CONVERGENCE ANALYSIS OF SPLIT LEARNING ON NON-IID DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Split Learning (SL) is one promising variant of Federated Learning (FL), where the AI model is split and trained at the clients and the server collaboratively. By offloading the computation-intensive portions to the server, SL enables efficient model training on resource-constrained clients. Despite its booming applications, SL still lacks rigorous convergence analysis on non-IID data, which is critical for hyperparameter selection. In this paper, we first prove that SL exhibits an $\mathcal{O}(1/\sqrt{T})$ convergence rate for non-convex objectives on non-IID data, where T is the number of total steps. By comparing the convergence analysis and experimental results, SL can outperform FL in terms of convergence rate (w.r.t. per-client training/communication rounds, and hence, the computation efficiency) and exhibit comparable accuracy to FL on mildly non-IID data. In contrast, FL prevails on highly non-IID data.

1 INTRODUCTION

Split Learning (SL) (Gupta & Raskar, 2018; Vepakomma et al., 2018) is one promising variant of Federated Learning (FL) (McMahan et al., 2017). To reduce the computational burden of each client in FL, SL splits the AI model to be trained at the clients and server separately (where the computation-intensive portions are typically offloaded to the server). The computation effectiveness of SL is critical for the model training on resource-constrained (e.g., IoT) devices. SL is regarded as one of the enabling technologies for edge intelligence in future networks (Zhou et al., 2019).

The comparisons of FL and SL are of practical interest for the design and deployment of intelligent networks. Existing studies have focused on various aspects for their comparisons (Thapa et al., 2020; Gao et al., 2020; 2021), e.g., in terms of learning performance (Gupta & Raskar, 2018), computation efficiency (Vepakomma et al., 2018), communication overhead (Singh et al., 2019), and privacy issues (Thapa et al., 2021). For example, with the emphasis on the learning performance comparison, Gao et al. (2020; 2021) find that SL exhibits (i) faster convergence speed than FL under IID data in terms of communication rounds; (ii) better learning performance under imbalanced data; (iii) worse learning performance under (extreme) non-IID data, etc. The difference arises from the distinct process of model updates of FL and SL. In particular, FL takes the average of the local model parameters at the end of each round; SL only trains the clients in sequence and does not average the client updates. Figure 1 plots the client drift (Karimireddy et al., 2020; Wang et al., 2020; Li et al., 2022) of FL and SL under both IID and non-IID data to visualize the update process. Under the IID setting, SL approach the global optima $\mathbf{x}^*$ faster than FL given the sequential training mechanism. In contrast, under the non-IID setting, SL may deviate from the global optima for the same reason.

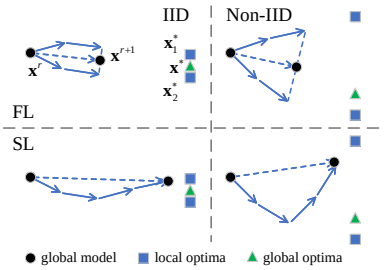


Figure 1: Illustration of the model updates of FL and SL for 2 clients and 2 local update steps during one communication round.

Convergence analysis is critical for the performance comparison between SL and FL. Specifically, a rigorous analysis is of paramount importance for the vital research questions raised by (Gao et al., 2020) (which are only empirically evaluated but remain unsolved in theory), i.e., RQ1-"*What factors affect SL performance?*" and RQ2-"*In which setting will the SL performance outperform FL?*". A

wealth of work has analyzed the convergence of FL in the cases of IID (Stich, 2018; Zhou & Cong, 2017; Khaled et al., 2020), non-IID (Li et al., 2020; 2019; Khaled et al., 2020; Karimireddy et al., 2020) and unbalanced data (Wang et al., 2020). However, with the distinct update process, the convergence analysis of SL has yet to be solved on non-IID data. To this end, this paper aims to derive a rigorous convergence results of SL itself and draw the comparison results of FL and SL theoretically.

Main contributions. The main contributions can be summarized with respect to the two research questions above:

- We prove the convergence of SL on non-IID data with the standard assumptions used in FL literature [1] with a convergence rate of $\mathcal{O}(1/\sqrt{T})$ in Section 4.2. By this, we find that the convergence of SL is affected by factors such as data heterogeneity and the number of local update steps, etc. Experimental results verify the analysis results empirically in Section 5.2. To the best of our knowledge, this work is the first to give the convergence analysis of SL on non-IID data.
- Combining the theory in Section 4.3 and experiments presented in Section 5.3, we compare FL[2] and SL in theory and in practice. The results demonstrates that as compared to FL, SL exhibits a higher convergence rate in terms of communication rounds and shows comparable accuracy on mildly non-IID data. In contrast, FL prevails on highly non-IID data. The results can bridge the gap between FL and SL and guide the deployment of these two in real-world applications.

2 Preliminaries and Algorithm of Split Learning

As two of the most popular distributed learning frameworks, both FL and SL aim to train the global model from distributed datasets. The optimization problem of FL and SL with N clients can be given by

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}) = \sum_{i=1}^{N} p_i f_i(\mathbf{x}) \right\}, \tag{1}$$

where $p_i = n_i/n$ is the ratio of local samples at client i (n and n_i are the sizes of the overall dataset $\mathcal{D}$ and local dataset $\mathcal{D}_i$ at client i, respectively), $\mathbf{x}$ is the model parameters, $f(\mathbf{x})$ is the global objective, $f_i(\mathbf{x})$ denotes the local objective function on client i. In particular, $f_i(\mathbf{x}) \overset{\text{def}}{=} \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(\mathbf{x};\xi_i)] = \frac{1}{n_i}\sum_{\xi_i\in\mathcal{D}_i} f_i(\mathbf{x};\xi_i)$, where ξ_i represents a data sample from the local dataset $\mathcal{D}_i$.

Different from FL, SL is required to update the local models of the clients in a relay-based manner at each communication round. The global model of SL is updated according to

$$\mathbf{x}^{r+1} = \mathbf{x}^r + \eta_g(\mathbf{x}_{|\mathcal{S}|}^{r,K} - \mathbf{x}^{r,0}) = \mathbf{x}^r - \eta_g \sum_{i=1}^{|\mathcal{S}|}\sum_{k=0}^{K-1} \eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k}), \tag{2}$$

where $\mathbf{g}_i(\mathbf{x})$ denotes the stochastic gradients over a mini-batch of b samples and parameter $\mathbf{x}$ of client i, $\mathbf{x}^r$ represents the global model in the r-th round, and $\mathbf{x}_i^{r,k}$ represents the local model (parameters) of client i after the k-th local update in the r-th communication round. $\mathcal{S}$ is a subset selected for model update of all clients, K is the number of local update steps ($K = En_i/b$, E is the number of local epochs) and η_l is the local learning rate. Note in Eq. (2) that we propose adding the global learning rate η_g against the vanilla SL algorithm (Gupta & Raskar, 2018; Vepakomma et al., 2018; Thapa et al., 2020; 2021). The global learning rate design mechanism is originally developed in the FL setting (Karimireddy et al., 2020; Reddi et al., 2020; Wang et al., 2020) to reduce the client drift, and can be readily adopted in SL for the same function.

The process of SL with the global learning rate is summarized in Algorithm 1. Note that Algorithm 1 can operate in both the centralized and peer-to-peer mode by having the server and the first client execute the global update separately in step 15. For brevity, we have omitted some unconcerned

[1] We only show the convergence for non-convex objective functions here since SL is now often used in large deep learning models whose objective functions are possibly non-convex. Nevertheless, similar methods can be used to get the convergence for general convex functions and strongly convex functions.

[2] FedAvg is used for comparison in this paper.

details of SL in Algorithm 1, e.g., using $\mathbf{x}$ to denote the parameters of the whole model without differentiating the parameters in clients and the server (step 6, 12, 15), the security and privacy settings and so on. More details about SL can be found in (Gupta & Raskar, 2018; Thapa et al., 2021).

Algorithm 1 Split Learning with the Global Learning Rate

1: Input: $\mathbf{x}^0$, η_g, η_l
2: **for** communication round $r = 0, \ldots, R-1$ **do**
3: Sample a subset $\mathcal{S}$ of clients and determine their update order
4: **for** client index $i = 1, \ldots, |\mathcal{S}|$ **do**
5: Client i requests the latest model
6: Initialize local model: $\mathbf{x}_i^{r,0} \leftarrow \mathbf{x}_{i-1}^{r,K}$ (or $\mathbf{x}^{r,0}$) ▷ client and server
7: **for** local update step $k = 0, \ldots, K-1$ **do**
8: Forward propagation and send activations to the server ▷ client
9: Forward propagation with activations from the client ▷ server
10: Back-propagation and send gradients to the client ▷ server
11: Back-propagation with gradients from the server ▷ client
12: Local model updates: $\mathbf{x}_i^{r,k+1} \leftarrow \mathbf{x}_i^{r,k} - \eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k})$ ▷ client and server
13: **end for**
14: **end for**
15: Global model updates: $\mathbf{x}^{r+1,0} \leftarrow \mathbf{x}^{r,0} + \eta_g(\mathbf{x}_{|\mathcal{S}|}^{r,K} - \mathbf{x}^{r,0})$ ▷ client and server
16: **end for**

3 Related work

Variants of SL. SL is deemed as a promising paradigm for distributed model training at resource-constrained devices (e.g., IoT), given its computational efficiency on the client side. Most existing works focus on reducing the training delay arising from the relay-based training manner in the multi-user scenario. SplitFed (Thapa et al., 2020) is one popular model parallel algorithm that combines the strengths of FL and SL, where each client has one corresponding instance of server-side model to form a pair, each pair trains the model in parallel and then aggregate the model parameters before the next round. SplitFedv2(Thapa et al., 2020), SplitFedv3(Gawali et al., 2021) and SFLG(Gao et al., 2021) are the variants of SplitFed. In particular, SFLG (Gao et al., 2021) is one generalized framework of SplitFed, which improves the flexibility and scalability of SL greatly. The clients are allocated to different servers to form multiple groups. The training in the same group is identical to SplitFedv2 and the training between the groups is in parallel as in SplitFedv1.

Convergence analysis of SL. The convergence analysis of SL is identical to standard Minibatch-SGD on IID data (Wang et al., 2022; Park et al., 2021), so some convergence properties of Minibatch SGD is applied to SL too. The algorithm in Han et al. (2021) reduces the latency and downlink communication on SplitFed by adding auxiliary networks at client-side to generate the local loss for model updating. Their convergence analysis combines the analysis of Belilovsky et al. (2020) and FedAvg. Wang et al. (2022) proposed FedLite to reduce the uplink communication overhead by compressing activations with product quantization and provided the convergence analysis of FedLite. However, their convergence recovers that of Minibatch SGD when there is no quantization. Furthermore, we find that (Woodworth et al., 2020a;b) compared the convergence of distributed Minibatch SGD and local SGD under homogeneous and heterogeneous settings, respectively. To differentiate our work, we show how these algorithms operate in Table 1. As we can see, FL and SL make K times more updates than Minibatch SGD; client i in SL starts local updates from $\mathbf{x}_{i-1}^{r,K}$ instead of $\mathbf{x}^r$ in FL. As a result, the convergence of SL on non-IID data is still lacking.

4 Convergence Analysis of SL

4.1 Assumptions

We make the following standard assumptions on the local objective functions $\{f_i(\mathbf{x})\}_{i=1}^N$.

Table 1: The update rules of Minibatch SGD, FL and SL. We use the descriptions of Local SGD and Minibatch SGD in (Woodworth et al., 2020b) and notations in Section 2. $\xi_i^{r,k} \sim \mathcal{D}_i$ denotes one random sample from client i at the $k+1$-th local update in the r-th communication round. We stress the difference between Minibatch SGD and FL, SL with blue; between FL and SL with red.

Algorithm	Global update	Local update
Minibatch SGD	$\mathbf{x}^{r+1} = \mathbf{x}^r - \eta_l \frac{1}{NK} \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}^r; \xi_i^{r,k})$	-
Local SGD (or FL)	$\mathbf{x}^{r+1} = \mathbf{x}^r - \eta_l \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k}; \xi_i^{r,k})$	$\mathbf{x}_i^{r,k+1} = \mathbf{x}^r - \eta_l \sum_{t=0}^k \nabla f_i(\mathbf{x}_i^{r,t}; \xi_i^{r,t})$
SL	$\mathbf{x}^{r+1} = \mathbf{x}^r - \eta_l \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k}; \xi_i^{r,k})$	$\mathbf{x}_i^{r,k+1} = \mathbf{x}_{i-1}^{r,K} - \eta_l \sum_{t=0}^k \nabla f_i(\mathbf{x}_i^{r,t}; \xi_i^{r,t})$

Assumption 1 (L-smooth)**.** *Each local objective function f_i is L-smooth, $i \in \{1, 2, \ldots, N\}$, i.e., there exists a constant $L > 0$ such that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$.*

Assumption 2 (Unbiased Gradient and Bounded Variance)**.** *For each client i, $i \in \{1, 2, \ldots, N\}$, the stochastic gradient $\mathbf{g}_i(\mathbf{x}) \stackrel{def}{=} \nabla f_i(\mathbf{x}; \xi_i)$ is unbiased $\mathbb{E}\mathbf{g}_i(\mathbf{x}) = \nabla f_i(\mathbf{x})$ and has bounded variance $\mathbb{E}_{\xi_i}[\|\mathbf{g}_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$.*

Assumption 3 (Bounded Dissimilarity)**.** *There exist constants $B \geq 1$ and $G \geq 0$ such that $\frac{1}{N} \sum_{i=1}^N \|\nabla f_i(\mathbf{x})\|^2 \leq B^2 \|\nabla f(\mathbf{x})\|^2 + G^2$. In the IID case, $B = 1$, $G = 0$. In the non-IID case, B and G measure the heterogeneity of data distribution.*

4.2 Convergence Result and Discussion

Without loss of generality, all the clients participate in the SL process ($|\mathcal{S}| = N$) and the unweighted global objective function $f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x})$ is adopted. Note that the results in the unweighted case can be readily extended to the weighted case in Eq.(1).

Client drift in SL is upper bounded, as shown in Lemma 4. However, as stated in Section 2, SL differs from FL that the clients need to update their models in sequence, which renders the drift from the initial model (of round r) $\mathbf{x}^r$ to the present local model $\mathbf{x}_i^{r,k}$ influenced by previous clients (see Table 1). By bounding the drift ($\mathbf{x}_i^{r,k} - \mathbf{x}^r$) and following the proof (Karimireddy et al., 2020; Wang et al., 2020), we bridge the gap between FL and SL, giving the convergence results of SL (details are in Appendix A), as follows:

Theorem 1. *Let Assumptions 1, 2 and 3 hold. Suppose that the local learning rate satisfies $\eta \leq \min\left\{\frac{\eta_g}{8KL}, \frac{1}{NKL}, \frac{\eta_g}{2\sqrt{3}NKLB}\right\}$. For Algorithm 1, it holds that*

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2] \leq \frac{4[f(\mathbf{x}^0) - f(\mathbf{x}^*)]}{\eta NKR} + \frac{12\eta^2 N^2 K^2 L^2 G^2}{\eta_g^2} + \frac{6\eta^2 N^2 K L^2 \sigma^2}{\eta_g^2} + 2\eta L \sigma^2, \quad (3)$$

where $\bar{\mathbf{x}}^R = \frac{1}{R} \sum_{r=0}^{R-1} \mathbf{x}^r$ and $\eta = \eta_g \eta_l$ are the averaged global model over the R rounds and the multiplication of local and global learning rates, respectively.

Corollary 1. *Choose $\eta = \frac{1}{L\sqrt{T}}$ and apply the result of Theorem 1. For sufficiently large $T = NKR$, it holds that*

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2] \leq \mathcal{O}\left(\frac{L[f(\mathbf{x}^0) - f(\mathbf{x}^*)]}{\sqrt{T}} + \frac{N^2 K^2 G^2}{\eta_g^2 T} + \frac{N^2 K \sigma^2}{\eta_g^2 T} + \frac{\sigma^2}{\sqrt{T}}\right). \quad (4)$$

In Theorem 1 (also in Corollary 1 and Theorem 2), the upper bound of $\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2]$ consists of three types of terms: (a) *optimization terms*, depending on the initial error $\|\mathbf{x}^0 - \mathbf{x}^*\|$, i.e., the first term on the right-hand side in Eq.(3); (b) *stochastic terms*, depending on σ, i.e., the third and forth term (Stich & Karimireddy, 2019); (c) *heterogeneity terms*, depending on the heterogeneity assumption (Assumption 3 in our setting), i.e., the second term.

We can see that the result demonstrates the relationships between convergence and factors such as the number of clients, the data heterogeneity, the global/local learning rates, the local update steps. According to the result, a large η means high rate that the optimization term decreases at but

causes large heterogeneity term and stochastic terms. Next, we discuss the convergence rate and the influence factors of SL in detail.

Convergence rate. By Corollary 1, for sufficiently large T, the convergence rate is determined by the optimization term and the last term (see Eq.(4)), resulting in a convergence rate of $\mathcal{O}(1/\sqrt{T})$.

Effect of η_g. The global learning rate η_g can be used to reduce the client drift without hurting the progress of SL. A large η_g relaxes the constraints on η and reduces the heterogeneity term, hence improving the convergence rate. For example, by Corollary 1, SL shows $\mathcal{O}(\frac{LF+\sigma^2}{\sqrt{T}} + \frac{N^2K^2G^2+N^2K\sigma^2}{T})$ convergence rate if $\eta_g = 1$ (F denotes $f(\mathbf{x}^0) - f(\mathbf{x}^*)$); while it shows a convergence rate of $\mathcal{O}(\frac{LF+\sigma^2}{\sqrt{T}} + \frac{K^2G^2+K\sigma^2}{T})$ if $\eta_g = N$.

Effect of data heterogeneity. According to Theorem 1, When on highly non-IID data (B and G are large), a small η is required for the convergence of SL (see the condition of Theorem 1). In addition, the heterogeneity term also increases. As a result, large data heterogeneity harms the convergence of SL, which is consistent with the intuition.

Effect of K. K is the maximal number of local update steps among clients. Two cases are considered: (i) R is fixed and (ii) T is fixed. In the first case, as K increases, the first term of Eq.(3) decreases and the second and third terms increase. This implies that one needs to meticulously design the value of K for the best performance. In fact, large value of K forces SL prone to converge to local optima, which is not preferable on significantly heterogeneous data. From Eq.(3), we can also get that large data heterogeneity makes the optimal K small. This property is analogous to FL (McMahan et al., 2017; Li et al., 2019). In the second case, we expect K is as large as possible to reduce the communication overhead without hurting the convergence rate. Thus we can get the optimal $K = \mathcal{O}(T^{\frac{1}{4}}N^{-1})$ ($\eta_g = 1$) that keeps the $\mathcal{O}(1/\sqrt{T})$ convergence rate by Corollary 1.

4.3 Comparison between FL and SL

In this section, we first give the convergence result of FL based on our setting (Theorem 2), and then compare the results of FL and SL to answer the second question "*In which setting will the SL performance outperform FL?*" theoretically.

Theorem 2. *Let Assumptions 1, 2 and 3 hold. Suppose that the local learning rate satisfies* $\eta \leq \min\left\{\frac{\eta_g}{8KL}, \frac{1}{KL}, \frac{\eta_g}{2\sqrt{3}KLB}\right\}$. *For FL*[3]*, it holds that*

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2] \leq \frac{4[f(\mathbf{x}^0) - f(\mathbf{x}^*)]}{\eta KR} + \frac{12\eta^2K^2L^2G^2}{\eta_g^2} + \frac{6\eta^2KL^2\sigma^2}{\eta_g^2} + \frac{2\eta L\sigma^2}{N}, \tag{5}$$

where $\bar{\mathbf{x}}^R = \frac{1}{R}\sum_{r=0}^{R-1}\mathbf{x}^r$ *and* $\eta = \eta_g\eta_l$.

Theorem 2 is a modified version of Theorems I and V in (Karimireddy et al., 2020), one of the best known convergence results for the non-convex functions of FL, which makes our comparison persuasive. Choosing $\eta = \frac{\sqrt{N}}{L\sqrt{T}}$ for Theorem 2 provides the $\mathcal{O}(1/\sqrt{NT})$ convergence rate, the linear speedup in FL. Then we can get the optimal $K = \Theta(T^{\frac{1}{4}}N^{-\frac{3}{4}})$ directly when ignoring the effect of η_g (i.e., setting $\eta_g = 1$). We summarize the results of Theorem 1 and Theorem 2 in Table 2. The comparisons are as follows:

SL shows worse guarantee than FL in terms of iterations/steps. The comparison indicates that SL has a worse convergence guarantee than FL in terms of iterations, especially when N is very large: the optimization term is equal but the heterogeneity term and stochastic terms of $\mathcal{B}_{SL}$ are larger if $\eta_{SL} = \eta_{FL}$ and both two satisfy the corresponding conditions. This is intuitive since models are trained in sequence in SL but in parallel in FL (Thapa et al., 2020; Han et al., 2021).

Comparisons in terms of communication rounds. We next compare FL and SL in terms of communication rounds. Note that this comparison (running for the same R) is fair given the same total computation costs (including the computation cost on client-side and server-side). Then we have:

[3]FedAvg with η_g (Karimireddy et al., 2020)

- **SL enforces stronger restrictions on learning rate η_{SL} than FL, especially on highly non-IID data.** η_{SL} has a tougher condition than η_{FL} (see the second and fourth row of Table 2), which indicates SL is more sensitive to the heterogeneity of data.
- **SL shows a higher convergence rate than FL on mildly non-IID data with a slight variance of the stochastic gradients σ.** To make a comparison, we need to choose appropriate η for both. Note that on mildly non-IID data, the restrictions of η_{SL} are relaxed and the difference between the upper bounds of η_{FL} and η_{SL} are small. Here we assume that $\eta_{SL} \approx \eta_{FL} = \frac{\sqrt{N}}{L\sqrt{KR}}$ can be chosen. Then considering small values of the variance and data heterogeneity (i.e., σ, B and G are small), the guarantees are determined by the optimization terms. With the same setting (K, R, F), the optimization term of SL $\mathcal{O}(\frac{LF}{N\sqrt{NKR}})$ is smaller than that of FL $\mathcal{O}(\frac{LF}{\sqrt{NKR}})$ considerably when N is large, leading to a higher convergence rate of SL. However, the stochastic and heterogeneity terms can render SL more unstable than FL. The convergence rate gap will reduce as R (or K) increases since the optimization term will have very limited reduction with big R (or K).
- **SL can match the guarantee of FL by choosing smaller η_{SL} on highly non-IID data.** First, on highly non-IID data, the tougher conditions of η_{SL} restrict its value heavily. Nevertheless, we find that the guarantee of SL can still match that of FL by choosing $\eta_{SL} = \Theta(\eta_{FL}/N)$.

Table 2: Comparison of convergence results between FL and SL for non-convex functions. The results of FL and SL are based on Theorem 2 and Theorem 1 respectively. The upper bounds of $\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2]$ are in the first and third rows. The corresponding conditions of η are in the second and forth rows. For a particular algorithm A, we use η_A and $\mathcal{B}_A$ to denote η and the upper bound of $\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^R)\|^2]$ respectively.

FL	$\frac{4[f(\mathbf{x}^0)-f(\mathbf{x}^*)]}{\eta KR} + \frac{12\eta^2K^2L^2G^2}{\eta_g^2} + \frac{6\eta^2KL^2\sigma^2}{\eta_g^2} + \frac{2\eta L\sigma^2}{N}$
	$\eta_{FL} \leq \min\left\{\frac{\eta_g}{8KL}, \frac{1}{KL}, \frac{\eta_g}{2\sqrt{3}KLB}\right\}$
SL	$\frac{4[f(\mathbf{x}^0)-f(\mathbf{x}^*)]}{\eta NKR} + \frac{12\eta^2N^2K^2L^2G^2}{\eta_g^2} + \frac{6\eta^2N^2KL^2\sigma^2}{\eta_g^2} + 2\eta L\sigma^2$
	$\eta_{SL} \leq \min\left\{\frac{\eta_g}{8KL}, \frac{1}{NKL}, \frac{\eta_g}{2\sqrt{3}NKLB}\right\}$

5 Experiments

With the emphasis on the convergence of SL, we conduct the SL experiments in a virtual split setup (i.e., the model portions are virtually split and stored in the memory). For fair comparisons on the resource-constrained devices, the convergence rate is evaluated in terms of the communication rounds (i.e., the per-client training rounds). In the following, we first demonstrate the detailed experimental setup in Section 5.1, evaluate the efficiency of SL in terms of data heterogeneity, local update steps in Section 5.2, and illustrate the comparison results of the SL and FL in Section 5.3. More experimental details can be found in Appendix B.

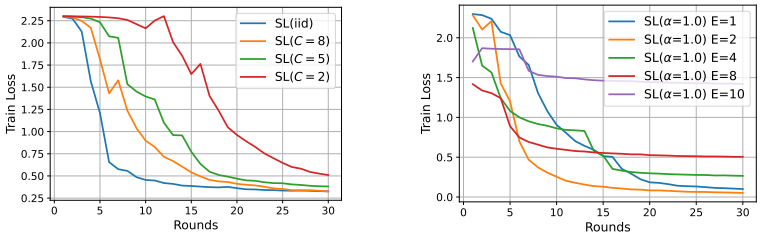
5.1 Experimental Setup

Datasets and models. We adopt the following setups: (i) training LeNet-5 (LeCun et al., 1998) on the MNIST dataset (LeCun et al., 1998); (ii) training LeNet-5 on the Fashion-MNIST dataset (Xiao et al., 2017); (iii) training VGG-11 (Simonyan & Zisserman, 2014) on the CIFAR-10 dataset (Krizhevsky et al., 2009). For SL, the LeNet-5 is split after the second 2D MaxPool layer, with 6% of the entire model size retained in client; the VGG-11 is split after the third 2D MaxPool layer, with 10% of the entire model size at the client. Ideally, the split layer position has no effect on the performance of SL (Wang et al., 2022).

Data distribution. Both IID and non-IID datasets are considered in the experiments. For the non-IID setting, we adopt two mechanisms to generate non-IID data: (i) using a Dirichlet distribution Dir(α) to generate mildly non-IID data, where a smaller α indicates higher data heterogeneity (Hsu

et al., 2019; Zhu et al., 2021); (ii) using a similar mechanism like (McMahan et al., 2017; Zhao et al., 2018) to generate the highly (pathological) non-IID data, where we first divide all the training samples into shards (all the shards have the same size and the most of them only contain samples from one single class), and then assign each client the same number of shards. For the pathological non-IID partition, we use three distributions, where most clients contain samples from 2, 5, 8 classes. $C = 2, 5, 8$ represent these three partitions respectively. The original test sets are used to evaluate the generalization performance of the global model (test accuracy) after each training round.

5.2 Experimental Results of SL



(a) Effect of data heterogeneity

(b) Effect of K

Figure 2: Results of SL on MNIST dataset. Training loss curves for (a) different data heterogeneity; (b) different number of local epochs.

In this section, we study the effects of data heterogeneity local update steps using MNIST dataset. The training samples are assigned to 10 clients. For data heterogeneity, we use four distribution: IID, $C = 8$, $C = 5$, $C = 2$. For K, we use five settings, $E = 1$, $E = 2$, $E = 4$, $E = 8$, $E = 10$ with $\text{Dir}_{10}(1.0)$ distribution. The results are shown in Figure 2. The details and results on Fashion-MNIST and CIFAR-10 are deferred to Appendix B.1.

Effect of data heterogeneity. As shown in Figure 2a, the training loss curve of IID distribution is the lowest and most stable. When the data heterogeneity increases (C decreases from 8 to 2), it has shown that: (i) SL converges slower; (ii) the training loss turns larger, especially the red curve SL($C = 2$); (iii) the curve becomes more unstable. The phenomenon is in accordance with our analysis that large data heterogeneity harms the convergence of SL in Section 4.2.

Effect of K. Figure 2b shows the training loss in terms of communication rounds R. Even though SL shows a higher speed at the early stage when E is large (see $E = 4, 8, 10$), it may not reach the smallest training loss (when $E = 2$, SL shows the best), which verifies that optimal K exists and suitable K can speed up the convergence rate. Over-large K can even harm the convergence rate (see the curves $E = 8$, $E = 10$).

5.3 Empirical Comparison between FL and SL

We have compared FL and SL *in the worst case* theoretically (see Section 4.2). The question arises that *how about the learning performance of SL in practice compared to FL?*. In the previous studies, the same learning rates are used to evaluate the performances of SL and FL (Gao et al., 2020; 2021). However, theoretical analyses in Section 4.2 show that the appropriate learning rate for SL may deviate from that of FL. To this end, we first test the performance of SL and FL under different (meticulously adjusted) learning rates for fair comparisons. The learning rates of FL and SL are selected from $\{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ (see Table 3, 4, and Appendix B.2 for details).

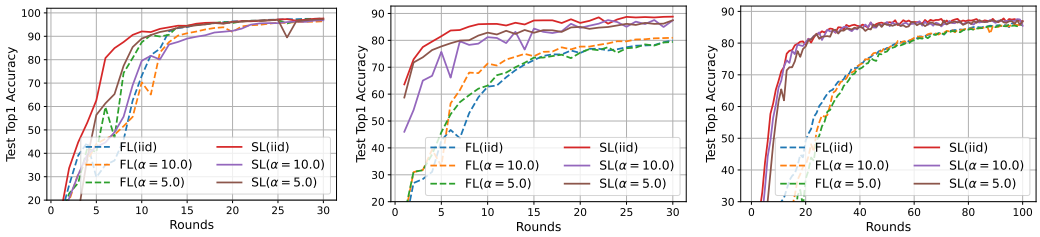
The best and threshold learning rate of SL is smaller than FL. We refer to the learning rate making the best test accuracy and minimal learning rate making the training die as the best learning rate and the threshold learning rate. According to our theory, the tougher condition indicates the smaller threshold learning rate of SL and the properties of Eq.3, Eq.5 show a smaller best learning rate. To verify this point, we use the "best" learning rate, which makes the best test accuracy among the learning rates we choose, to represent the best test accuracy. The "threshold" learning rate is similar. As shown in Table 3 and 4, the "best" and "threshold" learning rates of SL are smaller than FL (see "Best" and "Min" columns in Table 3, Table 4), especially on highly non-IID data, which

verify our intuition. Besides, we note that the "best" and "threshold" learning rate turn small as the heterogeneity of data becomes large, which is also in accordance with our theory.

Performance comparison on mildly non-IID data. Here we use Dirichlet distribution $\text{Dir}_{10}(10.0)$ and $\text{Dir}_{10}(5.0)$ to generate the mildly non-IID data. The results of FL and SL on mildly non-IID data are presented in Table 3. We can observe that (i) the "best" learning rate of SL can be large, even equal to FL in some cases; (ii) SL can obtain comparable test accuracy to FL. And we see that SL converges faster than FL at the early stage but slow down later in Figure 3. So we have the finding that SL can outperform FL in terms of convergence rate (w.r.t. communication rounds) while preserving comparable test accuracy to FL on mildly non-IID data. In fact, SL on IID data is equivalent to Minibatch-SGD, the studies (McMahan et al., 2017; Zhao et al., 2018) on them can also verify our results.

Table 3: Performance comparisons of FL and SL on mildly non-IID data. We put the "best" learning rate on the *Best* column (the "best" test accuracy (%) is in the parenthesis) and the "threshold" learning rate on the *Min* column. Note that ># means the threshold learning rate larger than #.

	Dataset	Model	N	b		IID		$\alpha = 10.0$		$\alpha = 5.0$	
						Best	Min	Best	Min	Best	Min
1	MNIST	LeNet-5	10	100	FL	0.05 (99.2)	>0.1	0.05 (99.1)	>0.1	0.05 (99.1)	>0.1
					SL	5e-3 (98.9)	>0.1	0.01 (99.1)	>0.1	5e-3 (98.8)	>0.1
2	MNIST	LeNet-5	10	1000	FL	0.1 (98.2)	>0.1	0.05 (97.3)	>0.1	0.1 (98.1)	>0.1
					SL	0.01 (97.8)	>0.1	5e-3 (97.6)	0.05	0.01 (98.1)	0.05
3	Fashion-MNIST	LeNet-5	10	100	FL	0.05 (89.6)	>0.1	0.05 (89.6)	>0.1	0.05 (89.5)	>0.1
					SL	0.01 (90.4)	>0.1	0.01 (88.1)	>0.1	5e-3 (89.6)	>0.1
4	Fashion-MNIST	LeNet-5	10	1000	FL	0.1 (83.0)	>0.1	0.1 (83.6)	>0.1	0.1 (82.6)	>0.1
					SL	0.1 (89.8)	>0.1	0.1 (87.5)	>0.1	0.1 (87.6)	>0.1
5	CIFAR-10	VGG-11	10	100	FL	0.1 (88.9)	>0.1	0.1 (88.5)	>0.1	0.05 (88.4)	>0.1
					SL	0.01 (88.5)	0.1	0.01 (88.2)	0.1	0.01 (88.1)	0.1



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 3: Convergence rate comparison on mildly non-IID data w.r.t. communication rounds. The "best" learning rates are used. (a) $N = 10$, $b = 1000$; (b) $N = 10$, $b = 1000$; (c) $N = 10$, $b = 100$.

Performance comparison on highly (pathological) non-IID data. Even though the theory proves that guarantee of SL can match that of FL by choosing $\eta_{SL} = \Theta(\eta_{FL}/N)$, FL preforms better than SL on highly non-IID data in most cases in our experimental setting as shown in Table 4, where we stress the higher "best" test accuracy between FL and SL under the same condition with different color, i.e., blue for FL and red for SL. Besides, we find that using large mini-batch can improve the generalization performance of SL (see the difference between $b = 10$ and $b = 100$ in Table 3). By the theory, enlarging the mini-batch size b will reduce the variance σ (Stich, 2018) and local update steps K for training through the entire dataset, which decreases the stochastic and heterogeneity terms while preserving the optimization term still small.

6 Conclusion

In this work, we first derived the convergence guarantee of SL for non-convex objectives on non-IID data. The results reveal that the convergence of SL is affected by the factors such as data heterogeneity and local update steps, etc. Furthermore, we compare SL against FL theoretically and empirically, ending up with the conclusions that (i) SL exhibits higher convergence rate than FL with the comparable test accuracy to FL on mildly non-IID data; (ii) FL prevails on highly non-IID data. Our work can bridge the gap between FL and SL and provide deep understanding of these two approaches.

Table 4: Performance comparisons of FL and SL on highly non-IID data with the same annotation as Table 3.

	Dataset	Model	N	b		$C = 8$		$C = 5$		$C = 2$	
						Best	Min	Best	Min	Best	Min
1	MNIST	LeNet-5	10	10	FL	5e-3 (99.2)	0.05	0.01 (99.1)	0.05	5e-3 (98.4)	0.05
					SL	1e-3 (98.8)	0.05	5e-3 (97.8)	0.05	5e-4 (96.3)	0.01
2	MNIST	LeNet-5	10	100	FL	0.05 (99.2)	>0.1	0.05 (99.0)	>0.1	0.01 (97.3)	0.1
					SL	5e-3 (98.7)	0.1	1e-3 (98.0)	0.1	1e-3 (97.2)	5e-3
3	Fashion-MNIST	LeNet-5	10	10	FL	5e-3 (89.3)	0.05	5e-3 (88.5)	0.05	1e-3 (74.8)	0.05
					SL	1e-3 (70.1)	0.05	1e-3 (53.3)	0.05	5e-4 (49.8)	5e-3
4	Fashion-MNIST	LeNet-5	10	100	FL	0.05 (89.3)	>0.1	0.05 (86.3)	>0.1	0.01 (64.3)	0.1
					SL	5e-3 (75.1)	>0.1	1e-3 (57.3)	0.1	5e-4 (66.2)	0.01
5	CIFAR-10	VGG-11	10	10	FL	0.01 (88.4)	0.05	0.01 (86.0)	0.05	0.01 (68.3)	0.05
					SL	5e-4 (81.3)	0.01	5e-4 (72.2)	1e-3	1e-4 (20.0)	5e-4
6	CIFAR-10	VGG-11	10	100	FL	0.05 (88.1)	>0.1	0.1 (85.4)	>0.1	0.05 (69.5)	0.1
					SL	5e-3 (83.8)	0.01	1e-3 (79.0)	5e-3	1e-3 (27.5)	5e-3

REFERENCES

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Decoupled greedy learning of CNNs. In *International Conference on Machine Learning*, pp. 736–745. PMLR, 2020.

James Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

Yansong Gao, Minki Kim, Sharif Abuadbba, Yeonjae Kim, Chandra Thapa, Kyuyeon Kim, Seyit A Camtepe, Hyoungshick Kim, and Surya Nepal. End-to-end evaluation of federated learning and split learning for internet of things. *arXiv preprint arXiv:2003.13376*, 2020.

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.

Manish Gawali, CS Arvind, Shriya Suryavanshi, Harshit Madaan, Ashrika Gaikwad, KN Bhanu Prakash, Viraj Kulkarni, and Aniruddha Pant. Comparison of privacy-preserving distributed deep learning methods in healthcare. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 457–471. Springer, 2021.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

Dong-Jun Han, Hasnain Irshad Bhatti, Jungmoon Lee, and Jaekyun Moon. Accelerating federated learning with split learning on locally generated losses. In *ICML 2021 Workshop on Federated Learning for User Privacy and Data Confidentiality. ICML Board*, 2021.

Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33: 14068–14080, 2020.

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

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

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*, pp. 4519–4529. PMLR, 2020.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978. IEEE, 2022.

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

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

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

Jihong Park, Sumudu Samarakoon, Anis Elgabli, Joongheon Kim, Mehdi Bennis, Seong-Lyun Kim, and Mérouane Debbah. Communication-efficient and distributed learning over wireless networks: Principles and applications. *Proceedings of the IEEE*, 109(5):796–819, 2021.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

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

Abhishek Singh, Praneeth Vepakomma, Otkrist Gupta, and Ramesh Raskar. Detailed comparison of communication efficiency of split learning and federated learning. *arXiv preprint arXiv:1909.09145*, 2019.

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

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Chandra Thapa, Mahawaga Arachchige Pathum Chamikara, Seyit Camtepe, and Lichao Sun. Splitfed: When federated learning meets split learning. *arXiv preprint arXiv:2004.12088*, 2020.

Chandra Thapa, Mahawaga Arachchige Pathum Chamikara, and Seyit A Camtepe. Advancements of federated learning towards privacy preservation: from federated learning to split learning. In *Federated Learning Systems*, pp. 79–109. Springer, 2021.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Jianyu Wang, Hang Qi, Ankit Singh Rawat, Sashank Reddi, Sagar Waghmare, Felix X Yu, and Gauri Joshi. Fedlite: A scalable approach for federated learning on resource-constrained clients. *arXiv preprint arXiv:2201.11865*, 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*, pp. 10334–10343. PMLR, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 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.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.

Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8): 1738–1762, 2019.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 12878–12889. PMLR, 2021.

A PROOF OF RESULTS

A.1 BASIC TECHNICAL LEMMAS

Lemma 1. *$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are N vectors, then*

$$\|\boldsymbol{x}_i + \boldsymbol{x}_j\|^2 \leq 2\|\boldsymbol{x}_i\|^2 + 2\|\boldsymbol{x}_j\|^2 \tag{6}$$

$$\|\boldsymbol{x}_i + \boldsymbol{x}_j\|^2 \leq (1+a)\|\boldsymbol{x}_i\|^2 + (1+\frac{1}{a})\|\boldsymbol{x}_j\|^2 \quad \textit{for any } a > 0, \tag{7}$$

$$\left\|\sum_{i=1}^N \boldsymbol{x}_i\right\|^2 \leq N\sum_{i=1}^N \|\boldsymbol{x}_i\|^2. \tag{8}$$

Lemma 2 (Jensen's inequality)**.** *For any convex function f and any vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ we have*

$$f\left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i\right) \leq \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}_i). \tag{9}$$

As a special case with $f(x) = \|x\|^2$, we obtain

$$\left\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i\right\|^2 \leq \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i\|^2. \tag{10}$$

Lemma 3. *$\Xi_1, \Xi_2, \ldots, \Xi_N$ are N random vectors in $\mathbb{R}^d$.*

$$\mathrm{Var}\left(\sum_{i=1}^N \Xi_i\right) = \sum_{i,j=1}^N \mathrm{Cov}(\Xi_i, \Xi_j) = \sum_{i=1}^N \mathrm{Var}(\Xi_i) + \sum_{i\neq j} \mathrm{Cov}(\Xi_i, \Xi_j) \tag{11}$$

Ξ_i satisfies $\mathbb{E}[\Xi_i] = \xi_i$, $\forall i \in [N]$. There are two cases:

1. *Suppose that $\{\Xi_i\}_{i=1}^N$ are independent from each other, it follows that:*

$$\mathrm{Cov}(\Xi_i - \xi_i, \Xi_j - \xi_j) = \mathbb{E}\langle \Xi_i - \xi_i, \Xi_j - \xi_j\rangle - \langle \mathbb{E}(\Xi_i - \xi_i), \mathbb{E}(\Xi_j - \xi_j)\rangle = 0, \ \forall i \neq j. \tag{12}$$

$$\textit{So} \qquad \mathrm{Var}\left(\sum_{i=1}^N (\Xi_i - \xi_i)\right) = \sum_{i=1}^N \mathrm{Var}(\Xi_i - \xi_j). \tag{13}$$

2. *Suppose that the random vectors $\{\Xi_i - \xi_i\}_{i=1}^N$ form a martingale difference sequence, then it follows that:*

$$\mathrm{Cov}(\Xi_i - \xi_i, \Xi_j - \xi_j) = \mathbb{E}\langle \Xi_i - \xi_i, \Xi_j - \xi_j\rangle - \langle \mathbb{E}(\Xi_i - \xi_i), \mathbb{E}(\Xi_j - \xi_j)\rangle = 0, \ \forall i \neq j. \tag{14}$$

$$\textit{So} \qquad \mathrm{Var}\left(\sum_{i=1}^N (\Xi_i - \xi_i)\right) = \sum_{i=1}^N \mathrm{Var}(\Xi_i - \xi_j). \tag{15}$$

Proof. The proof of the first case is vanilla. The second case is due to the property of martingale difference sequence. The proof of the second case can be found in Lemma 4 (separating mean and variance) of (Karimireddy et al., 2020), Lemma 2 of (Wang et al., 2020) and Chapter 15 (Martingales) in (Davidson, 1994). □

A.2 USEFUL LEMMAS

Lemma 4 (Bounded Drift (Karimireddy et al., 2020))**.** *For any local learning rate satisfying $\eta_l \leq \frac{1}{8Lk}$, the client drift caused by local updates (local training) is bounded for any $k \in \mathbb{N}^+$, as given by:*

$$\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \mathbf{x}^{r,0}\right\|^2 \leq 6k^2\eta_l^2\left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + 3k\eta_l^2\sigma^2 \tag{16}$$

Proof. This is a modified version of Lemma 8 (bounded drift) in (Karimireddy et al., 2020). When $k > 1$, taking expectation conditional on $\mathbf{x}_i^{r,k-1}$, we have

$$\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \mathbf{x}^{r,0}\right\|^2$$

$$= \mathbb{E}\left\|\mathbf{x}_i^{r,k-1} - \eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \mathbf{x}^{r,0}\right\|^2 \tag{17}$$

$$= \mathbb{E}\left\|\left[\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0} - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right] - \left[\eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right]\right\|^2 \tag{18}$$

$$= \left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0} - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right\|^2 + \mathbb{E}\left\|\eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right\|^2 \tag{19}$$

$$\leq \left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0} - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right\|^2 + \eta_l^2 \sigma^2 \tag{20}$$

$$\overset{(7)}{\leq} \left(1 + \tfrac{1}{k-1}\right)\left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + k\eta_l^2 \left\|\nabla f_i(\mathbf{x}_i^{r,k-1})\right\|^2 + \eta_l^2 \sigma^2 \tag{21}$$

$$\overset{(6)}{\leq} \left(1 + \tfrac{1}{k-1}\right)\left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + 2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}_i^{r,k-1}) - \nabla f_i(\mathbf{x}^{r,0})\right\|^2 + 2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2 \tag{22}$$

$$\overset{\text{Lem. 1}}{\leq} \left(1 + \tfrac{1}{k-1}\right)\left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + 2kL^2\eta_l^2 \left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + 2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2 \tag{23}$$

$$\leq \left(1 + \tfrac{1}{k-1} + 2kL^2\eta_l^2\right)\left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + 2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2 \tag{24}$$

$$\leq \left(1 + \tfrac{33}{32(k-1)}\right)\left\|\mathbf{x}_i^{r,k-1} - \mathbf{x}^{r,0}\right\|^2 + 2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2, \tag{25}$$

where the last inequality uses $\eta_l \leq \frac{1}{8Lk}$. Then, taking expectation conditional on $\mathbf{x}^{r,0}$ (or $\mathbf{x}_i^{r,0}$) and unrolling the recursion above, we get

$$\begin{aligned}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \mathbf{x}^{r,0}\right\|^2 &\leq (1 + \tfrac{33}{32(k-1)})^k \left\|\mathbf{x}_i^{r,0} - \mathbf{x}^{r,0}\right\|^2 \\ &\quad + \sum_{\tau=0}^{k-1} (1 + \tfrac{33}{32(k-1)})^\tau \left(2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2\right) \qquad (26) \\ &\leq \sum_{\tau=0}^{k-1} (1 + \tfrac{33}{32(k-1)})^\tau \left(2k\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2\right) \qquad (27)\end{aligned}$$

Using the facts that $(1 + \frac{c}{n})^n (c \in \mathbb{R}^+)$ is a monotonically increasing sequence and $\lim_{n\to\infty}(1 + \frac{c}{n})^n = e^c$, we have

$$\sum_{\tau=0}^{k-1} (1 + \tfrac{33}{32(k-1)})^\tau \leq \left[\sum_{\tau=0}^{k-2} (1 + \tfrac{33}{32(k-1)})^\tau + (1 + \tfrac{33}{32(k-1)})^{k-1}\right] \tag{28}$$

$$\leq \tfrac{32}{33}(k-1)\left\{\left[\left(1 + \frac{1}{\frac{32}{33}(k-1)}\right)^{\frac{32}{33}(k-1)}\right]^{\frac{33}{32}} - 1\right\} + \left[\left(1 + \frac{1}{\frac{32}{33}(k-1)}\right)^{\frac{32}{33}(k-1)}\right]^{\frac{33}{32}} \tag{29}$$

$$\leq \tfrac{32}{33}\left(e^{\frac{33}{32}} - 1\right)(k-1) + e^{\frac{33}{32}} \tag{30}$$

$$\leq 3k. \tag{31}$$

Plugging (31) back into (27), we get

$$\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \mathbf{x}^{r,0}\right\|^2 \leq 6k^2\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + 3k\eta_l^2\sigma^2. \tag{32}$$

When $k = 1$, taking the expectation conditional on $\mathbf{x}^{r,0}$, we have:

$$\mathbb{E}\left\|\mathbf{x}_i^{r,1} - \mathbf{x}^{r,0}\right\|^2 = \mathbb{E}\left\|\mathbf{x}_i^{r,0} - \eta_l \mathbf{g}_i(\mathbf{x}_i^{r,0}) - \mathbf{x}^{r,0}\right\|^2 \tag{33}$$

$$= \left\|\eta_l \nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \mathbb{E}\left\|\eta_l \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \eta_l \nabla f_i(\mathbf{x}_i^{r,k-1})\right\|^2 \tag{34}$$

$$\leq \eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + \eta_l^2 \sigma^2 \leq 6\eta_l^2 \left\|\nabla f_i(\mathbf{x}^{r,0})\right\|^2 + 3\eta_l^2 \sigma^2 \tag{35}$$

Combining the two cases when $k = 1$ and $k > 1$ completes the proof of Lemma 4. □

Lemma 5 (Bounded Accumulated Stochastic Gradient)**.**

$$\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k})\right\|^2 + NK\sigma^2 \tag{36}$$

Proof. We give the bound of the accumulated stochastic gradient of FL and SL based on Lemma 4 (separating mean and variance) of (Karimireddy et al., 2020) and the proof of Theorem 1 (Convergence of Surrogate Objective) of (Wang et al., 2020).

The proof is on the expectation on $\mathbf{x}^{r,0}$. Beginning with $\mathbb{E}[\|\mathbf{x}\|^2] = \mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|^2] + \|\mathbb{E}[\mathbf{x}]\|^2$, we get

$$\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \left[\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla f_i(\mathbf{x}_i^{r,k})\right]\right\|^2 + \mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k})\right\|^2$$

For brevity, we use $\mathbf{d}_i^{r,k}$ to denote $\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla f_i(\mathbf{x}_i^{r,k})$. Then we continue in FL and SL two cases:

1. **FL**: Note that the clients are independent during each round in FL, so $\mathrm{Cov}(\mathbf{d}_i^r, \mathbf{d}_j^r) = 0, \forall i \neq j$. Then we get:

$$\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{d}_i^{r,k}\right\|^2 \overset{(13)}{=} \sum_{i=1}^{N} \mathbb{E}\left\|\sum_{k=0}^{K-1} \mathbf{d}_i^{r,k}\right\|^2 \overset{(15)}{=} \sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbb{E}\left\|\mathbf{d}_i^{r,k}\right\|^2 \overset{\text{Asm. 2}}{=} NK\sigma^2, \tag{37}$$

   where the second equality is due to the properties of martingale difference sequence.

2. **SL**: In SL, we refer to (15) and have

$$\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbf{d}_i^{r,k}\right\|^2 \overset{(15)}{=} \sum_{i=1}^{N}\sum_{k=0}^{K-1} \mathbb{E}\left\|\mathbf{d}_i^{r,k}\right\|^2 \overset{\text{Asm. 2}}{=} NK\sigma^2. \tag{38}$$

This concludes the proof. □

A.3 Proof of Theorem 1

Proof. Beginning with Assumption 1, we have

$$f(\mathbf{x}^{r+1,0}) - f(\mathbf{x}^{r,0}) \leq \left\langle \nabla f(\mathbf{x}^{r,0}), \mathbf{x}^{r+1,0} - \mathbf{x}^{r,0} \right\rangle + \frac{L}{2}\left\|\mathbf{x}^{r+1,0} - \mathbf{x}^{r,0}\right\|^2. \tag{39}$$

Noting that the global model update can be written as: $\mathbf{x}^{r+1} - \mathbf{x}^r = -\eta_g \eta_l \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k})$ in round r, for the expectation on $\mathbf{x}^{r,0}$, we get

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{r+1,0})] - f(\mathbf{x}^{r,0}) \\
&\leq \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{r,0}), -\eta_g\eta_l \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\rangle\right] + \frac{L}{2}\mathbb{E}\left\| \eta_g\eta_l \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\|^2 && (40)\\
&= -N\eta_g\eta_l \sum_{k=0}^{K-1} \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{r,0}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\rangle\right] + \frac{L}{2}\eta_g^2\eta_l^2 \mathbb{E}\left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\|^2 && (41)\\
&\overset{\text{Lem. 5}}{\leq} -N\eta_g\eta_l \sum_{k=0}^{K-1} \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{r,0}), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\rangle\right] + \frac{L}{2}\eta_g^2\eta_l^2 \mathbb{E}\left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 \\
&\quad + \frac{1}{2}NKL\eta_g^2\eta_l^2\sigma^2 && (42)\\
&= -N\eta_g\eta_l \sum_{k=0}^{K-1} \left[\frac{1}{2}\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2 + \frac{1}{2}\mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 - \frac{1}{2}\mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) - \nabla f(\mathbf{x}^{r,0}) \right\|^2\right] \\
&\quad + \frac{L}{2}\eta_g^2\eta_l^2 \mathbb{E}\left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 + \frac{1}{2}NKL\eta_g^2\eta_l^2\sigma^2, && (43)
\end{aligned}
$$

where we use Lemma (5) in the second inequality and the fact that $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ in the second equality. Note that

$$
\begin{aligned}
&-\frac{1}{2}N\eta_g\eta_l \sum_{k=0}^{K-1} \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 + \frac{L}{2}\eta_g^2\eta_l^2 \mathbb{E}\left\| \sum_{i=1}^N \sum_{k=0}^{K-1} \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 \\
&\overset{(8)}{\leq} -\frac{1}{2}\frac{1}{N}\eta_g\eta_l \sum_{k=0}^{K-1} \mathbb{E}\left\| \sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 + \frac{L}{2}\eta_g^2\eta_l^2 K \sum_{k=0}^{K-1} \mathbb{E}\left\| \sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 \\
&\leq -\frac{1}{2N}\eta_g\eta_l(1 - NKL\eta_g\eta_l) \sum_{k=0}^{K-1} \mathbb{E}\left\| \sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) \right\|^2 && (44)
\end{aligned}
$$

$$
\begin{aligned}
\text{and} \quad \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{r,k}) - \nabla f(\mathbf{x}^{r,0}) \right\|^2 &= \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N [\nabla f_i(\mathbf{x}_i^{r,k}) - \nabla f_i(\mathbf{x}^{r,0})] \right\|^2 \\
&\overset{(10)}{\leq} \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\| \nabla f_i(\mathbf{x}_i^{r,k}) - \nabla f_i(\mathbf{x}^{r,0}) \right\|^2 \\
&\overset{\text{Asm. 1}}{\leq} \frac{1}{N}\sum_{i=1}^N L^2 \mathbb{E}\left\| \mathbf{x}_i^{r,k} - \mathbf{x}^{r,0} \right\|^2, && (45)
\end{aligned}
$$

where the first equality of (45) results from the fact that $\nabla f(\mathbf{x}^{r,0}) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}^{r,0})$. By plugging (44) and (45) into (43) and using $\eta_g\eta_l \leq \frac{1}{NKL}$, we get

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{r+1,0})] - f(\mathbf{x}^{r,0}) \leq -\frac{1}{2}NK\eta_g\eta_l \left\|\nabla f(\mathbf{x}^{r,0})\right\|^2 + \frac{1}{2}L^2\eta_g\eta_l \sum_{i=1}^N \sum_{k=0}^{K-1} \mathbb{E}\left\| \mathbf{x}_i^{r,k} - \mathbf{x}^{r,0} \right\|^2 \\
+ \frac{1}{2}NKL\eta_g^2\eta_l^2\sigma^2 && (46)
\end{aligned}
$$

Here we can not use Lemma 4 directly, but we have:

$$\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\mathbf{x}^{r,0}\right\|^2=\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\mathbf{x}_i^{r,0}+\mathbf{x}_i^{r,0}-\mathbf{x}_{i-1}^{r,0}+\cdots+\mathbf{x}_2^{r,0}-\mathbf{x}_1^{r,0}\right\|^2 \tag{47}$$

$$\overset{(8)}{\leq} i\mathbb{E}\left[\left\|\mathbf{x}_i^{r,k}-\mathbf{x}_i^{r,0}\right\|^2+\left\|\mathbf{x}_i^{r,0}-\mathbf{x}_{i-1}^{r,0}\right\|^2+\cdots+\left\|\mathbf{x}_2^{r,0}-\mathbf{x}_1^{r,0}\right\|^2\right] \tag{48}$$

$$\overset{\text{Lem. 4}}{\leq} i\left(6K^2\eta_l^2\sum_{t=1}^{i}\left\|\nabla f_t(\mathbf{x}^{r,0})\right\|^2+3iK\eta_l^2\sigma^2\right), \tag{49}$$

$$\leq N\left(6K^2\eta_l^2\sum_{t=1}^{N}\left\|\nabla f_t(\mathbf{x}^{r,0})\right\|^2+3NK\eta_l^2\sigma^2\right) \tag{50}$$

$$\overset{\text{Asm. 3}}{\leq} 6N^2K^2\eta_l^2B^2\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2+6N^2K^2\eta_l^2G^2+3N^2K\eta_l^2\sigma^2 \tag{51}$$

where we use the Lemma 4 in the second inequality. Then, we have:

$$\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{x}_i^{r,k}-\mathbf{x}^{r,0}\|^2\leq 6N^3K^3\eta_l^2B^2\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2+6N^3K^3\eta_l^2G^2+3N^3K^2\eta_l^2\sigma^2. \tag{52}$$

Plugging (52) into (46), we get

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}^{r+1,0})]-f(\mathbf{x}^{r,0})\leq&-\frac{1}{2}NK\eta_g\eta_l\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2+\frac{1}{2}NK\eta_g\eta_l\left(6N^2K^2L^2B^2\eta_l^2\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2\right)\\&+\frac{1}{2}NK\eta_g\eta_l\left(6N^2K^2L^2\eta_l^2G^2+3N^2KL^2\eta_l^2\sigma^2\right)\\&+\frac{1}{2}NKL\eta_g^2\eta_l^2\sigma^2.\end{aligned} \tag{53}$$

Dividing both sides by $NK\eta_g\eta_l$ and using $\eta_l\leq\frac{1}{2\sqrt{3}NKLB}$ (i.e., $\frac{1}{2}-3N^2K^2L^2B^2\eta_l^2\geq\frac{1}{4}$), we get

$$\frac{\mathbb{E}[f(\mathbf{x}^{r+1,0})]-f(\mathbf{x}^{r,0})}{NK\eta_g\eta_l}\leq-\frac{1}{4}\left\|\nabla f(\mathbf{x}^{r,0})\right\|^2+3N^2K^2L^2\eta_l^2G^2+\frac{3}{2}N^2KL^2\eta_l^2\sigma^2+\frac{1}{2}L\eta_g\eta_l\sigma^2$$

Taking unconditional expectation, rearranging the terms and then averaging the above equation over $r=\{0,\cdots,R-1\}$, we have

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(\mathbf{x}^r)\right\|^2\leq&\frac{4[f(\mathbf{x}^0)-f(\mathbf{x}^R)]}{\eta_g\eta_lNKR}+12\eta_l^2N^2K^2L^2G^2+6\eta_l^2N^2KL^2\sigma^2+2\eta_g\eta_lL\sigma^2\\\leq&\frac{4[f(\mathbf{x}^0)-f(\mathbf{x}^*)]}{\eta_g\eta_lNKR}+12\eta_l^2N^2K^2L^2G^2+6\eta_l^2N^2KL^2\sigma^2+2\eta_g\eta_lL\sigma^2\end{aligned} \tag{54}$$

Finally, using the fact that $\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^R)\right\|^2\leq\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(\mathbf{x}^r)\right\|^2$ where $\bar{\mathbf{x}}^R=\frac{1}{R}\sum_{r=0}^{R-1}\mathbf{x}^r$ and $\eta=\eta_g\eta_l$, we completes the proof of Theorem 1. □

A.4 Proof of Theorem 2

Proof. The proof is similar to that of Theorem 1, except the three points:

i The global model update in FL is $\mathbf{x}^{r+1}-\mathbf{x}^r=-\eta_g\eta_l\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})$ in round r;

ii The first case of Lemma 5 should be used to bound the term $\mathbb{E}\left\|\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2$;

iii We can bound the term $\mathbb{E}\|\mathbf{x}_i^{r,k}-\mathbf{x}^{r,0}\|^2$ with Lemma 4 directly.

□

B MORE EXPERIMENTAL DETAILS

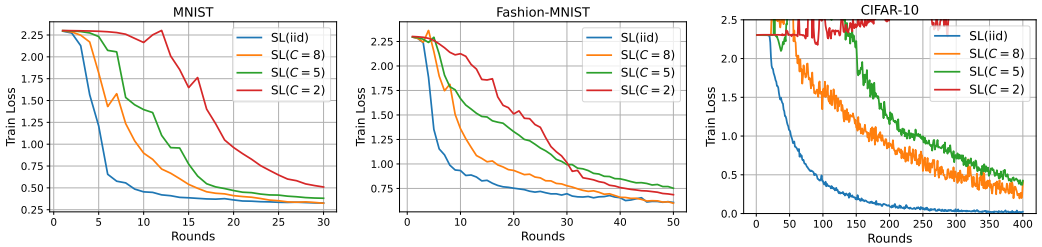
Platform. We train LeNet-5 on MNIST and Fashion-MNIST with Nvidia GeForce RTX 3070 Ti, VGG-11 on CIFAR-10 with Nvidia Tesla V100. The algorithms are implemented by PyTorch. We use the random seed "1234" by default. We use vanilla SGD algorithm with momentem = 0.9 and weight decay = 1e-4 as (He et al., 2020). The detailed information of the models and other information can be found in our code.

We use the same setups as the main body by default: (i) training LeNet-5 on MNIST; (ii) training LeNet-5 on Fashion-MNIST; (iii) training VGG-11 on CIFAR-10.

B.1 MORE RESULTS OF SL

This section is complimentary to Section 5.2 to study the factors that affects the performance of SL. We report the details of the experiments, such as η_l, η_g, b and so on. Besides, we give the additional study of the global learning rate η_g in SL.

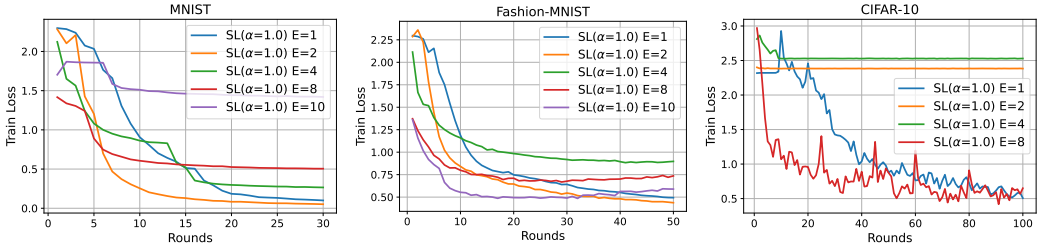
Effect of data heterogeneity.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 4: Effect of data heterogeneity. (a) MNIST, $N = 10$, $b = 1000$, $\eta_l = 0.01$, $\eta_g = 1.0$; (b) Fashion-MNIST, $N = 10$, $b = 1000$, $\eta_l = 0.01$, $\eta_g = 1.0$; (c) CIFAR-10, $N = 10$, $b = 100$, $\eta_l = 0.001$, $\eta_g = 1.0$.

Effect of K.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 5: Effect of K. $\text{Dir}_{10}(1.0)$ is used. (a) MNIST, $N = 10$, $b = 1000$, $\eta_l = 0.01$, $\eta_g = 1.0$; (b) Fashion-MNIST, $N = 10$, $b = 1000$, $\eta_l = 0.01$, $\eta_g = 1.0$; (c) CIFAR-10, $N = 10$, $b = 100$, $\eta_l = 0.005$, $\eta_g = 1.0$.

Effect of η_g. The effect of η_g in FL has been empirically studied in (Reddi et al., 2020). We follow a similar method to study η_g of SL, by choosing different combinations of η_g and η_l to train. We pick a sequence of η_l and then select η_g carefully to train the model, yielding the Figure 6. As we can see, the dark green grids with high test accuracy concentrate on the left bottom triangular regions, which shows that $\eta = \eta_g\eta_l$ should avoid to be too large. Our theory also suggests that η should be no more than $\min\{\frac{\eta_g}{8KL}, \frac{1}{NKL}, \frac{\eta_g}{2\sqrt{3}NKLB}\}$ to ensure the convergence. Also note that η_g cannot be set infinitely too large and has a limited range too (the training process plateaus or even diverge when η_g is over-large), which is also shown in FL (Reddi et al., 2020). Thus tuning η_g in the limited range is suggested. However, except the results on MNIST when $\eta_l = 5e-3$ shows η_g can accelerate the convergence rate evidently, the results on Fashion-MINST and CIFAR-10 show no

distinct improvement by setting $\eta_g > 1$ as the theory suggests ($\eta_g = 1$ shows the best generalization performance in many cases). This is also observed in FL (Reddi et al., 2020). Further researches on η_g are needed, to address the issues such as "*which settings are η_g ($\eta_g > 1$) appropriate for*" and "*how to tune η_g to gain improvement*".

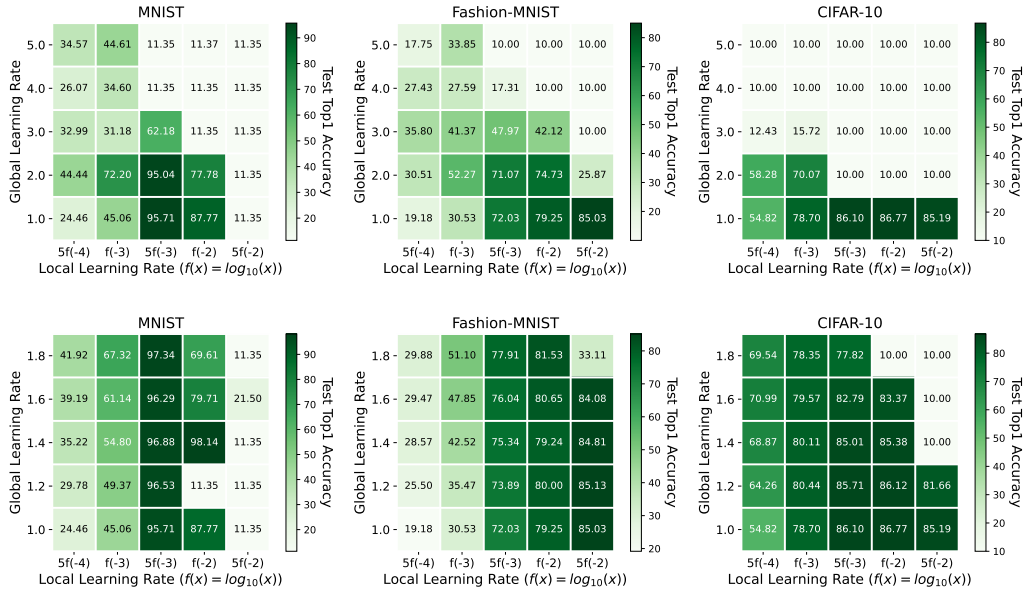


Figure 6: Test accuracy for various local/global learning rates combination. $\text{Dir}_{10}(10.0)$ is used. For MNIST, $N = 10$, $b = 1000$, $E = 1$; We average test accuracy over the last 10 rounds from 30 total rounds; (b) For Fashion-MNIST, $N = 10$, $b = 1000$, $E = 1$; We average test accuracy over the last 10 rounds from 30 total rounds; (c) For CIFAR-10, $N = 10$, $b = 100$, $E = 1$. We average test accuracy over the last 20 rounds from 100 total rounds.

B.2 More comparisons between FL and SL

The learning rates of FL and SL are selected from $\{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$. We report the overall results of comparisons between FL and SL on MNIST, Fashion-MNIST and CIFAR-10 datasets with different learning rates in Table 5, Table 6 and Table 7 respectively. Table 3 and Table 4 in the main body are based on these three tables here. "L-M", "L-M" and "V-10" denote LeNet-5 on MNIST, LeNet-5 on Fashion-MNIST and VGG-11 on CIFAR-10 respectively. We highlight the "best" test accuracy among all chosen learning rates with blue for FL and red for SL. We underline the test accuracy of the "threshold" learning rate with blue for FL and red for SL. Moreover, we find that the performance of SL when $N = 100$ is better than that when $N = 10$, which is required for further research.

Table 5: The detailed results of FL and SL with different learning rates on MNIST dataset. We average the test accuracy over the last 10 rounds from 50 total rounds on MNIST dataset.

Tag	N	Dist.	Alg	b	0.0005		0.001		0.005		0.01		0.05		0.1	
					FL	SL	FL	SL	FL	SL	FL	SL	FL	SL	FL	SL
L-M	10	IID	sgd	10	98.8	99.0	99.0	99.0	99.2	99.0	99.2	98.5	11.3	11.3	11.3	11.3
L-M	10	$C = 8$	sgd	10	98.5	98.7	99.0	98.8	99.2	98.4	99.2	97.1	10.1	11.3	11.3	11.3
L-M	10	$C = 5$	sgd	10	98.2	88.7	98.7	88.4	99.0	97.8	99.1	96.8	11.3	11.3	11.3	11.3
L-M	10	$C = 2$	sgd	10	95.9	96.3	97.1	96.1	98.4	20.0	98.3	11.3	10.1	11.3	10.2	11.3
L-M	10	IID	sgd	100	85.3	98.7	94.8	98.8	98.6	98.9	99.0	98.9	99.2	89.2	79.2	49.9
L-M	10	$\alpha = 10.0$	sgd	100	86.4	98.4	94.9	98.6	98.6	98.9	98.9	99.1	99.1	98.9	89.1	98.3
L-M	10	$\alpha = 5.0$	sgd	100	85.8	97.0	94.2	97.7	98.3	98.8	98.9	89.3	99.1	77.8	99.1	96.7
L-M	10	$\alpha = 1.0$	sgd	100	87.4	97.4	94.3	78.5	98.2	98.6	98.8	40.1	99.1	97.3	99.1	11.3
L-M	10	$C = 8$	sgd	100	89.0	97.8	95.1	98.3	98.4	98.7	98.9	89.4	99.2	98.3	99.0	11.3
L-M	10	$C = 5$	sgd	100	88.6	97.3	94.1	98.0	98.2	90.1	98.7	80.8	99.0	79.5	98.4	11.3
L-M	10	$C = 2$	sgd	100	70.8	95.6	79.1	97.2	96.5	11.4	97.3	11.3	96.5	11.4	11.3	11.3
L-M	10	IID	sgd	1000	9.7	40.2	9.8	62.3	39.6	96.5	58.8	97.8	78.0	89.3	98.2	61.0
L-M	10	$\alpha = 10.0$	sgd	1000	9.7	42.5	9.8	73.7	42.9	97.6	72.1	75.3	97.3	11.4	60.4	51.2
L-M	10	$\alpha = 5.0$	sgd	1000	9.8	41.3	9.9	68.2	44.7	97.1	73.4	98.1	97.0	11.4	98.1	11.4
L-M	10	$\alpha = 1.0$	sgd	1000	9.7	40.8	12.4	63.6	37.8	96.9	70.3	98.0	97.3	60.5	88.5	11.4
L-M	10	$C = 8$	sgd	1000	9.7	40.2	9.8	56.5	40.8	96.6	63.4	97.7	97.1	11.4	98.2	89.1
L-M	10	$C = 5$	sgd	1000	9.7	36.2	9.8	54.9	40.0	94.1	72.6	96.6	96.6	60.9	97.3	11.4
L-M	10	$C = 2$	sgd	1000	9.7	37.3	9.8	53.9	39.5	87.6	52.8	93.5	93.2	11.4	85.2	11.4
L-M	100	IID	sgd	10	85.9	98.9	95.7	99.1	98.4	99.3	98.8	98.8	11.3	11.3	11.3	11.3
L-M	100	$C = 8$	sgd	10	87.4	98.7	95.5	99.0	98.4	99.2	98.6	97.9	11.3	11.3	11.3	11.3
L-M	100	$C = 5$	sgd	10	89.1	98.8	94.5	99.0	98.5	98.4	98.6	97.2	98.2	11.3	10.8	11.3
L-M	100	$C = 2$	sgd	10	79.6	98.9	89.3	99.0	97.1	11.3	98.1	11.3	24.2	11.3	9.1	11.3

Table 6: The detailed results of FL and SL with different learning rates on Fashion-MNIST dataset. We average the test accuracy over the last 10 rounds from 50 total rounds on Fashion-MNIST dataset.

Tag	N	Dist.	Alg	b	0.0005		0.001		0.005		0.01		0.05		0.1	
					FL	SL	FL	SL	FL	SL	FL	SL	FL	SL	FL	SL
L-F	10	IID	sgd	100	61.2	86.4	74.1	87.8	86.0	89.9	87.7	90.4	89.6	88.7	89.2	87.4
L-F	10	$\alpha = 10.0$	sgd	100	61.8	82.7	74.2	85.2	86.0	88.1	87.8	88.1	89.5	85.4	89.0	82.9
L-F	10	$\alpha = 5.0$	sgd	100	62.1	85.8	76.4	78.8	85.9	89.6	87.6	89.2	89.5	88.0	88.9	86.8
L-F	10	$\alpha = 1.0$	sgd	100	68.2	76.3	74.3	79.3	83.7	83.3	86.4	74.7	88.1	81.8	87.2	59.5
L-F	10	$C = 8$	sgd	100	68.6	67.5	75.9	71.4	85.3	75.1	87.4	74.9	89.3	67.7	88.5	53.0
L-F	10	$C = 5$	sgd	100	64.4	52.4	69.7	57.3	79.6	19.2	84.6	33.7	86.3	15.7	84.1	10.5
L-F	10	$C = 2$	sgd	100	47.9	66.2	52.1	62.2	59.0	37.8	64.3	10.0	38.3	10.0	10.0	10.0
L-F	10	IID	sgd	1000	10.0	29.5	10.0	53.5	29.5	79.3	52.1	83.7	77.4	81.3	83.0	89.8
L-F	10	$\alpha = 10.0$	sgd	1000	10.0	29.8	10.0	51.8	30.1	77.5	56.1	82.1	78.8	87.0	83.6	87.5
L-F	10	$\alpha = 5.0$	sgd	1000	10.0	30.9	10.0	53.8	31.1	77.3	55.2	82.6	79.6	87.0	82.6	87.6
L-F	10	$\alpha = 1.0$	sgd	1000	10.0	31.1	10.0	46.9	35.2	74.7	55.0	80.0	75.8	79.8	79.6	15.7
L-F	10	$C = 8$	sgd	1000	10.0	29.4	10.0	42.6	29.4	66.2	56.4	69.9	77.7	79.1	75.2	62.0
L-F	10	$C = 5$	sgd	1000	10.0	29.2	10.0	41.7	29.6	59.6	48.8	64.7	70.9	56.7	74.5	10.0
L-F	10	$C = 2$	sgd	1000	10.0	29.2	10.0	37.4	30.6	64.9	39.5	71.7	56.9	10.0	51.5	10.0
L-F	100	IID	sgd	10	66.6	89.9	75.0	89.8	83.5	88.5	85.3	87.1	84.6	10.0	10.0	10.0
L-F	100	$C = 8$	sgd	10	68.3	79.6	74.9	79.7	83.4	77.8	85.5	68.5	84.5	10.0	25.2	10.0
L-F	100	$C = 5$	sgd	10	68.7	81.9	73.9	80.8	82.7	78.9	85.2	70.7	10.0	10.0	10.0	10.0
L-F	100	$C = 2$	sgd	10	61.5	68.1	69.1	69.3	76.2	10.0	78.2	10.0	10.0	10.0	10.0	10.0

Table 7: The detailed results of FL and SL with different learning rates on CIFAR-10 dataset. We average the test accuracy over the last 40 rounds from 400 total rounds on CIFAR-10 dataset. We don not execute the experiments whose learning rates are larger than the "threshold" learning rate ("-" in the table).

Tag	N	Dist.	Alg	b	0.0005		0.001		0.005		0.01		0.05		0.1	
					FL	SL	FL	SL	FL	SL	FL	SL	FL	SL	FL	SL
V-10	10	IID	sgd	10	86.7	88.3	87.8	88.7	89.4	87.0	89.3	10.0	10.0	-	-	-
V-10	10	$C = 8$	sgd	10	85.4	81.3	86.7	80.4	88.2	70.9	88.4	10.0	10.0	-	-	-
V-10	10	$C = 5$	sgd	10	80.1	72.2	83.4	10.0	85.9	10.0	86.0	-	10.0	-	-	-
V-10	10	$C = 2$	sgd	10	28.4	10.0	24.7	10.0	66.0	10.0	68.3	-	10.0	-	-	-
V-10	10	IID	sgd	100	13.0	84.5	46.1	86.1	86.3	88.0	87.1	88.5	88.9	86.9	88.7	10.0
V-10	10	$\alpha = 10.0$	sgd	100	15.5	84.3	48.6	85.9	85.5	87.9	86.8	88.2	88.5	86.3	88.5	10.0
V-10	10	$\alpha = 5.0$	sgd	100	17.4	84.2	49.8	85.8	85.8	88.1	86.9	88.1	88.4	86.3	88.4	10.0
V-10	10	$\alpha = 1.0$	sgd	100	23.0	78.1	47.6	81.5	83.9	84.7	85.2	10.0	87.3	-	87.3	-
V-10	10	$\alpha = 0.5$	sgd	100	21.4	72.1	45.4	76.9	81.3	10.0	83.8	10.0	-	-	-	-
V-10	10	$C = 8$	sgd	100	22.4	74.3	49.6	80.9	84.8	83.8	86.0	10.0	88.1	-	87.9	-
V-10	10	$C = 5$	sgd	100	23.4	69.6	40.2	79.0	79.8	10.0	82.2	-	84.6	-	85.4	-
V-10	10	$C = 2$	sgd	100	15.2	22.4	21.6	27.5	43.8	10.0	47.3	-	69.5	-	10.0	-
V-10	100	IID	sgd	10	12.9	88.1	38.7	88.5	81.2	87.2	84.1	10.0	10.0	-	-	-
V-10	100	$C = 5$	sgd	10	23.2	82.2	35.4	80.6	78.3	10.0	81.4	-	10.0	-	-	-
V-10	100	$C = 2$	sgd	10	13.4	79.0	26.2	10.0	39.1	10.0	50.1	-	10.0	-	-	-