UNIFIED ANALYSES FOR HIERARCHICAL FEDERATED LEARNING: TOPOLOGY SELECTION UNDER DATA HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical Federated Learning (HFL) addresses critical scalability limitations in conventional federated learning by incorporating intermediate aggregation layers, yet optimal topology selection across varying data heterogeneity conditions and network conditions remains an open challenge. This paper establishes the first unified convergence framework for all four HFL topologies (Star-Star, Star-Ring, Ring-Star, and Ring-Ring) under non-convex objectives and different intra/inter-group data heterogeneity. Our theoretical analysis reveals three fundamental principles for topology selection: (1) The top-tier aggregation topology exerts greater influence on convergence than the intra-group topology, with ring-based top-tier configurations generally outperforming star-based alternatives; (2) Optimal topology strongly depends on client grouping characteristics, where Ring-Star excels with numerous small groups while Star-Ring is superior for large, client-dense clusters; and (3) Inter-group heterogeneity dominates convergence dynamics across all topologies, necessitating clustering strategies that minimize inter-group divergence. Extensive experiments on CIFAR-10/CINIC-10/Fashion-MNIST with ResNet-18/VGG-9/ResNet-10 validate these insights, and provide practitioners with theoretically grounded guidance for HFL system design in real-world deployments.

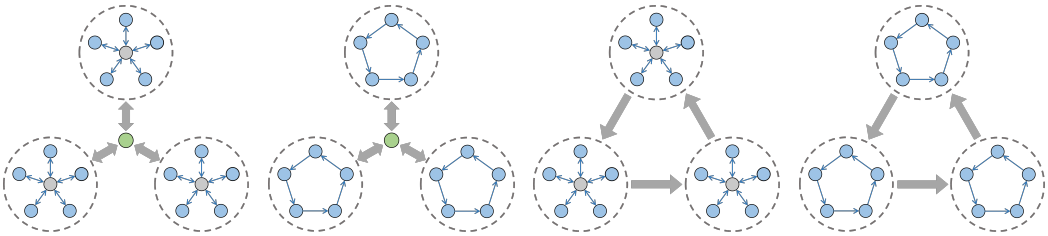
1 INTRODUCTION

Federated Learning (FL)(McMahan et al., 2017) has revolutionized collaborative machine learning by enabling distributed model training across decentralized devices while preserving data privacy. However, conventional single-tier FL faces critical scalability challenges in large-scale deployments, including communication bottlenecks, synchronization latency, and vulnerability to single-point failures. Hierarchical Federated Learning (HFL)(Liu et al., 2020; Deng et al., 2021) has emerged as a promising paradigm, introducing intermediate aggregation layers (such as edge servers or cluster heads) to form a two/multi-tier architecture that distributes the coordination burden for massive deployment. Despite its promise, the theoretical understanding of HFL remains nascent, particularly under realistic conditions of data heterogeneity and diverse hierarchical topologies.

In two-tier HFL frameworks, each level of aggregation can adopt either *star (parallel) or ring (sequential)* topology, yielding four distinct configurations: Star-Star, Star-Ring, Ring-Star, and Ring-Ring (see Figure 1). These topological choices fundamentally influence the convergence dynamics, robustness to data heterogeneity, and communication efficiency. For instance, star aggregation enables parallel client updates but may suffer from abrupt synchronization of divergent models, while ring aggregation propagates updates sequentially, potentially mitigating client drift in non-IID settings through incremental alignment (Li & Lyu, 2023; 2025).

Literature Review. Existing theoretical analyses of HFL have largely focused on the Star-Star topology on non-convex functions (Zhou & Cong, 2019; Wang et al., 2022; Castiglia et al., 2021), data heterogeneity(Wang et al., 2022), partial client participation (Jiang & Zhu, 2024), and other variants (Liu et al., 2022; Yang et al., 2023). Some recent works explore Star-Ring topology (Lee et al., 2020; Ding et al., 2024; Fang et al., 2022) and the Ring-Star topology (Chaoyang et al., 2020; Huang et al., 2024). However, a unified convergence analysis that compares all four topologies is

still lacking. This theoretical gap impedes informed topology selection in practical deployments, where system performance is highly sensitive to data distribution and network conditions.



(a) Star-Star HFL (b) Star-Ring HFL (c) Ring-Star HFL (d) Ring-Ring HFL

Figure 1: Different topology configurations of HFL

Research Question. The central research problem is:

How should practitioners select the optimal HFL topology configuration when facing varying degrees of intra-group and inter-group data heterogeneity, diverse client grouping characteristics, and constrained network conditions?

This research problem is of practical importance for system convergence. For example, in systems with high inter-group heterogeneity (such as clients clustered by geographic region with distinct data distributions), selecting star topology at the top tier may amplify inter-group divergence through abrupt parallel synchronization. Conversely, ring aggregation at the top tier enables gradual, sequential alignment that may better accommodate distributional differences. This topology selection problem is further complicated when considering diverse client grouping characteristics in HFL. Without principled guidance for this selection, system designers face a critical "topology lottery", where deployment success depends on unguided architectural choices rather than theoretically-grounded decisions.

Analytical Challenges. Establishing a comprehensive convergence framework for HFL that encompasses all four topology configurations presents three critical challenges, stemming from the intricate hierarchical structure of HFL and the complex interactions between topology choices and data heterogeneity:

(1) Hierarchical Heterogeneity Interdependence. Unlike single-tier FL, HFL exhibits a cascading heterogeneity relationship where intra-group and inter-group data distributions interact in non-trivial ways. Specifically, intra-group client drift destabilizes lower-tier aggregations, which then amplifies inter-group divergence during upper-tier synchronization. This creates a feedback loop where local model divergence within groups directly exacerbates global model inconsistency. The mathematical consequence is that heterogeneity cannot be decomposed into independent terms; rather, convergence bounds must account for the multiplicative interaction between intra/inter-group divergence, requiring novel analytical techniques beyond conventional FL frameworks.

(2) Cross-Tier Dynamic Coupling. The two-tier aggregation architecture creates a bidirectional dependency where updates at one tier directly influence the error propagation at the other tier. This coupling means that convergence behavior emerges from the interaction of hierarchical layers rather than from the sum of individual tier performances. For example, the effective learning rate at the global tier depends on the accumulated error from lower-tier aggregations, while the stability of lower-tier updates is conditioned on the quality of the global model. This interdependence invalidates standard approaches that analyze hierarchical systems through sequential single-tier approximations.

(3) Compounded Topology-Specific Biases. Different topologies introduce distinct statistical properties that compound across hierarchical layers in topology-dependent ways. Star topology provides unbiased parallel updates but suffers from high variance, while ring topology reduces variance through sequential updates but introduces temporal bias that accumulates along the update chain. Critically, in HFL these effects compound across tiers: ring-based lower tiers accumulate inter-client gradient biases, while upper-tier ring aggregation propagates outdated global estimates. Such topology-specific error propagation patterns require sophisticated cross-client error term analysis

Table 1: Convergence Rates for Different Hierarchical Topologies

Topology	Convergence Rate[4]
Star-Star	$\mathcal{O}\big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{PMKR}} + \frac{(\boldsymbol{L\sigma^2 A})^{1/2}}{\sqrt{\boldsymbol{P^2K^2R}}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\big)$
	$\mathcal{O}\big(\frac{LA}{R} + \Big(\frac{LA\sigma^2}{KPMGR}\Big)^{1/2} + \Big(\frac{LA\sigma}{R}\sqrt{\frac{M}{KP}}\Big)^{1/2} + \Big(\frac{LA\zeta}{R}\Big)^{2/3} + \Big(\frac{LA\hat{\zeta}}{PR}\Big)^{2/3}\big)$[1]
	$\mathcal{O}\big(\frac{LA}{\sqrt{PKR}} + \frac{\sigma^2}{MR} + \frac{\sigma^2}{PR} + \frac{1}{\sqrt{PKR}}\frac{\sigma^2}{GM}\big)$[2]
	$\mathcal{O}\big(\frac{LA}{\sqrt{GPMKR}} + \frac{L\sigma^2}{\sqrt{GPMKR}} + \frac{\hat{\zeta}^2}{K^2R} + \frac{\zeta^2}{K^2R} + \frac{\zeta^2}{R}\big)$[3]
Star-Ring	$\mathcal{O}\big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{PMKR}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\big)$
Ring-Star	$\mathcal{O}\big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{\boldsymbol{G}PMKR}} + \frac{(\boldsymbol{L\sigma^2 A})^{1/2}}{\sqrt{\boldsymbol{G^2P^2K^2R}}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{\boldsymbol{G^{2/3}}P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\big)$
Ring-Ring	$\mathcal{O}\big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{\boldsymbol{G}PMKR}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{\boldsymbol{G^{2/3}}P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\big)$

[1] HSGD under Non-IID data on non-convex case Wang et al. (2022).

[2] Hier-Local-QSGD under IID data on non-convex case Liu et al. (2022).

[3] HFL with a partial client participation under Non-IID data on non-convex case Jiang & Zhu (2024).

[4] We omit absolute constants and polylogarithmic factors. R denotes the number of global rounds. G denotes the number of groups. P denotes the number of group update steps. M denotes the number of clients in a group. K denotes the number of local steps. σ denotes SGD variance. ζ denotes inter-group heterogeneity. $\hat{\zeta}$ denotes intra-group heterogeneity. L denotes L-smoothness constant. $A \coloneqq F(\mathbf{x}^{(0)}) - F^*$.

that explicitly tracks how topology choices modulate bias-variance tradeoffs across hierarchical layers.

Contributions. This paper establishes the first unified theoretical framework that analyzes and compares all four HFL topology configurations under non-convex optimization objective and different intra/inter-group data heterogeneity (see Table 1). Our convergence bounds explicitly quantify the effects of key system parameters, including number of groups G , clients per group M, local steps K, and group rounds P, and reveal how topology choices interact with intra/inter-group data heterogeneity to shape convergence behavior. Our theoretical analysis formalizes the tradeoff between ring and star aggregation mechanisms across hierarchical tiers and provides principled guidance for topology selection in practical deployments.

(1) HFL prioritizes scalability over convergence acceleration. Counterintuitively, HFL is primarily valuable for enabling large-scale deployments where single-tier FL becomes impractical, rather than inherently accelerating convergence. Crucially, HFL with ring aggregation at the top tier (Ring-Star, Ring-Ring) consistently outperforms star-based counterparts under data heterogeneity. This reveals that carefully selected single-tier FL configurations may actually converge faster than two-tier HFL, positioning HFL as a solution for scalability constraints rather than a convergence accelerator.

(2) Inter-group heterogeneity dominates convergence dynamics. We establish that inter-group data divergence (ζ) exerts a more significant impact on convergence than intra-group heterogeneity ($\hat{\zeta}$) across all four topologies. This finding fundamentally reshapes client clustering strategies, indicating that minimizing inter-group distributional differences should take precedence over optimizing intra-group homogeneity. Effective grouping, such as forming clusters with approximately IID inter-group distributions, would converge faster than fine-tuning intra-group training dynamics.

(3) Optimal topology selection depends critically on group structure. Our analysis reveals that the optimal topology selection depends critically on group structural characteristics. Ring-Star excels when numerous small groups exist, as sequential inter-group updates benefit from increased par-

allelism at the lower tier and fine-grained global alignment. Star-Ring is preferable for few large, client-dense clusters, where intra-group ring aggregation enables deep local refinement before global synchronization.

We validate these theoretical insights through extensive experiments on CIFAR-10, CINIC-10, and Fashion-MNIST using ResNet-18, VGG-9, and ResNet-10 under four distinct heterogeneity scenarios. The results consistently demonstrate accuracy gains from informed topology selection, with ring-based top-tier configurations showing particular advantages in heterogeneous environments. By establishing this unified analytical framework, our work bridges a critical gap in HFL theory and provides actionable, theoretically grounded guidance for system design.

2 CONVERGENCE THEORY

This section presents a unified convergence analysis of HFL under non-convex optimization objectives. In the following, we formalize the setup of HFL with four different topologies, introduce general assumptions, derive the convergence bounds, and extract actionable insights for topology selection in practical deployments.

2.1 SETUP

We begin by formalizing the HFL framework and the update mechanisms for each topology configuration. In two-tier HFL, the global objective is to minimize:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left\{F(\mathbf{x})=\frac{1}{G}\sum_{g=1}^{G}F_g(\mathbf{x})=\frac{1}{G}\sum_{g=1}^{G}\frac{1}{M}\sum_{m=1}^{M}F_{g,m}(\mathbf{x})\right\} \tag{1}$$

where F_g represents the average local objective function over all clients in group g ($g \in [G]$), and $F_{g,m}$ denotes the local objective function of client m ($m \in [M]$) in group g, defined as $F_{g,m}(x) = \mathbb{E}_{\xi\sim\mathcal{D}_m}[f_m(x;\xi)]$, where $\mathcal{D}_m$ is the local dataset of client m.

The process of HFL with four topology configurations operate according to distinct update rules (see detailed algorithms in the Appendix):

(1) Star-Star. Each group g initializes its model as $\mathbf{x}_{g,0}^{(r)} = \mathbf{x}^{(r)}$. Within each group, clients initialize their models as $\mathbf{x}_{g,p,m,0}^{(r)} = \mathbf{x}_{g,p}^{(r)}$, perform K parallel local updates, and send updates to the group server for aggregation. After P group updates, the global server aggregates group parameters to generate the next global parameters $\mathbf{x}^{(r+1)}$.

(2) Star-Ring. Each group g initializes its model as $\mathbf{x}_{g,0}^{(r)} = \mathbf{x}^{(r)}$. Within each group, clients initialize their models from the previous client in sequence and perform K local updates. The group server aggregates the latest parameters from the last client. After P group updates, group servers send their updated parameters to the global server for aggregation.

(3) Ring-Star. Each group g initializes its model with the latest parameters from the previous group. Within each group, clients initialize their models as $\mathbf{x}_{g,p,m,0}^{(r)} = \mathbf{x}_{g,p}^{(r)}$, perform K parallel local updates, and send updates to the group server for aggregation. After P group updates, group servers send their updated parameters to the next group in sequence.

(4) Ring-Ring. Each group g initializes its model with the latest parameters from the previous group. Within each group, clients initialize their models from the previous client and perform K local updates. The group server aggregates the latest parameters from the last client. After P group updates, group servers send their updated parameters to the next group in sequence.

2.2 ASSUMPTIONS

Assumption 1. *(L-Smoothness). Each local objective function $F_{g,m}$ is L-smooth, $g \in \{1, 2, ..., G\}$, $m \in \{1, 2, ..., M\}$, i.e., there exists one constant L such that*

$$\|\nabla F_{g,m}(\mathbf{x}) - \nabla F_{g,m}(\mathbf{y})\| \le L\|x - y\| \quad \textit{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2}$$

Assumption 2. *(Bounded variance). For the local objective function $F_{g,m}$ in any client, the local stochastic gradient $\nabla F_{g,m}(\mathbf{s}, \xi_m)$ computed using a mini-batch ξ_m, sampled uniformly at random from local dataset, has bounded variance, that is $\|\nabla F_{g,m}(\mathbf{x}, \xi_m) - \nabla F_{g,m}(\mathbf{x}', \xi_m)\| \leq \sigma^2$, for all clients.*

Assumption 3. *(Bounded Inter-Group Heterogeneity). There exists one constant ζ^2 , $g \in \{1, 2, ..., G\}$, such that*

$$\frac{1}{G}\sum_{g=1}^{G} \|\nabla F(\mathbf{x}) - \nabla F_g(\mathbf{x})\|^2 \leq \zeta^2 \tag{3}$$

Assumption 4. *(Bounded Intra-Group Heterogeneity). There exists one constant ζ_g^2 , $g \in \{1, 2, ..., G\}$, $m \in \{1, 2, ..., M\}$, such that*

$$\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_{g,m}(\mathbf{x}) - \nabla F_g(\mathbf{x})\|^2 \leq \zeta_g^2 \tag{4}$$

Furthermore, we define the average intra-group heterogeneity as $\hat{\zeta}^2 := \frac{1}{G}\sum_{g=1}^{G} \zeta_g^2$.

The first two assumptions are standard in non-convex optimization(Ghadimi & Lan, 2013; Bottou et al., 2018). Assumptions 3 and 4 extend standard FL analysis to the hierarchical setting, explicitly modeling both inter/intra-group data heterogeneity (Wang & Ji, 2022). In particular, Assumption 3 (bounded inter-group heterogeneity) and Assumption 4 (bounded intra-group heterogeneity) measures the data heterogeneity across or within client groups, respectively. For example, $\zeta^2 = 0$ when the data is IID across all groups. This means that the statistical distributions of data across different groups are identical. As a result, the average gradients of the groups do not deviate from the global gradient, eliminating inter-group divergence.

2.3 Convergence Analysis

Theorem 1. *Under Assumptions 2– 4, the following convergence bounds hold for each HFL topology, where $A = F(x^{(0)}) - F^*$ represents the initial optimality gap.*

***Star-Star**: There exists $\tilde{\eta} = PK\eta$, and $\tilde{\eta} \leq \frac{1}{12L}$, such that*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] \lesssim \frac{A}{\tilde{\eta}R} + \frac{L\tilde{\eta}\sigma^2}{PMK} + \frac{L\tilde{\eta}\sigma^2}{P^2K^2} + \frac{L^2\tilde{\eta}^2\hat{\zeta}^2}{P^2} + L^2\tilde{\eta}^2\zeta^2. \tag{5}$$

***Star-Ring**: There exists $\tilde{\eta} = PMK\eta$, and $\tilde{\eta} \leq \frac{1}{12L}$, such that*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] \lesssim \frac{A}{\tilde{\eta}R} + \frac{L\tilde{\eta}\sigma^2}{PMK} + \frac{L^2\tilde{\eta}^2\hat{\zeta}^2}{P^2} + L^2\tilde{\eta}^2\zeta^2. \tag{6}$$

***Ring-Star**: There exists $\tilde{\eta} = GPK\eta$, and $\tilde{\eta} \leq \frac{1}{12L}$, such that*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] \lesssim \frac{A}{\tilde{\eta}R} + \frac{L\tilde{\eta}\sigma^2}{GPMK} + \frac{L^2\tilde{\eta}^2\sigma^2}{G^2P^2K^2} + \frac{L^2\tilde{\eta}^2\hat{\zeta}^2}{G^2P^2} + L^2\tilde{\eta}^2\zeta^2. \tag{7}$$

***Ring-Ring**: There exists $\tilde{\eta} = GPMK\eta$, and $\tilde{\eta} \leq \frac{1}{12L}$, such that*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] \lesssim \frac{A}{\tilde{\eta}R} + \frac{L\tilde{\eta}\sigma^2}{GPMK} + \frac{L^2\tilde{\eta}^2\hat{\zeta}^2}{G^2P^2} + L^2\tilde{\eta}^2\zeta^2. \tag{8}$$

Effective Learning Rate. Theorem 1 introduces a topology-dependent effective learning rate, denoted by $\tilde{\eta}$, which incorporates key architectural parameters: the number of groups G, group rounds P, clients per group M, local update steps K, and global rounds R. This effective learning rate captures the cumulative impact of the hierarchical update structure on convergence dynamics. The derived bounds in Theorem consist of two components: an *optimization term* that decreases with R, and *error terms* arising from stochastic noise and data heterogeneity. While a larger $\tilde{\eta}$ accelerates optimization (i.e., reducing optimization term), it also magnifies the error term. To balance the tradeoff, Corollary 1 prescribes an appropriate $\tilde{\eta}$ that minimizes the overall convergence bound.

Corollary 1. *(Convergence under effective learning rate). By choosing learning rate $\tilde{\eta} \leq 1/(12L)$, the convergence rate satisfies the following, where $\mathcal{O}(\cdot)$ hides absolute constants:*

***Star-Star**:*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] = \mathcal{O}\Big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{PMKR}} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{P^2K^2R}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\Big). \tag{9}$$

***Star-Ring**:*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] = \mathcal{O}\Big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{PMKR}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\Big). \tag{10}$$

***Ring-Star**:*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] = \mathcal{O}\Big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{GPMKR}} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{G^2P^2K^2R}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{G^{2/3}P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\Big). \tag{11}$$

***Ring-Ring**:*

$$\mathbb{E}[\|\nabla F(\bar{x}^{(R)})\|^2] = \mathcal{O}\Big(\frac{LA}{R} + \frac{(L\sigma^2 A)^{1/2}}{\sqrt{GPMKR}} + \frac{(L^2A^2\hat{\zeta}^2)^{1/3}}{G^{2/3}P^{2/3}R^{2/3}} + \frac{(L^2A^2\zeta^2)^{1/3}}{R^{2/3}}\Big). \tag{12}$$

2.4 Key Implications

The Top-Tier Dominance Principle. Contrary to intuitive expectations, the aggregation mechanism at the global tier exerts a significantly stronger influence on convergence than the intra-group topology. This principle is quantitatively demonstrated in Corollary 1. The error terms for ring-based top-tier topologies contain additional scaling factors of G in their denominators for both the SGD variance and intra-group heterogeneity terms. This means that ring-based global aggregation is inherently more robust to both stochastic noise and data heterogeneity. This manifests in two crucial ways: (i) Ring-based top-tier configurations (Ring-Star, Ring-Ring) consistently outperform star-based alternatives under data heterogeneity, with the gap widening as inter-group divergence increases. (ii) The performance difference between top-tier topologies exceeds that between lower-tier configurations. For example, Ring-Star typically outperforms Star-Star by a larger margin than Star-Ring outperforms Star-Star, despite both differing only in the lower tier.

Inter-Group Heterogeneity as the Fundamental Bottleneck. Our analysis quantitatively establishes that inter-group heterogeneity (ζ) is the primary convergence bottleneck across all topologies. While all topologies share the same asymptotic convergence rate of $\mathcal{O}(1/\sqrt{R})$, the practical convergence speed is dominated by inter-group divergence, which decays slowly at $\mathcal{O}\frac{(L^2A^2)^{1/3}}{R^{2/3}}$ regardless of topology choice. Intra-group heterogeneity ($\hat{\zeta}$) decays significantly faster—particularly in ring-based top-tier configurations ($\mathcal{O}\frac{(L^2A^2)^{1/3}}{R^{2/3}G^{2/3}P^{2/3}}$)—making it a secondary concern compared to inter-group divergence. This insight provides a principled foundation for system design: to accelerate convergence in heterogeneous environments, minimizing inter-group divergence should should be prioritized. Practical strategies such as intelligent client clustering, e.g., grouping clients with statistically similar data distributions (Zeng et al., 2022), are therefore more impactful than optimizing local training dynamics within groups.

Topology-Structure Compatibility Principle. The optimal topology selection depends critically on the underlying client grouping structure, creating a fundamental design trade-off. Ring-Star excels with numerous small groups. When clients naturally form many small clusters (e.g., IoT devices, retail outlets), Ring-Star leverages parallelism at the lower tier while benefiting from the smoothing effect of sequential global updates. Its convergence rate improves dramatically with increasing G, making it ideal for deployments with abundant but sparse client clusters. Star-Ring dominates with few large clusters. In settings with limited but data-rich clusters, Star-Ring's intra-group ring aggregation enables deeper local refinement before global synchronization, producing higher-quality group models. This topology shows diminishing returns as G increases beyond a certain point. Star-Star consistently underperforms. Despite its conceptual simplicity, the double averaging in Star-Star significantly dampens the effective learning rate, making it the least efficient configuration across all heterogeneity scenarios.

3 EXPERIMENTS

To validate the theoretical insights derived from our convergence analysis, this section conducts a comprehensive set of experiments on three benchmark datasets: CIFAR-10, CINIC-10, and Fashion-MNIST. We utilize several widely adopted neural network architectures, including ResNet-18, VGG-9, and ResNet-10, to assess performance.

3.1 EXPERIMENTAL SETTINGS

Our evaluation spans three distinct datasets to ensure a comprehensive analysis. We use CIFAR-10 (Krizhevsky et al., 2009), a primary benchmark in federated learning; CINIC-10 (Darlow et al., 2018), which serves as a more challenging natural image dataset; and Fashion-MNIST (Xiao et al., 2017), a common grayscale image classification task. To isolate the effects of aggregation topology and data heterogeneity without interference from normalization dynamics, we remove all batch normalization layers from the network architectures applied to each dataset: ResNet-18 (Lin et al., 2020) and VGG-9 (Acar et al., 2021) for CIFAR-10, ResNet-18 for CINIC-10, and ResNet-10 for Fashion-MNIST. This adjustment ensures a cleaner validation of our convergence bounds, as batch normalization can introduce non-linear and data-dependent behavior that complicates gradient analysis. We fix the mini-batch size to 20 and employ SGD as the local optimizer, with a constant learning rate, zero momentum, and gradient clipping applied to stabilize training. The global model is updated over R communication rounds, with each group performing P group-level updates and each client conducting K local steps per update.

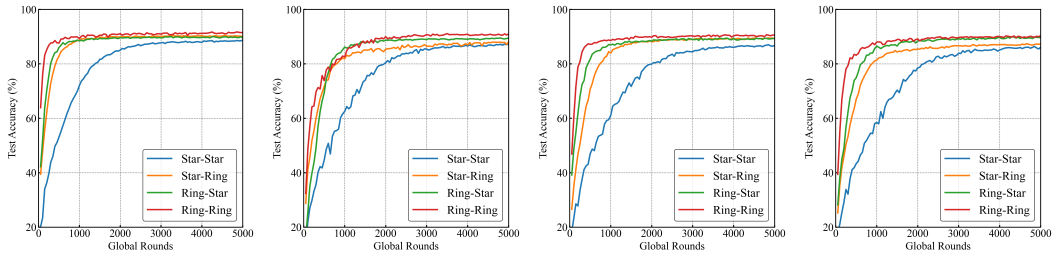
We simulate a hierarchical setup with $N = 100$ clients evenly distributed across $G = 10$ groups, unless otherwise specified, and examine four data partitioning schemes (Fang et al., 2024): (1) IID within groups & IID between groups, where data is uniformly and randomly partitioned at both group and client levels; (2) Non-IID within groups & IID between groups, where groups receive statistically similar data distributions, but clients within each group are assigned non-IID partitions via a Dirichlet distribution with parameter $\alpha = 0.1$; (3) IID within groups & Non-IID between groups, where clients within a group share IID data, but group-level distributions differ significantly, again using a Dirichlet split across groups; and (4) Non-IID within groups & Non-IID between groups, where the entire dataset is partitioned using a Dirichlet distribution, resulting in heterogeneous data at both intra- and inter-group levels.

3.2 EFFECT OF TOPOLOGY

Figure 2 presents the test accuracy curves for the four HFL topologies (Star-Star, Star-Ring, Ring-Star, and Ring-Ring) under the four heterogeneity settings. The results consistently show that topologies with a ring-based top-tier aggregation (i.e., Ring-Star and Ring-Ring) achieve superior convergence speed and higher final accuracy compared to their star-based counterparts. Notably, the classical Star-Star configuration (equivalent to standard HFedAvg) performs the worst across all settings. This is attributed to its conservative update mechanism, i.e., the double averaging at both group and global levels dampens the effective learning rate, slowing convergence. In contrast, ring-based top-tier updates propagate changes sequentially, enabling more aggressive and continuous model refinement. This allows the global model to traverse the loss landscape more rapidly, especially in heterogeneous environments. However, the aggressive nature of ring topologies also introduces sensitivity to biased clients and hyperparameter choices. A single client with a skewed data distribution can steer the entire chain, potentially degrading performance. Therefore, careful tuning of the learning rate and other hyperparameters is essential when deploying ring-based topologies to avoid instability.

3.3 EFFECT OF DATA HETEROGENEITY

Table 2 reports the final test accuracy for all topology and data partition combinations. A key observation is that inter-group heterogeneity has a more detrimental effect on model performance than intra-group heterogeneity. For instance, on CIFAR-10 with ResNet-18 under the Star-Ring topology, shifting to Non-IID group distributions causes a 2.08% accuracy drop (from 90.30% to 88.22%), whereas Non-IID client distributions lead to a smaller drop of only 0.75% (to 89.55%). The effect is even more pronounced on CINIC-10, where in the same setup, inter-group heterogeneity



(a) Groups IID and Clients IID (b) Groups IID and Clients Non-IID (c) Groups Non-IID and Clients IID (d) Groups Non-IID and Clients Non-IID

Figure 2: Comparison of the four HFL topologies on CIFAR-10 Dataset

Table 2: Test accuracy (%) on CIFAR-10, CINIC-10 and Fashion-MNIST under various HFL topologies and data partitioning approaches. The non-IID partitions are generated using a Dirichlet distribution with $\alpha = 0.1$.

Dataset	Model	Heterogeneity		Topology			
		Inter	Intra	Star-Star	Star-Ring	Ring-Star	Ring-Ring
CIFAR-10	ResNet-18	IID	IID	88.48	90.30	90.40	**91.53**
			Non-IID	87.01	89.55	89.75	**91.10**
		Non-IID	IID	87.03	88.22	89.15	**90.94**
			Non-IID	86.78	87.40	90.01	**90.33**
	VGG-9	IID	IID	84.83	87.30	87.77	**89.10**
			Non-IID	84.21	85.33	87.81	**88.17**
		Non-IID	IID	85.00	85.42	86.16	**88.12**
			Non-IID	83.80	85.04	87.05	**87.63**
CINIC-10	ResNet-18	IID	IID	76.88	78.59	78.70	**79.56**
			Non-IID	74.25	76.09	78.23	**78.35**
		Non-IID	IID	74.20	72.53	75.83	**76.23**
			Non-IID	73.63	74.21	**77.11**	76.78
Fashion-MNIST	ResNet-10	IID	IID	89.70	92.59	92.67	**93.01**
			Non-IID	87.45	92.33	92.76	**93.07**
		Non-IID	IID	88.21	89.41	91.40	**91.33**
			Non-IID	88.04	92.18	92.27	**93.33**

results in a substantial 6.06% performance degradation (from 78.59% to 72.53%), compared to a 2.50% drop for intra-group heterogeneity. This trend also holds for Fashion-MNIST with ResNet-10, where inter-group heterogeneity causes a significant 3.18% accuracy drop (from 92.59% to 89.41%), while the impact of intra-group heterogeneity is a negligible 0.26% decrease.

This empirical finding strongly supports our theoretical conclusion that inter-group divergence (ζ) is the dominant bottleneck in HFL convergence. It suggests that system designers should prioritize clustering strategies that minimize distributional differences between groups even at the expense of increased intra-group heterogeneity. For example, grouping clients by semantic similarity of data (e.g., geographic region, user demographics) rather than arbitrary network proximity can significantly improve convergence.

3.4 Effect of Groups

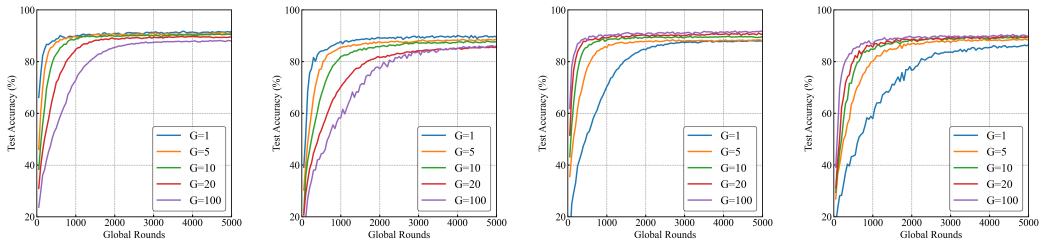
We further investigate how the number of groups G influences performance, focusing on the hybrid topologies (i.e., Star-Ring and Ring-Star) as they offer a practical balance between convergence efficiency and stability. With the total number of clients $N = 100$, we vary the number of groups $G \in \{1, 5, 10, 20, 100\}$ and tune the learning rate for each configuration to ensure optimal perfor-

mance. Figure 3 illustrate the convergence of Star-Ring and Ring-Star under both IID and Non-IID settings. We can find two distinct patterns in Figure 3:

(1) Star-Ring performs best with fewer, larger groups, i.e., small values of G. This is because intra-group ring aggregation benefits from longer update chains: within a large group, sequential updates allow for deeper local refinement before global synchronization, producing higher-quality group models.

(2) Ring-Star, in contrast, excels with more, smaller groups, i.e., large values of G. Here, parallel intra-group aggregation (star) is less effective in large groups due to the averaging of divergent local updates, which can dilute valuable gradients. Smaller groups reduce this averaging effect, and the sequential inter-group updates in Ring-Star enable fine-grained global alignment.

It is worth noting that our conclusion continues to hold even in the extreme cases—Ring-Star with $G = N$ and Star-Ring with $G = 1$, both degenerating to a pure ring topology. This does not conflict with the notion of "catastrophic forgetting" in sequential federated learning, because the step size is not fixed. We scale it with the number of groups to keep the effective learning rate constant. Under the same effective learning rate, selecting an appropriate G therefore yields optimal performance. These results highlight a critical design principle: *optimal topology selection depends on the underlying group structure*. In applications with a few large, data-rich clusters (e.g., hospital networks), Star-Ring is preferable. In contrast, systems with many small or independent units (e.g., IoT devices, retail outlets) benefit more from the Ring-Star topology.



(a) Star-Ring (IID) (b) Star-Ring (Non-IID) (c) Ring-Star (IID) (d) Ring-Star (Non-IID)

Figure 3: Comparsion of Star-Ring and Ring-Star topologies with different numbers of groups on CIFAR10 Dataset

4 Conclusion

This paper presents the first unified convergence analysis for all four HFL topologies under non-convex objectives and intra/inter-group data heterogeneity. Our results reveal that: (1) top-tier topology dictates convergence behavior, and ring-based top-tier aggregation generally converges faster than star-based methods; (2) inter-group heterogeneity is the dominant bottleneck, outweighing intra-group effects; and (3) optimal topology depends on group structure, where Ring-Star suits many small groups, while Star-Ring excels with few large clusters. These findings enable system designers to move beyond heuristic topology choices and instead make informed, theoretically grounded decisions based on deployment-specific constraints such as network scale, client distribution, and data heterogeneity profiles.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Timothy Castiglia, Anirban Das, and Stacy Patterson. Multi-level local sgd: Distributed sgd for heterogeneous hierarchical networks. In *International Conference on Learning Representations*, 2021.

HE Chaoyang, M ANNAVARAM, and S AVESTIMEHR. Group knowledge transfer: Collaborative training of large cnns on the edge [j]. *arXiv preprint arXiv:2007.14513*, 2020.

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

Yongheng Deng, Feng Lyu, Ju Ren, Yongmin Zhang, Yuezhi Zhou, Yaoxue Zhang, and Yuanyuan Yang. Share: Shaping data distribution at edge for communication-efficient hierarchical federated learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, pp. 24–34. IEEE, 2021.

Yucheng Ding, Chaoyue Niu, Yikai Yan, Zhenzhe Zheng, Fan Wu, Guihai Chen, Shaojie Tang, and Rongfei Jia. Distributed optimization over block-cyclic data. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pp. 1–6, 2024.

Qingze Fang, Zhiwei Zhai, Shuai Yu, Qiong Wu, Xiaowen Gong, and Xu Chen. Olive branch learning: A topology-aware federated learning framework for space-air-ground integrated network. *IEEE Transactions on Wireless Communications*, 22(7):4534–4551, 2022.

Wenzhi Fang, Dong-Jun Han, Evan Chen, Shiqiang Wang, and Christopher Brinton. Hierarchical federated learning with multi-timescale gradient correction. *Advances in Neural Information Processing Systems*, 37:78863–78904, 2024.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Jianjun Huang, Lixin Ye, and Li Kang. Fedsr: A semi-decentralized federated learning algorithm for non-iidness in iot system. *arXiv preprint arXiv:2403.14718*, 2024.

Xiaohan Jiang and Hongbin Zhu. On the convergence of hierarchical federated learning with partial worker participation. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

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

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

Yipeng Li and Xinchen Lyu. Convergence analysis of sequential federated learning on heterogeneous data. *Advances in Neural Information Processing Systems*, 36:56700–56755, 2023.

Yipeng Li and Xinchen Lyu. Sharp bounds for sequential federated learning on heterogeneous data. *Journal of Machine Learning Research*, 26(70):1–55, 2025.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.

Lumin Liu, Jun Zhang, SH Song, and Khaled B Letaief. Client-edge-cloud hierarchical federated learning. In *ICC 2020-2020 IEEE international conference on communications (ICC)*, pp. 1–6. IEEE, 2020.

Lumin Liu, Jun Zhang, Shenghui Song, and Khaled B Letaief. Hierarchical federated learning with quantization: Convergence analysis and system design. *IEEE Transactions on Wireless Communications*, 22(1):2–18, 2022.

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.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. Demystifying why local aggregation helps: Convergence analysis of hierarchical sgd. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8548–8556, 2022.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. *Advances in neural information processing systems*, 35:19124–19137, 2022.

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

Zhengjie Yang, Sen Fu, Wei Bao, Dong Yuan, and Albert Y Zomaya. Hierarchical federated learning with momentum acceleration in multi-tier networks. *IEEE Transactions on Parallel and Distributed Systems*, 34(10):2629–2641, 2023.

Shenglai Zeng, Zonghang Li, Hongfang Yu, Yihong He, Zenglin Xu, Dusit Niyato, and Han Yu. Heterogeneous federated learning via grouped sequential-to-parallel training. In *International Conference on Database Systems for Advanced Applications*, pp. 455–471. Springer, 2022.

Fan Zhou and Guojing Cong. A distributed hierarchical sgd algorithm with sparse global reduction. *arXiv preprint arXiv:1903.05133*, 2019.

A Algorithm Details

For clarity and completeness, this appendix provides the detailed pseudocode for the four HFL topologies mentioned in the main body of our paper. Each algorithm outlines a different communication pattern for both inter-group and intra-group model aggregation.

The Star-Star topology (Algorithm 1) represents a fully parallel framework. Both the groups at the server level and the clients within each group perform their training and updates in parallel, synchronizing with their respective servers before aggregation.

Algorithm 1 Star-Star Hierarchical FL

for global rounds $r = 0, 1, \ldots, R-1$ **do**
  for groups $g = 1, 2, \ldots, G$ **in parallel do**
    Initialize group model: $\mathbf{x}_{g,0}^{(r)} = \mathbf{x}^{(r)}$
    for group rounds $p = 0, 1, \ldots, P-1$ **do**
      for clients $m = 1, 2, \ldots, M$ **in parallel do**
        Initialize local model: $\mathbf{x}_{g,p,m,0}^{(r)} = \mathbf{x}_{g,p}^{(r)}$
        for local steps $k = 0, 1, \ldots, K-1$ **do**
          $\mathbf{x}_{g,p,m,k+1}^{(r)} = \mathbf{x}_{g,p,m,k}^{(r)} - \eta \mathbf{g}_{g,p,m,k}^{(r)}$
        end for
      end for
      Group aggregation: $\mathbf{x}_{g,p+1}^{(r)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_{g,p,m,K}^{(r)}$
    end for
  end for
  Global aggregation: $\mathbf{x}^{(r+1)} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{x}_{g,P}^{(r)}$
end for

The Star-Ring topology (Algorithm 2) combines parallel inter-group communication with sequential intra-group updates. While groups update in parallel with the global server, clients within each group form a ring, passing the model sequentially from one client to the next.

Conversely, the Ring-Star topology (Algorithm 3) employs sequential communication among groups and parallel updates within them. The groups form a ring at the global level, while clients inside each group operate in a standard star configuration.

The Ring-Ring topology (Algorithm 4) implements a fully sequential communication protocol. Both the groups at the global level and the clients within each group update their models in a sequential, ring-based manner.

Algorithm 2 Star-Ring Hierarchical FL

for global rounds $r = 0, 1, \ldots, R-1$ **do**
  for groups $g = 1, 2, \ldots, G$ **in parallel do**
    Initialize group model: $\mathbf{x}_{g,0}^{(r)} = \mathbf{x}^{(r)}$
    for group rounds $p = 0, 1, \ldots, P-1$ **do**
      for clients $m = 1, 2, \ldots, M$ **in sequence do**
        Initialize local model: $\mathbf{x}_{g,p,m,0}^{(r)} = \begin{cases} \mathbf{x}_{g,p}^{(r)} & \text{if } m = 1 \\ \mathbf{x}_{g,p,m-1,K}^{(r)} & \text{if } m > 1 \end{cases}$
        for local steps $k = 0, 1, \ldots, K-1$ **do**
          $\mathbf{x}_{g,p,m,k+1}^{(r)} = \mathbf{x}_{g,p,m,k}^{(r)} - \eta \mathbf{g}_{g,p,m,k}^{(r)}$
        end for
      end for
      Group model: $\mathbf{x}_{g,p+1}^{(r)} = \mathbf{x}_{g,p,M,K}^{(r)}$
    end for
  end for
  Global aggregation: $\mathbf{x}^{(r+1)} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{x}_{g,P}^{(r)}$
end for

Algorithm 3 Ring-Star Hierarchical FL

for global rounds $r = 0, 1, \ldots, R-1$ **do**
  for groups $g = 1, 2, \ldots, G$ **in sequence do**
    Initialize group model: $\mathbf{x}_{g,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)} & \text{if } g = 1 \\ \mathbf{x}_{g-1,P}^{(r)} & \text{if } g > 1 \end{cases}$
    for group rounds $p = 0, 1, \ldots, P-1$ **do**
      for clients $m = 1, 2, \ldots, M$ **in parallel do**
        Initialize local model: $\mathbf{x}_{g,p,m,0}^{(r)} = \mathbf{x}_{g,p}^{(r)}$
        for local steps $k = 0, 1, \ldots, K-1$ **do**
          $\mathbf{x}_{g,p,m,k+1}^{(r)} = \mathbf{x}_{g,p,m,k}^{(r)} - \eta \mathbf{g}_{g,p,m,k}^{(r)}$
        end for
      end for
      Group aggregation: $\mathbf{x}_{g,p+1}^{(r)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_{g,p,m,K}^{(r)}$
    end for
  end for
  Global model: $\mathbf{x}^{(r+1)} = \mathbf{x}_{G,P}^{(r)}$
end for

Algorithm 4 Ring-Ring Hierarchical FL

for global rounds $r = 0, 1, \ldots, R-1$ **do**
  for groups $g = 1, 2, \ldots, G$ **in sequence do**
    Initialize group model: $\mathbf{x}_{g,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)} & \text{if } g = 1 \\ \mathbf{x}_{g-1,P}^{(r)} & \text{if } g > 1 \end{cases}$
    for group rounds $p = 0, 1, \ldots, P-1$ **do**
      for clients $m = 1, 2, \ldots, M$ **in sequence do**
        Initialize local model: $\mathbf{x}_{g,p,m,0}^{(r)} = \begin{cases} \mathbf{x}_{g,p}^{(r)} & \text{if } m = 1 \\ \mathbf{x}_{g,p,m-1,K}^{(r)} & \text{if } m > 1 \end{cases}$
        for local steps $k = 0, 1, \ldots, K-1$ **do**
          $\mathbf{x}_{g,p,m,k+1}^{(r)} = \mathbf{x}_{g,p,m,k}^{(r)} - \eta \mathbf{g}_{g,p,m,k}^{(r)}$
        end for
      end for
      Group model: $\mathbf{x}_{g,p+1}^{(r)} = \mathbf{x}_{g,p,M,K}^{(r)}$
    end for
  end for
  Global model: $\mathbf{x}^{(r+1)} = \mathbf{x}_{G,P}^{(r)}$
end for

B NOTATIONS

Table 3 summarizes the notations appearing in this paper.

Table 3: Key notations for HFL algorithm.

Symbol	Description
R, r	number, index of training rounds
G, g	number, index of groups
M, m	number, index of clients in each group
K, k	number, index of local update steps
η	learning rate (or stepsize)
$\tilde{\eta}$	effective learning rate
L	L-smoothness constant (Assumption 1)
σ	upper bound on variance of stochastic gradients at each client (Assumption 2)
ζ	constants in Assumption 3 to bound inter-group heterogeneity
ζ_g	constants in Assumption 4 to bound intra-group heterogeneity
$F/F_g/F_g, m$	global objective/group p objective/local objective of client m in group p
$\mathbf{x}^{(r)}$	global model parameters in the r-th round
$\mathbf{x}_{g,p,m,k}^{(r)}$	local model parameters of the m-th client after k local steps in the g-th group after p group steps in the r-th round
$\mathbf{g}_{g,p,m,k}^{(r)}$	$\mathbf{g}_{g,p,m,k}^{(r)} := \nabla f_{g,m}(\mathbf{x}_{g,p,m,k}^{(r)}; \xi)$ denotes the stochastic gradients of $F_{g,m}$ regarding $\mathbf{x}_{g,p,m,k}^{(r)}$

C PROOF OF THEMERM1

Lemma 1. *Let Assumptions 2, 1 hold. If the learning rate satisfies* $\eta \leq \frac{1}{2L\mathcal{K}GPMK}$*, then*

$$\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right)-F\left(\mathbf{x}\right)\right] \leq -\frac{1}{2}\mathcal{K}GPMK\eta\|\nabla F\left(\mathbf{x}\right)\|^2 + L\mathcal{K}^2GPMK\eta^2\sigma^2 + \frac{1}{2}L^2\mathcal{K}\eta\sum_{g,p,m,k}\mathbb{E}\|\mathbf{x}_{g,p,m,k}-\mathbf{x}\|^2$$

where $\sum_{g,p,m,k}$ *is a shorthand for the quadruple summation* $\sum_{g=1}^{G}\sum_{p=0}^{P-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}$.

Proof. In the following, we focus on a single training round, and hence we drop the superscripts r for a while, e.g., writing $\mathbf{x}_{g,p,m,k}$ to replace $\mathbf{x}_{g,p,m,k}^{(r)}$. Specially, we would like to use $\mathbf{x}$ to replace $\mathbf{x}_{1,0,1,0}^{(r)}$. Unless otherwise stated, the expectation is conditioned on $\mathbf{x}^{(r)}$.

Starting from the smoothness of F (applying Assumption 1, $\|\nabla F(\mathbf{x})-\nabla F(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$), we have

$$\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right)-F\left(\mathbf{x}\right)\right] \leq \underbrace{\mathbb{E}\left\langle\nabla F\left(\mathbf{x}\right),\Delta\mathbf{x}\right\rangle}_{A_1} + \underbrace{\frac{L}{2}\mathbb{E}\|\Delta\mathbf{x}\|^2}_{A_2}$$

For the four topologies of HFL, the model udpates of one global round is shown as Table 4.

Table 4: The model udpates of one global round for the four topologies of HFL.

	$\Delta\mathbf{x}$
Star-Star	$-\eta\frac{1}{G}\sum_{g=1}^{G}\sum_{p=0}^{P-1}\frac{1}{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$
Star-Ring	$-\eta\frac{1}{G}\sum_{g=1}^{G}\sum_{p=0}^{P-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$
Ring-Star	$-\eta\sum_{g=1}^{G}\sum_{p=0}^{P-1}\frac{1}{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$
Ring-Ring	$-\eta\sum_{g=1}^{G}\sum_{p=0}^{P-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$

Consequently, the model updates within a single global round for the four topologies can be represented by a unified format,

$$\Delta\mathbf{x} = \mathbf{x}^{(R+1)}-\mathbf{x}^{(R)} = -\mathcal{K}\eta\sum_{g,p,m,k}\mathbf{g}_{g,p,m,k},$$

$$\text{where}\quad \mathcal{K} = \begin{cases} 1/(GM) & \text{for Star-Star topology,}\\ 1/G & \text{for Star-Ring topology,}\\ 1/M & \text{for Ring-Star topology,}\\ 1 & \text{for Ring-Ring topology.}\end{cases}$$

After substituting the overall updates $\Delta\mathbf{x}$, we can get A_1

$$\mathbb{E}\left\langle\nabla F\left(\mathbf{x}\right),\Delta\mathbf{x}\right\rangle = -\mathcal{K}GPMK\eta\mathbb{E}\left\langle\nabla F\left(\mathbf{x}\right),\frac{1}{GPMK}\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\right\rangle$$

$$
\begin{aligned}
&= -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\left[\|\nabla F\left(\mathbf{x}\right)\|^2 + \|\frac{1}{GPMK}\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\right.\\
&\quad \left.-\|\frac{1}{GPMK}\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right) - \nabla F\left(\mathbf{x}\right)\|^2\right]\\
&= -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 - \frac{1}{2}\mathcal{K}\frac{1}{GPMK}\eta\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\quad + \frac{1}{2}\mathcal{K}\frac{1}{GPMK}\eta\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right) - \sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}\right)\|^2\\
&\leq -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 - \frac{1}{2}\mathcal{K}\frac{1}{GPMK}\eta\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\quad + \frac{1}{2}\mathcal{K}\eta\sum_{g,p,m,k}\mathbb{E}\|\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right) - \nabla F_{g,m}\left(\mathbf{x}\right)\|^2\\
&\leq -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 - \frac{1}{2}\mathcal{K}\frac{1}{GPMK}\eta\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\quad + \frac{1}{2}L^2\mathcal{K}\eta\sum_{g,p,m,k}\mathbb{E}\|\mathbf{x}_{g,p,m,k} - \mathbf{x}\|^2
\end{aligned}
$$

Bounding the term A_2,

$$
\begin{aligned}
\frac{L}{2}\mathbb{E}\left\|-\mathcal{K}\eta\sum_{g,p,m,k}\mathbf{g}_{g,p,m,k}\right\|^2 &\leq L\mathcal{K}^2\eta^2\mathbb{E}\|\sum_{g,p,m,k}\left(\mathbf{g}_{g,p,m,k} - \nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\right)\|^2\\
&\quad + L\eta^2\mathbb{E}\|\mathcal{K}\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\leq L\mathcal{K}^2\eta^2\sum_{g,p,m,k}\mathbb{E}\|\mathbf{g}_{g,p,m,k} - \nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\quad + L\mathcal{K}\eta^2\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2\\
&\leq L\mathcal{K}^2GPMK\eta^2\sigma^2 + L\mathcal{K}^2\eta^2\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2
\end{aligned}
$$

Substitute A_1 and A_2 to $\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right) - F\left(\mathbf{x}\right)\right]$, we have

$$
\begin{aligned}
\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right) - F\left(\mathbf{x}\right)\right] &\leq -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 + L\mathcal{K}^2GPMK\eta^2\sigma^2\\
&\quad + \frac{1}{2}L^2\mathcal{K}\eta\sum_{g,p,m,k}\mathbb{E}\|\mathbf{x}_{g,p,m,k} - \mathbf{x}\|^2\\
&\quad - \frac{\frac{1}{2} - L\mathcal{K}GPMK\eta)\mathcal{K}\eta}{GPMK}\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2
\end{aligned}
$$

We have $\frac{(\frac{1}{2}-L\mathcal{K}GPMK\eta)\mathcal{K}\eta}{GPMK}\mathbb{E}\|\sum_{g,p,m,k}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k}\right)\|^2 \geq 0$, when $\frac{1}{2} - L\mathcal{K}GPMK\eta \geq 0$, thus

$$
\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right) - F\left(\mathbf{x}\right)\right] \leq -\frac{1}{2}\mathcal{K}GPMK\eta\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 + L\mathcal{K}^2GPMK\eta^2\sigma^2
$$

$$+\frac{1}{2}L^2\mathcal{K}\eta\underbrace{\sum_{g,p,m,k}\mathbb{E}\|\mathbf{x}_{g,p,m,k}-\mathbf{x}\|^2}_{\text{client drift}}$$

C.1 Bounding the client drift with Assumptions 3 and 4

We define the client drift in HFL:

$$E_r:=\sum_{g=1}^{G}\sum_{p=0}^{P-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{x}_{g,p,m,k}^{(r)}-\mathbf{x}^{(r)}\|^2.$$

Lemma 2. *Let Assumptions 2, 1 hold. If the learning rate satisfies $\eta\leq\frac{1}{12LK\mathcal{G}PMK}$, then the client drift is bounded*
***Star-Star**:*

$$\begin{aligned}E_r\leq{}&11\eta^2GPMK^2\sigma^2+11\eta^2GP^2K^2\sigma^2+11\eta^2GPMK^3\hat{\zeta}^2+11\eta^2GPMK^3\zeta^2\\&+11\eta^2GP^3K^3\zeta^2+11\eta^2GP^3K^3\mathbb{E}\|\nabla F(\mathbf{x})\|^2\end{aligned}$$

***Star-Ring**:*

$$\begin{aligned}E_r\leq{}&16\eta^2GPMK^2\sigma^2+16\eta^2GPM^2K^2\sigma^2+16\eta^2GP^2M^2K^2\sigma^2+16\eta^2GPMK^3\hat{\zeta}^2\\&+16\eta^2GPM^3K^3\hat{\zeta}^2+16\eta^2GPMK^3\zeta^2+16\eta^2GPM^3K^3\zeta^2+16\eta^2GP^3M^3K^3\zeta^2\\&+16\eta^2GP^3M^3K^3\mathbb{E}\|\nabla F(\mathbf{x})\|^2\end{aligned}$$

***Ring-Star**:*

$$\begin{aligned}E_r\leq{}&16\eta^2GPMK^2\sigma^2+16\eta^2GP^2K^2\sigma^2+16\eta^2G^2P^2K^2\sigma^2+16\eta^2GPMK^3\hat{\zeta}^2\\&+16\eta^2GPMK^3\zeta^2+16\eta^2GP^3MK^3\zeta^2+16\eta^2G^3P^3MK^3\zeta^2+16\eta^2G^3P^3MK^3\mathbb{E}\|\nabla F(\mathbf{x})\|^2\end{aligned}$$

***Ring-Ring**:*

$$\begin{aligned}E_r\leq{}&21\eta^2GPMK^2\sigma^2+21\eta^2GPM^2K^2\sigma^2+21\eta^2GP^2M^2K^2\sigma^2+21\eta^2G^2P^2M^2K^2\sigma^2\\&+21\eta^2GPMK^3\hat{\zeta}^2+21\eta^2GPM^3K^3\hat{\zeta}^2+21\eta^2GPMK^3\zeta^2+21\eta^2GPM^3K^3\zeta^2\\&+21\eta^2GP^3M^3K^3\zeta^2+21\eta^2G^3P^3M^3K^3\zeta^2+21\eta^2G^3P^3M^3K^3\mathbb{E}\|\nabla F(\mathbf{x})\|^2\end{aligned}$$

Proof. The overall updates of HFL from $\mathbf{x}^{(r)}$ to $\mathbf{x}_{g,p,m,k}^{(r)}$ are shown in the following table.

Table 5: The factor "$-\eta$" is omitted for all cells.

$\mathbf{x}_{g,p,m,k}-\mathbf{x}$	$\mathbf{x}_{g,p,m,k}-\mathbf{x}_{g,p,m,0}$	$\mathbf{x}_{g,p,m,0}-\mathbf{x}_{g,p,1,0}$	$\mathbf{x}_{g,p,1,0}-\mathbf{x}_{g,0,1,0}$	$\mathbf{x}_{g,0,1,0}-\mathbf{x}_{1,0,1,0}$
Star-Star	$\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}$		$\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p',m,k}$	
Star-Ring	$\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}$	$\sum_{m'=1}^{m-1}\sum_{k'=0}^{K-1}\mathbf{g}_{g,p,m',k}$	$\sum_{p'=0}^{p-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p',m,k}$	
Ring-Star	$\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}$		$\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p',m,k}$	$\sum_{g'=1}^{g-1}\sum_{p'=0}^{P-1}\frac{1}{M}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$
Ring-Ring	$\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}$	$\sum_{m'=1}^{m-1}\sum_{k'=0}^{K-1}\mathbf{g}_{g,p,m',k}$	$\sum_{p'=0}^{p-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p',m,k}$	$\sum_{g'=1}^{g-1}\sum_{p'=0}^{P-1}\sum_{m=1}^{M}\sum_{k=0}^{K-1}\mathbf{g}_{g,p,m,k}$

To bound E_r, we first bound $\mathbb{E}\|\mathbf{x}_{g,p,m,k}-\mathbf{x}\|$. The derivation processes for the four topologies are similar. Here, we take the star-star topology as an illustrative example.

$$
\begin{aligned}
\mathbb{E}\|\mathbf{x}_{g,p,m,k}-\mathbf{x}\|^2 &= \eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbf{g}_{g,p',m',k'}\|^2\\
&\leq 5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbf{g}_{g,p',m',k'}\\
&\quad-\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right)-\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right)\|^2\\
&\quad+5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right)+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right)\\
&\quad-\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}\right)-\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}\right)\|^2\\
&\quad+5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}\right)+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}\right)\\
&\quad-\sum_{k'=0}^{k-1}\nabla F_{g}\left(\mathbf{x}\right)-\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g}\left(\mathbf{x}\right)\|^2\\
&\quad+5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\nabla F_{g}\left(\mathbf{x}\right)+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g}\left(\mathbf{x}\right)\\
&\quad-\sum_{k'=0}^{k-1}\nabla F\left(\mathbf{x}\right)-\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F\left(\mathbf{x}\right)\|^2\\
&\quad+5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\nabla F\left(\mathbf{x}\right)+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

Bounding the first term in the left-hand inequality,

$$
\begin{aligned}
&5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\mathbf{g}_{g,p,m,k'}+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbf{g}_{g,p',m',k'}-\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right)\\
&-\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right)\|^2\\
&\leq 10\eta^2\sum_{k'=0}^{k-1}\mathbb{E}\|\mathbf{g}_{g,p,m,k'}-\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right)\|^2\\
&\quad+10\eta^2\sum_{p'=0}^{p-1}\frac{1}{M^2}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbb{E}\|\mathbf{g}_{g,p',m',k'}-\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right)\|^2\\
&\leq 10\eta^2k\sigma^2+10\eta^2\frac{pK}{M}\sigma^2
\end{aligned}
$$

Bounding the second term in the left-hand inequality,

$$
5\eta^2\mathbb{E}\|\sum_{k'=0}^{k-1}\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right)+\sum_{p'=0}^{p-1}\frac{1}{M}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right)
$$

$$
\begin{aligned}
&- \sum_{k'=0}^{k-1} \nabla F_{g,m}\left(\mathbf{x}\right) - \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F_{g,m'}\left(\mathbf{x}\right) \|^2 \\
&\leq 10\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\nabla F_{g,m}\left(\mathbf{x}_{g,p,m,k'}\right) - \nabla F_{g,m}\left(\mathbf{x}\right)\|^2 \\
&\quad + 10\eta^2 pMK \sum_{p'=0}^{p-1} \frac{1}{M^2} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\nabla F_{g,m'}\left(\mathbf{x}_{g,p',m',k'}\right) - \nabla F_{g,m'}\left(\mathbf{x}\right)\|^2 \\
&\leq 10L^2\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\mathbf{x}_{g,p,m,k'} - \mathbf{x}\|^2 + 10L^2\eta^2 \frac{pK}{M} \sum_{p'=0}^{p-1} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\mathbf{x}_{g,p',m',k'} - \mathbf{x}\|^2
\end{aligned}
$$

Bounding the third term in the left-hand inequality,

$$
\begin{aligned}
&5\eta^2\mathbb{E}\| \sum_{k'=0}^{k-1} \nabla F_{g,m}\left(\mathbf{x}\right) + \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F_{g,m'}\left(\mathbf{x}\right) - \sum_{k'=0}^{k-1} \nabla F_g\left(\mathbf{x}\right) \\
&- \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F_g\left(\mathbf{x}\right) \|^2 \\
&\leq 10\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\nabla F_{g,m}\left(\mathbf{x}\right) - \nabla F_g\left(\mathbf{x}\right)\|^2 \\
&\quad + 10\eta^2 \frac{pK}{M} \sum_{p'=0}^{p-1} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\nabla F_{g,m'}\left(\mathbf{x}\right) - \nabla F_g\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

Bounding the fourth term in the left-hand inequality,

$$
\begin{aligned}
&5\eta^2\mathbb{E}\| \sum_{k'=0}^{k-1} \nabla F_g\left(\mathbf{x}\right) + \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F_g\left(\mathbf{x}\right) - \sum_{k'=0}^{k-1} \nabla F\left(\mathbf{x}\right) - \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F\left(\mathbf{x}\right) \|^2 \\
&\leq 10\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\nabla F_g\left(\mathbf{x}\right) - \nabla F\left(\mathbf{x}\right)\|^2 + 10\eta^2 \frac{pK}{M} \sum_{p'=0}^{p-1} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\nabla F_g\left(\mathbf{x}\right) - \nabla F\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

Bounding the fifth term in the left-hand inequality,

$$
\begin{aligned}
&5\eta^2\mathbb{E}\| \sum_{k'=0}^{k-1} \nabla F\left(\mathbf{x}\right) + \sum_{p'=0}^{p-1} \frac{1}{M} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \nabla F\left(\mathbf{x}\right)\|^2 \\
&\leq 10\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 + pMK \sum_{p'=0}^{p-1} \frac{1}{M^2} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

Substitute these terms into E_r,

$$
\begin{aligned}
E_r \leq &\sum_{g,p,m,k} \left(10\eta^2 k\sigma^2 + 10\eta^2 \frac{pK}{M}\sigma^2\right) \\
&+ \sum_{g,p,m,k} \left(10L^2\eta^2 k \sum_{k'=0}^{k-1} \mathbb{E}\|\mathbf{x}_{g,p,m,k'} - \mathbf{x}\|^2 + 10L^2\eta^2 \frac{pK}{M} \sum_{p'=0}^{p-1} \sum_{m'=1}^{M} \sum_{k'=0}^{K-1} \mathbb{E}\|\mathbf{x}_{g,p',m',k'} - \mathbf{x}\|^2\right)
\end{aligned}
$$

Table 6: The client drift for the four topologies of HFL.

	Φ_1	Φ_2
Star-Star	$11\eta^2GP^3K^3$	$11\eta^2GPMK^2\sigma^2+11\eta^2GP^2K^2\sigma^2+11\eta^2GPMK^3\hat{\zeta}^2+11\eta^2GPMK^3\zeta^2$ $+11\eta^2GP^3K^3\zeta^2$
Star-Ring	$16\eta^2GP^3M^3K^3$	$16\eta^2GPMK^2\sigma^2+16\eta^2GPM^2K^2\sigma^2+16\eta^2GP^2M^2K^2\sigma^2+16\eta^2GPMK^3\hat{\zeta}^2$ $+16\eta^2GPM^3K^3\hat{\zeta}^2+16\eta^2GPMK^3\zeta^2+16\eta^2GPM^3K^3\zeta^2+16\eta^2GP^3M^3K^3\zeta^2$
Ring-Star	$16\eta^2G^3P^3MK^3$	$16\eta^2GPMK^2\sigma^2+16\eta^2GP^2K^2\sigma^2+16\eta^2G^2P^2K^2\sigma^2+15\eta^2GPMK^3\hat{\zeta}^2$ $+16\eta^2GPMK^3\zeta^2+16\eta^2GP^3MK^3\zeta^2+16\eta^2G^3P^3MK^3\zeta^2$
Ring-Ring	$21\eta^2G^3P^3M^3K^3$	$21\eta^2GPMK^2\sigma^2+21\eta^2GPM^2K^2\sigma^2+21\eta^2GP^2M^2K^2\sigma^2+21\eta^2G^2P^2M^2K^2\sigma^2$ $+21\eta^2GPMK^3\hat{\zeta}^2+21\eta^2GPM^3K^3\hat{\zeta}^2+21\eta^2GPMK^3\zeta^2+21\eta^2GPM^3K^3\zeta^2$ $+21\eta^2GP^3M^3K^3\zeta^2+21\eta^2G^3P^3M^3K^3\zeta^2$

$$
\begin{aligned}
&+\sum_{g,p,m,k}\left(10\eta^2k\sum_{k'=0}^{k-1}\mathbb{E}\|\nabla F_{g,m}\left(\mathbf{x}\right)-\nabla F_g\left(\mathbf{x}\right)\|^2\right.\\
&\left.+10\eta^2\frac{pK}{M}\sum_{p'=0}^{p-1}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbb{E}\|\nabla F_{g,m'}\left(\mathbf{x}\right)-\nabla F_g\left(\mathbf{x}\right)\|^2\right)\\
&+\sum_{g,p,m,k}\left(10\eta^2k\sum_{k'=0}^{k-1}\mathbb{E}\|\nabla F_g\left(\mathbf{x}\right)-\nabla F\left(\mathbf{x}\right)\|^2\right.\\
&\left.+10\eta^2\frac{pK}{M}\sum_{p'=0}^{p-1}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1}\mathbb{E}\|\nabla F_g\left(\mathbf{x}\right)-\nabla F\left(\mathbf{x}\right)\|^2\right)\\
&+\sum_{g,p,m,k}\left(10\eta^2(k\sum_{k'=0}^{k-1}\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2+\frac{pK}{M}\sum_{p'=0}^{p-1}\sum_{m'=1}^{M}\sum_{k'=0}^{K-1})\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2\right)\\
&\leq 10\eta^2GPMK^2\sigma^2+10\eta^2GP^2K^2\sigma^2\\
&\quad+10\eta^2GPMK^3\hat{\zeta}^2+10\eta^2GPMK^3\zeta^2\\
&\quad+10\eta^2GP^3K^3\zeta^2+10\eta^2GPMK^3\|\nabla F\left(\mathbf{x}\right)\|^2\\
&\quad+10\eta^2GP^3K^3\|\nabla F\left(\mathbf{x}\right)\|^2+10L^2P^2K^2\eta^2E_r
\end{aligned}
$$

$$
\begin{aligned}
(1-10L^2P^2K^2\eta^2)E_r\leq\ &10\eta^2GPMK^2\sigma^2+10\eta^2GP^2K^2\sigma^2+10\eta^2GPMK^3\hat{\zeta}^2\\
&+10\eta^2GPMK^3\zeta^2+10\eta^2GP^3K^3\zeta^2\\
&+10\eta^2GPMK^3\|\nabla F\left(\mathbf{x}\right)\|^2+10\eta^2GP^3K^3\|\nabla F\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

With $\eta\leq\frac{1}{12LPK}$, $1-10L^2P^2K^2\eta^2\geq\frac{10}{11}$, we have

$$
\begin{aligned}
E_r\leq\ &11\eta^2GPMK^2\sigma^2+11\eta^2GP^2K^2\sigma^2+11\eta^2GPMK^3\hat{\zeta}^2\\
&+11\eta^2GPMK^3\zeta^2+11\eta^2GP^3K^3\zeta^2\\
&+11\eta^2GPMK^3\|\nabla F\left(\mathbf{x}\right)\|^2+11\eta^2GP^3K^3\|\nabla F\left(\mathbf{x}\right)\|^2
\end{aligned}
$$

E_r can be unified into a common format for the four topologies

$$
E_r\leq\Phi_1\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2+\Phi_2.
$$

C.2 PROOF OF THEOREM1

Proof. Substitute E_r into $\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right)-F\left(\mathbf{x}\right)\right]$, we can simplify the recursion as follows:

$$\begin{aligned}\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right)-F\left(\mathbf{x}\right)\right] \leq -(\frac{1}{2}\mathcal{K}GPMK\eta - \frac{1}{2}L^2\mathcal{K}\eta\Phi_1)\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 \\ + L\mathcal{K}^2GPMK\eta^2\sigma^2 + \frac{1}{2}L^2\mathcal{K}\eta\Phi_2\end{aligned}$$

Letting $\tilde{\eta} := \mathcal{K}GPMK\eta$, subtracting F^* from both sides and then rearranging the terms, we have

$$\begin{aligned}\mathbb{E}\left[F\left(\mathbf{x}+\Delta\mathbf{x}\right)-F^*\right] \leq \mathbb{E}\left[F\left(\mathbf{x}\right)-F^*\right] - \tilde{\eta}(\frac{1}{2} - \frac{L^2\Phi_1}{2GPMK})\mathbb{E}\|\nabla F\left(\mathbf{x}\right)\|^2 \\ + \frac{L\tilde{\eta}^2}{GPMK}\sigma^2 + \frac{L^2\tilde{\eta}\Phi_2}{2GPMK}\end{aligned}$$

Then applying Lemma 2, we have

$$\begin{aligned}\min_{0\leq r\leq R}\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2 \leq \frac{F\left(\mathbf{x}^0\right)-F^*}{(\frac{1}{2}-\frac{L^2\Phi_1}{2GPMK})\tilde{\eta}R} + \frac{L\tilde{\eta}}{GPMK(\frac{1}{2}-\frac{L^2\Phi_1}{2GPMK})}\sigma^2 \\ + \frac{L^2\Phi_2}{2GPMK(\frac{1}{2}-\frac{L^2\Phi_1}{2GPMK})}\end{aligned}$$

where we use

$$\min_{0\leq r\leq R}\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2 \leq \frac{1}{R+1}\sum_{r=0}^{R}\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2.$$

With $\tilde{\eta} \leq \frac{1}{12L}$, $\frac{1}{2} - \frac{L^2\Phi_1}{2GPMK} \geq \frac{5}{12}$, $\min_{0\leq r\leq R}\mathbb{E}\|\nabla F(\mathbf{x}^{(r)})\|^2$satisfies the following upper bounds:

Star-Star:

$$\begin{aligned}\min_{0\leq r\leq R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{12L\tilde{\eta}\sigma^2}{5GPMK} + \frac{12L^2\tilde{\eta}^2\sigma^2}{P^2K^2} + \frac{12L^2\tilde{\eta}^2\sigma^2}{PMK} + \frac{12L^2\tilde{\eta}^2\hat{\zeta}^2}{P^2} \\ + \frac{12L^2\tilde{\eta}^2\zeta^2}{P^2} + 12L^2\tilde{\eta}^2\zeta^2 + \frac{12A}{5\tilde{\eta}R}.\end{aligned}$$

Star-Ring:

$$\begin{aligned}\min_{0\leq r\leq R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{18L\tilde{\eta}\sigma^2}{5GPMK} + \frac{18L^2\tilde{\eta}^2\sigma^2}{P^2M^2K} + \frac{18L^2\tilde{\eta}^2\sigma^2}{P^2MK} + \frac{18L^2\tilde{\eta}^2\sigma^2}{PMK} \\ + \frac{18L^2\tilde{\eta}^2\hat{\zeta}^2}{P^2M^2} + \frac{18L^2\tilde{\eta}^2\hat{\zeta}^2}{P^2} + \frac{18L^2\tilde{\eta}^2\zeta^2}{P^2M^2} + \frac{18L^2\tilde{\eta}^2\zeta^2}{P^2} \\ + 18L^2\tilde{\eta}^2\zeta^2 + \frac{18A}{5\tilde{\eta}R}.\end{aligned}$$

Ring-Star:

$$\begin{aligned}\min_{0\leq r\leq R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{18L\tilde{\eta}\sigma^2}{5GPMK} + \frac{18L^2\tilde{\eta}^2\sigma^2}{G^2P^2K^2} + \frac{18L^2\tilde{\eta}^2\sigma^2}{G^2PMK} + \frac{18L^2\tilde{\eta}^2\sigma^2}{GPMK} \\ + \frac{18L^2\tilde{\eta}^2\hat{\zeta}^2}{G^2P^2} + \frac{18L^2\tilde{\eta}^2\zeta^2}{G^2P^2} + \frac{18L^2\tilde{\eta}^2\zeta^2}{G^2} + 18L^2\tilde{\eta}^2\zeta^2 \\ + \frac{18A}{5\tilde{\eta}R}.\end{aligned}$$

Ring-Ring:

$$\min_{0\le r\le R}\mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \le \frac{24L\tilde{\eta}\sigma^2}{5GPMK}+\frac{24L^2\tilde{\eta}^2\sigma^2}{G^2P^2M^2K}+\frac{24L^2\tilde{\eta}^2\sigma^2}{G^2P^2MK}+\frac{24L^2\tilde{\eta}^2\sigma^2}{G^2PMK} + \frac{24L^2\tilde{\eta}^2\sigma^2}{GPMK}+\frac{24L^2\tilde{\eta}^2\hat{\zeta}^2}{G^2P^2M^2}+\frac{24L^2\tilde{\eta}^2\hat{\zeta}^2}{G^2P^2}+\frac{24L^2\tilde{\eta}^2\zeta^2}{G^2P^2M^2} + \frac{24L^2\tilde{\eta}^2\zeta^2}{G^2P^2}+\frac{24L^2\tilde{\eta}^2\zeta^2}{G^2}+24L^2\tilde{\eta}^2\zeta^2+\frac{24A}{5\tilde{\eta}R}.$$

Here $A \coloneqq F(\mathbf{x}^{(0)}) - F^*$.