leq K-1)$. Then,*

$$\frac{1}{20}\sqrt{m-1+a_k^2} \leq \mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \leq \sqrt{m-1+a_k^2}.$$

Notably, the lower bound holds for $1 \leq m \leq \frac{M}{2}+1$ $(M \geq 4)$ and the upper bound holds for $1 \leq m \leq M$ $(M \geq 2)$.

Proof We consider the upper and lower bounds as follows. For the upper bound, similar to Rajput et al. (2020)'s Lemma 12 and Cha et al. (2023)'s Lemma B.5, we have

$$\begin{aligned} \mathbb{E}\left[|\mathcal{A}_{m,k}|\right] &= \mathbb{E}\left[\left|\sum_{i=1}^{m-1}\tau_i + a_k\tau_m\right|\right] \\ &\leq \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{m-1}\tau_i + a_k\tau_m\right)^2\right]} \\ &\leq \sqrt{\sum_{i=1}^{m-1}\mathbb{E}[(\tau_i)^2] + 2\sum_{i<j\leq m-1}\mathbb{E}[\tau_i\tau_j] + a_k^2 + 2a_k\sum_{i=1}^{m-1}\mathbb{E}(\tau_i\tau_m)} \\ &\leq \sqrt{m-1+a_k^2}. \qquad (\because \mathbb{E}[\tau_i\tau_j] < 0,\ \forall i \neq j) \end{aligned}$$

For the lower bound, suppose that $3 \le m \le \frac{M}{2}+1$ (i.e., $2 \le m-1 \le \frac{M}{2}$ and $M \ge 4$). Then,

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] &= \mathbb{E}\left[|\mathcal{E}_{m-1} + a_k\tau_m|\right] \\
&= \mathbb{E}\left[|\mathcal{E}_{m-1} + a_k\tau_m| \mid \mathcal{E}_{m-1}\tau_m \ge 0\right] \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m \ge 0) \\
&\quad + \mathbb{E}\left[|\mathcal{E}_{m-1} + a_k\tau_m| \mid \mathcal{E}_{m-1}\tau_m < 0\right] \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0) \\
&= \mathbb{E}\left[|\mathcal{E}_{m-1}| \mid \mathcal{E}_{m-1}\tau_m \ge 0\right] \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m \ge 0) + a_k \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m \ge 0) \\
&\quad + \mathbb{E}\left[|\mathcal{E}_{m-1}| \mid \mathcal{E}_{m-1}\tau_m < 0\right] \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0) - a_k \cdot \mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0) \\
&= \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m \ge 0) - a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0) \\
&= \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m = 0) + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m > 0) - a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0) \\
&= \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m = 0) \\
&\quad + a_k \sum_{i=1}^{m-1} \mathbb{P}(\mathcal{E}_{m-1}\tau_m > 0 \mid |\mathcal{E}_{m-1} = i|)\mathbb{P}(|\mathcal{E}_{m-1} = i|) \\
&\quad - a_k \sum_{i=1}^{m-1} \mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0 \mid |\mathcal{E}_{m-1} = i|)\mathbb{P}(|\mathcal{E}_{m-1} = i|).
\end{aligned}$$

Since $\mathbb{P}(\mathcal{E}_{m-1}\tau_m > 0 \mid |\mathcal{E}_{m-1} = i|) = \frac{(M-m+1-i)/2}{M-m+1}$ and $\mathbb{P}(\mathcal{E}_{m-1}\tau_m < 0 \mid |\mathcal{E}_{m-1} = i|) = \frac{(M-m+1+i)/2}{M-m+1}$, we get

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] &= \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m = 0) - a_k \cdot \frac{1}{M-m+1} \sum_{i=1}^{m-1} i \cdot \mathbb{P}(|\mathcal{E}_{m-1}| = i) \\
&= \left(1 - \frac{a_k}{M-m+1}\right) \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] + a_k\mathbb{P}(\mathcal{E}_{m-1}\tau_m = 0) \\
&\ge \left(1 - \frac{1}{M-m+1}\right) \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] && (\because 0 \le a_k < 1) \\
&\ge \left(1 - \frac{2}{M}\right) \mathbb{E}\left[|\mathcal{E}_{m-1}|\right] && (\because m-1 \le \tfrac{M}{2}) \\
&\ge \frac{1}{2}\mathbb{E}\left[|\mathcal{E}_{m-1}|\right], && (\because M \ge 4)
\end{aligned}$$

where we use $\mathbb{E}\left[|\mathcal{E}_{m-1}|\right] = \sum_{i=1}^{m-1} i \cdot \mathbb{P}(|\mathcal{E}_{m-1}| = i)$ in the second equality.

For the convenience of subsequent proofs, we need a tighter lower bound for $\mathbb{E}\left[|\mathcal{E}_{m-1}|\right]$, which can be achieved with a few modifications to Cha et al. (2023)'s Lemma B.5.

Let us start from Ineq. (21) in Cha et al. (2023)'s Lemma B.5.

For the even integers $m \ge 4$, we have

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{E}_m|\right] &\ge \left(\frac{M-2}{M-1} \cdot \frac{\sqrt{m}}{\sqrt{m-2}}\right) \cdot \left(\frac{M}{M+2m}\right) \cdot \left(\frac{\sqrt{m}}{5}\right) \\
&= \left(\frac{M-2}{M-1} \cdot \frac{\sqrt{m-1}}{\sqrt{m-2}}\right) \cdot \left(\frac{M}{M+2m}\right) \cdot \left(\frac{\sqrt{m+1}}{5}\right)
\end{aligned}$$

It can be shown that $\frac{M-2}{M-1}\cdot\frac{\sqrt{m-1}}{\sqrt{m-2}} \geq 1 \iff (2M-3)m \leq M^2-2$. Since $m \leq \frac{M}{2}$ (Note that the constraint $m \leq \frac{M}{2}$ is for $\mathbb{E}\left[|\mathcal{E}_m|\right]$ in Cha et al. (2023)'s Lemma B.5), it follows that $(2M-3)m \leq M^2 - \frac{3}{2}M \leq M^2-2$ when $M \geq 8$. Then, we can get $\mathbb{E}\left[\mathcal{E}_m\right] \geq \frac{\sqrt{m+1}}{10}$ for $4 \leq m \leq \frac{M}{2}$.

For the even integers $m = 2$, we have $\mathbb{E}\left[|\mathcal{E}_2|\right] = 1 - \frac{1}{M-1} \geq \frac{2}{3} \geq \frac{\sqrt{3}}{10}$ $(M \geq 2m \geq 4)$. Now, we complete the proof for the even cases $2 \leq m \leq \frac{M}{2}$.

In fact, the lower bound holds in odd cases $1 \leq m \leq \frac{M}{2}$ in Cha et al. (2023)'s Lemma B.5 without any modification (see their last inequality $\mathbb{E}\left[\mathcal{E}_m\right] \geq \frac{M-m}{M-m-1}\cdot\frac{\sqrt{m+1}}{10} \geq \frac{\sqrt{m+1}}{10}$). We can also prove it with the same steps as Cha et al. (2023)'s Lemma B.5.

Note that the lower bound does not hold in the last case $m = 0$.

As a summary, we can get a tighter bound $\mathbb{E}\left[|\mathcal{E}_m|\right] \geq \frac{\sqrt{m+1}}{10}$ for $1 \leq m \leq \frac{M}{2}$.

Returning to $\mathbb{E}\left[|\mathcal{A}_{m,k}|\right]$ and using the tighter lower bound for $\mathbb{E}\left[|\mathcal{E}_{m-1}|\right]$, we have

$$\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \geq \frac{1}{2}\mathbb{E}\left[|\mathcal{E}_{m-1}|\right] \geq \frac{\sqrt{m}}{20}.$$

The above lower bound does not hold for $m = 1$ since it requires the false argument $\mathcal{E}_0 = 0 \geq \frac{\sqrt{1}}{10}$. To incorporate the case where $m = 1$, we consider a looser bound

$$\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \geq \frac{\sqrt{m}}{20} \geq \frac{1}{20}\sqrt{m-1+a_k^2}.$$

At last, let us verify whether this lower bound holds for the remaining cases where $m = 1, 2$. When $m = 1$, $\mathbb{E}\left[|\mathcal{A}_{1,k}|\right] = a_k \geq \frac{a_k}{20\sqrt{2}}$. When $m = 2$ $(M \geq 2m \geq 4)$, it follows that

$$\begin{aligned}
\mathbb{E}\left[|\mathcal{A}_{2,k}|\right] = \mathbb{E}\left[|\tau_1 + a_k\tau_2|\right] &= (1+a_k)\mathbb{P}(\tau_1\tau_2 = +1) + (1-a_k)\mathbb{P}(\tau_1\tau_2 = -1) \\
&= (1+a_k)\cdot\frac{2\cdot\binom{\frac{M}{2}}{2}}{\binom{M}{2}} + (1-a_k)\cdot\frac{\binom{\frac{M}{2}}{1}\cdot\binom{\frac{M}{2}}{1}}{\binom{M}{2}} \\
&= 1 - \frac{a_k}{M-1}.
\end{aligned}$$

Here we adopt $M \geq 4$ for $m = 2$, and then $\mathbb{E}\left[|\mathcal{A}_{2,k}|\right] = 1 - \frac{a_k}{M-1} \geq 1 - \frac{a_k}{2} \geq \frac{1}{2} = \frac{1}{2\sqrt{2}}\sqrt{1+1^2} \geq \frac{1}{20}\sqrt{1+a_k^2}$. Now we complete the proof of the lower bound of $\mathbb{E}\left[|\mathcal{A}_{2,k}|\right]$, which holds for $1 < m \leq \frac{M}{2}+1$ and $M \geq 4$. ■

Lemma 13 *Let $\tau = (\tau_1, \tau_2, \ldots, \tau_M)$ be a random permutation of $\frac{M}{2}$ $+1$'s and $\frac{M}{2}$ -1's. Let $\mathcal{A}_{m,k} \coloneqq \mathcal{E}_{m-1} + a_k\tau_m$, where $\mathcal{E}_{m-1} = \sum_{i=1}^{m-1}\tau_i$ and $a_k = k/K$ $(0 \leq k \leq K-1)$. The probability distribution of $\mathcal{A}_{m,k}$ is symmetric with respect to 0. And For $1 \leq m \leq M$ and $0 \leq k \leq K-1$ (excluding the case $m = 1, k = 0$), it holds that*

$$\frac{1}{6} \leq \mathbb{P}(\mathcal{A}_{m,k} > 0) = \mathbb{P}(\mathcal{A}_{m,k} < 0) \leq \frac{1}{2}.$$

Proof When $m = 1$ and $k = 0$, $\mathcal{A}_{1,0} = \mathcal{E}_0 = 0$ (defined). When $m = M + 1$ and $k = 0$, $\mathcal{A}_{M+1,0} = 0$. In these two cases, $\mathbb{P}(\mathcal{A}_{m,k} = 0) = 1$ and $\mathbb{P}(\mathcal{A}_{m,k} < 0) = \mathbb{P}(\mathcal{A}_{m,k} > 0) = 0$.

When $2 \le m \le M$ and $k = 0$, we get $\mathbb{P}(\mathcal{A}_{m,k} > 0) = \mathbb{P}(\mathcal{A}_{m,k} < 0) \ge \frac{1}{6}$ according to Yun et al. (2022)' Lemma 14.

When $1 \le m \le M$ and $0 < k \le K - 1$, similarly, we can first prove that $\mathcal{A}_{m,k}$ is symmetric and then compute $\mathbb{P}(\mathcal{A}_{m,k} = 0)$. As shown in Table 7, we conclude all cases into four categories $\mathcal{A}_{m,k} = -p - a_k$, $\mathcal{A}_{m,k} = -p + a_k$, $\mathcal{A}_{m,k} = p - a_k$ and $\mathcal{A}_{m,k} = p + a_k$. We can get that the probability distribution of $\mathcal{A}_{m,k}$ is symmetric. Furthermore, since $0 < a_k < 1$, we can get that $\mathbb{P}(\mathcal{A}_{m,k} = 0) = 0$, and thus $\mathbb{P}(\mathcal{A}_{m,k} > 0) = \mathbb{P}(\mathcal{A}_{m,k} < 0) = \frac{1}{2}$.

Value	Probability
$-p - a_k$	$\frac{\binom{\frac{M}{2}}{\frac{m-1-p}{2}}\binom{\frac{M}{2}}{\frac{m-1+p}{2}}}{\binom{M}{m-1}} \cdot \frac{\frac{M-m+1-p}{2}}{M-m+1}$
$-p + a_k$	$\frac{\binom{\frac{M}{2}}{\frac{m-1-p}{2}}\binom{\frac{M}{2}}{\frac{m-1+p}{2}}}{\binom{M}{m-1}} \cdot \frac{\frac{M-m+1+p}{2}}{M-m+1}$
$p - a_k$	$\frac{\binom{\frac{M}{2}}{\frac{m-1-p}{2}}\binom{\frac{M}{2}}{\frac{m-1+p}{2}}}{\binom{M}{m-1}} \cdot \frac{\frac{M-m+1+p}{2}}{M-m+1}$
$p + a_k$	$\frac{\binom{\frac{M}{2}}{\frac{m-1-p}{2}}\binom{\frac{M}{2}}{\frac{m-1+p}{2}}}{\binom{M}{m-1}} \cdot \frac{\frac{M-m+1-p}{2}}{M-m+1}$

Table 7: Probability distribution of $\mathcal{A}_{m,k}$.

In summary, we have proved that $\mathbb{P}(\mathcal{A}_{m,k} > 0) = \mathbb{P}(\mathcal{A}_{m,k} < 0) \ge \frac{1}{6}$ for $1 \le m \le M$ and $0 \le k \le K - 1$ (except the case $m = 1, k = 0$). Note that for all cases, the probability distribution is symmetric, we can get $\mathbb{P}(\mathcal{A}_{m,k} > 0) = \mathbb{P}(\mathcal{A}_{m,k} < 0) \le \frac{1}{2}$. ■

Lemma 14 *Supposing that* $x_{1,0} \ge 0$, $\lambda_0/\lambda \ge 1010$ *and* $\eta \le \frac{1}{101\lambda MK}$, *then for* $1 \le m \le M$, $0 \le k \le K - 1$, *we have*

$$\mathbb{E}\left[|x_{m,k} - x_{1,0}|\right] \le \frac{1}{100} x_{1,0} + \frac{101}{100} K\eta\zeta\sqrt{m - 1 + a_k^2}$$

Proof According to Eq. (3),

$$x_{m,k} = x_{1,0} - \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \left((\lambda \mathbb{1}_{x_{b_1(i),b_2(i)} < 0} + \lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)} \ge 0}) x_{b_1(i),b_2(i)} \right) - K\eta\zeta\mathcal{A}_{m,k}$$

(we have dropped the superscript r), we can get

$$\begin{aligned}\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right] &\leq \eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left|(\lambda \mathbb{1}_{x_{b_1(i),b_2(i)}<0}+\lambda_0 \mathbb{1}_{x_{b_1(i),b_2(i)}\geq 0})x_{b_1(i),b_2(i)}\right|\right] + K\eta\zeta\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \\ &\leq \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left|x_{b_1(i),b_2(i)}\right|\right] + K\eta\zeta\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \qquad (\because \lambda_0 \leq \lambda) \\ &\leq \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left|x_{b_1(i),b_2(i)}-x_{1,0}\right|\right] + K\eta\zeta\mathbb{E}\left[|\mathcal{A}_{m,k}|\right].\end{aligned}$$

For any integers $m \geq 1$, $k \geq 0$ satisfying $(m-1)K+k \leq MK$, using Lemma 12, $\mathbb{E}\left[|\mathcal{A}_{m,k}|\right] \leq \sqrt{m-1+a_k^2}$, we can get

$$\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right] \leq \lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left|x_{b_1(i),b_2(i)}-x_{1,0}\right|\right] + K\eta\zeta\sqrt{m-1+a_k^2}.$$

Let $h_{m,k} \coloneqq \lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta\sum_{i=0}^{\mathcal{B}_{m,k}-1} h_{b_1(i),b_2(i)} + K\eta\zeta\sqrt{m-1+a_k^2}$ and $h_{1,0}=0$. It can be verified that the sequence $h_{1,0},\ldots,h_{m,0},h_{m,1},h_{m,2},\ldots,h_{m,K-1},\ h_{m+1,0},\ldots,h_{M+1,0}$ is monotonically increasing. When $k=0$, then

$$h_{m,0}-h_{m-1,K-1} = \lambda\eta\mathbb{E}\left[x_{1,0}\right] + \lambda\eta h_{m-1,K-1} + K\eta\zeta\left(\sqrt{m-1}-\sqrt{m-2+a_{K-1}^2}\right) > 0.$$

When $1 \leq k \leq K-1$, then

$$h_{m,k}-h_{m,k-1} = \lambda\eta\mathbb{E}\left[x_{1,0}\right] + \lambda\eta h_{m,k-1} + K\eta\zeta\left(\sqrt{m-1+a_k^2}-\sqrt{m-1+a_{k-1}^2}\right) > 0.$$

This means that $h_{b_1(i),b_2(i)} < h_{m,k}$ for any integer $i < \mathcal{B}_{m,k}$. So we can get

$$\begin{aligned}&h_{m,k} \leq \lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta\mathcal{B}_{m,k}h_{m,k} + K\eta\zeta\sqrt{m-1+a_k^2} \\ \Longrightarrow\ &h_{m,k} \leq \frac{\lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right]}{1-\lambda\eta\mathcal{B}_{m,k}} + \frac{K\eta\zeta\sqrt{m-1+a_k^2}}{1-\lambda\eta\mathcal{B}_{m,k}} \quad (\because \lambda\eta\mathcal{B}_{m,k} \leq \lambda\eta\mathcal{B}_{M,K} = \lambda MK\eta \leq \tfrac{1}{101})\end{aligned}$$

By mathematical induction, we can get $\mathbb{E}\left[|x_{m,k}-x_{1,0}|\right] \leq h_{m,k} \leq \frac{\lambda\eta\mathcal{B}_{m,k}x_{1,0}}{1-\lambda\eta\mathcal{B}_{m,k}} + \frac{K\eta\zeta\sqrt{m-1+a_k^2}}{1-\lambda\eta\mathcal{B}_{m,k}}$. When $m=1$ and $k=0$, $\mathbb{E}\left[|x_{1,0}-x_{1,0}|\right] = h_{1,0} = 0$. Then, suppose that $\mathbb{E}\left[|x_{i,j}-x_{1,0}|\right] \leq h_{i,j}$ for all i,j satisfying $i(K-1)+j \leq m(K-1)+k$.

- When $k=0$, it follows that

$$\begin{aligned}\mathbb{E}\left[|x_{m,0}-x_{1,0}|\right] &\leq \lambda\eta\mathcal{B}_{m,0}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,0}-1} \mathbb{E}\left[\left|x_{b_1(i),b_2(i)}-x_{1,0}\right|\right] + K\eta\zeta\sqrt{m-1} \\ &\leq \lambda\eta\mathcal{B}_{m,0}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,0}-1} h_{b_1(i),b_2(i)} + K\eta\zeta\sqrt{m-1} \leq h_{m,0}.\end{aligned}$$

- When $1 \leq k \leq K-1$, it follows that

$$\begin{aligned}\mathbb{E}\left[|x_{m,k} - x_{1,0}|\right] &\leq \lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} \mathbb{E}\left[\left|x_{b_1(i),b_2(i)} - x_{1,0}\right|\right] + K\eta\zeta\sqrt{m + a_k^2} \\ &\leq \lambda\eta\mathcal{B}_{m,k}\mathbb{E}\left[x_{1,0}\right] + \lambda\eta \sum_{i=0}^{\mathcal{B}_{m,k}-1} h_{b_1(i),b_2(i)} + K\eta\zeta\sqrt{m + a_k^2} \leq h_{m,k}.\end{aligned}$$

Then considering that $\lambda_0/\lambda \geq 1010$ and $\lambda\eta\mathcal{B}_{m,k} \leq \lambda MK\eta \leq \frac{1}{101}$ (it implies $\frac{\lambda\eta\mathcal{B}_{m,k}}{1-\lambda\eta\mathcal{B}_{m,k}} \leq \frac{1}{100}$ and that $\frac{1}{1-\lambda\eta\mathcal{B}_{m,k}} \leq \frac{101}{100}$), we get $\mathbb{E}\left[|x_{m,k} - x_{1,0}|\right] \leq \frac{1}{100}x_{1,0} + \frac{101}{100}K\eta\zeta\sqrt{m - 1 + a_k^2}$. ■

Appendix E. The Mechanism of "Two Learning Rates" in SFL

This section shows that the mechanism of "two learning rates" can also be applied to SFL. In theory, it can achieve a similar improvement to (almost the same as) that in PFL Karimireddy et al. (2020). Next, we show how to use the mechanism of "two learning rates" in SFL, and compare it with that of PFL.

E.1 The Mechanism of "Two Learning Rates" in SFL

The mechanism of "two learning rates" includes two learning rates, the global/server learning rate and the local/client learning rate. The client learning rate is on the client-side for local updates; the server learning rate is on the server-side for global updates. The modified algorithms are provided in Algorithm 3 and Algorithm 4. The modified lines are marked in a pink color box. Here γ denotes the server learning rate; η denotes the client learning rate. For SFL, one simple and practical implementation is illustrated in Figure 5.

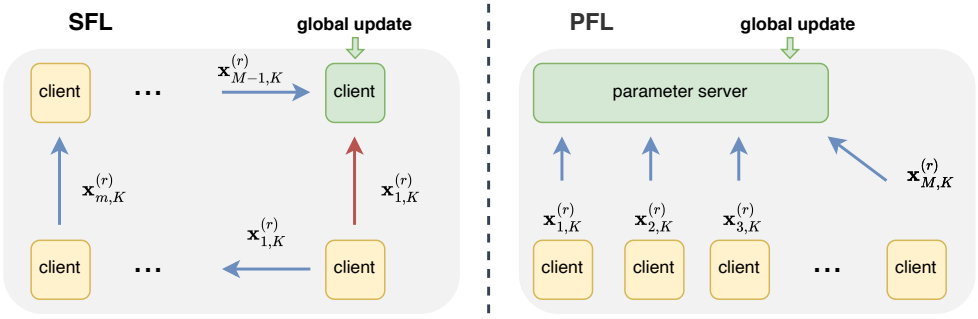

Figure 5: The mechanism of "two learning rates" in SFL and PFL. The global updates of SFL are performed at the last client. It performs the global updates with its parameters $\mathbf{x}_{M,K}^{(r)}$ and the initial parameters $\mathbf{x}^{(r)}$ received from the first client.

Algorithm 3: Sequential FL

for $r = 0, \ldots, R-1$ **do**
  Sample a permutation $\pi_1, \pi_2, \ldots, \pi_M$ of $\{1, 2, \ldots, M\}$
  for $m = 1, \ldots, M$ **in sequence do**
    $\mathbf{x}_{m,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)}, & m = 1 \\ \mathbf{x}_{m-1,K}^{(r)}, & m > 1 \end{cases}$
    for $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{\pi_m,k}^{(r)}$
  $\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} - \gamma \left(\mathbf{x}^{(r)} - \mathbf{x}_{M,K}^{(r)} \right)$

Algorithm 4: Parallel FL

for $r = 0, \ldots, R-1$ **do**
  for $m = 1, \ldots, M$ **in parallel do**
    $\mathbf{x}_{m,0}^{(r)} = \mathbf{x}^{(r)}$
    for $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{m,k}^{(r)}$
  $\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} - \gamma \left(\mathbf{x}^{(r)} - \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_{m,K}^{(r)} \right)$

According to the new update rule of SFL, we can get the upper bounds of SFL and PFL in Theorem 15 and Theorem 16. We only consider the non-convex case for convenience. We see that the server learning rate γ has the similar effect on the upper bounds for SFL to those for PFL. The advantage of $\frac{1}{M^{1/3}}$ of SFL still exists.

Theorem 15 (SFL, non-convex) *Under the same conditions as those of the non-convex case in Theorem 1, there exists* $\tilde{\eta} = \gamma \eta M K \lesssim \frac{1}{L(1+\beta^2/M)}$ $(\gamma \geq 1)$*, such that*

$$\min_{0 \leq r \leq R} \mathbb{E} \left[\left\| \nabla F(\mathbf{x}^{(r)}) \right\|^2 \right] \lesssim \frac{A}{\tilde{\eta} R} + \frac{\tilde{\eta} L \sigma^2}{MK} + \frac{\tilde{\eta}^2 L^2 \sigma^2}{\gamma^2 MK} + \frac{\tilde{\eta}^2 L^2 \zeta^2}{\gamma^2 M}.$$

After tuning the learning rate, we get

$$\min_{0 \leq r \leq R} \mathbb{E} \left[\left\| \nabla F(\mathbf{x}^{(r)}) \right\|^2 \right] = \mathcal{O} \left(\frac{LA \left(1 + \frac{\beta^2}{M}\right)}{R} + \frac{\left(L \sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2 \sigma^2 A^2\right)^{1/3}}{\gamma^{2/3} M^{1/3} K^{1/3} R^{2/3}} + \frac{\left(L^2 \zeta^2 A^2\right)^{1/3}}{\gamma^{2/3} M^{1/3} R^{2/3}} \right).$$

Proof See Appendix E.2. ■

Theorem 16 (PFL, non-convex) *Under the same conditions as those of the non-convex case in Theorem 1, there exists* $\tilde{\eta} = \gamma \eta K \lesssim \frac{1}{L(1+\beta^2)}$ $(\gamma \geq 1)$*, such that*

$$\min_{0 \leq r \leq R} \mathbb{E} \left[\left\| \nabla F(\mathbf{x}^{(r)}) \right\|^2 \right] \lesssim \frac{A}{\tilde{\eta} R} + \frac{\tilde{\eta} L \sigma^2}{MK} + \frac{\tilde{\eta}^2 L^2 \sigma^2}{\gamma^2 K} + \frac{\tilde{\eta}^2 L^2 \zeta^2}{\gamma^2}.$$

After tuning the learning rate, we get

$$\min_{0 \leq r \leq R} \mathbb{E} \left[\left\| \nabla F(\mathbf{x}^{(r)}) \right\|^2 \right] = \mathcal{O} \left(\frac{LA(1+\beta^2)}{R} + \frac{\left(L \sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2 \sigma^2 A^2\right)^{1/3}}{\gamma^{2/3} K^{1/3} R^{2/3}} + \frac{\left(L^2 \zeta^2 A^2\right)^{1/3}}{\gamma^{2/3} R^{2/3}} \right).$$

Proof See Karimireddy et al. (2020)'s Theorem 1 or Yang et al. (2021)'s Theorem 1. Note that Karimireddy et al. (2020) set $\gamma = \sqrt{M}$ and Yang et al. (2021) set $\gamma = \sqrt{MK}$. Here we keep $\gamma = \gamma$ for comparison. ■

E.2 Proofs of Theorem 15

Proof We consider the full client participation for simplicity. Since the global update is performed at the end of one training round, the client drift bound (that is, Li and Lyu (2023)'s Lemma 10) is unaffected. We can focus on Li and Lyu (2023)'s Lemma 9. Substituting the overall updates $\Delta \mathbf{x} = -\eta \sum_{m=1}^{M} \sum_{k=0}^{K-1} \mathbf{g}_{\pi_m}(\mathbf{x}_{m,k})$ with $\Delta \mathbf{x} = -\eta\gamma \sum_{m=1}^{M} \sum_{k=0}^{K-1} \mathbf{g}_{\pi_m}(\mathbf{x}_{m,k})$ in Li and Lyu (2023)'s Lemma 9, we get the recursion

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{x}^{(r+1)}) - F(\mathbf{x}^{(r)})\right] \\
&\leq -\frac{1}{6}\eta\gamma MK\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2 + 2\eta^2\gamma^2 LMK\sigma^2 + \frac{5}{6}\eta\gamma L^2 \sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{x}_{m,k}^{(r)} - \mathbf{x}^{(r)}\|^2.
\end{aligned}$$

Plugging Li and Lyu (2023)'s Lemma 10 into it, and using $\eta\gamma \leq \frac{1}{6LMK(1+\beta^2/M)}$, we get

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{x}^{(r+1)}) - F(\mathbf{x}^{(r)})\right] \\
&\leq -\frac{1}{10}\eta\gamma MK\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2 + 2L\eta^2\gamma^2 MK\sigma^2 + \frac{15}{8}\eta^3\gamma L^2M^2K^2\sigma^2 + \frac{15}{8}\eta^3\gamma L^2M^2K^3\zeta^2.
\end{aligned}$$

Letting $\tilde{\eta} = \eta\gamma MK$, we can get

$$\mathbb{E}\left[F(\mathbf{x}^{(r+1)}) - F(\mathbf{x}^{(r)})\right] \leq -\frac{1}{10}\tilde{\eta}\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2 + \frac{2L\tilde{\eta}^2\sigma^2}{MK} + \frac{15}{8}\frac{\tilde{\eta}^3L^2\sigma^2}{\gamma^2MK} + \frac{15}{8}\frac{\tilde{\eta}^3L^2\zeta^2}{\gamma^2M}.$$

Then, we get

$$\min_{0\leq r\leq R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{10A}{\tilde{\eta}R} + \frac{20\tilde{\eta}L\sigma^2}{MK} + \frac{75}{4}\frac{\tilde{\eta}^2L^2\sigma^2}{\gamma^2MK} + \frac{75}{4}\frac{\tilde{\eta}^2L^2\zeta^2}{\gamma^2M}.$$

Since $\eta\gamma \leq \frac{1}{6LMK(1+\beta^2/M)}$, using Li and Lyu (2023)'s Lemma 8, we can get

$$\begin{aligned}
&\min_{0\leq r\leq R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \\
&= \mathcal{O}\left(\frac{LA\left(1+\frac{\beta^2}{M}\right)}{R} + \frac{\left(L\sigma^2A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2\sigma^2A^2\right)^{1/3}}{\gamma^{2/3}M^{1/3}K^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2A^2\right)^{1/3}}{\gamma^{2/3}M^{1/3}R^{2/3}}\right).
\end{aligned}$$

■

Appendix F. Comparison with Lower Bounds of SGD-RR

In this section, we compare the lower bounds of SFL with those of SGD-RR. The lower bounds of SFL are stated in Theorem 5 and Theorem 6; Theorem 5 is for arbitrary learning rates $\eta > 0$ and Theorem 6 is for small learning rates $0 < \eta \lesssim \frac{1}{LMK}$. By comparing them with the corresponding theorems of SGD-RR in Cha et al. (2023), we have the following comparison results: (1) for arbitrary learning rates $\eta > 0$, the lower bound of SFL is worse than that of SGD-RR by a factor of κ; (2) for small learning rates $0 < \eta \lesssim \frac{1}{LMK}$, the lower bounds of SFL match those of SGD-RR. See Theorems 17, 5, 18 and 6.

Be attention that *SGD-RR is a special case of SFL, where one single step of GD (Gradient Descent) step is performed on each local objective function (that is, $\sigma = 0$ and $K = 1$).*

The lower bounds for arbitrary learning rates $\eta > 0$. The lower bounds are restated in Theorem 17 (SGD-RR, Cha et al. 2023) and Theorem 5 (SFL) with our notations. First, we see that there are more components (including stochasticity and heterogeneity) in the lower bounds of SFL. Given that SGD-RR is a special case of SFL, by letting $\sigma = 0$ and $K = 1$, we next focus on the most noteworthy heterogeneity term (the last term, with ζ) in Theorem 5, and compare it with those of SGD-RR in Theorem 17.

We consider two cases $R \gtrsim \kappa$ and $R \lesssim \kappa$. When $R \gtrsim \kappa$, the lower bound $\Omega\left(\frac{L\zeta^2}{\mu^2 MR^2}\right)$ of SGD-RR is better than $\Omega\left(\frac{\zeta^2}{\mu MR^2}\right)$ of SFL with an advantage of κ. When $R \lesssim \kappa$, the lower bound $\Omega\left(\frac{\zeta^2}{\mu MR}\right)$ of SGD-RR is also better than $\Omega\left(\frac{\zeta^2}{\mu MR^2}\right)$ of SFL. Thus, we get that the lower bounds for SGD-RR in Cha et al. (2023) are better than ours for SFL for $\eta > 0$. It is still open whether the lower bounds of SFL can be better for arbitrary $\eta > 0$.

The lower bounds for small learning rates $0 < \eta \lesssim \frac{1}{LMK}$. The lower bounds are restated in Theorem 18 (SGD-RR, Cha et al. 2023) and Theorem 6 (SFL) with our notations. First, we see that there are more components (including stochasticity and heterogeneity) in the lower bounds of SFL. Similarly, we next focus on the most noteworthy heterogeneity term (the last term, with ζ) in Theorem 6. It can be shown that the lower bounds of SFL completely match those of SGD-RR.

Theorem 17 (Theorem 3.1 in Cha et al. (2023)) *For any $M \geq 2$ and $\kappa \geq 2415$, there exist a 3-dimensional function, whose local objective functions are μ-strongly convex (Definition 1) and L-smooth (Definition 2), and satisfy Assumption 4 (heterogeneity), and an initialization point $\mathbf{x}^{(0)}$ such that for any constant learning rate $\eta > 0$, the last-round global parameter $\mathbf{x}^{(R)}$ satisfy*

$$\mathbb{E}\left[F(\mathbf{x}^{(R)}) - F^*\right] = \begin{cases} \Omega\left(\frac{L\zeta^2}{\mu^2 MR^2}\right) & \text{if } R \geq 161\kappa, \\ \Omega\left(\frac{\zeta^2}{\mu MR}\right) & \text{if } R < 161\kappa. \end{cases}$$

Theorem 18 (Theorem 3.3 and Corollary 3.5 in Cha et al. (2023)) *Under the same conditions of Theorem 17 (unless explicitly stated), there exist a multi-dimensional global objective function and an initialization point, such that for $\eta \leq \frac{1}{161LM}$, the arbitrary weighted average global parameters $\bar{\mathbf{x}}^{(R)}$ satisfy the lower bounds:*

***Strongly convex**: If $R \geq 161\kappa$ and $\kappa \geq 2415$, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{L\zeta^2}{\mu^2 MR^2}\right).$$

***General convex**: If $R \geq 161^3 \max\left\{\frac{\zeta}{LM^{1/2}D}, \frac{L^2MD^2}{\zeta^2}\right\}$, then*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \Omega\left(\frac{\left(L\zeta^2 D^4\right)^{1/3}}{M^{1/3}R^{2/3}}\right).$$

Appendix G. Comparison with Malinovsky et al. (2023)

In this section, we compare the results in Malinovsky et al. (2023) with ours. Notably, the local solver in Malinovsky et al. (2023) is SGD-RR, so they assume that each local component functions f_m is L-smooth, and use the technique of Shuffling Variance (Mishchenko et al., 2020). Since the bounds in Malinovsky et al. (2023) are for strongly convex cases, we next compare their bound with ours in the strongly convex case. For comparison, we restate their Theorem 6.1 and our Theorem 3 in Corollary 19 and Corollary 20. Since Malinovsky et al. (2023)'s local solver is SGD-RR, while our local solver is SGD, for convenience and fairness, we next only compare the heterogeneity terms. As shown in Corollary 19 and Corollary 20, the upper bounds of Malinovsky et al. (2023) almost match ours, except an advantage of MK on the first term. This is because they use the advanced technique of Shuffling Variance.

Corollary 19 (Corollary of Malinovsky et al. (2023)'s Theorem 6.1) *Letting $R = M$, $T = R$, $\gamma = \eta$, $N = K$, $\tilde{\sigma}_\star = \zeta_*$, $\sigma_\star = 0$ and $C = 1$ in Malinovsky et al. (2023)'s Theorem 6.1, we can get the upper bounds for SFL in our notations:*

$$\mathbb{E}\|\mathbf{x}^{(R)} - \mathbf{x}^*\|^2 = \mathcal{O}\left(D^2 \exp\left(-\mu\tilde{\eta}R\right) + \frac{\tilde{\eta}^2 L\zeta_*^2}{\mu M}\right),$$
$$\mathbb{E}\|\mathbf{x}^{(R)} - \mathbf{x}^*\|^2 = \tilde{\mathcal{O}}\left(D^2 \exp\left(\frac{-\mu \boldsymbol{MK} R}{L}\right) + \frac{L\zeta_*^2}{\mu^3 MR^2}\right),$$

where $\tilde{\eta} = \eta MK \lesssim \frac{1}{L} \cdot MK$. Here we set $N = K$, since the number of local steps equals the size of the local data set in Malinovsky et al. (2023).

Corollary 20 (Corollary of Theorem 3) *For the purpose of comparison, we consider the bound of $\mathbb{E}\|\mathbf{x}^{(R)} - \mathbf{x}^*\|^2$ here instead of $\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right]$:*

$$\mathbb{E}\|\mathbf{x}^{(R)} - \mathbf{x}^*\|^2 = \mathcal{O}\left(D^2 \exp\left(-\mu\tilde{\eta}R\right) + \frac{\tilde{\eta}^2 L\zeta_*^2}{\mu M}\right),$$
$$\mathbb{E}\|\mathbf{x}^{(R)} - \mathbf{x}^*\|^2 = \tilde{\mathcal{O}}\left(D^2 \exp\left(\frac{-\mu R}{L}\right) + \frac{L\zeta_*^2}{\mu^3 MR^2}\right),$$

where $\tilde{\eta} = \eta MK \lesssim \frac{1}{L}$.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: optimal rates without component convexity and large epoch requirements. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *International Conference on Machine Learning (ICML)*, 2023.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2011.

Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 2018.

Cheng Chen, Ziyi Chen, Yi Zhou, and Bhavya Kailkhura. FedCluster: Boosting the convergence of federated learning via cluster-cycling. In *IEEE International Conference on Big Data (Big Data)*, 2020.

Zhikun Chen, Daofeng Li, Rui Ni, Jinkang Zhu, and Sihai Zhang. FedSeq: A hybrid federated learning framework based on sequential in-cluster training. *IEEE Systems Journal*, 2023.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *International Conference on Machine Learning (ICML)*, 2023.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. CINIC-10 is not ImageNet or CIFAR-10. *arXiv preprint arXiv:1810.03505*, 2018.

Hubert Eichner, Tomer Koren, Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2019.

Yansong Gao, Minki Kim, Chandra Thapa, Sharif Abuadbba, Zhi Zhang, Seyit Camtepe, Hyoungshick Kim, and Surya Nepal. Evaluation and optimization of distributed machine learning techniques for internet of things. *IEEE Transactions on Computers*, 2021.

Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local SGD) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 2018.

Samuel Horváth, Maziar Sanjabi, Lin Xiao, Peter Richtárik, and Michael Rabbat. FedShuffle: Recipes for better use of local work in federated learning. *Transactions on Machine Learning Research (TMLR)*, 2022.

Yixing Huang, Christoph Bert, Ahmed Gomaa, Rainer Fietkau, Andreas Maier, and Florian Putz. An experimental survey of incremental transfer learning for multicenter collaboration. *IEEE Access*, 2024.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up federated averaging via extrapolation. In *International Conference on Learning Representations (ICLR)*, 2023.

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 (ICML)*, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, 2020.

Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *International Conference on Machine Learning (ICML)*, 2024.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Technical report*, 2009.

Jin-woo Lee, Jaehoon Oh, Sungsu Lim, Se-Young Yun, and Jae-Gil Lee. Tornadoaggregate: Accurate and scalable federated learning via the ring-based architecture. *arXiv preprint arXiv:2012.03214*, 2020.

Hanmin Li and Peter Richtárik. On the convergence of FedProx with extrapolation and inexact prox. In *International OPT Workshop on Optimization for Machine Learning at NeurIPS 2024*, 2024.

Hanmin Li, Kirill Acharya, and Peter Richtárik. The power of extrapolation in federated learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations (ICLR)*, 2020.

Yipeng Li and Xinchen Lyu. Convergence analysis of sequential federated learning on heterogeneous data. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Yucheng Lu, Si Yi Meng, and Christopher De Sa. A general analysis of example-selection for stochastic gradient descent. In *International Conference on Learning Representations (ICLR)*, 2022.

Grigory Malinovsky, Samuel Horváth, Konstantin Pavlovich Burlachenko, and Peter Richtárik. Federated learning with regularized client participation. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities workshop at ICML 2023*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning (ICML)*, 2022a.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning (ICML)*, 2022b.

Cristinel Mortici. On Gospers formula for the Gamma function. *Journal of Mathematical Inequalities*, 2011.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning (ICML)*, 2019.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience, 1983.

Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten Van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research (JMLR)*, 2021.

Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities workshop at ICML 2023*, 2023.

Kumar Kshitij Patel, Margalit Glasgow, Ali Zindari, Lingxiao Wang, Sebastian U Stich, Ziheng Cheng, Nirmit Joshi, and Nathan Srebro. The limits and potentials of local SGD for distributed heterogeneous learning with intermittent communication. In *Conference on Learning Theory (COLT)*, 2024.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning (ICML)*, 2020.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Pavlovich Burlachenko, and Peter Richtárik. Federated optimization algorithms with random reshuffling and gradient compression. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities workshop at ICML 2023*, 2023.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory (COLT)*, 2020.

Itay Safran and Ohad Shamir. Random shuffling beats SGD only after many epochs on ill-conditioned problems. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Chandra Thapa, Pathum Chamikara Mahawaga Arachchige, Seyit Camtepe, and Lichao Sun. SplitFed: When federated learning meets split learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *Transactions on Machine Learning Research (TMLR)*, 2024.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning (ICML)*, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xingrun Yan, Shiyuan Zuo, Rongfei Fan, Han Hu, Li Shen, Puning Zhao, and Yong Luo. Sequential federated learning in hierarchical architecture on non-IID datasets. *arXiv preprint arXiv:2408.09762*, 2024.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Liangqi Yuan, Yunsheng Ma, Lu Su, and Ziran Wang. Peer-to-peer federated continual learning for naturalistic driving action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Liangqi Yuan, Ziran Wang, Lichao Sun, S Yu Philip, and Christopher G Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, 2024.

Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations (ICLR)*, 2022.

Riccardo Zaccone, Andrea Rizzardi, Debora Caldarola, Marco Ciccone, and Barbara Caputo. Speeding up heterogeneous federated learning with sequentially trained superclients. In *International Conference on Pattern Recognition (ICPR)*, 2022.