On the Convergence of Cyclic Hierarchical Federated Learning with Heterogeneous Data

Jingjing Fu[1], Haibo Yang[2], Xiaonan Zhang[3], Linke Guo[1]

[1]Electrical and Computer Engineering, Clemson University, Clemson, SC, USA
[2]Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA
[3]Computer Science, Florida State University, Tallahassee, FL, USA
jfu@clemson.edu, hbycis@rit.edu, xzhang14@fsu.edu, linkeg@clemson.edu

***Abstract*—Hierarchical Federated Learning (HFL) advances the classic Federated Learning (FL) by introducing the multi-layer architecture between clients and the central server, in which edge servers aggregate models from respective clients and further send to the central server. Instead of directly uploading each update from clients for aggregation, the HFL not only reduces the communication and computational overhead but also greatly enhances the scalability of supporting a massive number of clients. When HFL operates for applications having a large-scale clients, edge servers train their models in a cyclic pattern (a ring architecture) as opposed to the star-type of architecture where each edge develops their own models independently.We refer it as Cyclic HFL(CHFL). Driven by its promising feature of handling data heterogeneity and resiliency, CHFL has a great potential to be deployed in practice. Unfortunately, the thorough convergence analysis on CHFL remains lacking, especially considering the widely-existing data heterogeneity issue among clients. To the best of our knowledge, we are the first to provide a theoretical convergence analysis for CHFL in strongly convex, general convex, and non-convex objectives. In particular, CHFL achieves a $\tilde{\mathcal{O}}(1/(MNRKT))$ convergence rate for strongly convex objectives and an $\mathcal{O}(1/\sqrt{MNRKT})$ rate for general convex objectives. For smooth non-convex objectives, we prove convergence to an ε-stationary point, with the dominant term scaling as $\mathcal{O}(1/\sqrt{MNRKT})$, where M is the number of edges, N is clients per edge, K is the local update steps, and R is the edge rounds. Extensive experiments on real-world datasets corroborate our theory and demonstrate that CHFL performs competitively under both inter- and intra-edge heterogeneity.**

***Index Terms*—Hierarchical Federated Learning, Convergence Rate, Cyclic Pattern, Client Clustering.**

I. Introduction

Federated Learning (FL) [1] has emerged as a promising distributed learning framework in which a large number of distributed devices collaborate to train a joint model without sharing their data. Although FL has attracted a significant interest on theoretical research, the standard FL architecture does not always perform well in practical scenarios. When the number of clients (devices) participate in the FL, the communication burden in between the central server and each client for model update and aggregation may significantly impact the FL performance. As a solution, the edge-based FL has been proposed [2] to replace a single central server with multiple edge servers. In particular, each edge server acts as a parameter server responsible for a smaller set of clients, who can meet specific communication or data requirements. However, as stated in [3], this edge-based FL still suffers the performance drop due to the limited number of participating clients managed by each edge.

To alleviate the communication burden on a central server, Hierarchical Federated Learning (HFL) [4] introduces a three-layer architecture, i.e., a central server, multiple edge servers, and clients. Each edge server, along with its associated clients, forms an edge. The HFL has two levels of model updates including edge model update and global model update, in which many edge model updates follow the star architecture [5] to aggregate local model updates from clients. On the other hand, most HFL works [4], [6] also assume the global model update follows the star architecture, i.e., the centralized HFL, and provides convergence analysis based on various assumptions. Rather than the star architecture, for many FL applications across a large geographic regions [7]–[9], the ring architecture is a better fit in terms of the participation pattern, where each edge server updates their models to another edge server rather than the central server, namely, the Cyclic HFL (CHFL). Compared to the star-like architecture with centralized HFL, CHFL improves scalability by accommodating more clients without being affected by central server dropout issues when the communication burden increases. Unfortunately, existing convergence analyses for cyclic training are limited to FL, and a comprehensive study for HFL under standard assumptions is still missing. Our main contributions are summarized as follows:

- We derive convergence guarantees for CHFL on heterogeneous data for strongly convex, general convex, and non-convex objectives. Our analysis generalizes several existing FL methods as special cases [9], [10], further echoing the correctness of our results, and achieves improved convergence rates compared with centralized HFL baselines.
- We show that achieving optimal convergence improvements depends on both data heterogeneity and system settings. Unlike current strategies that focus solely on reducing data heterogeneity(e.g.,grouping clients with similar data [14] or ensuring edges share similar data [15]), our approach also determines the best policy based on system configuration for various objectives. For example, in the

TABLE I: Convergence rates for FL and HFL variants.

System	Method	Cyclic Pattern	Convexity	Convergence Rate[a]
FL	SFL [9]	✓	Strongly Convex	$\tilde{\mathcal{O}}\left(\frac{1}{NKT}+\frac{1}{MT^2}\right)$
			General Convex	$\mathcal{O}\left(\frac{1}{NKT}+\frac{1}{(NK)^{1/3}T^{2/3}}\right)$
			Non-Convex	$\mathcal{O}\left(\frac{1}{\sqrt{NKT}}+\frac{1}{(NK)^{1/3}T^{2/3}}\right)$
	Scaffold [10]	×	Strongly Convex	$\mathcal{O}\left(\frac{1}{NKT}+\frac{1}{T^2}\right)$
	DSGD [11]	×	Strongly Convex	$\mathcal{O}\left(\frac{1}{NKT}+\frac{1}{T^2}\right)$
	CFL [12]	✓	Non-Convex	$\tilde{\mathcal{O}}\left(\frac{M}{KNT}+\frac{ML}{NT}\left(\frac{MN/M-N}{MN/M-1}\right)\right)$
HFL	Hier-Local-QSGD [4]	×	Non-Convex	$\mathcal{O}\left(\frac{1}{\sqrt{RKTMN}}+\frac{1}{RKT}\right)$
	Group-FEL [6]	×	Non-Convex	$\mathcal{O}\left(\frac{1}{\sqrt{RKT}}\right)$
	HSFL [13]	×	Strongly Convex	$\mathcal{O}\left(\frac{1}{T}\right)$
	Centralized HFL [14]	×	Strongly Convex	$\mathcal{O}\left(\frac{1}{TG(R,K)}\right)$[b]
	This paper	✓	Strongly Convex	$\tilde{\mathcal{O}}\left(\frac{1}{MNRKT}\right)$
			General Convex	$\mathcal{O}\left(\frac{1}{\sqrt{MNRKT}}\right)$
			Non-Convex	$\mathcal{O}\left(\frac{1}{\sqrt{MNRKT}}\right)$

[a] Absolute constants and polylog factors omitted for clarity.
[b] $G(R,K)$ is a function of edge rounds R and local steps K.

general convex case, having all edges sharing similar data will lead to an optimal convergence improvement when the number of edges is relatively small.

- We conduct experiments on standard benchmark datasets. The experimental results show that CHFL can achieve comparable or superior performance in terms of accuracy and convergence speed measured by local model updates. Meanwhile, we show that the inter-edge heterogeneity has more effect on convergence speed than the intra-edge heterogeneity in specific conditions.

II. Related Work

A. Hierarchical Federated Learning

In addressing the challenges of high communication overhead and latency in vanilla FL, HFL [14], [16] was proposed to add a layer of edge servers, which simplifies the communication process to occur only between edge servers and central server. Liu *et. al* in [14] prove HFL could achieve convergence amidst inter-edge data heterogeneity for strongly convex objective , yet the impact of intra-edge data heterogeneity on convergence remains unknown. Similarly, Abad *et.al* [17] show the reduced communication latency by deploying HFL in a real mobile edge computing system. Xu *et al.* in [18] introduce an adaptive HFL approach, focusing on optimal resource allocation and control of edge intervals to enhance training accuracy. OUEA [15] and SHARE [19] consider the clients cluster problem in HFL. OUEA cluster clients with similar data into one edge to improve performance in HFL but they only consider the convex objective. SHARE ensures each edge shares similar data to reduce data heterogeneity among edges to improve performance. Both discuss cluster policies in the context of centralized HFL, but these policies cannot be directly applied to CHFL to enhance performance. Our convergence analysis reveals that the effectiveness of cluster policies depends on system settings and their specific objectives.

B. Cyclic/Sequential Federated Learning

In vanilla Federated Learning, clients exhibit system heterogeneity. Hence, it is advisable to select qualified devices suited for FL, ensuring they have a stable network for efficient model updates, sufficient charging to manage energy use, and idle status to avoid disruptions. Compared with the vanilla FL setting with random device selection [20], [21], those qualified devices usually participate in FL at specific time and follow a cyclic pattern [7], [8]. Cho *et al.* in [12] explores various gradient update methods in FL under a cyclic pattern. However, their convergence analysis is based on the assumption that the local client's objective conforms to the Polyak-Łojasiewicz condition [22], a limitation considering that objectives in FL are often general non-convex [23]. Li *et al.* in [9] offers the convergence analysis for both Parallel Federated Learning (PFL) and Sequential Federated Learning (SFL) with convex and non-convex objectives. For non convex case, the convergence rate can be achieved $\mathcal{O}\left(\frac{1}{(NK)^{1/3}T^{2/3}}\right)$. They have the result that SFL has a better guarantee than PFL in specific conditions. However, they both discuss the convergence analysis on the two-layer FL, for which the cyclic pattern on HFL is still missing. To address this gap, we extend the application of the cyclic pattern to HFL and provide a convergence analysis for both convex and non-convex objectives. Furthermore, we delve into the impact of data heterogeneity on convergence speed across various client participant patterns in HFL. We attain optimal convergence rate compared to other studies [6], [18]. Some related algorithms convergence rates are summarized in Table.I to compare.

III. Preliminary on Cyclic HFL

A. Notations

We consider a CHFL system having a set of edge servers (interchangeably with edges) $\mathcal{M}$ with $M = |\mathcal{M}|$. Each edge server $i \in \mathcal{M}$ will serve clients $\mathcal{N}_i$ with $N = |\mathcal{N}_i|$. For each client $j \in \mathcal{N}_i$, it has the local empirical loss function $F_j(\mathbf{x}) = \frac{1}{|\mathcal{D}_j|}\sum_{\xi\in\mathcal{D}_j} \ell(\mathbf{x},\xi)$ where $\mathcal{D}_j$ is the training dataset and $\ell(\mathbf{x},\xi)$ is the loss value of the model $\mathbf{x}$ at data batch ξ. For each edge server i, it optimizes following objective, $f_i(\mathbf{x}) := \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} F_j(\mathbf{x})$ The global optimization task is identical to that of standard FL $f(\mathbf{x}) := \frac{1}{M}\sum_{i\in\mathcal{M}} f_i(\mathbf{x})$ and the model can be founded by achieving $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

B. Assumptions

Assumption 1 (Bounded variance)**.** *For the local objective $F_j(\mathbf{x})$ in any client, the local stochastic gradient $\nabla F_j(\mathbf{x},\xi_j)$ computed using a mini-batch ξ_j, sampled uniformly at random from local dataset $\mathcal{D}_j$, has bounded variance, that is $\mathbb{E}\|\nabla F_j(\mathbf{x},\xi_j) - \nabla F_j(\mathbf{x})\| \le \sigma^2$, for all clients.*

Assumption 2 (Smoothness)**.** *Smoothness of $F_j(\mathbf{x}), \forall j \in \mathcal{N}_i, \forall i \in \mathcal{M}$. The clients' local objective functions are all L-smooth, i.e., $\|\nabla F_j(\mathbf{x}) - \nabla F_j(\mathbf{x}')\| \le L\|\mathbf{x}-\mathbf{x}'\|$ for all $\mathbf{x}$ and $\mathbf{x}'$.*

Assumption 3 (Intra-Edge & Inter-Edge Data Heterogeneity for Non-Convex objective)**.** *There exist constants $\sigma_c, \sigma_g \ge 0$, such that for all $\mathbf{x}$, for all $i \in \mathcal{M}$ and for all $j \in \mathcal{N}_i$, $\left\|\nabla F_j(\mathbf{x}) - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\nabla F_j(\mathbf{x})\right\| \le \sigma_c$, and $\left\|\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\nabla F_j(\mathbf{x}) - \nabla f(\mathbf{x})\right\| \le \sigma_g$.*

Assumption 4 (Intra-Edge & Inter-Edge Data Heterogeneity for Convex objective)**.** *There exists two constants $\sigma_c, \sigma_g \ge 0$, such that $\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x^*)\|^2 \le \sigma_g^2$ and $\frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\|\nabla F_j(x^*) - \nabla f_i(x^*)\|^2 \le \sigma_c^2$ where $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ is one global minimizer.*

The first two assumptions are standard in both convex and non-convex optimization [9]. For Assumption 3, the bounded data heterogeneity is also a standard assumption in FL with different architectures [12], which is used for non-convex case. If all clients in one edge train model on Independent and Identically Distributed (IID) data, i.e. all clients in one edge share similar data and all edge may share different data, then $\sigma_c \simeq 0$. If all edges train model on IID data, i.e., all client in one edge may share different data and all edge share similar data, then $\sigma_g \simeq 0$. A larger σ_c or σ_g indicates a higher level of data heterogeneity. We take similar data as data with same label set and different data as data with different label set. Following Li *et al.* [9] and Koloskova *et al.* [11], Assumption 4 uses one weaker assumption to bound the diversity on intra-edge and inter-edge only at the optima for the convex case.

C. Description of Cyclic HFL

We assume all edge servers participate in a natural cyclic pattern without the guidance from the central server, as shown in Fig.1. In each global update $t = 0, ..., T-1$, the edge servers randomly forms a participated queue, $\mathcal{Q}$, with $|\mathcal{Q}| = M$. Hence, the cyclic process is as follows, the initial model $\mathbf{x}_0$ is randomly initialized or pre-trained on a public dataset and sent to the first edge server in $\mathcal{Q}$. At each edge, standard FL is performed: N clients run K local steps and upload updates to the edge server, and this process repeats for R edge rounds. The aggregated edge model is then passed to the next edge in $\mathcal{Q}$ as initialization. After all selected edges finish, one global round is completed. In the next global round, the first edge starts from the model produced by the last edge in the previous round. Details are provided in Alg. 1.This cyclic design reduces reliance on frequent central server coordination, improves robustness to single-point failures, and scales better under heterogeneous data in large deployments.

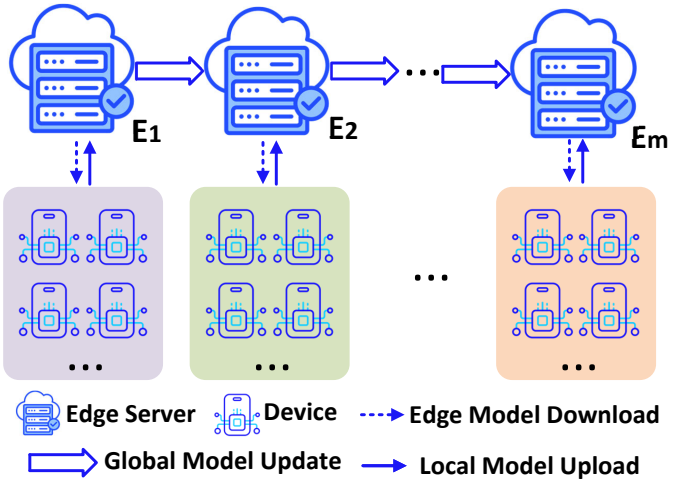


Fig. 1: System architecture of CHFL

Algorithm 1 Cyclic HFL(CHFL)

Initialization: $\mathbf{x}_0$
In each global round $t \in \{0, 1, \ldots, T-1\}$, sample an edge server permutation $\mathcal{Q}$
for edge server $i \in \mathcal{Q}$ **do**
  for *edge round* $r = 0$ to $R-1$ **do**
    for client $j \in \mathcal{N}_i$ **do**
      for *local step* $k = 0$ to $K-1$ **do**
        $\mathbf{x}_{r,k+1}^{i,j} = \mathbf{x}_{r,k}^{i,j} - \eta \mathbf{g}_{r,k}^{i,j\,a}$
      end for
    end for
    $\mathbf{x}_{r+1}^{i} = \mathbf{x}_{r}^{i} - \frac{\eta}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} \mathbf{g}_{r,j}^{\ b}$
  end for
  Transmit $\mathbf{x}^i$ to next edge server.
end for

$^a\mathbf{g}_{r,k}^{i,j} = \nabla F_j(\mathbf{x}_{r,k}^{i,j}, \xi_{r,k}^{i,j})$
$^b\mathbf{g}_{r,j} = \sum_{k=0}^{K-1}\nabla F_j(\mathbf{x}_{r,k}^{i,j}, \xi_{r,k}^{i,j})$

IV. Convergence Theory

In this section, we conduct the convergence analysis on the strongly convex, general convex, and non-convex cases for

CHFL (See proof details in Appendix). By comparing with the convergence rate of other state-of-the-art HFL algorithms, our convergence rate is the optimal. To be more insightful for practical applications, we also present the convergence rate when adopting partial edge/client participate in the CHFL.

Theorem 1. *With Assumptions 1, 2, 4 for strongly convex and general convex, Assumptions 1, 2, and 3 for non-convex case, $\tilde{\eta} := MNRK\eta$, we can obtain the convergence bounds with appropriate learning rates for CHFL as follows:*

a) Strongly convex: With $\Pi_{sc} := \mathbb{E}\left[f\left(\mathbf{x}^T\right) - f\left(\mathbf{x}^*\right)\right]$, and under the following learning rate condition, $\frac{1}{\mu T} \leq \tilde{\eta} \leq \frac{1}{35L}$, we have the convergence rate:

$$\Pi_{sc} \leq \tilde{\mathcal{O}}\Big(\frac{\sigma^2}{\mu MNRKT} + \frac{L\left(M^2NR^2 + NR^2 + 1\right)\sigma_c^2}{M^2N^2K^2\mu^2T^2} + \frac{L\,\sigma_g^2}{M\,\mu^2T^2} + \mu D^2\exp\big(-\frac{\mu T}{70L}\big)\Big) \quad (1)$$

b) General convex: With $\Pi_{gc} := \mathbb{E}\left[f\left(\mathbf{x}^T\right) - f\left(\mathbf{x}^*\right)\right]$, and under the following learning rate condition, $\tilde{\eta} \leq \frac{1}{35L}$, we have the convergence rate:

$$\Pi_{gc} \leq \mathcal{O}\Big(\frac{\sigma D}{\sqrt{MNRKT}} + \frac{\left(L\,\sigma_g^2\,D^4\right)^{1/3}}{(MT^2)^{1/3}} + \frac{LD^2}{T} + \frac{\left(L\left(M^2NR^2 + NR^2 + 1\right)D^4\,\sigma_c^2\right)^{1/3}}{(MNRT)^{2/3}}\Big) \quad (2)$$

c) Non-convex: Let $\Pi_{nc} := \min_{0\leq t\leq T}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{(t)})\|^2\right]$. Under the following learning rate condition, $\tilde{\eta} \leq \frac{1}{35L}$, we have the convergence rate:

$$\Pi_{nc} \leq \mathcal{O}\Big(\frac{(L\,\sigma^2\,H)^{1/2}}{\sqrt{MNRKT}} + \frac{L\,H}{T} + \frac{\left(L^2\,q_\sigma(M,N,R,K)\,H^2\right)^{1/3}}{(MNRKT^2)^{1/3}}\left(\sigma_g^2 + \sigma_c^2\right)^{1/3}\Big) \quad (3)$$

where $\mathcal{O}$ omits absolute constants and $\tilde{\mathcal{O}}$ omits absolute constants and polylogarithmic factors. $H := f(\mathbf{x}^0) - f(\mathbf{x}^*)$ and $q_\sigma(M,N,R,K) = \frac{RNK(M-1)}{3} + \frac{2R+3(R-1)(M-1)NK+3(M-1)(K-1)+\frac{2(K-1)}{RN}+3(K-1)}{6M}$ for the non-convex case. $D := ||\mathbf{x}^0 - \mathbf{x}^*||$ for the convex case.

A. Convergence rate

By Theorem 1, when the stochastic gradient variance is non-negligible, the variance term dominates the bound for sufficiently large T, yielding rates of $\widetilde{\mathcal{O}}(1/(MNRKT))$ for strongly convex objectives and $\mathcal{O}(1/\sqrt{MNRKT})$ for general convex objectives. For smooth non-convex objectives, the result is stated in terms of stationarity, with the leading term scaling as $\mathcal{O}(1/\sqrt{MNRKT})$. Depending on the relative magnitude of the variance and heterogeneity parameters (σ, σ_c, and σ_g), different terms may dominate the bound and lead to different asymptotic behaviors. In particular, when the stochastic variance is negligible (e.g., full-batch gradients or variance-reduced updates), the bound improves and can exhibit faster decays such as $\mathcal{O}(1/T^{2/3})$ in the general convex and non-convex cases, and $\tilde{\mathcal{O}}(1/T^2)$ in the strongly convex case under the corresponding dominance conditions. Our convergence rate improves upon SFL [9], which leverages only N, K to accelerate convergence across the three objective classes. In addition, our rate is faster than the conventional HFL [14], where the acceleration for the strongly convex case depends only on R and K.

1) Effect of data heterogeneity: For strongly and general convex objectives, σ_c and σ_g appear with different coefficients in the bounds, and thus may dominate under different regimes. For non-convex objectives, the heterogeneity enters through the symmetric term $(\sigma_g^2 + \sigma_c^2)^{1/3}$.

• **General convex case.** The decay rate of σ_c is $\mathcal{O}(\frac{(M^2NR^2+NR^2+1)^{\frac{1}{3}}}{(MNRT)^{2/3}})$ and the decay rate of σ_g is $\mathcal{O}(\frac{1}{(MT^2)^{\frac{1}{3}}})$. When $\frac{(M^2NR^2+NR^2+1)^{\frac{1}{3}}}{(MNRT)^{2/3}} < \frac{1}{(MT^2)^{\frac{1}{3}}}$, i.e., $M < N$, the inter-edge data heterogeneity has more effect on convergence speed than intra-edge data heterogeneity. This finding gives us the insights that **when the number of clients per edge exceeds the total number of edge servers (a common case in practice), reducing inter-edge data heterogeneity σ_g can enhance the convergence speed**. For example, if we train a next word prediction model with CHFL for all ages people, and each age has its own word typing habit, it is better to have one edge train the data with all ages to reduce σ_g rather than one edge covering one age.

• **Strongly convex case.** Our scheme can still achieve a faster convergence speed by reducing σ_g when $M < N$ with appropriate settings of K and R to satisfy $MR^2 < NK^2$. This condition ensures that the decay rate of intra-edge data heterogeneity is faster than inter-edge data heterogeneity, i.e. $\frac{M^2NR^2+NR^2+1}{M^2N^2K^2T^2} < \frac{1}{MT^2}$. These results align with the cluster policy where the data distribution among edges tends to be IID, i.e.,$\sigma_g \simeq 0$, as suggested by [15], [19]. However, other works like [14], [24] propose a completely opposite cluster policy by grouping clients with similar data to reduce intra-edge data heterogeneity σ_c and improve convergence speed. This opposing approach is also effective when $MR^2 > NK^2$, such as when the intra-edge data heterogeneity decays more slowly than the inter-edge heterogeneity.

• **Non-convex case** The decay rates of intra-edge and inter-edge data heterogeneity are the same, meaning that changes in system parameters M, R, K, and N have an equivalent impact on the influence of σ_c and σ_g with respect to convergence speed. Moreover, to mitigate the effect of data heterogeneity, Theorem 1 shows that increasing any of M, R, K, or N decreases the factor $q_\sigma(M,N,R,K)$, thereby accelerating convergence. In the extreme case where $N = 1$, CHFL reduces to standard FL with $R \times K$ local steps. In this setting, only inter-edge data heterogeneity σ_g remains, and increased local training along with more frequent aggregation (i.e., larger T) helps reduce its impact.

Based on our convergence analysis, **reducing any form of data heterogeneity, whether intra-edge or inter-edge, can enhance convergence speed. The challenge lies in determin-**

ing which reduction yields the optimal improvement. This largely depends on the settings of M, N, R, K and objectives.

B. Effect of edge round R

Based on Theorem 1, with non-zero σ, increasing R speeds up the convergence with sufficient large T, but it cannot always benefit convergence speed in other scenarios. For example, as in the strongly convex case, larger edge round R has a negative effect on the convergence speed when the intra-edge data heterogeneity dominates convergence. Moreover, when the dominant term depends on inter-edge data heterogeneity for both strongly convex and general convex case, increasing R can not affect the convergence speed. For the non-convex case, increasing R can accelerate convergence when the variance term dominates. In more heterogeneous settings, however, a larger R may amplify the drift-related error, leading to a trade-off.

C. Participation Pattern

Since partial edge/device participation [9] has more practical interest than full edge/device participation, we also derive the bound for partial participation for strongly convex and general convex cases. Consider only $S \leq M$ edges are randomly selected for training in each global round, in which each edge selects $P \leq N$ clients for local training. Fully participation can achieve a better convergence rate than partial participation. There are additional terms caused by partial participation on both strongly convex and general convex cases, which is consistent with [9]. The difference lies in the additional term caused by the edge sampling, i.e., the second term in Eq. (4) and Eq. (5). The inclusion of the two middle terms shows that the number of selected clients and edges can still enhance the convergence speed. Moreover, sampling P clients affects the intra-edge heterogeneity, while sampling S edges affects the inter-edge heterogeneity. Due to the limited space, we take $\phi_{sc}(S, P, K, R, T)$ and $\phi_{gc}(S, P, K, R, T)$ as the last three terms of Eq. 1 and Eq. 2, where M and N are replaced by S and P.

$$\Pi_{\text{sc}} = \tilde{\mathcal{O}}\Big(\frac{\sigma^2}{\mu SPRKT} + \frac{(N-P)\sigma_c^2}{\mu TP(N-1)} + \frac{(M-S)\sigma_g^2}{\mu TS(M-1)} + \phi_{\text{sc}}(S, P, K, R, T)\Big) \tag{4}$$

$$\Pi_{\text{gc}} = \mathcal{O}\Big(\frac{\sigma D}{\sqrt{SPRKT}} + \sqrt{\frac{(N-P)D^2\sigma_c^2}{TP(N-1)}} + \sqrt{\frac{(M-S)D^2\sigma_g^2}{TS(M-1)}} + \phi_{\text{gc}}(S, P, K, R, T)\Big) \tag{5}$$

V. Experiments

A. Experimental Settings

To validate our theoretical findings, we evaluated CNN on non-IID MNIST, ResNet-32 on non-IID CIFAR-10, and LSTM on the natural non-IID Shakespeare dataset [1]. For MNIST, we generate label-based non-IID partitions following [1]. We consider intra-edge and inter-edge heterogeneity, controlled by p_c (labels per client) and p_g (labels per edge), smaller p_c or p_g indicates stronger heterogeneity. We compare SFL [9], FL [1], CFL [12], HFL [14], and CHFL, where CFL is CHFL with $R = 1$ and SFL is CHFL with $N = 1, R = 1$. We use $M = 10$ edges and 500 clients, with $p_c, p_g \in \{1, 2, 5, 10\}$, learning rate $\eta = 0.01$, batch size $b = 32$, $R = 2$, $K = 2$, and $P = 2$. All methods run the same total number of client local steps instead of communication rounds.Our experiment is conducted with one NVIDIA A100 GPU, 4 CPU cores, and 128 GB memory.

B. Numerical Results

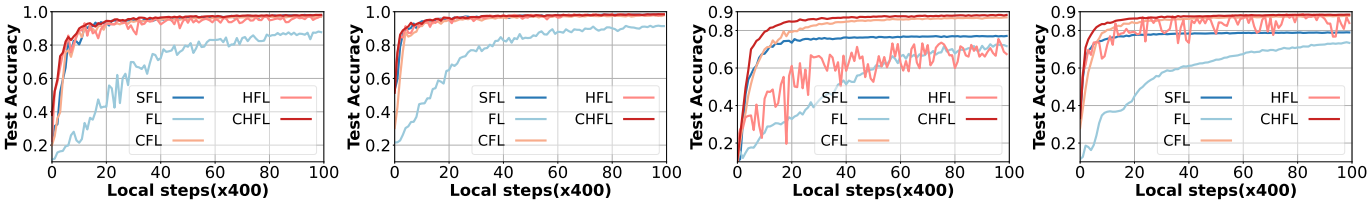
We evaluate test accuracy and convergence speed for the MNIST, CIFAR10 and Shakespeare datasets using different algorithms with various data heterogeneity and edge rounds.

1) Comparisons of Performance: CHFL demonstrates comparable performance under all p_c and p_g settings compared to other algorithms. We have test accuracy results on Table.II. For example, on the MNIST dataset with $p_c = 1$, CHFL achieves an accuracy of 97.84%, outperforming SFL's 97.5%, which is already the highest among the baseline methods. Furthermore, CHFL converges at a similar speed to CFL and SFL, both of which adopt a cyclic pattern, but with better accuracy. As shown in Fig. 2c, all other three methods converge after approximately 20×400 local steps, but CHFL converges at around 10×400 local steps, similar to SFL, while achieving a higher final accuracy of 70.32%, compared to 68.11% for SFL. For the CIFAR-10 dataset, Fig. 3 illustrates that CHFL converges faster than all other methods in terms of training steps, completing convergence after approximately 20×40 local steps and attaining the highest accuracy. On the Shakespeare dataset, CHFL also converges within about 30×16 local steps and achieves the highest accuracy among all compared methods, as shown in Fig. 4a.

2) Effect of Intra-Edge Data Heterogeneity: According to our convergence theory, data heterogeneity directly affects the performance of CHFL, where the more extreme the heterogeneity, the worse the performance. For the MNIST dataset, in the absence of data heterogeneity, CHFL achieves the best accuracy of 99.28%. Under extreme intra-edge heterogeneity, as in Fig.2a, CHFL requires around 30×100 local steps to converge with an accuracy drop of 1.16%. With reduced intra-edge heterogeneity, both convergence speed and accuracy improve. For instance, in Fig.2b, CHFL converges in approximately 20×400 local steps for $p_c = 5$, achieving a 0.32% accuracy improvement compared to the extreme cases of $p_c = 2$. Other algorithms exhibit similar trends that extreme intra-edge heterogeneity leads to slower convergence and lower accuracy, which is consistent with prior studies [14]. However, CHFL is more robust than FL and SFL under extreme intra-edge heterogeneity. This is due to its edge aggregation component with speed R, which aids generalization and accelerates convergence. For the CIFAR-10 dataset, under IID conditions, CHFL converges with the highest accuracy of 77.83%. When intra-edge heterogeneity

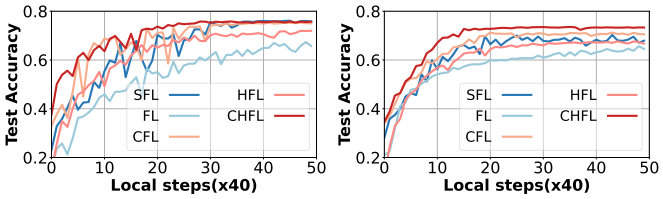
TABLE II: Test accuracy for comparison of various algorithms. Bold indicates the best.

Setting	SFL		FL		CFL		HFL		CHFL	
	MNIST	CIFAR-10	MNIST	CIFAR-10	MNIST	CIFAR-10	MNIST	CIFAR-10	MNIST	CIFAR-10
$p_c = 1$	0.975	0.100	0.8144	0.100	0.9254	0.100	0.8234	0.1139	**0.9784**	**0.1493**
$p_c = 2$	0.9778	0.3425	0.8773	0.5511	0.9693	0.4801	0.9613	0.5857	**0.9812**	**0.5857**
$p_c = 5$	0.9823	0.7573	0.9138	0.552	0.9784	0.7503	**0.9845**	0.720	0.9844	**0.758**
$p_c = 10$	0.991	0.7503	0.9509	0.7168	0.9878	0.7729	0.8266	0.7445	**0.9928**	**0.7783**
$p_g = 1$	0.6811	0.156	0.6285	0.100	0.5824	0.100	0.5111	0.156	**0.7032**	**0.1518**
$p_g = 2$	0.7708	0.3874	0.5309	0.2875	0.7168	0.1967	0.8270	0.2113	**0.9699**	**0.4596**
$p_g = 5$	0.8896	0.6525	0.8339	0.5765	0.9775	0.7059	0.9719	0.6579	**0.9845**	**0.7334**
$p_c = 1, p_g = 1$ (Shakespeare)	0.4516		0.4316		0.4308		0.4379		**0.4743**	



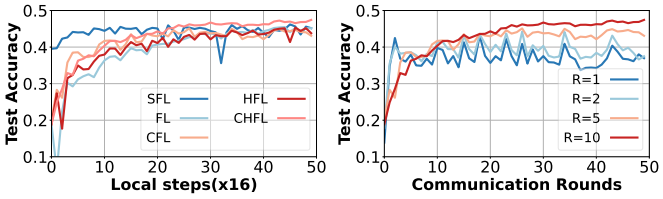
(a) p_c=2 (b) p_c=5 (c) $p_g = 2$ (d) p_g=5

Fig. 2: Test accuracy with data heterogeneity on MNIST Dataset



(a) $p_c = 5$ (b) $p_g = 5$

Fig. 3: Test accuracy with data heterogeneity on CIFAR10 Dataset



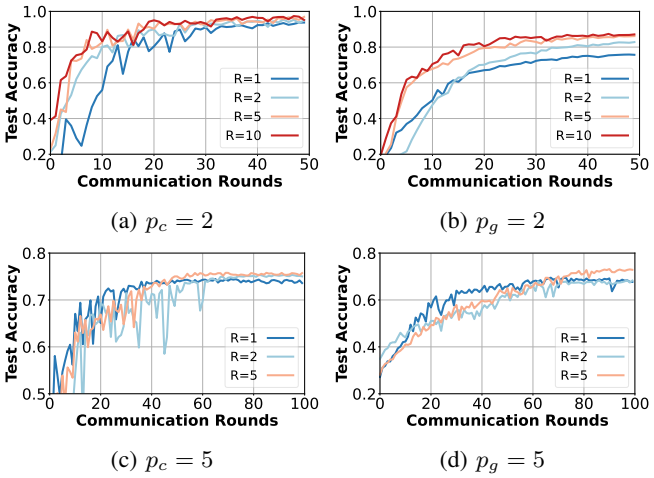
(a) $p_g = 1, p_c = 1$ (b) Effect pf R

Fig. 4: Test accuracy with data heterogeneity and edge round on Shakespeare Dataset

is present in Fig.3a, it reduces accuracy to 75.8%. For the Shakespeare dataset, CHFL converges in about 30×16 local steps, achieving a higher accuracy of 47.43% compared to other algorithms, as shown in Fig.4a.

3) Effect of Inter-Edge Data Heterogeneity: Under our hyperparameter setting, inter-edge data heterogeneity poses a more significant challenge to performance than intra-edge data heterogeneity. For the MNIST dataset, under extreme inter-edge heterogeneity, all algorithms exhibit a decrease in accuracy shown in Fig. 2c. However, CHFL still achieves comparable performance. For instance, CHFL with $p_c = 2$ achieves higher accuracy with 98.12% but 96.99% with $p_g = 2$. These results validate the insight that minimizing inter-edge data heterogeneity, i.e., ensuring data similarity across edges is more beneficial to overall model performance. Compared to CHFL, standard FL and HFL struggle under extreme heterogeneity, particularly with inter-edge heterogeneity. Due to their relatively infrequent model aggregation, especially in non-cyclic architectures, their updates are more susceptible to bias, causing greater fluctuations and instability, as observed in Fig.3a and Fig.3b.In comparison, CHFL achieves similar or even faster convergence than SFL [9] and CFL [12], both of which also adopt cyclic patterns. Importantly, CHFL reduces wall-clock training time compared to SFL by supporting parallel client updates within each edge, unlike SFL's sequential client updates. Compared to CFL, CHFL's additional edge layer with edge aggregation rounds R contributes to faster convergence.

4) Effect of edge round: We evaluate the effect of the edge-round parameter R in CHFL on MNIST, CIFAR-10, and Shakespeare. Overall, a larger R can accelerate convergence, depending on data heterogeneity. On MNIST, increasing R consistently improves convergence speed and accuracy, matching our theoretical results. For example, under severe inter-edge heterogeneity, $R = 10$ converges faster and achieves higher accuracy than $R = 1$. On CIFAR-10, increasing R provides little benefit, with similar convergence speed and accuracy. This suggests that larger R does not always help, as it also appears in the error term of the bound. On Shakespeare, a larger R both accelerates convergence and improves final accuracy. These results highlight that the impact of R is dataset-dependent and closely tied to the level of data heterogeneity.



(a) $p_c = 2$ (b) $p_g = 2$

(c) $p_c = 5$ (d) $p_g = 5$

Fig. 5: Test accuracy with data heterogeneity and edge round on MNIST (a,b) and CIFAR10 (c,d) Datasets.

VI. Conclusion

In this paper, we establish convergence guarantees for CHFL under heterogeneous data across strongly convex, general convex, and non-convex objectives. Our rates match the best-known order-optimal results in FL and HFL. We further analyze the impact of update schemes, heterogeneity, and key hyperparameters, showing that clustering clients based solely on data similarity does not necessarily improve convergence. Instead, its effectiveness depends on system settings and learning objectives. These insights provide guidance for deploying CHFL in practice.

VII. Acknowledgment

The work of J. Fu and L. Guo was partially supported by National Science Foundation under grant CCF-2312616. The work of H. Yang was partially supported by RIT CHAI Faculty Seed Grant, NIH award R16GM159671 and NSF grant CNS-2112471. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The work of X. Zhang was partially supported by the National Science Foundation under Grant CCF-2312617.

References

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

[2] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent iot applications: A cloud-edge based framework," IEEE Open Journal of the Computer Society, vol. 1, pp. 35–44, 2020.

[3] T. Wang, Y. Liu, X. Zheng, H.-N. Dai, W. Jia, and M. Xie, "Edge-based communication optimization for distributed federated learning," IEEE Transactions on Network Science and Engineering, vol. 9, no. 4, pp. 2015–2024, 2021.

[4] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical federated learning with quantization: Convergence analysis and system design," IEEE Transactions on Wireless Communications, vol. 22, no. 1, pp. 2–18, 2022.

[5] J.-w. Lee, J. Oh, S. Lim, S.-Y. Yun, and J.-G. Lee, "Tornadoaggregate: Accurate and scalable federated learning via the ring-based architecture," arXiv preprint arXiv:2012.03214, 2020.

[6] J. Liu, X. Wei, X. Liu, H. Gao, and Y. Wang, "Group-based hierarchical federated learning: Convergence, group formation, and sampling," in Proceedings of the 52nd International Conference on Parallel Processing, 2023, pp. 264–273.

[7] C. Zhu, Z. Xu, M. Chen, J. Konečnỳ, A. Hard, and T. Goldstein, "Diurnal or nocturnal? federated learning of multi-branch networks from periodically shifting distributions," in International Conference on Learning Representations, 2021.

[8] M. Paulik, M. Seigel, H. Mason, D. Telaar, J. Kluivers, R. van Dalen, C. W. Lau, L. Carlson, F. Granqvist, C. Vandevelde et al., "Federated evaluation and tuning for on-device personalization: System design & applications," arXiv preprint arXiv:2102.08503, 2021.

[9] Y. Li and X. Lyu, "Convergence analysis of sequential federated learning on heterogeneous data," Advances in Neural Information Processing Systems, vol. 36, 2024.

[10] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in International conference on machine learning. PMLR, 2020, pp. 5132–5143.

[11] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in International Conference on Machine Learning. PMLR, 2020, pp. 5381–5393.

[12] Y. J. Cho, P. Sharma, G. Joshi, Z. Xu, S. Kale, and T. Zhang, "On the convergence of federated averaging with cyclic client participation," arXiv preprint arXiv:2302.03109, 2023.

[13] L. U. Khan, M. Guizani, A. Al-Fuqaha, C. S. Hong, D. Niyato, and Z. Han, "A joint communication and learning framework for hierarchical split federated learning," IEEE Internet of Things Journal, 2023.

[14] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020, pp. 1–6.

[15] N. Mhaisen, A. A. Abdellatif, A. Mohamed, A. Erbad, and M. Guizani, "Optimal user-edge assignment in hierarchical federated learning based on statistical properties and network topology constraints," IEEE Transactions on Network Science and Engineering, vol. 9, no. 1, pp. 55–66, 2021.

[16] H. Zhou, J. Fu, Y. Yuan, L. Guo, and X. Zhang, "Similarity-guided rapid deployment of federated intelligence over heterogeneous edge computing," in IEEE INFOCOM 2025-IEEE Conference on Computer Communications. IEEE, 2025, pp. 1–10.

[17] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 8866–8870.

[18] B. Xu, W. Xia, W. Wen, P. Liu, H. Zhao, and H. Zhu, "Adaptive hierarchical federated learning over wireless networks," IEEE Transactions on Vehicular Technology, vol. 71, no. 2, pp. 2070–2083, 2021.

[19] Y. Deng, F. Lyu, J. Ren, Y. Zhang, Y. Zhou, Y. Zhang, and Y. Yang, "Share: Shaping data distribution at edge for communication-efficient hierarchical federated learning," in 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). IEEE, 2021, pp. 24–34.

[20] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," arXiv preprint arXiv:1811.03604, 2018.

[21] D. Huba, J. Nguyen, K. Malik, R. Zhu, M. Rabbat, A. Yousefpour, C.-J. Wu, H. Zhan, P. Ustinov, H. Srinivas et al., "Papaya: Practical, private, and scalable federated learning," Proceedings of Machine Learning and Systems, vol. 4, pp. 814–832, 2022.

[22] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16. Springer, 2016, pp. 795–811.

[23] R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon, and U. Topcu, "Faster non-convex federated learning via global and local momentum," in Uncertainty in Artificial Intelligence. PMLR, 2022, pp. 496–506.

[24] Z. Wang, H. Xu, J. Liu, Y. Xu, H. Huang, and Y. Zhao, "Accelerating federated learning with cluster construction and hierarchical aggrega-

tion," IEEE Transactions on Mobile Computing, vol. 22, no. 7, pp. 3805–3822, 2022.

APPENDIX

Due to space limitations, detailed proofs are available in https://github.com/fuJennifer/chfl:

A. Strongly convex case

Let Assumptions 1, 2, 4 hold, with each client doing a SGD update, at t global round, the global model update of CHFL after one complete training round,

$$\Delta \mathbf{x} = \mathbf{x}^{t+1} - \mathbf{x}^{t} = -\eta \sum_{i=0}^{S-1} \sum_{j=0}^{P-1} \sum_{r=0}^{R-1} \sum_{k=0}^{K-1} \mathbf{g}_{r,k}^{i,j} \tag{6}$$

where $\mathbf{g}_{r,k}^{i,j} = \nabla F_j(\mathbf{x}_{r,k}^{i,j}, \xi_{r,k}^{i,j})$. Based on Lemma 2 and Jensen's inequality in C.2 [9], we have optimality gap,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x} + \Delta\mathbf{x} - \mathbf{x}^*\|^2\right] \leq (1 - \frac{\mu MPKR\eta}{2})\|\mathbf{x} - \mathbf{x}^*\|^2 \\
&+ 7SPRK\eta^2\sigma^2 + 7S^2P^2R^2K^2\eta^2\frac{(N-P)}{P(N-1)}\sigma_c^2 \\
&+ \frac{21}{S(M-1)}\eta^2\sigma_g^2 - \frac{8}{5}SPKR\eta D_F(\mathbf{x}, \mathbf{x}^*) \\
&+ 42LS^2P^2R^2K^2 D_f(\mathbf{x}, \mathbf{x}^*) \\
&+ \left(2L\eta + 7L^2SPRK\eta^2\right) \underbrace{\sum_{i,j,r,k} \mathbb{E}\left[\left\|\mathbf{x}_{r,k}^{i,j} - \mathbf{x}\right\|^2\right]}_{\text{client drift } \mathbb{E}_t}
\end{aligned}$$

Follow Lemma 4 and Lemma 6 in [9] to bound client drift:

$$\begin{aligned}
\mathbb{E}_t &= \mathbb{E}\left[\left\|-\eta \sum\nolimits_{i',r',j',k'} g_{r',k'}^{i',j'}\right\|^2\right] \leq \frac{71}{10}\eta^2 q_B \sigma^2 + \frac{71}{10} q_c \eta^2 \sigma_c^2 \\
&+ \frac{71}{5} L q_{B^2} \eta^2 D_F(\mathbf{x}, \mathbf{x}^*) + 22 q_g \eta^2 \sigma_g^2 + 43 L q_{B^2} \eta^2 D_f(\mathbf{x}, \mathbf{x}^*)
\end{aligned}$$

Where $D_F(\mathbf{x}, \mathbf{x}^*) = \mathbb{E}\left[\|\nabla F_j(\mathbf{x}) - \nabla F_j(\mathbf{x}^*)\|^2\right]$ and $D_f(\mathbf{x}, \mathbf{x}^*) = \mathbb{E}\left[\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|^2\right]$.

Substitute $\mathbb{E}_t$ into $\mathbb{E}\left[\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|^2\right]$ with S, P, R, K replacement for one global round,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|^2\right] \leq 5\mu\left\|\mathbf{x}^0 - \mathbf{x}^*\right\|^2 \exp\left(-\frac{\mu}{2}\tilde{\eta}T\right) \\
&+ \left\{\frac{27\sigma^2}{SPKR} + \frac{24(N-P)}{P(N-1)}\sigma_c^2 + \frac{70(M-S)}{S(M-1)}\sigma_g^2\right\}\tilde{\eta} \\
&+ \left\{18L\left(\frac{1}{P} + \frac{1}{S^2P} + \frac{1}{S^2P^2R^2})\sigma_c^2 + \frac{53L}{S}\sigma_g^2\right)\right\} \times \tilde{\eta}^2
\end{aligned} \tag{7}$$

Based on Lemma 7 in [9] and L-smooth assumption,

$$\begin{aligned}
&\mathbb{E}\left[f\left(\mathbf{x}^T\right) - f\left(\mathbf{x}^*\right)\right] \\
&\leq \tilde{\mathcal{O}}\left(\mu D \exp(-\frac{\mu T}{70L}) + \frac{\sigma^2}{SPKR\mu T}\frac{(N-P)\sigma_c^2}{\mu TP(N-1)}\right. \\
&\left. + \frac{(M-S)}{S(M-1)\mu T}\sigma_g^2 + \frac{L\sigma_g^2}{S\mu^2T^2} + \frac{\left(\frac{1}{P} + \frac{1}{S^2P} + \frac{1}{S^2P^2P^2}\right)L\sigma_c^2}{\mu^2T^2}\right)
\end{aligned} \tag{8}$$

When setting $P = N$ and $S = M$, we get Eq. (1).

B. General convex case

For general case, when $\mu = 0$, the following bound based on Eq.(7) can be given,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|^2\right] \leq \left\|\mathbf{x}^t - \mathbf{x}^*\right\|^2 + \frac{79}{10}\frac{\tilde{\eta}^2}{SPKR}\sigma^2 \\
&+ \left[\frac{7\tilde{\eta}^2(N-P)}{P(N-1)} + \frac{53}{10}\tilde{\eta}^3 L\left(\frac{1}{P} + \frac{1}{S^2P} + \frac{1}{S^2P^2R^2}\right)\right]\sigma_c^2 \\
&+ \left[\frac{21\tilde{\eta}^2(M-S)}{S(M-1)} + \frac{3L\tilde{\eta}^3}{S}\right]\sigma_g^2 - \frac{3}{10}\tilde{\eta}D_F(\mathbf{x}, \mathbf{x}^*)
\end{aligned}$$

Applying Lemma 8 in [9], we get,

$$\begin{aligned}
&\mathbb{E}\left[f\left(\mathbf{x}^T\right) - f\left(\mathbf{x}^*\right)\right] \leq \mathcal{O}\left(\frac{\sigma D}{\sqrt{SPRKT}} + \sqrt{\frac{N-P}{P(N-1)T}}D\sigma_c\right. \\
&+ \sqrt{\frac{M-S}{(M-1)ST}}\sigma_g + \frac{\left(LD^4\sigma_c^2\right)^{\frac{1}{3}}}{P^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\left(L\sigma_c^2D^4\right)^{\frac{1}{3}}}{(ST)^{\frac{2}{3}}P^{\frac{1}{3}}} \\
&\left. + \frac{\left(L\sigma_c^2D^4\right)^{\frac{1}{3}}}{(SPRT)^{\frac{2}{3}}}\frac{\left(L\sigma_g^2D^4\right)^{\frac{1}{3}}}{M^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{LD^2}{T}\right)
\end{aligned}$$

When setting $P = N$ and $S = M$, we get Eq. (2).

C. Non-convex case

We focus on a single training round and drop t:

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})] \leq \mathbb{E}\langle\nabla F(\mathbf{x}), \Delta\mathbf{x}\rangle + \frac{L}{2}\mathbb{E}\|\Delta\mathbf{x}\|^2 \\
&\leq -\eta\sum_{i,j,r,k}\mathbb{E}\left[\left\langle\nabla F(\mathbf{x}), \nabla F_j(\mathbf{x}_{r,k}^{i,j})\right\rangle\right] + \frac{L\eta^2}{2}\mathbb{E}\left\|\sum_{i,j,r,k}\mathbf{g}_{r,k}^{i,j}\right\|^2
\end{aligned}$$

Bounding the first term in above equation,

$$\begin{aligned}
&-\eta\sum_{i,j,r,k}\mathbb{E}\left[\left\langle\nabla F(\mathbf{x}), \nabla F_j\left(\mathbf{x}_{r,k}^{i,j}\right)\right\rangle\right] \leq -\frac{\eta SPRK}{2}\|\nabla F(\mathbf{x})\|^2 \\
&+ \frac{L^2\eta}{2}\sum_{i,j,r,k}\mathbb{E}\left[\left\|\mathbf{x}_{r,k}^{i,j} - \mathbf{x}\right\|^2\right] - \frac{\eta}{2}\sum_{i,j,r,k}\mathbb{E}\left\|\nabla F_j\left(\mathbf{x}_{r,k}^{i,j}\right)\right\|^2
\end{aligned}$$

Bounding the second term,

$$\begin{aligned}
\frac{L\eta^2}{2}\mathbb{E}\left\|\sum_{i,j,r,k}\mathbf{g}_{r,k}^{i,j}\right\|^2 &\leq L\eta^2 SPRK\sigma^2 \\
&+ LSPRK\eta^2\sum_{i,j,r,k}\mathbb{E}\left[\left\|\nabla F_j\left(\mathbf{x}_{r,k}^{i,j}\right)\right\|^2\right]
\end{aligned}$$

When $\eta \leq \frac{1}{35LSPKR}$, we have $-\frac{\eta}{2}(1 - 2L\eta SKRP)\sum_{i,j,r,k}\mathbb{E}\left[\left\|\nabla F_j\left(\mathbf{x}_{r,k}^{i,j}\right)\right\|\right] < 0$. Thus, we bound $\mathbb{E}_t$ and substitute $\mathbb{E}_t$ into $\mathbb{E}\left[F\left(\mathbf{x}^{t+1}\right) - F\left(\mathbf{x}^t\right)\right]$. At this time, we subtracting F^* for both side and follow Lemma 8 in [9], The convergence rate of non-convex case is followed and $H := f(\mathbf{x}^0) - f(\mathbf{x}^*)$,

$$\begin{aligned}
&\min_{0\leq t\leq T}\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}^t\right)\right\|^2\right] \\
&\leq \mathcal{O}\left(\frac{\left(L\sigma^2H\right)^{\frac{1}{2}}}{\sqrt{SRKPT}} + \frac{\left(L^2H^2q_\sigma\left(\sigma_g^2 + \sigma_c^2\right)\right)^{\frac{1}{3}}}{(SPRKT^2)^{\frac{1}{3}}} + \frac{LH}{T}\right)
\end{aligned}$$

When setting $P = N$ and $S = M$, we get Eq. (3).