Hierarchical Federated Learning with Multi-Timescale Gradient Correction

Wenzhi Fang
Purdue University
fang375@purdue.edu

Dong-Jun Han
Yonsei University
djh@yonsei.ac.kr

Evan Chen
Purdue University
chen4388@purdue.edu

Shiqiang Wang
IBM Research
wangshiq@us.ibm.com

Christopher G. Brinton
Purdue University
cgb@purdue.edu

Abstract

While traditional federated learning (FL) typically focuses on a star topology where clients are directly connected to a central server, real-world distributed systems often exhibit hierarchical architectures. Hierarchical FL (HFL) has emerged as a promising solution to bridge this gap, leveraging aggregation points at multiple levels of the system. However, existing algorithms for HFL encounter challenges in dealing with *multi-timescale model drift*, i.e., model drift occurring across hierarchical levels of data heterogeneity. In this paper, we propose a multi-timescale gradient correction (MTGC) methodology to resolve this issue. Our key idea is to introduce distinct control variables to (i) correct the client gradient towards the group gradient, i.e., to reduce *client model drift* caused by local updates based on individual datasets, and (ii) correct the group gradient towards the global gradient, i.e., to reduce *group model drift* caused by FL over clients within the group. We analytically characterize the convergence behavior of MTGC under general non-convex settings, overcoming challenges associated with couplings between correction terms. We show that our convergence bound is immune to the extent of data heterogeneity, confirming the stability of the proposed algorithm against multi-level non-i.i.d. data. Through extensive experiments on various datasets and models, we validate the effectiveness of MTGC in diverse HFL settings. The code for this project is available at https://github.com/wenzhifang/MTGC.

1 Introduction

In the past several years, federated learning (FL) has emerged as a prevalent approach for distributed training [17, 22, 11, 52, 24]. Conventional FL has typically considered a star topology training architecture, where clients directly communicate with a central server for model synchronization [35, 26]. Scaling this architecture to large numbers of clients becomes problematic, however, given the heterogeneity in FL resource availability and dataset statistics manifesting over large geographies [14, 17, 10]. In practice, such communication networks are often comprised of a *hierarchical architecture* from clients to the main server, as observed in edge/fog computing [25, 38] and software-defined networks (SDN) [20], where devices are supported by intermediate edge servers that are in turn connected to the cloud.

To bridge this gap, researchers have proposed *hierarchical federated learning* (HFL) which integrates *group aggregations* into FL frameworks [30, 5, 47, 15]. In HFL (see Fig. 1), clients are segmented into multiple groups, and the training within each group is coordinated by a group aggregator node (e.g., an edge server coordinating a cell). Meanwhile, the central server orchestrates the training globally by periodically aggregating models across all client groups, facilitated by the group aggregators.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Fundamental challenges. One of the key objectives in FL is to reduce communication overhead while maintaining model performance. Research in conventional FL has established how the global aggregation period, i.e., the number of local iterations during two consecutive communications between clients and the server, impacts FL performance according to the degree of non-i.i.d. (non-independent or non-identically distributed) across client datasets: when local datasets are more heterogeneous, longer aggregation periods cause client models to drift further apart. In HFL, the situation becomes more complex, and is not yet well studied. There are multiple levels of aggregations within/across client groups, and the frequency of these aggregations diminishes further up the hierarchy (since the communication costs become progressively more expensive). As a result, *model drift occurs across multiple levels of non-i.i.d., at different timescales*. In the canonical two-level case from Fig. 1, we have (i) intra-group non-i.i.d., similar to conventional FL, and (ii) inter-group non-i.i.d., arising from data heterogeneity across different groups. This introduces (i) *client model drift* caused by local updates on individual datasets, usually at a shorter timescale, as well as (ii) *group model drift* caused by FL over clients within the group, usually at a longer timescale.

In conventional star-topology FL, algorithms like ProxSkip [36], SCAFFOLD [18], and FedDyn [1] have shown promise for correcting client model drift through local regularization and gradient tracking/correction. However, *these approaches are not easily extendable to the HFL scenario due to its multi-timescale communication architecture*. Specifically, when integrating these methods into HFL, control variables introduced to handle data heterogeneity, such as gradient tracking or dynamic regularization, need to be carefully injected at each level of the hierarchy, taking into account their coupled effects in taming non-i.i.d. Convergence analysis elucidating the impact of different updating frequencies for such control variables remains an unsolved challenge. Existing works on HFL have also not aimed to directly correct for multi-timescale model drift. This can be seen by the fact that the convergence bounds in existing HFL methods [30, 5, 47, 13] become worse as the extent of non-i.i.d. in the system increases (e.g., gradient divergence between hierarchy levels in [47]). Some works have proposed adaptive control of the aggregation period in HFL [13, 31], but they require frequent model aggregations to prevent excessive drift. We thus pose the following question:

How can we tame multi-timescale model drift in non-i.i.d. hierarchical federated learning to provably enhance model convergence performance while not introducing frequent model aggregations?

1.1 Contributions

In this paper, we propose *multi-timescale gradient correction* (MTGC), a methodology which can effectively address multi-level model drift over the topology of HFL with a theoretical guarantee. As depicted in Fig. 1, our key idea is to introduce coupled gradient correction terms – client-group correction and group-global correction – to (i) correct the client gradient towards the group gradient, i.e., to reduce client model drift caused by local updates based on their individual datasets, and (ii) correct the group gradient towards the global gradient, i.e., to reduce group model drift caused by FL across clients within the group, respectively. MTGC thus assists each client model to evolve towards improvements in global performance during HFL. We propose a strategy for updating these gradient correction terms after every group aggregation and global aggregation, respectively, and analyze the convergence behavior of MTGC. Due to the coupling of correction terms and their updates being performed at different timescales, additional challenges arise for theoretical analysis compared to prior work. We thoroughly investigate this problem and make the following contributions:

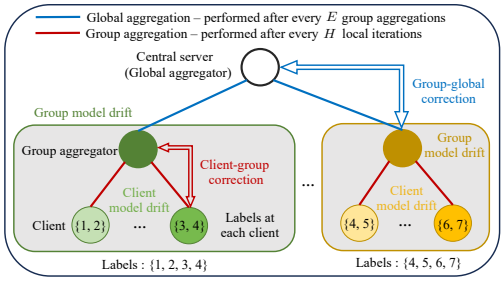


Figure 1: Illustration of multi-timescale gradient correction (MTGC) for multi-level non-i.i.d. in HFL.

- We develop the multi-timescale gradient correction (MTGC) algorithm for taming leveled model drift in HFL. MTGC incorporates coupled control variables for correcting client gradients and group gradients, effectively tackling model biases arising from various levels of non-i.i.d. data at different timescales. The estimation and update procedures for these control variables rely solely on the model updates, ensuring that no significant additional communication overhead is introduced.
- We characterize the convergence rate for MTGC under the non-convex setup. This rate is immune to the extent of intra and inter-group data heterogeneity, confirming the stability of our approach against multi-level non-i.i.d. statistics. Our theoretical result also demonstrates that MTGC achieves

linear speedup in the number of local iterations, group aggregations, and clients. Also, we show that the convergence rate of MTGC recovers that of SCAFFOLD, i.e., the non-hierarchical case, when the number of groups and group aggregation period reduces to one.

- We conduct extensive experiments using various datasets and models across different parameter settings, which demonstrate the superiority of MTGC in diverse non-i.i.d. HFL environments.

1.2 Related Works

Algorithms for conventional FL. The seminal work [35] developed the FedAvg algorithm, incorporating multiple local updates into distributed SGD [58] to relieve communication bottlenecks within conventional star-topology FL. However, FedAvg convergence analysis makes assumptions such as bounded gradients [28, 56, 41] or bounded gradient dissimilarity [45, 12], showing it is not resistant to non-i.i.d. data. To tackle this issue, numerous techniques have been proposed in the literature, including incorporating static/dynamic regularizers [27, 1, 57, 16], adaptive control variables [29, 8, 51, 50], and/or gradient tracking methods [18, 32]. Despite these efforts, existing FL algorithms are not easily extendable to HFL due to the timescale mismatch of multi-level model aggregations induced by the hierarchical system topology. Optimizing these algorithms and ensuring their theoretical convergence in the presence of hierarchical model drift remains unsolved. Our paper addresses these issues through a principled multi-timescale gradient correction method.

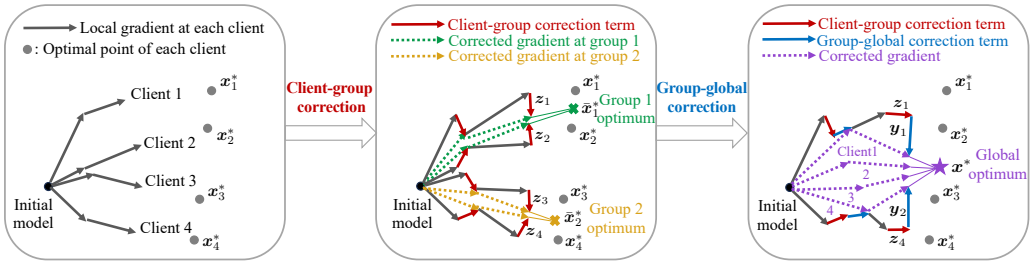
Hierarchical FL. The authors of [5, 30, 13, 49, 55, 47] explored a new FL branch, HFL, tailored for hierarchical systems consisting of a central server, group aggregators, and clients. To tackle the issue of limited communication resources, the authors of [5] developed a FedAvg-like algorithm called hierarchical FedAvg tailored to HFL, and analyzed its convergence behavior. However, their algorithm is built upon an assumption of i.i.d. data. Another work [47] investigated the convergence behavior of hierarchical FedAvg under the non-i.i.d. setup. However, the convergence bound becomes worse as the extent of data heterogeneity increases, making the algorithm vulnerable to non-i.i.d. data characteristics. In [34], ProxSkip-HUB is introduced, but requires clients to compute full batch gradients and upload them to group aggregators after every iteration, which is impractical especially when training large-scale models. Overall, there is still a lack of an algorithm that fully addresses the unique challenge of HFL, i.e., the multi-timescale model drift problem, with theoretical guarantees. We fill this gap by introducing multi-timescale gradient correction and providing theoretical insights.

Gradient tracking/correction. Both gradient tracking and gradient correction aim to fix the local updating directions of clients to mitigate the impact of model drift caused by data heterogeneity. The gradient tracking concept was originally proposed and analyzed in [9] and then extended to consider various factors like time-vary graphs and asynchronous updates [37, 40, 42, 54]. Subsequently, SCAFFOLD [18] applied gradient tracking in FL to mitigate the impact of data heterogeneity across clients, ensuring convergence and stability in non-i.i.d. settings. More recently, in [32, 2], the authors demonstrate the effectiveness of gradient tracking in fully decentralized FL, where clients conduct model aggregations through local client-to-client communications. In [6, 43], gradient tracking is further studied in a semi-decentralized FL setup. Compared to all prior research, our work is the earliest attempt to design an algorithm specifically tailored to multi-timescale model drift in HFL and its training process with periodic local/global aggregations. This presents new challenges in our algorithm design and convergence analysis due to the coupling of our correction terms through their updates at different timescales. In Section 5, we empirically validate the effectiveness of our approach over the prior gradient correction method.

2 Background and Motivation

2.1 Problem Setup: Hierarchical FL

We consider the hierarchical system depicted in Fig. 1. The central server is connected to N group aggregators, each linked to the clients within its region, defined as a group. Each group $j \in \{1, 2, \ldots, N\}$ consists of a set of n_j non-overlapping clients, denoted $\mathcal{C}_j$, resulting in a total of $\sum_{j=1}^{N} n_j$ clients within the system. Each client i has its own local data distribution $\mathcal{D}_i$. The goal of HFL is to construct an optimal global model $\boldsymbol{x}^*$ considering the data distributions of all clients in the system. The role of each group aggregator j involves coordinating the training for the n_j clients within its region, while the central server orchestrates the training across all N groups through



(a) No gradient correction (b) Only client-group gradient correction (c) Multi-timescale gradient correction (Ours)

Figure 2: Visualization of the local update process using multi-timescale gradient correction (MTGC) with 4 clients and 2 groups. **(a)** Without any gradient correction (e.g., hierarchical FedAvg), each client model moves towards its respective optimal point, denoted by $\boldsymbol{x}_i^*$. **(b)** When only client-group correction term $\boldsymbol{z}_i$ is applied, the model of client $i \in \mathcal{C}_j$ moves towards the group optimum $\bar{\boldsymbol{x}}_j^*$. **(c)** In MTGC, the gradient of client $i \in \mathcal{C}_j$ is adjusted by both the client-group correction term $\boldsymbol{z}_i$ and the group-global correction variable $\boldsymbol{y}_j$, assisting each client model to converge towards the global optimum $\boldsymbol{x}^*$ during local iterations.

interaction with the group aggregators. We can formally state the HFL learning objective as follows:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{N}\sum_{j=1}^{N} f_j(\boldsymbol{x}), \text{ where } f_j(\boldsymbol{x}) := \frac{1}{n_j}\sum_{i\in\mathcal{C}_j} F_i(\boldsymbol{x}) \text{ and } F_i(\boldsymbol{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[F_i(\boldsymbol{x},\xi_i)\right]. \quad (1)$$

Here, $f:\mathbb{R}^d\to\mathbb{R}$ denotes the global loss function, $f_j:\mathbb{R}^d\to\mathbb{R}$ is the loss specific to group j, and $F_i:\mathbb{R}^d\to\mathbb{R}$ represents the local loss for client i. In addition, ξ_i is the data point sampled from distribution $\mathcal{D}_i$. Note that our analysis can be easily extended to a weighted average form of (1) by incorporating positive coefficients for each $f_j(\boldsymbol{x})$ or $F_i(\boldsymbol{x})$. For simplicity, these coefficients are assumed to be included in $F_i(\boldsymbol{x})$ as in previous works [18, 48].

2.2 Limitation of Existing Works

In HFL algorithms, group aggregations are conducted after every H local client updates, while global aggregations are performed after every E group aggregations, introducing different timescales. Moreover, different forms of data heterogeneity exist in HFL: (i) intra-group non-i.i.d., due to data heterogeneity across different clients $i\in\mathcal{C}_j$, and (ii) inter-group non-i.i.d., arising from data heterogeneity across different groups $\mathcal{C}_1,...,\mathcal{C}_N$. These lead to *client model drift* and *group model drift*, respectively. The model drifts induced by multi-level data heterogeneity at different timescales hinder hierarchical FedAvg from converging. In Fig. 2(a), we see that during local training, each local model gradually converges towards the optimal point of its respective client's objective function. Hence, to guarantee theoretical convergence, existing HFL works either assume an i.i.d. setup [5] or rely on a bounded gradient dissimilarity assumption similar to the following [47, 15]:

$$\frac{1}{N}\sum_{j=1}^{N}\|\nabla f_j(\boldsymbol{x})-\nabla f(\boldsymbol{x})\|^2\le\delta_1^2,\forall\boldsymbol{x} \text{ and } \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\|\nabla F_i(\boldsymbol{x})-\nabla f_j(\boldsymbol{x})\|^2\le\delta_2^2,\forall\boldsymbol{x},\forall j. \quad (2)$$

The first inequality is employed to limit the group drift, i.e., the deviation of group gradients from the global gradient, while the second one bounds the client drift, i.e., the divergence of client gradients from their group gradient. As a result, the convergence bounds of algorithms in these works become worse as data heterogeneity increases (i.e., as δ_1 or δ_2 increase) [19]. Our approach, developed next, does not require these assumptions and remains stable regardless of the extent of data heterogeneity.

3 Algorithm

3.1 Intuition: Gradient Correction in Hierarchical FL

When relying on multiple local SGD iterations as in hierarchical FedAvg, the model update process is not stable even when the model has reached the optimal $\boldsymbol{x}^*$ satisfying $\nabla f(\boldsymbol{x}^*)=\boldsymbol{0}$. Specifically, with γ as the learning rate, we have $\boldsymbol{x}^*\neq\boldsymbol{x}^*-\gamma\nabla F_i(\boldsymbol{x}^*)$, as the global optimum $\boldsymbol{x}^*$ may not necessarily be optimal for each client's local loss due to data heterogeneity, i.e., $\nabla F_i(\boldsymbol{x}^*)\neq\boldsymbol{0}$ [39]. Correcting the client gradient $\nabla F_i(\boldsymbol{x}^*)$ to the global gradient $\nabla f(\boldsymbol{x}^*)$ is thus necessary to stabilize the process.[1]

Motivation and idea. In HFL, however, due to multi-level aggregations occurring at different timescales, it is infeasible to directly correct the client gradient to the global gradient. In particular,

[1]We motivate our approach here using full batch gradients, though our subsequent algorithm and analysis will support stochastic gradients.

Algorithm 1: HFL with Multi-Timescale Gradient Correction (MTGC)

Input: Initial model $\bar{\boldsymbol{x}}^0$, global aggregation period E, group aggregation period H, learning rate γ, and group-global correction $\boldsymbol{y}_j^0 = -\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{0,0}) + \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{0,0},\xi_{i,0}^{0,0}), \forall j$

each global round $t = 0, 1, \ldots, T-1$ **do**

  Group model initialization: $\boldsymbol{x}_j^{t,0} = \bar{\boldsymbol{x}}^t, \forall j$

  Client-group correction initialization: $\boldsymbol{z}_i^{t,0} = -\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{t,0}) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{t,0}), \forall i\in\mathcal{C}_j, \forall j$

  each group communication round $e = 0, 1, \ldots, E-1$ **do**

    Local model initialization: $\boldsymbol{x}_{i,0}^{t,e} = \bar{\boldsymbol{x}}_j^{t,e}, \forall i, j$

    each local iteration $h = 0, 1, \ldots, H-1$ **do**

      $\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\left(\nabla F_i(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t\right), \forall i\in\mathcal{C}_j, \forall j$ ◇ Clients do in parallel

    Group aggregation: $\bar{\boldsymbol{x}}_j^{t,e+1} = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,H}^{t,e}$

    Client-group corr. update: $\boldsymbol{z}_i^{t,e+1} = \boldsymbol{z}_i^{t,e} + \frac{1}{H\gamma}\left(\boldsymbol{x}_{i,H}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e+1}\right), \forall i\in\mathcal{C}_j, \forall j$ ◇ Clients do in parallel

  Global aggregation: $\bar{\boldsymbol{x}}^{t+1} = \frac{1}{N}\sum_{j=1}^{N}\bar{\boldsymbol{x}}_j^{t,E}$

  Group-global corr. update: $\boldsymbol{y}_j^{t+1} = \boldsymbol{y}_j^t + \frac{1}{HE\gamma}\left(\bar{\boldsymbol{x}}_j^{t,E} - \bar{\boldsymbol{x}}^{t+1}\right), \forall j$ ◇ Group aggregators in parallel

clients are not able to communicate with the central server directly, and there are multiple group aggregation steps before the group aggregators communicate with the main server. Our idea is thus to inject two gradient correction terms: client-group correction and group-global correction. Specifically, the desired iteration to obtain the updated model $\boldsymbol{x}_{\text{new}}$ at the optimal point $\boldsymbol{x}^*$ can be written as

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x}^* - \gamma\Big\{\nabla F_i(\boldsymbol{x}^*) + \underbrace{(\nabla f_j(\boldsymbol{x}^*) - \nabla F_i(\boldsymbol{x}^*))}_{\text{client-group correction}} + \underbrace{(\nabla f(\boldsymbol{x}^*) - \nabla f_j(\boldsymbol{x}^*))}_{\text{group-global correction}}\Big\}, \tag{3}$$

where $\nabla f_j(\boldsymbol{x}^*) - \nabla F_i(\boldsymbol{x}^*)$ and $\nabla f(\boldsymbol{x}^*) - \nabla f_j(\boldsymbol{x}^*)$ represent client-group and group-global correction terms, respectively. Since $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, the two correction terms will enable the model to remain at the optimal point. Given this intuition, the ideal local iteration at client $i \in \mathcal{C}_j$ can be written as

$$\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\Big\{\nabla F_i(\boldsymbol{x}_{i,h}^{t,e}) + \Big(\nabla f_j(\bar{\boldsymbol{x}}_{j,h}^{t,e}) - \nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\Big) + \Big(\nabla f(\bar{\boldsymbol{x}}_h^{t,e}) - \nabla f_j(\bar{\boldsymbol{x}}_{j,h}^{t,e})\Big)\Big\}, \tag{4}$$

where t, e, and h represent global communication rounds, group communication rounds, and client local iterations, respectively, $\bar{\boldsymbol{x}}_{j,h}^{t,e} = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,h}^{t,e}$ is the averaged model within group j, and $\bar{\boldsymbol{x}}_h^{t,e} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,h}^{t,e}$ is the averaged model across the system. Based on (4), we expect to bring each client model closer to the global optima during local updates, as illustrated in Fig. 2(c).

Challenge encountered in HFL. However, it is important to note that the update process in (4) still cannot be directly used in HFL. This is because client-group communication and group-global communication do not occur at every iteration of HFL training; instead, they happen at different timescales, and clients are not able to obtain the current group information $\nabla f_j(\bar{\boldsymbol{x}}_{i,h}^{t,e})$ and global information $\nabla f(\bar{\boldsymbol{x}}_{i,h}^{t,e})$ at every local iteration. We next propose a strategy that mimics the gradient correction described above while ensuring theoretical convergence.

3.2 Multi-Timescale Gradient Correction (MTGC)

Tackling multi-timescale model drifts. To approximate (4) during HFL training, we introduce two control variables $\boldsymbol{z}$ and $\boldsymbol{y}$ that track/approximate $\nabla f_j - \nabla F_i$ and $\nabla f - \nabla f_j$, respectively. The variables $\boldsymbol{z}$ and $\boldsymbol{y}$ are then employed to correct the local gradients to prevent model drifts. The challenge here is to keep updating $\boldsymbol{z}$ and $\boldsymbol{y}$ appropriately in the multi-timescale communication scenario, given that communications between the clients and group aggregator, and between the group aggregators and global aggregator, are not always feasible. We propose a strategy to update $\boldsymbol{z}$ after every H local iterations, i.e., whenever each client is able to communicate with the group aggregator, allowing the group information to be updated and shared among the clients within the same group. Similarly, we propose a strategy to update $\boldsymbol{y}$ after every E group aggregations, i.e.,

whenever the group aggregators are able to communicate with the global aggregator, enabling the global information to be refreshed and shared across all clients in the system. We name our strategy multi-timescale gradient correction (MTGC) due to the updates of $\boldsymbol{z}$ and $\boldsymbol{y}$ occurring in different timescales, to tackle the issue of multi-level model drift coupled across the hierarchy in HFL.

In Fig. 2(c), we illustrate MTGC during client-side model updates. In particular, at each local iteration h of group round e of global round t, each client $i \in \mathcal{C}_j$ updates its local model as follows:

$$\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma \left(\nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t \right). \tag{5}$$

(i) Client-group correction term. In (5), $\boldsymbol{z}_i^{t,e}$ is responsible for correcting the gradient of client $i \in \mathcal{C}_j$ towards the gradient of group j at the e-th group aggregation of global round t. After every group aggregation e at global round t, this term is updated at each client i as follows:

$$\boldsymbol{z}_i^{t,e+1} = \frac{1}{H} \sum_{h=0}^{H-1} \left(\left(\frac{1}{n_j} \sum_{i \in \mathcal{C}_j} \nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) \right) - \nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) \right). \tag{6}$$

(ii) Group-global correction term. $\boldsymbol{y}_j^t$ in (5) aims to correct the gradient of group j towards the global gradient. At the end of global round t, this term is updated at group aggregator j as follows:

$$\boldsymbol{y}_j^{t+1} = \frac{1}{HE} \sum_{e=0}^{E-1} \sum_{h=0}^{H-1} \left(\left(\frac{1}{N} \sum_{j=1}^{N} \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} \nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) \right) - \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} \nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) \right). \tag{7}$$

Key remarks. The updating policies for $\boldsymbol{z}_i^{t,e}$ and $\boldsymbol{y}_j^t$ follow similar patterns to the ideal corrections outlined in (4). Here, we observe that $\sum_{i \in \mathcal{C}_j} \boldsymbol{z}_i^{t,e} = \boldsymbol{0}, \forall j$ and $\sum_{j=1}^{N} \boldsymbol{y}_j^t = \boldsymbol{0}$, indicating that the correction terms do not have an impact on the per-iteration model averages. Instead, the introduction of $\boldsymbol{z}_i^{t,e}$ and $\boldsymbol{y}_j^t$ eliminates model drifts of clients and groups, respectively, during local iterations. Intuitively, as the iteration approaches the global optimal point, we expect $\boldsymbol{z}_i^{t,e} \to \nabla f_j(\boldsymbol{x}^*) - \nabla F_i(\boldsymbol{x}^*)$ and $\boldsymbol{y}_j^t \to \nabla f(\boldsymbol{x}^*) - \nabla f_j(\boldsymbol{x}^*)$ so that the update in (5) stabilizes at the global optimal point. We also see that $\boldsymbol{z}_i^{t,e}$ and $\boldsymbol{y}_j^t$ are coupled (5), i.e., the update of one of the terms affects $\boldsymbol{x}$ which in turn affects the other one, raising challenges for theoretical analysis. In Section 4, we will guarantee convergence of MTGC in general non-convex settings without relying on bounded data heterogeneity assumptions.

MTGC algorithm. The overall procedure of our training strategy is summarized in Algorithm 1, where we rewrite the updates of $\boldsymbol{z}_i^{t,e}$ and $\boldsymbol{y}_j^t$ in a different but equivalent manner to facilitate practical implementation of MTGC. Compared to hierarchical FedAvg, which does not consider any correction terms, we see that no additional communication is required for MTGC within each group round e. Additional communication is introduced only after E group aggregations for initializing $\boldsymbol{z}_i^{t,0}$ (Line 4)[2] and broadcasting $\boldsymbol{y}_j^{t+1}$ (obtained in Line 14) to the clients in $\mathcal{C}_j$. We will see in Section 5 that these marginal additional costs lead to significant performance enhancements for HFL settings.

Generalization to arbitrary number of levels. The proposed MTGC algorithm can be extended to an HFL system architecture with an arbitrary number of levels. Further discussions and experimental results for a three-level case are provided in Appendix E.

3.3 Connection with SCAFFOLD

When the number of groups reduces to $N = 1$ with $E = 1$, we have $\boldsymbol{y}_j^t = 0$ (no group-global correction), and thus MTGC reduces to SCAFFOLD [18]. In SCAFFOLD, at each round t, clients perform local updates according to $\boldsymbol{x}_{i,h+1}^t = \boldsymbol{x}_{i,h}^t - \gamma \left(\nabla F_i(\boldsymbol{x}_{i,h}^t, \xi_{i,h}^t) - \boldsymbol{c}_i^t + \boldsymbol{c}^t \right), h = 0, 1, \ldots, H-1$, where $\boldsymbol{c}_i^{t+1} = \boldsymbol{c}_i^t - \boldsymbol{c}^t + \frac{1}{H\gamma} \left(\bar{\boldsymbol{x}}^t - \boldsymbol{x}_{i,H}^t \right)$, and the server aggregates local models and controlling variables as $\bar{\boldsymbol{x}}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_{i,H}^t$ and $\boldsymbol{c}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{c}_i^{t+1}$. We can show that $\boldsymbol{c}_i^t - \boldsymbol{c}^t$ in SCAFFOLD plays the same role as $\boldsymbol{z}_i^{t,e}$ in MTGC. However, the additional term $\boldsymbol{y}_j^t$ introduced in MTGC for the multi-level setting makes the convergence guarantee more challenging, as $\boldsymbol{y}_j^t$ is coupled with $\boldsymbol{z}_i^{t,e}$ and both are updated at different time scales. These aspects will be thoroughly examined next.

4 Convergence Analysis

In this section, we establish a convergence guarantee for the proposed MTGC algorithm. Our theoretical analysis relies on the following standard assumptions commonly used in the literature on stochastic optimization and FL under non-convex settings [18, 44, 4].

[2] This is only required for theoretical analysis. In the experiments, we initialize $\boldsymbol{z}_i^{t,0} = \boldsymbol{0}, \forall i$.

Assumption 1. *Each local loss function F_i is differentiable and L-smooth, i.e., there exists a positive constant L such that for any $\boldsymbol{x}$ and $\boldsymbol{y}$, $\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{y})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|, \forall i$.*

Assumption 2. *The stochastic gradient $\nabla F_i(\boldsymbol{x}, \xi_i)$ is an unbiased estimate of the true gradient, i.e., $\mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\nabla F_i(\boldsymbol{x}, \xi_i)] = \nabla F_i(\boldsymbol{x}), \forall \boldsymbol{x}$ and the variance of the stochastic gradient $\nabla F_i(\boldsymbol{x}, \xi_i)$ is uniformly bounded as $\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\|\nabla F_i(\boldsymbol{x}, \xi_i) - \nabla F_i(\boldsymbol{x})\|^2 \leq \sigma^2, \forall \boldsymbol{x}$.*

Note that (i) global aggregation, (ii) the update of upper-level correction variable $\boldsymbol{y}$ and local aggregation, and (iii) the update of lower-level correction variable $\boldsymbol{z}$ are performed at different timescales in MTGC. If we directly consider $\{\nabla f(\bar{\boldsymbol{x}}^t)\}$ as in SCAFFOLD, it is difficult to capture the effects of group aggregation and correction variable $\boldsymbol{z}$. Moreover, it is hard to establish a tight connection between $\nabla f(\bar{\boldsymbol{x}}^t)$ and $\boldsymbol{x}_{i,h}^{t,e}$, $\forall i, h, \tau$ since there is a large lag between $\boldsymbol{x}_{i,h}^{t,e}$ and $\bar{\boldsymbol{x}}^t$. To tackle this, we introduce a new metric, which is the gradient $\nabla f(\hat{\boldsymbol{x}}^{t,e})$ at virtual sequence $\{\hat{\boldsymbol{x}}^{t,e} = \frac{1}{N}\sum_{j=1}^{N} \bar{\boldsymbol{x}}_j^{t,e}\}$, to characterize the convergence of MTGC.

We next state our main theoretical results. All the proofs are provided in Appendix F:

Theorem 4.1. *Suppose Assumptions 1 and 2 hold and the learning rate satisfies $\gamma \leq \frac{1}{40EHL}$. Then the iterates $\{\hat{\boldsymbol{x}}^{t,e}\}$ obtained by the MTGC algorithm satisfy*

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 = \mathcal{O}\left(\frac{f(\bar{\boldsymbol{x}}^0) - f^*}{\gamma TEH} + \frac{\gamma}{\tilde{N}}L\sigma^2 + \gamma^2 E^2 H^2 L^2 \sigma^2\right), \tag{8}$$

where $\tilde{N} = \left(\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\right)^{-1}$, and f^* is the lower bound of $f(\boldsymbol{x})$, i.e., $f(\boldsymbol{x}) \geq f^*$.

There are two key steps in our proof. The first is the characterization of the evolution of $\left\|\boldsymbol{z}_i^{t,e} + \nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right\|^2$ and $\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2$. By bounding these values that capture the error between each control variable and the ideal correction, we are able to establish a connection between the local updating direction and the global gradient without relying on the bounded gradient dissimilarity assumption, laying the foundation for the whole proof. The second is that we extracted a recursive relationship for the accumulation of group-level and client-level model drifts, and designed a novel Lyapunov function to mitigate the interplay impact between these drifts. Further details are provided in the appendix.

Applying an appropriate learning rate γ to Algorithm 1 yields the following corollary:

Corollary 4.1. *Under the assumptions of Theorem 4.1, let $\mathcal{F}_0 = f(\bar{\boldsymbol{x}}^0) - f^*$. Then there exists a learning rate $\gamma \leq \frac{1}{40EHL}$ such that the iterates $\{\hat{\boldsymbol{x}}^{t,e}\}$ satisfy*

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right). \tag{9}$$

Discussions. Corollary 4.1 provides the convergence upper bound of the MTGC algorithm. It shows that the error approaches zero as $T \to \infty$. If $\sigma \neq 0$, the upper bound is dominated by the first term in the right-hand side of (9), which characterizes the speed of convergence of MTGC to a stationary point in the stochastic case. This reveals MTGC achieves linear speedup in the number of group aggregations E and local updates H. In other words, we can attain the same level of performance with less global communication rounds, i.e., a smaller value of T, by increasing the number of local iterations, i.e., H, and group aggregations, i.e., E. When considering the special case $n_{j'} = n$, $\forall j' \in \{1, 2, \ldots, N\}$ with uniform client numbers, the rate becomes $\mathcal{O}(\sqrt{\mathcal{F}_0 L\sigma^2}/\sqrt{NnTEH})$. This implies that MTGC attains linear speedup in the number of clients as well.

Moreover, we also see that our convergence rate recovers the results of SCAFFOLD when the number of groups reduces to $N = 1$ and the number of group aggregations reduces to $E = 1$ (see Appendix G for more discussions). We also highlight that, different from prior works on HFL where the convergence bound becomes worse as the extent of data heterogeneity increases, our bound is stable against multi-level non-i.i.d. data due to the multi-timescale gradient correction approach.

5 Experimental Results

5.1 Setup

Dataset, model, hyperparameters, and compute setting. In our experiments, we consider four widely used datasets: EMNIST-Letters (EMNIST-L) [7], Fashion-MNIST [53], CIFAR-10 [23], and

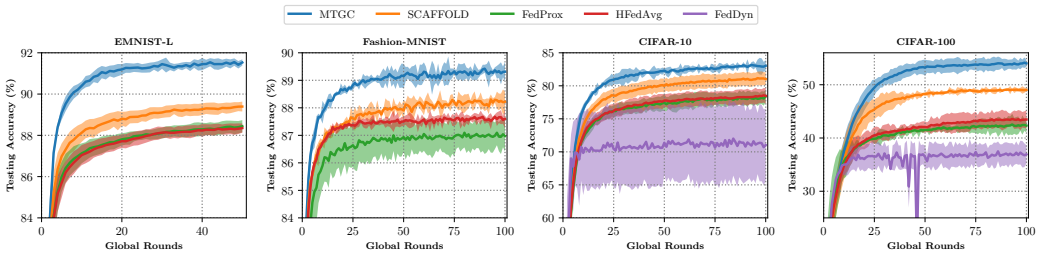


Figure 3: **Comparison with FL baselines.** In this experiment, popular FL algorithms are extended to the HFL setup for comparison with MTGC. We consider four datasets in the group non-i.i.d. & client non-i.i.d. setting. Experiments are conducted over 3 random trials. We see that MTGC obtains the best testing accuracy in each case, validating our multi-level approach for correcting multi-timescale model drifts.

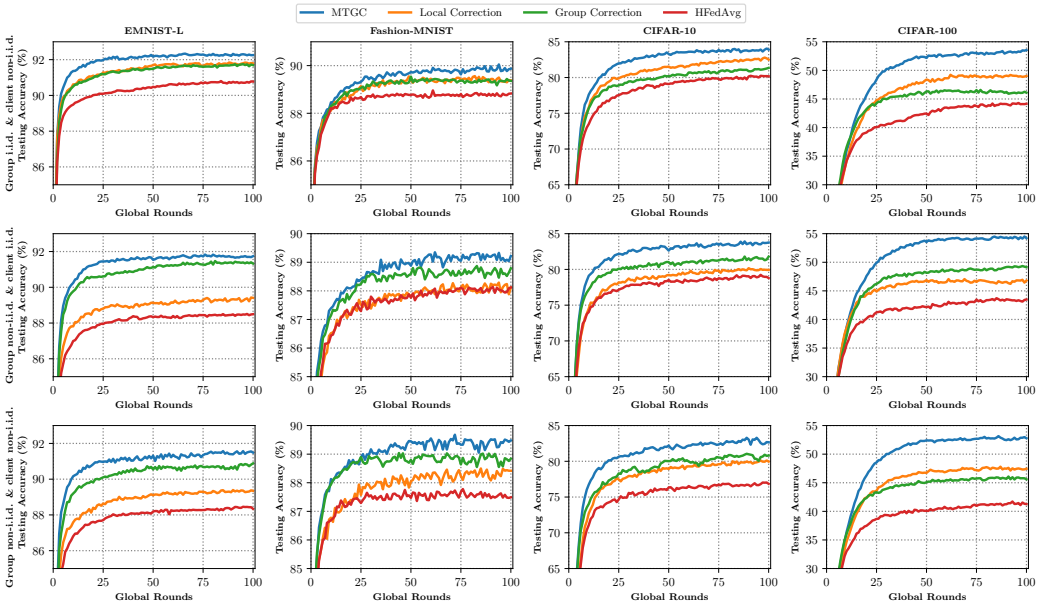


Figure 4: **Comparison with gradient correction baselines.** Three different data distribution scenarios are considered. We see that the local correction method is effective for handling client non-i.i.d. within each group (top row), while the group correction method is effective for handling non-i.i.d. across groups (middle row). MTGC obtains the most stable performance (all rows) by combining multiple correction levels.

CIFAR-100 [23]. The former two are processed through a multi-layer perceptron (MLP) model, featuring two hidden layers, each comprising 200 neurons, and ending with a softmax layer. For the CIFAR-10 classification task, we employ a convolutional neural network (CNN) following the architecture outlined in seminal work [35]. For CIFAR-100, we adopt a ResNet-18 model with batch normalization layers substituted by group normalization layers. Across all algorithms considered, we maintain a consistent learning rate $\eta = 0.1$ and batch size 50. We conduct the experiments based on a cluster of 3 NVIDIA A100 GPUs with 40 GB memory. Our code is based on the framework of [1].

FL data distribution. We set the total number of clients as 100, evenly distributed over $N = 10$ groups. We also study the effect of N in Appendix B. We consider three different data distribution settings: (i) group i.i.d. & client non-i.i.d., (ii) group non-i.i.d. & client i.i.d., and (iii) group non-i.i.d. & client non-i.i.d. scenarios. First, in the group i.i.d. & client non-i.i.d. case, the training dataset is initially divided uniformly and randomly into N segments corresponding to N groups. Subsequently, each segment is further divided into $100/N$ partitions for the clients using a Dirichlet distribution [1]. Second, in the group non-i.i.d. & client i.i.d. case, the dataset is first segmented into N partitions for the groups using a Dirichlet distribution, followed by a uniform random distribution of each segment to $100/N$ clients. Finally, when both groups and clients are non-i.i.d., the dataset is split into N segments for the groups using a Dirichlet distribution, and then, each group's segment is distributed among $100/N$ clients through a Dirichlet distribution. The Dirichlet parameter is set to 0.1.

Table 5.1: The *number of global rounds* required by different algorithms to attain the testing accuracy of 80% for CIFAR-10 under different settings. Taking HFedAvg as the benchmark, we show the speedup achieved by MTGC and other baselines as we vary aggregation periods E and H. MTGC consistently outperforms baselines, and the speedup gets more significant as E and H increase. Standard deviation is based on 3 random trials.

Data distribution	Params (E, H)	HFedAvg	Local Correction	Group Correction	MTGC
Group i.i.d. & client non-i.i.d.	(10, 20)	144.3±3.4 (1×)	57.0±0.8 (2.5×)	72.0±1.6 (2.0×)	**45.7**±0.9 (3.2×)
	(10, 30)	169.0±2.4 (1×)	51.0±1.4 (3.3×)	81.3±1.2 (2.1×)	**37.3**±1.9 (4.5×)
	(10, 40)	214.0±5.9 (1×)	45.3±1.2 (4.7×)	85.7±1.2 (2.5×)	**32.0**±0.8 (6.7×)
	(10, 20)	144.3±3.4 (1×)	57.0±0.8 (2.5×)	72.0±1.6 (2.0×)	**45.7**±0.9 (3.2×)
	(20, 20)	105.0±4.1 (1×)	34.0±0.8 (3.0×)	53.0±0.8 (2.0×)	**22.3**±0.9 (4.7×)
	(30, 20)	82.7±2.1 (1×)	25.7±0.9 (3.2×)	44.7±0.5 (1.8×)	**16.3**±1.2 (5.1×)
Group non-i.i.d. & client i.i.d.	(10, 20)	246.0±3.7 (1×)	92.3±1.7 (2.7×)	53.7±1.2 (4.6×)	**37.7**±0.5 (6.5×)
	(10, 30)	302.7±4.9 (1×)	88.0±1.6 (3.4×)	44.3±1.2 (6.8×)	**27.3**±1.2 (11.1×)
	(10, 40)	320.0±2.4 (1×)	94.7±1.2 (3.5×)	43.0±1.6 (7.2×)	**21.3**±1.2 (15.4×)
	(10, 20)	246.0±3.7 (1×)	92.3±1.7 (2.7×)	53.7±1.2 (4.6×)	**37.7**±0.5 (6.5×)
	(20, 20)	308.7±3.3 (1×)	74.3±1.7 (4.2×)	31.0±0.8 (10.0×)	**18.7**±1.7 (16.5×)
	(30, 20)	344.7±4.6 (1×)	85.7±2.1 (4.0×)	25.7±1.7 (13.4×)	**13.0**±0.8 (26.5×)
Group non-i.i.d. & client non-i.i.d.	(10, 20)	363.0±7.3 (1×)	141.7±2.9 (2.6×)	83.7±1.2 (4.3×)	**52.7**±1.2 (6.7×)
	(10, 30)	>500 (1×)	127.7±2.9 (>3.9×)	79.7±1.7 (>6.3×)	**38.7**±0.9 (>12.9×)
	(10, 40)	>500 (1×)	169.7±5.2 (>2.9×)	106.3±1.9 (>4.7×)	**31.7**±0.5 (>15.8×)
	(10, 20)	363.0±7.3 (1×)	141.7±2.9 (2.6×)	83.7±1.2 (4.3×)	**52.7**±1.2 (6.7×)
	(20, 20)	>500 (1×)	113.3±3.4 (>4.4×)	45.3±1.2 (>11.0×)	**25.0**±1.6 (>20.0×)
	(30, 20)	>500 (1×)	86.7±2.4 (>5.8×)	50.7±2.4 (>9.9×)	**19.6**±0.5 (>25.5×)

5.2 Results and Discussion

Comparison with conventional FL algorithms. For comparison, we first apply the well-known FL methods, FedProx [27], SCAFFOLD [18], and FedDyn [1], to HFL, by running their training algorithms within each group of the hierarchical system. We also consider HFedAvg [47] as a baseline. Fig. 3 compares MTGC with these baselines in the group non-i.i.d. & client non-i.i.d. case. We observe that MTGC outperforms all the considered conventional algorithms, achieving the highest testing accuracy, especially for the complicated CIFAR-100 dataset. FedDyn achieves the lowest performance, demonstrating significant variance and instability. The significant performance gap between MTGC and FedDyn, in particular, can be attributed to the hierarchical setup disrupting the special structure of FedDyn. This result reveals that some algorithms designed for the conventional star-topology FL may be non-trivial to be extended to hierarchical setups. The overall results confirm the effectiveness of our approach that effectively tackles the multi-timescale drift problem in HFL.

Comparison with gradient correction baselines. In Fig. 4, we compare MTGC with the gradient correction baselines. Specifically, we apply local correction ($\boldsymbol{z}_i^{t,e}$) to HFedAvg, and group correction ($\boldsymbol{y}_j^t$) to HFedAvg. These baselines can be viewed as schemes applying SCAFFOLD [18] within each group and across groups, respectively. We also report the results of the original HFedAvg to see the effects of gradient correction clearly. We make the following key observations. First, the testing accuracy achieved by HFedAvg decreases as the extent of data heterogeneity increases, e.g., from the first or second row to the third row in Fig. 4. This shows that data heterogeneity hinders the convergence of HFedAvg. Second, with the assistance of local or group correction, the algorithm attains a higher accuracy. In the case of group i.i.d. & client non-i.i.d., HFedAvg augmented with client local correction performs better than the variant with group correction. Conversely, in the scenario where groups are non-i.i.d. and clients are i.i.d., the opposite holds. This can be explained by the dominance of data heterogeneity in each case. In the former scenario, because the heterogeneity is primarily at the client-level, local client correction becomes more beneficial. On the other hand, in the latter scenario, where the heterogeneity shifts to the group level, group correction becomes more advantageous. Finally, we see that MTGC consistently outperforms baselines under all settings, where the performance gains brought by the multi-timescale gradient correction become more significant when it comes to the group non-i.i.d.& client non-i.i.d. case.

Speedup in H and E. In Table 5.1, we investigate the effects H and E, which determine the periods of group aggregation and global aggregation in HFL. We report the number of global rounds required to attain the desired testing accuracy of 80% for CIFAR-10 under different settings. We have the following observations: As E or H increases, the required number of global rounds of MTGC for achieving the desired accuracy decreases. This demonstrates the speedup of the proposed algorithm in the number of local iterations and group aggregations, which fits well with our theory discussed in Section 4. In addition, the speedup achieved by MTGC compared to HFedAvg gets more significant

as E or H increases. For instance, in the group i.i.d. & client non-i.i.d. case, MTGC attains $3.3\times$ speedup when $E = 10$, $H = 20$, which increases to $4.7\times$ when $E = 10$, $H = 40$. This reveals that MTGC utilizes local iterations better compared with the baselines.

Impact of data heterogeneity. Consistent with the results in Fig. 4, we see from Table 5.1 that the required number of global rounds of HFedAvg increases as data heterogeneity increases, while MTGC is more stable against non-i.i.d. data. The gain of MTGC over HFedAvg becomes evident as data heterogeneity increases, confirming the effectiveness of our multi-timescale gradient correction approach for addressing the unique challenges of HFL.

Further experiments. Additional experimental results including the impacts of hierarchical system parameters and the performance in 3-level HFL are provided in Appendices B and E.

6 Conclusion and Limitation

We have proposed MTGC, a multi-timescale gradient correction approach for HFL. Embedded with control variables updated in different timescales, MTGC effectively corrects gradient biases and alleviates both client model drift and group model drift in hierarchical setups. We established the convergence bound of MTGC in the non-convex setup and showed its stability against multi-level data heterogeneity. Finally, we confirmed the advantage of our MTGC through extensive experiments in different non-i.i.d. HFL settings. A limitation of our work is that despite providing experiments for HFL systems with more than two levels (in Appendix E), our convergence analysis focused on the two-level case, which provides an interesting future direction of investigation.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under grants CNS-2146171 and CPS-2313109, and by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-24-1-0083.

References

[1] Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: International Conference on Learning Representations (2020)

[2] Alghunaim, S.A.: Local exact-diffusion for decentralized optimization and learning. IEEE Transactions on Automatic Control (2024)

[3] Bao, W., Wang, H., Wu, J., He, J.: Optimizing the collaboration structure in cross-silo federated learning. In: International Conference on Machine Learning. pp. 1718–1736. PMLR (2023)

[4] Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning (2018)

[5] Castiglia, T., Das, A., Patterson, S.: Multi-level local SGD: Distributed SGD for heterogeneous hierarchical networks. In: International Conference on Learning Representations (2020)

[6] Chen, E., Wang, S., Brinton, C.G.: Taming subnet-drift in D2D-enabled fog learning: A hierarchical gradient tracking approach. arXiv preprint arXiv:2312.04728 (2023)

[7] Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: EMNIST: an extension of mnist to handwritten letters (2017)

[8] Condat, L., Agarskỳ, I., Malinovsky, G., Richtárik, P.: TAMUNA: Doubly accelerated federated learning with local training, compression, and partial participation. arXiv preprint arXiv:2302.09832 (2023)

[9] Di Lorenzo, P., Scutari, G.: Next: In-network nonconvex optimization. IEEE Transactions on Signal and Information Processing over Networks **2**(2), 120–136 (2016)

[10] Fang, W., Han, D.J., Brinton, C.G.: Submodel partitioning in hierarchical federated learning: Algorithm design and convergence analysis. In: ICC 2024-IEEE International Conference on Communications. pp. 268–273. IEEE (2024)

[11] Fang, W., Yu, Z., Jiang, Y., Shi, Y., Jones, C.N., Zhou, Y.: Communication-efficient stochastic zeroth-order optimization for federated learning. IEEE Transactions on Signal Processing **70**, 5058–5073 (2022)

[12] Haddadpour, F., Mahdavi, M.: On the convergence of local descent methods in federated learning. arXiv preprint arXiv:1910.14425 (2019)

[13] Hosseinalipour, S., Azam, S.S., Brinton, C.G., Michelusi, N., Aggarwal, V., Love, D.J., Dai, H.: Multi-stage hybrid federated learning over large-scale D2D-enabled fog networks. IEEE/ACM transactions on networking **30**(4), 1569–1584 (2022)

[14] Hosseinalipour, S., Brinton, C.G., Aggarwal, V., Dai, H., Chiang, M.: From federated to fog learning: Distributed machine learning over heterogeneous wireless networks. IEEE Communications Magazine **58**(12), 41–47 (2020)

[15] Jiang, X., Zhu, H.: On the convergence of hierarchical federated learning with partial worker participation. In: The 40th Conference on Uncertainty in Artificial Intelligence (2024)

[16] Jiang, X., Rodomanov, A., Stich, S.U.: Federated optimization with doubly regularized drift correction. In: Forty-first International Conference on Machine Learning (2024)

[17] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and trends® in machine learning **14**(1–2), 1–210 (2021)

[18] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: SCAFFOLD: Stochastic controlled averaging for federated learning. In: International conference on machine learning. pp. 5132–5143. PMLR (2020)

[19] Khaled, A., Mishchenko, K., Richtárik, P.: Tighter theory for local SGD on identical and heterogeneous data. In: International Conference on Artificial Intelligence and Statistics. pp. 4519–4529. PMLR (2020)

[20] Kim, H., Feamster, N.: Improving network management with software defined networking. IEEE Communications Magazine **51**(2), 114–119 (2013)

[21] Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., Stich, S.: A unified theory of decentralized SGD with changing topology and local updates. In: International Conference on Machine Learning. pp. 5381–5393. PMLR (2020)

[22] Konecnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 **8** (2016)

[23] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Toronto, ON, Canada (2009)

[24] Lan, G., Liu, X.Y., Zhang, Y., Wang, X.: Communication-efficient federated learning for resource-constrained edge devices. IEEE Transactions on Machine Learning in Communications and Networking (2023)

[25] Li, H., Ota, K., Dong, M.: Learning iot in edge: Deep learning for the internet of things with edge computing. IEEE network **32**(1), 96–101 (2018)

[26] Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)

[27] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine learning and systems. vol. 2, pp. 429–450 (2020)

[28] Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of FedAvg on Non-IID data. In: International Conference on Learning Representations (2019)

[29] Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., Cheng, Y.: Variance reduced local SGD with lower communication complexity. arXiv preprint arXiv:1912.12844 (2019)

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

[31] Liu, L., Zhang, J., Song, S., Letaief, K.B.: Hierarchical federated learning with quantization: Convergence analysis and system design. IEEE Transactions on Wireless Communications **22**(1), 2–18 (2022)

[32] Liu, Y., Lin, T., Koloskova, A., Stich, S.U.: Decentralized gradient tracking with local steps. Optimization Methods and Software pp. 1–28 (2024)

[33] Ma, J., Long, G., Zhou, T., Jiang, J., Zhang, C.: On the convergence of clustered federated learning. arXiv preprint arXiv:2202.06187 (2022)

[34] Malinovsky, G., Yi, K., Richtárik, P.: Variance reduced ProxSkip: Algorithm, theory and application to federated learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 15176–15189 (2022)

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

[36] Mishchenko, K., Malinovsky, G., Stich, S., Richtárik, P.: ProxSkip: Yes! local gradient steps provably lead to communication acceleration! finally! In: International Conference on Machine Learning. pp. 15750–15769. PMLR (2022)

[37] Nedic, A., Olshevsky, A., Shi, W.: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM Journal on Optimization **27**(4), 2597–2633 (2017)

[38] Nguyen, V.D., Chatzinotas, S., Ottersten, B., Duong, T.Q.: Fedfog: Network-aware optimization of federated learning over wireless fog-cloud systems. IEEE Transactions on Wireless Communications **21**(10), 8581–8599 (2022)

[39] Pathak, R., Wainwright, M.J.: Fedsplit: An algorithmic framework for fast federated optimization. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7057–7066 (2020)

[40] Scutari, G., Sun, Y.: Distributed nonconvex constrained optimization over time-varying digraphs. Mathematical Programming **176**, 497–544 (2019)

[41] Stich, S.U.: Local SGD converges fast and communicates little. In: International Conference on Learning Representations (2019)

[42] Tian, Y., Sun, Y., Scutari, G.: Achieving linear convergence in distributed asynchronous multiagent optimization. IEEE Transactions on Automatic Control **65**(12), 5264–5279 (2020)

[43] Wang, H., Chi, Y.: Communication-efficient federated optimization over semi-decentralized networks. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 13241–13245. IEEE (2024)

[44] Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H.B., y Arcas, B.A., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., Diggavi, S., Eichner, H., Gadhikar, A., Garrett, Z., Girgis, A.M., Hanzely, F., Hard, A., He, C., Horvath, S., Huo, Z., Ingerman, A., Jaggi, M., Javidi, T., Kairouz, P., Kale, S., Karimireddy, S.P., Konecny, J., Koyejo, S., Li, T., Liu, L., Mohri, M., Qi, H., Reddi, S.J., Richtarik, P., Singhal, K., Smith, V., Soltanolkotabi, M., Song, W., Suresh, A.T., Stich, S.U., Talwalkar, A., Wang, H., Woodworth, B., Wu, S., Yu, F.X., Yuan, H., Zaheer, M., Zhang, M., Zhang, T., Zheng, C., Zhu, C., Zhu, W.: A field guide to federated optimization (2021)

[45] Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7611–7623 (2020)

[46] Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: A novel framework for the analysis and design of heterogeneous federated learning. IEEE Transactions on Signal Processing **69**, 5234–5249 (2021)

[47] Wang, J., Wang, S., Chen, R.R., Ji, M.: Demystifying why local aggregation helps: Convergence analysis of hierarchical SGD. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8548–8556 (2022)

[48] Wang, S., Ji, M.: A lightweight method for tackling unknown participation statistics in federated averaging. In: International Conference on Learning Representations (2024)

[49] Wang, Z., Xu, H., Liu, J., Huang, H., Qiao, C., Zhao, Y.: Resource-efficient federated learning with hierarchical aggregation in edge computing. In: IEEE Conference on Computer Communications. pp. 1–10 (2021)

[50] Wu, F., Guo, S., Qu, Z., He, S., Liu, Z., Gao, J.: Anchor sampling for federated learning with partial client participation. In: International Conference on Machine Learning. pp. 37379–37416. PMLR (2023)

[51] Wu, F., Guo, S., Wang, H., Zhang, H., Qu, Z., Zhang, J., Liu, Z.: From deterioration to acceleration: A calibration approach to rehabilitating step asynchronism in federated optimization. IEEE Transactions on Parallel and Distributed Systems **34**(5), 1548–1559 (2023)

[52] Wu, F., Li, Z., Li, Y., Ding, B., Gao, J.: Fedbiot: Llm local fine-tuning in federated learning without full model. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3345–3355 (2024)

[53] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)

[54] Xu, J., Tian, Y., Sun, Y., Scutari, G.: Distributed algorithms for composite optimization: Unified framework and convergence analysis. IEEE Transactions on Signal Processing **69**, 3555–3570 (2021)

[55] Yang, H.: H-FL: A hierarchical communication-efficient and privacy-protected architecture for federated learning. arXiv preprint arXiv:2106.00275 (2021)

[56] Yu, H., Yang, S., Zhu, S.: Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 5693–5700 (2019)

[57] Zhang, X., Hong, M., Dhople, S., Yin, W., Liu, Y.: FedPD: A federated learning framework with adaptivity to Non-IID data. IEEE Transactions on Signal Processing **69**, 6055–6070 (2021)

[58] Zinkevich, M., Weimer, M., Li, L., Smola, A.: Parallelized stochastic gradient descent. In: Advances in Neural Information Processing Systems. vol. 23 (2010)

Appendix

A Connection Between HFL and Cluster FL

Our work focuses on HFL, employing a multi-layered structure consisting of local nodes, local aggregators, and a central server. Both clustered FL and HFL aim to improve FL learning efficiency by leveraging structured client groupings. The difference between them lies in the grouping criteria. HFL focuses on collaborative training over a given network topology, where clients are generally grouped based on their geographical location or network connection status, and aims to build a single global model under this setting. CFL groups clients to optimize model training, with different global models constructed depending on the group. [33] demonstrates how dynamic clustering based on data distributions can enhance model performance. [3] explores alleviating negative transfer from collaboration by clustering clients into non-overlapping coalitions based on their distribution distances and data quantities.

B Additional Experiments on CIFAR-10

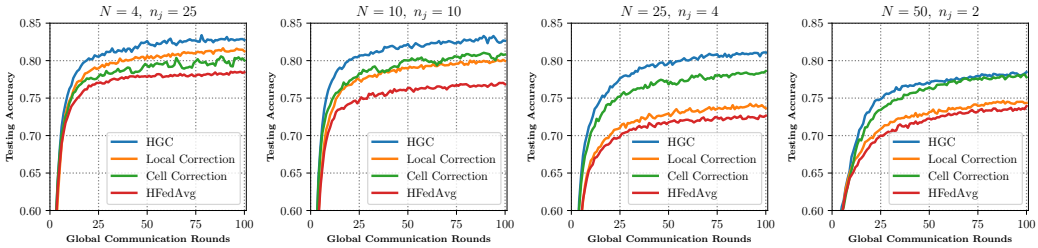


Figure 5: Comparison of testing accuracy versus global communication round across different system parameters under both group non-i.i.d. and client non-i.i.d. setup. E and H are set to 30 and 20, respectively.

The impact of system parameters. Fig. 5 shows how the performance of MTGC changes with different numbers of groups and clients in each group. From this figure, we observe that as the number of clients in each group, i.e., n_j increases, client correction becomes more important. On the other hand, as the number of groups increases, the algorithm with group correction performs better than that with client correction.

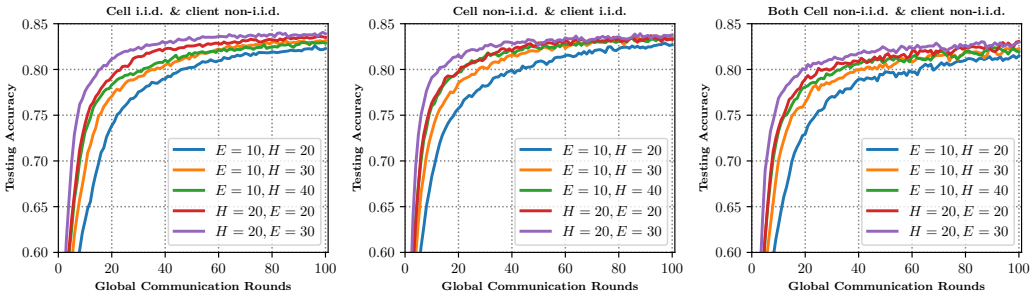
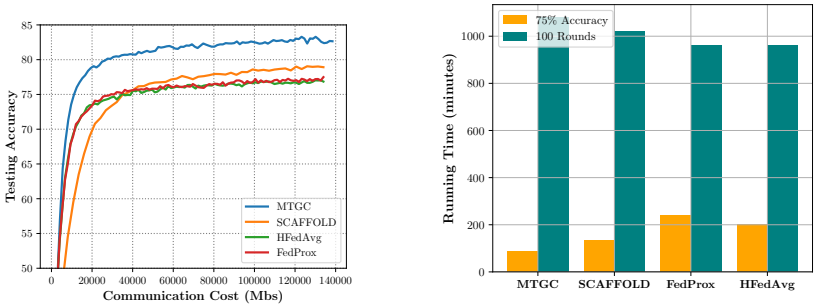


Figure 6: Performance of MTGC under a different number of local iterations, i.e., H, and a different number of group aggregations, i.e., E. The number of groups and clients in each group is set to $N = 10$ and $n_j = 10$, respectively.

The impact of local iteration and group aggregation: Fig. 6 depicts the performance of MTGC under a different number of local iterations, i.e., H, and a different number of group aggregations, i.e., E. It's clear that MTGC achieves speedup in the number of local iterations and group aggregations.

Communication cost comparison. Compared to HFedAvg, MTGC requires initializing the correction variables at the start of each global round, which adds additionaly communication overhead. Specifically, for every E steps of group aggregation, MTGC incurs an additional communication cost equivalent to one transmission of the model parameters. In other words, the per-aggregation communication complexity of MTGC is $\frac{E+1}{E}$ times that of HFedAvg. To show this impact, we have added experiments comparing the communication cost and testing accuracy at the client side. This



(a) Testing accuracy versus communication cost

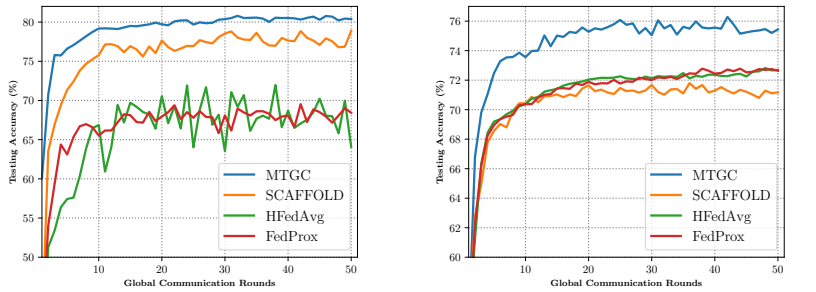
(b) Running time comparison

Figure 7: Comparison of communication cost (a); and running time for attaining 75% testing accuracy and finishing 100 global rounds on CIFAR-10 (b)

experiment was conducted on CIFAR-10 dataset with $E = 30$ and $H = 20$ under both client and group non-i.i.d. setup. The model and other parameters are the same as in the original manuscript. The results are shown in Fig. 7a. The results demonstrate that MTGC achieves higher testing accuracy for a given communication cost, highlighting the efficiency and effectiveness of our approach.

Running time comparison. We compared the computation time of our MTGC algorithm with the baselines. Using NVIDIA A100 GPUs with 40 GB memory, we conducted experiments on the CIFAR-10 dataset with $E = 30$ and $H = 20$ under both client and group non-i.i.d. setup. The model and other parameters are the same as in the original manuscript. We report the required time for attaining a preset accuracy of 75% and for running 100 global rounds in Fig. 7b of the attached pdf. The speedup in the convergence makes up for the introduced computation cost per iteration due to the extra operation induced by the correction variables. Actually, the computation cost incurred by the correction variable is relatively small compared to computing gradients in a neural network using backpropagation.

C Experiments on Distribution Shift Datasets



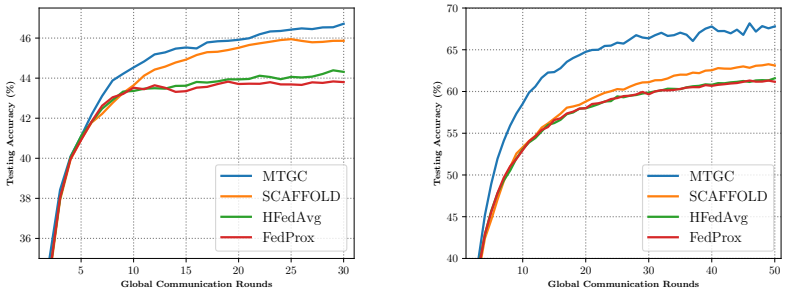
(a) Performance comparisons under label shift on Fashion-MNIST

(b) Performance comparisons under feature shift on Fashion-MNIST

Figure 8: Performance comparisons on Fashion-MNIST under label shift and Fashion-MNIST under feature shift

To further show the robustness, we studied the performance of MTGC under another two different non-i.i.d. scenarios: label shift and feature shift, as referenced in [3, 33]. These experiments were performed using the Fashion-MNIST dataset.

For label shift [3, 33], we randomly assign 3 classes out of 10 classes to each group with a relatively balanced number of instances per class, and then assign 2 classes to each client. As discussed in [3], label shift adds more heterogeneity to this system. According to the results shown in Fig. 8a, it is clear that the proposed algorithm is more robust against data heterogeneity. Specifically, there is less oscillation in MTGC compared with HFedAvg and the attained accuracy of MTGC in the given communication round is higher than all baselines.



(a) Performance comparison on the Shakespeare dataset based on LSTM model

(b) Performance comparison under CINIC-10 dataset

Figure 9: Performance evaluation on Shakespeare and CINIC-10

For feature shift [3], we first partition data following the group non-i.i.d. & client non-i.i.d. case as in our original manuscript, and then let clients at different groups rotate images for different angles. Concretely, for the clients at the i-th group, the angle is $-50 + 10 \times i$. Note that this rotation is only applied to the training set. The feature shift increases the diversity between the training set and the testing set, which thus adds difficulty to this classification task. In Fig. 8b, we see that MTGC attains the best performance among these baselines.

D Additional Experiments on CINIC-10 and Shakespear Datasets

We conducted additional experiments on the larger Shakespeare and CINIC-10 datasets. For the **Shakespeare dataset**, we randomly pick 100 characters (people) in Shakespeare's plays. We let each client have 1,500 samples, where each sample is a sequence of 80 characters (words). Considering that there are 100 clients in the system, there are 150,000 train samples in total. This means that the number of samples is 3 times that of CIFAR-10 (or CIFAR-100), which has 50,000 train samples. The performance comparison is presented in Fig. 9a, where we use the LSTM model, the same as [1], and set the learning rate 0.5, $H = 75$, and $E = 30$. It is seen that MTGC consistently outperforms the baseline methods in larger datasets.

The **CINIC-10 dataset** contains 90,000 training images, 90,000 validation images, and 90,000 test images, significantly larger than CIFAR-10 and CIFAR-100 with 60,000 images. It includes images from both CIFAR-10 and ImageNet, enhancing diversity. We believe that the larger size and diversity of CINIC-10 further confirm the validity of our experiments. The model and hyperparameters used for the CINIC-10 dataset are the same as those of the CIFAR-10 task shown in the original manuscript. As illustrated in Fig. 9b, MTGC maintains its superior performance on the CINIC-10 dataset, consistent with its performance on other tasks.

E Extension to the HFL System with Arbitrary Number of Levels

We extend MTGC for the HFL system with M levels in this subsection. For presentation ease, we adopt different notations than those used in the main text. Specifically, we denote the number of total iterations at clients as r. The aggregation periods for level m are denoted as P_m. This means that the m-th level aggregator aggregates the model from the clients within its coverage after every P_m local iterations. The global server is treated as the first level aggregator. Note that $P_m > P_{m+1}$ and $P_{m+1} \mid P_m, \forall m = 1, \ldots, M-1$. We denote the model maintained at the nodes connected to the m-th level aggregator $(k_1, k_2, \ldots, k_{m-1})$ as $\boldsymbol{x}^r_{k_1,\ldots,k_m}$, where $k_m \in \{1, \ldots, N_m\}$. The gradient correction term between nodes $(k_1, k_2, \ldots, k_{m-1})$ and $(k_1, k_2, \ldots, k_m)$ is denoted as $\boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_m}$. The overall procedures are summarized in Algorithm 2.

Algorithm 2: MTGC for the HFL System with Arbitrary Number of Levels

Input: $\gamma, \{P_i : i \in \{1, 2, \ldots, M\}\}$

Initialize: $\boldsymbol{\nu}^r_{k_1}, \boldsymbol{\nu}^r_{k_1,k_2}, \ldots, \boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_M}, \forall k_1, k_2, \ldots, k_M$

$r = 0, 1, \ldots, R-1$ **do**

  each client $(k_1, \ldots, k_M) \in \mathcal{V}$ **in parallel do**

    Compute stochastic gradient: $\boldsymbol{g}^r_{k_1,\ldots,k_M}$

    $\boldsymbol{x}^{r+1}_{k_1,\ldots,k_M} = \boldsymbol{x}^r_{k_1,\ldots,k_M} - \gamma \left(\boldsymbol{g}^r_{k_1,\ldots,k_M} + \boldsymbol{\nu}^r_{k_1} + \boldsymbol{\nu}^r_{k_1,k_2} + \ldots + \boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_M}\right)$

  level $i = M, \ldots, 1$ **do**

    if $P_i \mid r+1$ **then**

      i-th level model aggregate:

      $\boldsymbol{x}^{r+1}_{k_1,\ldots,k_i} = \frac{1}{N_i \ldots N_M} \sum_{k_i=1}^{N_i} \cdots \sum_{k_M=1}^{N_M} \boldsymbol{x}^{r+1}_{k_i,\ldots,k_M}, \forall k_1, \ldots, k_i$

      i-th level correction term update:

      $\boldsymbol{\nu}^{r+1}_{k_1,k_2,\ldots,k_i} = \boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_i} + \frac{1}{\gamma P_i}\left(\boldsymbol{x}^{r+1}_{k_1,\ldots,k_i} - \boldsymbol{x}^t_{k_1,\ldots,k_i}\right)$

      Model dissemination: $\boldsymbol{x}^{r+1}_{k_1,\ldots,k_M} \leftarrow \boldsymbol{x}^{r+1}_{k_1,\ldots,k_i}, \forall k_1, \ldots, k_M$

      Initialize $\boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_{i+1}}, \ldots, \boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_M}, \forall k_1, k_2, \ldots, k_M$

    else

      $\boldsymbol{\nu}^{r+1}_{k_1} = \boldsymbol{\nu}^r_{k_1}$, $\boldsymbol{\nu}^{r+1}_{k_1,k_2} = \boldsymbol{\nu}^r_{k_1,k_2}, \ldots, \boldsymbol{\nu}^{r+1}_{k_1,k_2,\ldots,k_i} = \boldsymbol{\nu}^r_{k_1,k_2,\ldots,k_i}, \forall k_1, k_2, \ldots, k_i$

      break

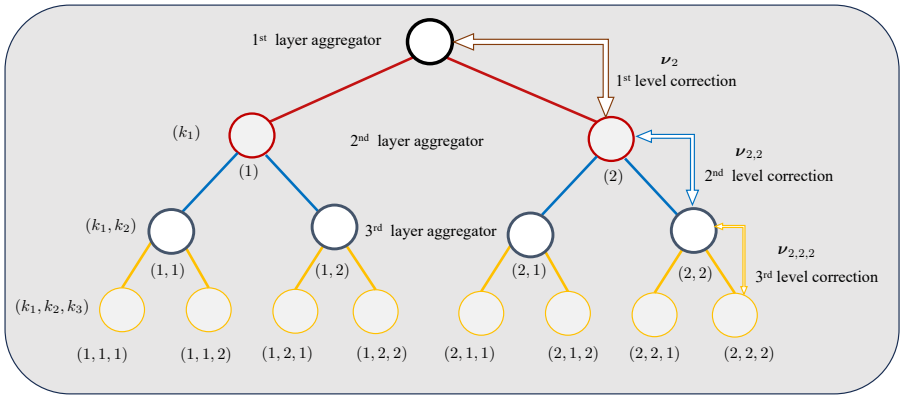


Figure 10: HFL system with 3-level topology

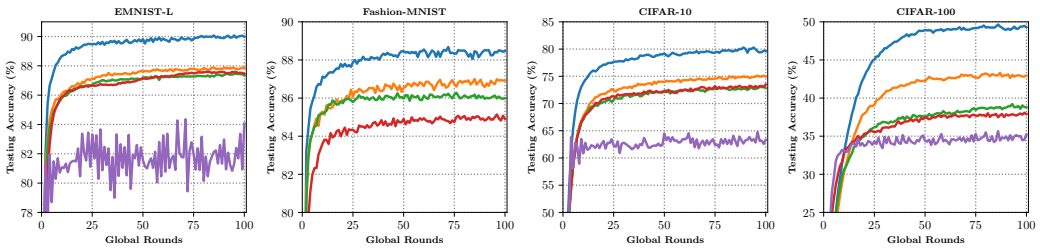


Figure 11: Performance of MTGC for three-level HFL with data non-i.i.d. across each level. Parameters are set to $N_1 = 4,\ N_2 = 5,\ N_3 = 5,\ P_1 = 500,\ P_2 = 100,\ P_3 = 10$.

We numerically validate the performance of MTGC by conducting experiments in the three-level case as shown in Fig. 10. The results are shown in Fig. 11. The total number of clients is set to be 100 while $N_1 = 4,\ N_2 = 5,\ N_3 = 5$. Additionally, the aggregation periods are set to be $P_1 = 500,\ P_2 = 100,\ P_3 = 10$. The data is non-i.i.d. distributed across each level.

F Proof of the Main Results

F.1 Preliminaries

Before proceeding to the proof of the main theorem. We introduce some basic inequalities in this subsection that will be frequently used in our proof.

Lemma F.1.1. *For any set of K vectors $\{\boldsymbol{p}_k\}_{k=1}^K$, $\left\|\frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k\right\|^2 \leq \frac{1}{K}\sum_{k=1}^K \|\boldsymbol{p}_k\|^2$, $\left\|\sum_{k=1}^K \boldsymbol{p}_k\right\|^2 \leq K\sum_{k=1}^K \|\boldsymbol{p}_k\|^2$, and*

$$\frac{1}{K}\sum_{k=1}^K \left\|\boldsymbol{p}_k - \frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k\right\|^2 = \frac{1}{K}\sum_{k=1}^K \|\boldsymbol{p}_k\|^2 - \left\|\frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k\right\|^2.$$

Lemma F.1.2. *For any two vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d$, $\|\boldsymbol{p}+\boldsymbol{q}\|^2 \leq (1+\alpha)\|\boldsymbol{p}\|^2 + \left(1+\frac{1}{\alpha}\right)\|\boldsymbol{q}\|^2$.*

Lemma F.1.3. *Suppose a sequence of random vectors $\{\boldsymbol{p}_k\}_{k=1}^K$ satisfy $\mathbb{E}[\boldsymbol{p}_k] = \boldsymbol{0}, \forall k$. Then,*

$$\mathbb{E}\left\|\frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k\right\|^2 = \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}\|\boldsymbol{p}_k\|^2.$$

Lemma F.1.4. [46, Lemma 2] *Suppose a sequence of random vectors $\{\boldsymbol{p}_k\}_{k=1}^K$ satisfy $\mathbb{E}[\boldsymbol{p}_k \,|\, \boldsymbol{p}_{k-1}, \boldsymbol{p}_{k-2}, \ldots, \boldsymbol{p}_1] = \boldsymbol{0}, \forall k$. Then,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^K \boldsymbol{p}_k\right\|^2\right] = \sum_{k=1}^K \mathbb{E}\left[\|\boldsymbol{p}_k\|^2\right].$$

Lemma F.1.5. [21, Lemma 17] *For any $a_0 \geq 0, b \geq 0, c \geq 0, d > 0$, there exist a constant $\eta \leq \frac{1}{d}$ such that*

$$\frac{a_0}{T\eta} + b\eta + c\eta^2 \leq 2\left(\frac{a_0 b}{T}\right)^{\frac{1}{2}} + 2c^{\frac{1}{3}}\left(\frac{a_0}{T}\right)^{\frac{2}{3}} + \frac{da_0}{T}. \tag{10}$$

F.2 Proofs of Theorem 4.1 and Corollary 4.1

For the convenience of presentation, we introduce the following notations

$$\begin{aligned}
\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N Z_j^{t,e} &= \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{z}_i^{t,e} + \nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right\|^2 \\
\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N Y_j^{t,e} &= \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N \mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \\
D_t &= \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N \mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2 \\
Q_t &= \sum_{e=0}^{E-1}\frac{1}{NH}\sum_{j=1}^N \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e} - \boldsymbol{x}_{i,h}^{t,e}\right\|^2 \\
\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N \Theta_j^{t,e} &= \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^N \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2,
\end{aligned} \tag{11}$$

where $Z_j^{t,e}$ and $Y_j^{t,e}$ characterize the biases between client-group correction term $\boldsymbol{z}_i^{t,e}$ and $\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)$ and between group-global correction term $\boldsymbol{y}_j^t$ and $\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)$, respectively, D_t and Q_t denote the group model drift and client model drift, respectively, and $\Theta_j^{t,e}$ represents model progress for group j.

To prove the convergence of MTGC, we start with characterizing the evolution of the global loss, i.e., $f(\boldsymbol{x})$ through the following lemma.

Lemma F.2.1. *Suppose that Assumptions 1 and 2 hold and $\gamma \leq \frac{1}{2HL}$, then the iterates generated by Algorithm 1 satisfy*

$$\mathbb{E}f(\bar{\boldsymbol{x}}^{t+1}) \leq \mathbb{E}f(\bar{\boldsymbol{x}}^{t}) - \frac{\gamma H}{2}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 + \gamma L^2 H\left(Q_t + D_t\right) + \gamma^2 LEH\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2. \tag{12}$$

Lemma (F.2.1) implies that we need further to study the evolution of Q_t and D_t. In particular, we establish upper bounds for Q_t and D_t in Lemmas F.2.2 and F.2.3, respectively.

Lemma F.2.2. *Suppose that Assumptions 1 and 2 hold and $\gamma \leq \frac{1}{8HL}$, then the client model drift Q_t, defined in (11), can be bounded as*

$$\begin{aligned} Q_t \leq & 24\gamma^2 H^2 L^2 D_t + 12\gamma^2 H^2 \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Z_t^{t,e} + 12\gamma^2 H^2 \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Y_j^{t,e} \\ & + 24\gamma^2 H^2 \sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 + 3EH\gamma^2\sigma^2. \end{aligned} \tag{13}$$

Lemma F.2.3. *Suppose that Assumptions 1 and 2 hold and $\gamma \leq \frac{1}{10EHL}$. Then the group model drift D_t, defined in (11), can be bounded as*

$$D_t \leq 24\gamma^2 E^2 H^2 L^2 Q_t + 12\gamma^2 E^2 H^2 \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Y_j^{t,e} + 3\gamma^2 E^3 H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2. \tag{14}$$

The results shown in Lemmas F.2.2 and F.2.3 suggest that $Z_j^{t,e}$ and $Y_j^{t,e}$ are crucial for understanding the dynamics of MTGC. Hence, we derive upper bounds for $Z_j^{t,e}$ and $Y_j^{t,e}$, which are presented in Lemmas F.2.4 and F.2.5, respectively.

Lemma F.2.4. *Suppose that Assumptions 1 and 2 hold. Then the bias between client-group correction term $\boldsymbol{z}_i^{t,e}$ and $\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)$, i.e., $Z_j^{t,e}$ can be bounded as*

$$\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Z_j^{t,e} \leq 4L^2 Q_t + 4L^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\Theta_j^{t,e} + 2\frac{E}{H}\sigma^2 + \sigma^2. \tag{15}$$

Lemma F.2.5. *Suppose that Assumptions 1 and 2 hold. Then the bias between group-global correction term $\boldsymbol{y}_j^t$ and $\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)$, i.e., $Y_j^{t,e}$ defined in (11), $t \geq 1$, can be bounded as*

$$\begin{aligned} \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Y_j^{t,e} \leq & \left(8L^2 + 48\gamma^2 L^4 E^2 H^2\right)\left(Q_{t-1} + D_{t-1}\right) + 48\gamma^2 L^4 E^2 H^2\left(Q_t + D_t\right) \\ & + 48\gamma^2 L^2 E^2 H^2 \sum_{\tau=0}^{E-1}\left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2 + \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right) \\ & + 32\gamma^2 L^2 E^2 H\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 + \frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2. \end{aligned} \tag{16}$$

Additionally, when $t = 0$, $Y_j^{0,e}$ can be bounded as

$$\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Y_j^{0,e} \leq E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 + L^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{0,e} - \bar{\boldsymbol{x}}^{0}\right\|^2. \tag{17}$$

In addition, the upper bound of $\Theta_j^{t,e}$ is presented in the following lemma.

Lemma F.2.6. *Suppose that Assumptions 1 and 2 hold. Then the group model progress at the (t,e)-th round, i.e., $\Theta_j^{t,e}$, defined in (11), can be bounded as*

$$\begin{aligned} \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\Theta_j^{t,e} \leq & 8\gamma^2 H^2 L^2 Q_t + 8\gamma^2 H^2 L^2 D_t + 8\gamma^2 H^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N} Y_j^{t,e} \\ & + 8\gamma^2 H^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 + 2\gamma^2 EH\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2. \end{aligned} \tag{18}$$

Recalling Lemma F.2.1, we can see that what we actually need is the evolution of $Q_t + D_t$. With Lemmas F.2.2-F.2.6, we can characterize this evolution which is formalized in Lemma F.2.7.

Lemma F.2.7. *Suppose that Assumptions 1 and 2 hold and* $\gamma \leq \frac{1}{33EHL}$*, the model deviation* $\Gamma_t = Q_t + D_t$ *satisfies*

$$\Gamma_t \leq \frac{1}{2}\Gamma_{t-1} + \left(1152\gamma^4 H^4 L^2 + 72\gamma^2 E^2 H^2\right) \sum_{\tau=0}^{E-1} \left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2 + \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right) + 294\gamma^2 E^3 H \sigma^2,$$

$$\Gamma_0 \leq \left(648\gamma^4 H^4 L^2 + 42\gamma^2 E^2 H^2\right) \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)\right\|^2 + 146\gamma^2 E^3 H^2 \sigma^2.$$

The proofs of Lemmas F.2.1-F.2.7 are provided in Appendix F.3. With these lemmas, we are ready to prove Theorem 4.1.

F.2.1 Proof of Theorem 4.1

Starting with Lemma F.2.1, i.e.,

$$\mathbb{E} f(\bar{\boldsymbol{x}}^{t+1}) - \gamma L^2 H \Gamma_t \leq \mathbb{E} f(\bar{\boldsymbol{x}}^t) - \frac{\gamma H}{2} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 + \gamma^2 LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2, \tag{19}$$

where $\Gamma_t = Q_t + D_t$. Adding $2\gamma L^2 H \Gamma_t$ on both sides of the above inequality and utilizing Lemma F.2.7, we have

$$\begin{aligned}
\mathbb{E} f(\bar{\boldsymbol{x}}^{t+1}) + \gamma L^2 H \Gamma_t \leq\ & \mathbb{E} f(\bar{\boldsymbol{x}}^t) + \gamma L^2 H \Gamma_{t-1} - \frac{\gamma H}{2} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 + \gamma^2 LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2 \\
& + 2\gamma L^2 H \times 294\gamma^2 E^3 H \sigma^2 + \left(2304\gamma^5 H^5 L^4 + 144\gamma^3 E^2 H^3 L^2\right) \sum_{\tau=0}^{E-1} \left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2 + \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right).
\end{aligned}$$

For notation ease, we denote $\Phi_{t+1} = \mathbb{E} f(\bar{\boldsymbol{x}}^{t+1}) - f^* + \gamma L^2 H \Gamma_t$, $\Phi_t \geq 0$, $\forall t \geq 0$. As long as $\gamma \leq \frac{E}{8HL}$ in Theorem 4.1, $2304\gamma^5 H^5 L^4 \leq 36\gamma^3 E^2 H^3 L^2$, we thus have

$$\begin{aligned}
\Phi_{t+1} \leq\ & \Phi_t - \frac{\gamma H}{2} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 + 180\gamma^3 E^2 H^3 L^2 \sum_{\tau=0}^{E-1} \left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2 + \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right) \\
& + 2\gamma L^2 H \times 294\gamma^2 E^3 H \sigma^2 + \gamma^2 LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2.
\end{aligned} \tag{20}$$

When $\gamma \leq \frac{1}{40EHL}$, we have $\frac{\gamma H}{2} - 360\gamma^3 E^2 H^3 L^2 \geq \frac{\gamma H}{4}$. Telescoping the above inequality from $t = 1$ to $T - 1$, we have

$$\begin{aligned}
\Phi_T \leq\ & \Phi_1 - \frac{\gamma H}{4} \sum_{t=1}^{T-1} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2 + 180\gamma^3 E^2 H^3 L^2 \sum_{\tau=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,\tau})\right\|^2 \\
& + 2\gamma(T-1) L^2 H \times 294\gamma^2 E^3 H \sigma^2 + \gamma^2 (T-1) LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2.
\end{aligned} \tag{21}$$

According to Lemma F.2.1, when $t = 0$,

$$\mathbb{E} f(\bar{\boldsymbol{x}}^1) \leq \mathbb{E} f(\bar{\boldsymbol{x}}^0) - \frac{\gamma H}{2} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,e})\right\|^2 + \gamma L^2 H \Gamma_0 + \gamma^2 LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2, \tag{22}$$

it follows that

$$\mathbb{E} f(\bar{\boldsymbol{x}}^1) + \gamma L^2 H \Gamma_0 \leq \mathbb{E} f(\bar{\boldsymbol{x}}^0) - \frac{\gamma H}{2} \sum_{e=0}^{E-1} \mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,e})\right\|^2 + 2\gamma L^2 H \Gamma_0 + \gamma^2 LEH \frac{1}{N^2} \sum_{j=1}^{N} \frac{1}{n_j} \sigma^2. \tag{23}$$

Combining (23) with (21) and plugging the upper bound of Γ_0, established in Lemma F.2.7, into the inequality, we have

$$\begin{aligned}\Phi_T \leq & \mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*+2\gamma L^2 H\times 146\gamma^2E^3H^2\sigma^2+2\gamma TL^2H\times 292\gamma^2E^3H\sigma^2+\gamma^2TLEH\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\ &-\left(\frac{\gamma H}{2}-180\gamma^3E^2H^3L^2-2\gamma L^2H\times\left(648\gamma^4H^4L^2+42\gamma^2E^2H^2\right)\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,e})\right\|^2\\ &-\frac{\gamma H}{4}\sum_{t=1}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2.\end{aligned}$$

The setup of γ presented in Theorem 4.1 is enough to guarantee $\frac{\gamma H}{2}-180\gamma^3E^2H^3L^2-2\gamma L^2H\times\left(648\gamma^4H^4L^2+42\gamma^2E^2H^2\right)\geq\frac{\gamma H}{4}$. Therefore, combining the last terms and taking some basic algebra operation, we have

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\leq 4\frac{\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*}{\gamma TEH}+4\gamma L\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+\frac{1168}{T}\gamma^2L^2E^2H^2\sigma^2+2352\gamma^2L^2E^2H\sigma^2.$$

The above bound can be further simplified to

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\leq 4\frac{\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*}{\gamma TEH}+4\frac{\gamma L\sigma^2}{\tilde{N}}+3520\gamma^2E^2H^2L^2\sigma^2. \tag{24}$$

This completes the proof of Theorem 4.1.

F.2.2 Proof of Corollary 4.1

Rewriting the bound in 24 as

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\leq\frac{4\left(\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*\right)}{T(\gamma EH)}+\frac{4L\sigma^2}{\tilde{N}EH}(\gamma EH)+3520L^2\sigma^2(\gamma EH)^2, \tag{25}$$

and recalling Lemma F.1.5, one can claim that there exists a learning rate $(\gamma EH)\leq\frac{1}{d}$ such that

$$\begin{aligned}\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\leq & 8\sqrt{\frac{(\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*)L\sigma^2}{\tilde{N}TEH}}+96\left(\frac{(\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*)L\sigma}{T}\right)^{\frac{2}{3}}+\frac{d(\mathbb{E}f(\bar{\boldsymbol{x}}^0)-f^*)}{T}\\ \sim & \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0L\sigma^2}{\tilde{N}TEH}}+\left(\frac{\mathcal{F}_0L\sigma}{T}\right)^{\frac{2}{3}}+\frac{d\mathcal{F}_0}{T}\right).\end{aligned} \tag{26}$$

Given that we need $\gamma\leq\frac{1}{40EHL}$ for Theorem 4.1, we can set $d=40L$. We thus can find a step size in the range of $(\gamma EH)\leq\frac{1}{40L}$, i.e., $\gamma\leq\frac{1}{40EHL}$ such that

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\leq\mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0L\sigma^2}{\tilde{N}TEH}}+\left(\frac{\mathcal{F}_0L\sigma}{T}\right)^{\frac{2}{3}}+\frac{L\mathcal{F}_0}{T}\right). \tag{27}$$

This completes the proof of Corollary 4.1.

F.3 Proofs of Lemmas F.2.1-F.2.7

F.3.1 Proof of Lemma F.2.1

Under the framework of MTGC, the virtual global model obeys the following iteration:

$$\hat{\boldsymbol{x}}^{t,e+1}=\hat{\boldsymbol{x}}^{t,e}-\gamma\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{z}_i^{t,e}+\boldsymbol{y}_j^t\right).$$

As $\sum_{i\in\mathcal{C}_j}\boldsymbol{z}_i^{t,e}=\mathbf{0}$ and $\sum_{j=1}^{N}\boldsymbol{y}_j^t=\mathbf{0}$, the the virtual global model iteration reduces to

$$\hat{\boldsymbol{x}}^{t,e+1}=\hat{\boldsymbol{x}}^{t,e}-\gamma\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right). \tag{28}$$

With Assumption 1, we have

$$f(\boldsymbol{y})\leq f(\boldsymbol{x})+<\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}>+\frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \tag{29}$$

Plugging (28) into (29), we have

$$\begin{aligned}\mathbb{E}f(\hat{\boldsymbol{x}}^{t,e+1})\leq&\mathbb{E}f(\hat{\boldsymbol{x}}^{t,e})\underbrace{-\gamma\mathbb{E}\left\langle\nabla f(\hat{\boldsymbol{x}}^{t,e}),\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\right\rangle}_{T_1}\\&\underbrace{+\gamma^2L\frac{1}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\right\|^2}_{T_2}.\end{aligned} \tag{30}$$

Utilizing $\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right]=\mathbf{0}$, we rewrite T_1 as follows

$$\begin{aligned}T_1=&-\gamma H\mathbb{E}\left\langle\nabla f(\hat{\boldsymbol{x}}^{t,e}),\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\rangle\\
=&\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\mp\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{\boldsymbol{x}}_j^{t,e})-\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2-\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2\\
&-\frac{\gamma H}{2}\mathbb{E}\left\|\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2\\
\leq&\gamma H\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})-\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{\boldsymbol{x}}_j^{t,e})\right\|^2+\gamma H\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{\boldsymbol{x}}_j^{t,e})-\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2\\
&-\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2-\frac{\gamma H}{2}\mathbb{E}\left\|\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2\\
\leq&\gamma H\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j(\hat{\boldsymbol{x}}^{t,e})-\nabla f_j(\bar{\boldsymbol{x}}_j^{t,e})\right\|^2+\gamma H\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\mathbb{E}\left\|\nabla F_i(\bar{\boldsymbol{x}}_j^{t,e})-\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2\\
&-\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2-\frac{\gamma H}{2}\mathbb{E}\left\|\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2\\
\leq&\gamma HL^2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2+\gamma HL^2\frac{1}{NH}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{h=0}^{H-1}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\boldsymbol{x}_{i,h}^{t,e}\right\|^2\\
&-\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2-\frac{\gamma H}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2,\end{aligned}$$

where the last inequality comes from Assumption 1.

On the other hand, we can bound T_2 as follows

$$
\begin{aligned}
T_2 \leq & \mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right)\right\|^2+\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2 \\
\leq & \frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j^2}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\sum_{h=0}^{H-1}\boldsymbol{g}_{i,h}^{t,e}-\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2+\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2 \\
\leq & \frac{H\sigma^2}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}+\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t,e})\right\|^2,
\end{aligned}
$$

where second inequality comes from Lemma F.1.3 and the last inequality follows Assumption 2 and Lemma F.1.4.

Plugging the derived upper bounds of T_1 and T_2 into 30 and utilized $\gamma \leq \frac{1}{2HL}$, we obtain

$$
\begin{aligned}
\mathbb{E}f(\hat{\boldsymbol{x}}^{t,e+1}) \leq & \mathbb{E}f(\hat{\boldsymbol{x}}^{t,e})-\frac{\gamma H}{2}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,e})\right\|^2+\gamma^2 L\frac{H\sigma^2}{N^2}\sum_{j=1}^{N}\frac{1}{n_j} \\
& +\gamma HL^2\frac{1}{NH}\sum_{j=1}^{N}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\boldsymbol{x}_{i,h}^{t,e}\right\|^2+\gamma HL^2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2.
\end{aligned}
$$

Telescoping the above inequality from $e=0$ to $H-1$ gives rise to Lemma F.2.1.

F.3.2 Proofs of Lemmas F.2.2 and F.2.6

Part I (Lemma F.2.2): Let $q_{j,h}^{t,e}=\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\boldsymbol{x}_{i,h}^{t,e}\right\|^2$, $q_{j,0}^{t,e}=0$. For $0\leq h\leq H-2$, we have

$$
\begin{aligned}
& n_j q_{j,h+1}^{t,e} \\
= & \sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e}-\gamma\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t+\boldsymbol{z}_i^{t,e}\right)-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2 \\
\leq & \left(1+\frac{1}{H-1}\right)\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2+H\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\gamma\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t+\boldsymbol{z}_i^{t,e}\right)\right\|^2+n_j\gamma^2\sigma^2 \\
= & \left(1+\frac{1}{H-1}\right)\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2+\gamma^2 H\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\mp\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\mp\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right. \\
& \left.\mp\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)\mp\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)+\boldsymbol{y}_j^t+\boldsymbol{z}_i^{t,e}\right\|^2+n_j\gamma^2\sigma^2 \\
\leq & \left(1+\frac{1}{H-1}+4\gamma^2HL^2\right)\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2+4\gamma^2H\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{z}_i^{t,e}+\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right)-\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right\|^2 \\
& +4\gamma^2Hn_j\mathbb{E}\left\|\boldsymbol{y}_i^t+\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)-\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+8\gamma^2HL^2n_j\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\| \\
& +8\gamma^2Hn_j\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+n_j\gamma^2\sigma^2,
\end{aligned}
$$

where the first inequality comes from Lemma F.1.2 and Assumption 2 and the second inequality follows Lemma F.1.1. Let $\rho_1=\left(1+\frac{1}{H-1}+4\gamma^2HL^2\right)$. As $\gamma\leq\frac{1}{8HL}$, we have $\rho_1^h\leq\rho_1^{H-1}\leq\left(1+\frac{1}{H-1}+\frac{1}{16(H-1)}\right)^{H-1}\leq e_0^{\frac{17}{16}}<3$ and $\sum_{h=0}^{H-1}\rho_1^h\leq 3H$, where e_0 denotes Euler's number. We

thus have

$$
\begin{aligned}
& n_j q_{j,h+1}^{t,e} \\
\leq & \left(\sum_{\tau=0}^{h} \rho_1^{\tau}\right)\left(4\gamma^2 H \sum_{i\in\mathcal{C}_j} \mathbb{E}\left\|\boldsymbol{z}_i^{t,e}+\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right)-\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right\|^2+8\gamma^2 H L^2 n_j \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|\right. \\
& \left.+4\gamma^2 H n_j \mathbb{E}\left\|\boldsymbol{y}_j^t+\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)-\nabla f\left(\bar{\boldsymbol{x}}^t\right)\right\|^2+8\gamma^2 H n_j \mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+n_j\gamma^2\sigma^2\right) \\
\leq & 12\gamma^2 H^2 \sum_{i\in\mathcal{C}_j} \mathbb{E}\left\|\boldsymbol{z}_i^{t,e}+\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e}\right)-\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\right\|^2+12\gamma^2 H^2 n_j \mathbb{E}\left\|\boldsymbol{y}_j^t+\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)-\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \\
& +24\gamma^2 H^2 L^2 n_j \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|+24\gamma^2 H^2 n_j \mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+3Hn_j\gamma^2\sigma^2,
\end{aligned}
$$

where the last inequality follows Assumption 1.

Plugging the derived upper bound of $n_j q_{j,h+1}^{t,e}$ into $Q_t = \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{H}\sum_{h=0}^{H-1} q_{j,h}^{t,e}\right)$ gives rise to Lemma F.2.2.

Part II (Lemma F.2.6): First, $\bar{\boldsymbol{x}}_j^{t,e+1} = \bar{\boldsymbol{x}}_j^{t,e} + \gamma\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{z}_i^{t,e}+\boldsymbol{y}_j^t\right)$. As $\sum_{i\in\mathcal{C}_j}\boldsymbol{z}_i^{t,e}=\mathbf{0}$, we can rewrite $\Theta_j^{t,e}$ as

$$
\Theta_j^{t,e} = \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2 = \mathbb{E}\left\|\gamma\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t\right)\right\|^2.
$$

Next, we establish an upper bound for $\Theta_j^{t,e}$ as follows

$$
\begin{aligned}
\Theta_j^{t,e} \leq & 2\gamma^2\mathbb{E}\left\|\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t\right)\right\|^2+2\gamma^2\mathbb{E}\left\|\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right)\right\|^2 \\
\leq & 2\gamma^2 H\sum_{h=0}^{H-1}\mathbb{E}\left\|\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t\right\|^2+2\frac{1}{n_j^2}\sum_{i\in\mathcal{C}_j}\gamma^2\mathbb{E}\left\|\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right)\right\|^2 \\
\leq & 2\gamma^2 H\sum_{h=0}^{H-1}\mathbb{E}\left\|\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\mp\nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e}\right)\mp\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)\mp\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)+\boldsymbol{y}_j^t\right\|^2+2\gamma^2\frac{H\sigma^2}{n_j} \\
\leq & 8\gamma^2 H^2 L^2\frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e}-\bar{\boldsymbol{x}}_j^{t,e}\right\|^2+8\gamma^2 H^2\mathbb{E}\left\|\boldsymbol{y}_j^t+\nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)-\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \\
& +8\gamma^2 H^2 L^2\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|+8\gamma^2 H^2\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+2\gamma^2\frac{H\sigma^2}{n_j},
\end{aligned}
$$

where the second inequality holds due to Lemmas F.1.1 and F.1.3, the second third inequality comes from Lemmas F.1.4 and Assumption 2, and the last inequality follows Assumption 1. Plugging this upper bound into $\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\Theta_j^{t,e}$ gives rise to Lemma F.2.6.

F.3.3 Proof of Lemma F.2.3

To bound $D_t = \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_t^e\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|^2$, we first rewrite $\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\hat{\boldsymbol{x}}^{t,e+1}\right\|^2$ as follows

$$
\begin{aligned}
& \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\hat{\boldsymbol{x}}^{t,e+1}\right\|^2 \\
= & \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\gamma\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t+\boldsymbol{z}_i^{t,e}\right)\right. \\
& \left.-\hat{\boldsymbol{x}}^{t,e}+\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\gamma\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\boldsymbol{y}_j^t+\boldsymbol{z}_i^{t,e}\right)\right\|^2 \\
= & \mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}-\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\gamma\left(\boldsymbol{y}_j^t+\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\right)+\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{h=0}^{H-1}\gamma\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\right\|^2,
\end{aligned}
$$

where we utilize $\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{y}_j^t=\mathbf{0}$ and $\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{z}_i^{t,e}=\mathbf{0}$. Next, we bound $\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\hat{\boldsymbol{x}}^{t,e+1}\right\|^2$ as follows

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\hat{\boldsymbol{x}}^{t,e+1}\right\|^2\\
\leq&\left(1+\frac{1}{E-1}\right)\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|^2+\gamma^2E\mathbb{E}\left\|\sum_{h=0}^{H-1}\boldsymbol{y}_j^t+\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\right)\right.\\
&\left.\mp\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)-\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)\mp\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2\\
\leq&\left(1+\frac{1}{E-1}\right)\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|^2\\
&+2\gamma^2E\mathbb{E}\left\|\sum_{h=0}^{H-1}\boldsymbol{y}_j^t+\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)-\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2\\
&+2\gamma^2E\mathbb{E}\left\|\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right.\\
&\left.-\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)+\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2,
\end{aligned}
$$

where the first inequality comes from Lemma F.1.2. We thus have

$$
\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1}-\hat{\boldsymbol{x}}^{t,e+1}\right\|^2\\
\leq&\left(1+\frac{1}{E-1}\right)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|^2\\
&+2\gamma^2EH^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{y}_j^t+\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)-\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2\\
&+2\gamma^2E\frac{N-1}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\|\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}\right)-\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2\\
\leq&\left(1+\frac{1}{E-1}\right)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e}-\hat{\boldsymbol{x}}^{t,e}\right\|^2+2\gamma^2EH\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
&+2\gamma^2EH^2\underbrace{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{y}_j^t+\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)-\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2}_{T_3},
\end{aligned}
\tag{31}
$$

where the first equality comes from Lemmas F.1.1 and the second inequality follows Lemmas F.1.3 and F.1.4. Additionally, we bound T_3 as

$$
\begin{aligned}
T_3 =& \frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t \mp \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) \mp \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right) - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right\|^2 \\
\leq& 2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 + 2\frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right) - \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)\right. \\
& \left. - \left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right)\right\|^2 \\
\leq& 2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 + 2\frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right) - \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \\
\leq& 4L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2 + 4L^2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e} - \hat{\boldsymbol{x}}^{t,e}\right\|^2 \\
& + 2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2,
\end{aligned}
$$

where the second inequality follows Lemma F.1.1 and the last inequality follows Assumption 1. Plugging the derived upper bound of T_3 into (31) gives rise to

$$
\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1} - \hat{\boldsymbol{x}}^{t,e+1}\right\|^2 \leq& \left(1 + \frac{1}{E-1} + 8\gamma^2EH^2L^2\right)\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e} - \hat{\boldsymbol{x}}^{t,e}\right\|^2 + 2\gamma^2EH\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 \\
& + 4\gamma^2EH^2\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 + 8\gamma^2EH^2L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2.
\end{aligned}
$$

Let $\rho_2 = 1 + \frac{1}{E-1} + 8\gamma^2EH^2L^2$. As $\gamma \leq \frac{1}{10EHL}$, we have $\rho_2 \leq \rho_2^{E-1} \leq \left(1 + \frac{1}{E-1} + \frac{1}{12(E-1)}\right)^{E-1} \leq e_0^{\frac{13}{12}} < 3$, where e_0 denotes Euler's number. Therefore, we have

$$
\begin{aligned}
& \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1} - \hat{\boldsymbol{x}}^{t,e+1}\right\|^2 \\
\leq& \left(\sum_{\nu=0}^{e}\rho_2^\nu\right)2\gamma^2EH\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 + 4\max\{\rho_2^\nu\}\gamma^2EH^2\sum_{\nu=0}^{e}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,\nu}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,\nu}\right)\right\|^2 \\
& + 8\max\{\rho_2^\nu\}\gamma^2EH^2L^2\sum_{\nu=0}^{e}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\nu} - \bar{\boldsymbol{x}}_j^{t,\nu}\right\|^2 \\
\leq& 6\gamma^2(e+1)EH\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 + 12\gamma^2EH^2\sum_{\nu=0}^{e}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,\nu}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,\nu}\right)\right\|^2 \\
& + 24\gamma^2EH^2L^2\sum_{\nu=0}^{e}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\nu} - \bar{\boldsymbol{x}}_j^{t,\nu}\right\|^2.
\end{aligned} \tag{32}
$$

Hence, for $D_t = \sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_t^e\left\|\bar{\boldsymbol{x}}_j^{t,e} - \hat{\boldsymbol{x}}^{t,e}\right\|^2$, we have

$$
\begin{aligned}
D_t \leq& 3\gamma^2E^3H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 + 12\gamma^2E^2H^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^t + \nabla f_j\left(\hat{\boldsymbol{x}}^{t,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \\
& + 24\gamma^2E^2H^2L^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2.
\end{aligned} \tag{33}
$$

This completes the proof of Lemma F.2.3.

F.3.4 Proof of Lemma F.2.4

For $e=0$, $\boldsymbol{z}_i^{t,0} = -\nabla F_i\left(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}\right) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}\right)$ where $\boldsymbol{x}_{i,0}^{t,0} = \bar{\boldsymbol{x}}_j^{t,0}$, we have

$$\begin{aligned} Z_j^{t,0} =& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|-\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,0}, \xi_{i,0}^{t,0}\right) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,0}, \xi_{i,0}^{t,0}\right) + \nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,0}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,0}\right)\right\|^2 \\ \leq& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,0}\right) - \nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,0}, \xi_{i,0}^{t,0}\right)\right\|^2 \leq \sigma^2, \end{aligned} \tag{34}$$

where the first inequality follows Lemma F.1.1 and the second inequality holds due to Assumption 2.

In addition, when $e \geq 0$, $\boldsymbol{z}_i^{t,e}$ obeys the following iteration

$$\begin{aligned} \boldsymbol{z}_i^{t,e+1} =& \boldsymbol{z}_i^{t,e} + \frac{1}{H\gamma}\left(\boldsymbol{x}_{i,H}^{t,e} - \boldsymbol{x}_{i,0}^{t,e} - \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\boldsymbol{x}_{i,H}^{t,e} - \boldsymbol{x}_{i,0}^{t,e}\right)\right) \\ =& \boldsymbol{z}_i^{t,e} - \frac{1}{H}\sum_{h=0}^{H-1}\left(\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) + \boldsymbol{y}_j^t + \boldsymbol{z}_i^{t,e}\right) - \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) + \boldsymbol{y}_j^t + \boldsymbol{z}_i^{t,e}\right)\right). \end{aligned}$$

As $\sum_{i\in\mathcal{C}_j}\boldsymbol{z}_i^{t,e} = \mathbf{0}$, we thus have

$$\boldsymbol{z}_i^{t,e+1} = \frac{1}{H}\sum_{h=0}^{H-1}\left(\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right)\right).$$

To establish an upper bound for $\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Z_j^{t,e+1}$, we start with bounding $Z_j^{t,e}$ as follows

$$\begin{aligned} Z_j^{t,e+1} =& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\frac{1}{H}\sum_{h=0}^{H-1}\left(\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right)\right) + \nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right)\right\|^2 \\ =& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \frac{1}{H}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) - \nabla f_j\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\frac{1}{H}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right)\right)\right\|^2 \\ \leq& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \frac{1}{H}\sum_{h=0}^{H-1}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right)\right\|^2 \\ =& \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\frac{1}{H}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) \mp \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right)\right\|^2 \\ \leq& 2\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\frac{1}{H}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right)\right\|^2 + 2\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\frac{1}{H}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right)\right)\right\|^2 \\ \leq& 2\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\left(\nabla F_i\left(\bar{\boldsymbol{x}}_j^{t,e+1}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}\right)\right)\right\|^2 + 2\frac{\sigma^2}{H} \\ \leq& 2L^2\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\frac{1}{H}\sum_{h=0}^{H-1}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1} \mp \bar{\boldsymbol{x}}_j^{t,e} - \boldsymbol{x}_{i,h}^{t,e}\right\|^2 + 2\frac{\sigma^2}{H} \\ \leq& 4L^2\frac{1}{H}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2 + 4L^2\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,e+1} - \bar{\boldsymbol{x}}_j^{t,e}\right\|^2 + 2\frac{\sigma^2}{H}, \end{aligned} \tag{35}$$

where the first inequality comes from Lemma F.1.1, the third inequality follows Lemma F.1.4 and Assumption 2, and the fourth inequality follows Assumption 1. Combining this bound with (34), we thus obtain Lemma F.2.4, i.e.,

$$\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Z_j^{t,e} \leq 4L^2Q_t + 4L^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\Theta_j^{t,e} + 2\frac{E}{H}\sigma^2 + \sigma^2. \tag{36}$$

F.3.5 Proof of Lemma F.2.5

Part I ($t \geq 1$)**:** As $\sum_{i\in\mathcal{C}_j} \boldsymbol{z}_i^{t,e} = \mathbf{0}$, the updating rule of $\boldsymbol{y}_j^{t+1}$ can be simplified as

$$\boldsymbol{y}_j^{t+1} = \frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right) - \frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right). \tag{37}$$

We next bound $Y_j^{t+1,e}$ as follows

$$\begin{aligned}
Y_j^{t+1,e} =& \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\boldsymbol{y}_j^{t+1} + \nabla f_j\left(\hat{\boldsymbol{x}}^{t+1,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t+1,e}\right)\right\|^2 \\
=& \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right) - \frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right)\right. \\
&\left. + \nabla f_j\left(\hat{\boldsymbol{x}}^{t+1,e}\right) - \nabla f\left(\hat{\boldsymbol{x}}^{t+1,e}\right)\right\|^2 \\
\leq& \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j\left(\hat{\boldsymbol{x}}^{t+1,e}\right) - \frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right) \mp \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau}\right)\right)\right\|^2 \\
\leq& \frac{2}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j\left(\hat{\boldsymbol{x}}^{t+1,e}\right) - \frac{1}{EH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau}\right)\right\|^2 \\
&+ \frac{2}{NE^2H^2}\sum_{j=1}^{N}\frac{1}{n_j^2}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau},\xi_{i,h}^{t,\tau}\right) - \nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau}\right)\right)\right\|^2 \\
\leq& \frac{2}{NEH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\mathbb{E}\left\|\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau}\right) - \nabla f_j\left(\hat{\boldsymbol{x}}^{t+1,e}\right)\right\|^2 + \frac{2}{EH}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 \\
\leq& \frac{2}{NEH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,\tau}\right) - \nabla F_i\left(\hat{\boldsymbol{x}}^{t+1,e}\right)\right\|^2 + \frac{2}{EH}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 \\
\leq& \frac{2L^2}{NEH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2 + \frac{2}{EH}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2,
\end{aligned} \tag{38}$$

where the first inequality holds due to Lemma F.1.1, the second inequality follows Lemmas F.1.1 and F.1.3, the third inequality follows Lemmas F.1.1 and F.1.4 and Assumption 2, the fourth inequality follows Lemma F.1.1, and the final one comes from Assumption 1.

Given the above inequality, it follows that

$$\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t+1,e} \leq 2L^2\sum_{e=0}^{E-1}\frac{1}{NEH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2 + \frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2. \tag{39}$$

Additionally,

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2 \leq& \mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau} \mp \bar{\boldsymbol{x}}_j^{t,\tau} \mp \hat{\boldsymbol{x}}^{t,\tau} \mp \bar{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2 \\
\leq& 4\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau} - \bar{\boldsymbol{x}}_j^{t,\tau}\right\|^2 + 4\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,\tau} - \hat{\boldsymbol{x}}^{t,\tau}\right\|^2 \\
&+ 4\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau} - \bar{\boldsymbol{x}}^{t+1}\right\|^2 + 4\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2,
\end{aligned} \tag{40}$$

which implies that we need further to bound $\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau} - \bar{\boldsymbol{x}}^{t+1}\right\|^2$ and $\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2$.

Under the framework of MTGC, the virtual global model obeys the following iteration:

$$\hat{\boldsymbol{x}}^{t+1,e} = \bar{\boldsymbol{x}}^{t+1} - \gamma\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t+1,\tau},\xi_{i,h}^{t+1,\tau}\right) + \boldsymbol{z}_i^{t+1,\tau} + \boldsymbol{y}_j^{t+1}\right).$$

As $\sum_{i\in\mathcal{C}_j} \boldsymbol{z}_i^{t+1,\tau} = \mathbf{0}$ and $\sum_{j=1}^N \boldsymbol{y}_j^{t+1} = \mathbf{0}$, the the virtual global model iteration reduces to

$$\hat{\boldsymbol{x}}^{t+1,e} = \bar{\boldsymbol{x}}^{t+1} - \gamma \sum_{j=1}^{N} \frac{1}{n_j} \sum_{i\in\mathcal{C}_j} \sum_{\tau=0}^{e-1} \sum_{h=0}^{H-1} \nabla F_i \left(\boldsymbol{x}_{i,h}^{t+1,\tau}, \xi_{i,h}^{t+1,\tau}\right).$$

For $\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2$, we have

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1} - \hat{\boldsymbol{x}}^{t+1,e}\right\|^2 \\
=&\gamma^2 \mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}, \xi_{i,h}^{t+1,\tau}) \mp \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \right\|^2 \\
\leq& 2\gamma^2 \mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}, \xi_{i,h}^{t+1,\tau}) - \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \right\|^2 \\
&+ 2\gamma^2 \mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \right\|^2 \\
=& 2\gamma^2 \frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j^2}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\| \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}, \xi_{i,h}^{t+1,\tau}) - \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \right\|^2 \\
&+ 2\gamma^2 \mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \right\|^2 \\
\leq& 2\gamma^2 eH \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) \mp \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{\boldsymbol{x}}_j^{t+1,\tau}) \mp \nabla f(\hat{\boldsymbol{x}}^{t+1,\tau}) \right\|^2 + 2\gamma^2 \frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 \\
\leq& 6\gamma^2 eH \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\| \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,h}^{t+1,\tau}) - \nabla f_j(\bar{\boldsymbol{x}}_j^{t+1,\tau}) \right\|^2 + 2\gamma^2 \frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2 \\
&+ 6\gamma^2 eH \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\mathbb{E}\left\| \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\bar{\boldsymbol{x}}_j^{t+1,\tau}) - \nabla f(\hat{\boldsymbol{x}}^{t+1,\tau}) \right\|^2 + 6\gamma^2 eH \sum_{\tau=0}^{e-1}\sum_{h=0}^{H-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t+1,\tau})\right\|^2 \\
\leq& 6\gamma^2 L^2 EH \sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t+1,\tau} - \bar{\boldsymbol{x}}_j^{t+1,\tau}\right\|^2 + 6\gamma^2 L^2 EH^2 \sum_{\tau=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t+1,\tau} - \hat{\boldsymbol{x}}^{t+1,\tau}\right\|^2 \\
&+ 6\gamma^2 EH^2 \sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t+1,\tau})\right\|^2 + 2\gamma^2 \frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2,
\end{aligned} \tag{41}$$

where the second equality holds due to Lemma F.1.3, the second inequality follows Lemma F.1.4 and Assumption 2, the third inequality comes from Lemma F.1.1, the last inequality follows Lemma F.1.1 and Assumption 1.

Similarly, we can bound $\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau} - \bar{\boldsymbol{x}}^{t+1}\right\|^2$ as

$$\begin{aligned}
&\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau} - \bar{\boldsymbol{x}}^{t+1}\right\|^2 \\
\leq& 6\gamma^2 L^2 EH \sum_{\tau'=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau'} - \bar{\boldsymbol{x}}_j^{t,\tau'}\right\|^2 + 6\gamma^2 L^2 EH^2 \sum_{\tau'=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,\tau'} - \hat{\boldsymbol{x}}^{t,\tau'}\right\|^2 \\
&+ 6\gamma^2 EH^2 \sum_{\tau'=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau'})\right\|^2 + 2\gamma^2 \frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned} \tag{42}$$

Given (41) and (42), it follows that

$$\begin{aligned}
&\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau}-\bar{\boldsymbol{x}}^{t+1}\right\|^2+\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1}-\hat{\boldsymbol{x}}^{t+1,e}\right\|^2\\
\leq&6\gamma^2L^2EH^2Q_{t+1}+6\gamma^2L^2EH^2D_{t+1}+6\gamma^2EH^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t+1,\tau})\right\|^2+2\gamma^2\frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
&+6\gamma^2L^2EH^2Q_t+6\gamma^2L^2EH^2D_t+6\gamma^2EH^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2+2\gamma^2\frac{EH}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}\tag{43}$$

Combining (39) and (40), we can obtain

$$\begin{aligned}
&\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t+1,e}\\
\leq&2L^2\sum_{e=0}^{E-1}\frac{1}{NEH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau}-\hat{\boldsymbol{x}}^{t+1,e}\right\|^2+\frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
\leq&8L^2\frac{1}{NH}\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{t,\tau}-\bar{\boldsymbol{x}}_j^{t,\tau}\right\|^2+8L^2\frac{1}{N}\sum_{\tau=0}^{E-1}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{t,\tau}-\hat{\boldsymbol{x}}^{t,\tau}\right\|^2\\
&+8L^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{t,\tau}-\bar{\boldsymbol{x}}^{t+1}\right\|^2+8L^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\bar{\boldsymbol{x}}^{t+1}-\hat{\boldsymbol{x}}^{t+1,e}\right\|^2+\frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}\tag{44}$$

Plugging (43) into (44), we have

$$\begin{aligned}
\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t+1,e}\leq&\ 8L^2Q_t+8L^2D_t+32\gamma^2L^2E^2H\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+\frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
&+48\gamma^2L^4E^2H^2Q_t+48\gamma^2L^4E^2H^2D_t+48\gamma^2L^2E^2H^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\\
&+48\gamma^2L^4E^2H^2Q_{t+1}+48\gamma^2L^4E^2H^2D_{t+1}+48\gamma^2L^2E^2H^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t+1,\tau})\right\|^2.
\end{aligned}$$

Part II ($t=0$)**:** On the other hand, when $t=0$, we have

$$\begin{aligned}
&\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{0,e}\\
=&\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|-\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\bar{\boldsymbol{x}}^0,\xi_{i,0}^{0,0})+\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\bar{\boldsymbol{x}}^0,\xi_{i,0}^{t,0})+\nabla f_j\left(\hat{\boldsymbol{x}}^{0,e}\right)-\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)\right\|^2\\
\leq&\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|-\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\bar{\boldsymbol{x}}^0,\xi_{i,0}^{0,0})+\nabla f_j\left(\bar{\boldsymbol{x}}^0\right)+\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\bar{\boldsymbol{x}}^0,\xi_{i,0}^{t,0})-\nabla f\left(\bar{\boldsymbol{x}}^0\right)\right\|^2\\
&+\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_j\left(\hat{\boldsymbol{x}}^{0,e}\right)-\nabla f_j\left(\bar{\boldsymbol{x}}^0\right)-\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)+\nabla f\left(\bar{\boldsymbol{x}}^0\right)\right\|^2\\
\leq&E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+L^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{0,e}-\bar{\boldsymbol{x}}^0\right\|^2.
\end{aligned}$$

Similar to (41), we have

$$\begin{aligned}
&\sum_{e=0}^{E-1}\mathbb{E}\left\|\hat{\boldsymbol{x}}^{0,e}-\bar{\boldsymbol{x}}^{0}\right\|^2\\
\leq&6\gamma^2L^2E^2H\sum_{\tau=0}^{E-1}\sum_{h=0}^{H-1}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\mathbb{E}\left\|\boldsymbol{x}_{i,h}^{0,\tau}-\bar{\boldsymbol{x}}_j^{0,\tau}\right\|^2+6\gamma^2L^2E^2H^2\sum_{\tau=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left\|\bar{\boldsymbol{x}}_j^{0,\tau}-\hat{\boldsymbol{x}}^{0,\tau}\right\|^2\\
&+6\gamma^2E^2H^2\sum_{\tau=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,\tau})\right\|^2+2\gamma^2\frac{E^2H}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}$$

Combining the above two inequalities gives rise to

$$\begin{aligned}
\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{0,e}\leq&E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+6\gamma^2L^4E^2H^2\left(Q_0+D_0\right)+6\gamma^2L^2E^2H^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,e})\right\|^2\\
&+2\gamma^2L^2\frac{E^2H}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}$$

To this end, we complete the proof of Lemma F.2.5.

F.3.6 Proof of Lemma F.2.7

Part I ($t\geq 1$)**:** Replacing $\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}\Theta_j^{t,e}$ in (15) with an upper bound established in Lemma F.2.6 (i.e., (18)), and then plugging it into (13), we have

$$\begin{aligned}
Q_t\leq&24\gamma^2H^2L^2D_t+12\gamma^2H^2\Bigg\{4L^2Q_t+4L^2\Bigg(8\gamma^2H^2L^2Q_t+8\gamma^2H^2L^2D_t+8\gamma^2H^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t,e}\\
&+8\gamma^2H^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+2\gamma^2EH\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\Bigg)+2\frac{E}{H}\sigma^2+\sigma^2\Bigg\}\\
&+12\gamma^2H^2\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t,e}+24\gamma^2H^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+3EH\gamma^2\sigma^2\\
=&\left(48\gamma^2H^2L^2+384\gamma^4H^4L^4\right)Q_t+\left(24\gamma^2H^2L^2+384\gamma^4H^4L^4\right)D_t\\
&+\left(384\gamma^4H^4L^2+24\gamma^2H^2\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2+\left(384\gamma^4H^4L^2+12\gamma^2H^2\right)\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t,e}\\
&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2.
\end{aligned}\tag{45}$$

Summing (45) with the inequality established in Lemma F.2.3 and utilizing the upper bound of $\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t,e}$ established in Lemma F.2.5, we have

$$\begin{aligned}
Q_t+D_t\leq&\left(48\gamma^2H^2L^2+384\gamma^4H^4L^4+24\gamma^2E^2H^2L^2\right)Q_t+\left(24\gamma^2H^2L^2+384\gamma^4H^4L^4\right)D_t\\
&+\left(384\gamma^4H^4L^2+24\gamma^2H^2\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2\\
&+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\Bigg\{\left(8L^2+48\gamma^2L^4E^2H^2\right)\left(Q_{t-1}+D_{t-1}\right)\\
&+48\gamma^2L^4E^2H^2\left(Q_t+D_t\right)+48\gamma^2L^2E^2H^2\sum_{\tau=0}^{E-1}\left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2+\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right)\\
&+32\gamma^2L^2E^2H\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+\frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\Bigg\}\\
&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2+3\gamma^2E^3H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}$$

Reorganizing the above inequality gives rise to

$$
\begin{aligned}
&Q_t+D_t\\
\leq&\left(48\gamma^2H^2L^2+384\gamma^4H^4L^4+24\gamma^2E^2H^2L^2+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\times 48\gamma^2L^4E^2H^2\right)\\
&\times\left(Q_t+D_t\right)+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\left(8L^2+48\gamma^2L^4E^2H^2\right)\left(Q_{t-1}+D_{t-1}\right)\\
&+\left(384\gamma^4H^4L^2+24\gamma^2H^2+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\times 48\gamma^2L^2E^2H^2\right)\\
&\times\sum_{\tau=0}^{E-1}\left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2+\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right)\\
&+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\left\{32\gamma^2L^2E^2H\frac{1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+\frac{2}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\right\}\\
&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2+3\gamma^2E^3H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}
$$

When $\gamma\leq\frac{1}{7EHL}$, we have $48\gamma^2L^2E^2H^2\leq 1$ and

$$
\begin{aligned}
&\left(1-\left(60\gamma^2H^2L^2+768\gamma^4H^4L^4+36\gamma^2E^2H^2L^2\right)\right)\left(Q_t+D_t\right)\\
\leq&\left(3840\gamma^4H^4L^4+120\gamma^2H^2L^2+120\gamma^2E^2H^2L^2\right)\left(Q_{t-1}+D_{t-1}\right)\\
&+\left(768\gamma^4H^4L^2+36\gamma^2H^2+12\gamma^2E^2H^2\right)\sum_{\tau=0}^{E-1}\left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2+\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right)\\
&+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\frac{3}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2+3\gamma^2E^3H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}
\tag{46}
$$

As $\gamma\leq\frac{1}{33EHL}$, we have $1-\left(60\gamma^2H^2L^2+768\gamma^4H^4L^4+36\gamma^2E^2H^2L^2\right)\geq\frac{2}{3}$ and $3840\gamma^4H^4L^4+120\gamma^2H^2L^2+120\gamma^2E^2H^2L^2\leq\frac{1}{3}$. Hence, (46) can be simplified as

$$
\begin{aligned}
Q_t+D_t\leq&\frac{1}{2}\left(Q_{t-1}+D_{t-1}\right)+\left(1152\gamma^4H^4L^2+72\gamma^2E^2H^2\right)\\
&\times\sum_{\tau=0}^{E-1}\left(\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t-1,\tau})\right\|^2+\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{t,\tau})\right\|^2\right)+\phi(\gamma,\sigma^2),
\end{aligned}
$$

where

$$
\begin{aligned}
\phi(\gamma,\sigma^2)=&\left(1728\gamma^4H^4L^2+216\gamma^2E^2H^2\right)\frac{1}{H}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+144\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\
&+42\gamma^2EH\sigma^2+18\gamma^2H^2\sigma^2+5\gamma^2E^3H\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.
\end{aligned}
$$

Further, we establish an upper bound for $\phi(\gamma,\sigma^2)$ as follows,

$$
\begin{aligned}
\phi(\gamma,\sigma^2)=&1728\gamma^4H^3L^2\frac{1}{\bar{n}}\sigma^2+144\gamma^4\frac{1}{\bar{n}}EH^3L^2\sigma^2+\gamma^2\sigma^2\left\{216\frac{E^2H}{\bar{n}}+42EH+18H^2+5\frac{E^3H}{\bar{n}}\right\}\\
\leq&1728\gamma^4EH^3L^2\frac{1}{\bar{n}}\sigma^2+144\gamma^4\frac{1}{\bar{n}}EH^3L^2\sigma^2+281\gamma^2E^3H\sigma^2\\
\leq&294\gamma^2E^3H\sigma^2,
\end{aligned}
$$

where the last inequality holds due to $\frac{1}{\bar{n}}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\leq 1$ and $\gamma\leq\frac{\sqrt{\bar{n}}E}{12HL}\leq\frac{1}{33EHL}$.

Part II ($t=0$)**:** Summing (45) with the inequality established in Lemma F.2.3 and utilizing the upper bound of $\sum_{e=0}^{E-1}\frac{1}{N}\sum_{j=1}^{N}Y_j^{t,e}$ established in Lemma F.2.5, we have

$$\begin{aligned}\Gamma_0 \leq& \left(48\gamma^2H^2L^2+384\gamma^4H^4L^4+24\gamma^2E^2H^2L^2\right)\Gamma_0+\left(384\gamma^4H^4L^2+24\gamma^2H^2\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)\right\|^2\\&+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\left\{E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+6\gamma^2L^4E^2H^2\Gamma_1\right.\\&\left.+6\gamma^2L^2E^2H^2\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f(\hat{\boldsymbol{x}}^{0,e})\right\|^2+2\gamma^2L^2\frac{E^2H}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\right\}\\&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2+3\gamma^2E^3H\frac{N-1}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.\end{aligned}$$

Reorganizing the above inequality gives rise to

$$\begin{aligned}\Gamma_0 \leq& \left(50\gamma^2H^2L^2+432\gamma^4H^4L^4+26\gamma^2E^2H^2L^2\right)\Gamma_0+\left(26\gamma^2H^2+432\gamma^4H^4L^2+2\gamma^2E^2H^2\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)\right\|^2\\&+\left(384\gamma^4H^4L^2+12\gamma^2H^2+12\gamma^2E^2H^2\right)\left\{E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+2\gamma^2L^2\frac{E^2H}{N^2}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\right\}\\&+96\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+27\gamma^2EH\sigma^2+12\gamma^2H^2\sigma^2+3\gamma^2E^3H\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.\end{aligned}$$

Following the same derivation as in Part I, we obtain

$$\Gamma_0 \leq \left(648\gamma^4H^4L^2+42\gamma^2E^2H^2\right)\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{0,e}\right)\right\|^2+\psi(\gamma,\sigma^2).$$

where

$$\begin{aligned}\psi(\gamma,\sigma^2)=&\left(1044\gamma^4H^4L^2+72\gamma^2E^2H^2\right)E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+144\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\&+42\gamma^2EH\sigma^2+18\gamma^2H^2\sigma^2+5\gamma^2E^3H\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2.\end{aligned}$$

Utilizing $\frac{1}{\bar{n}}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\leq 1$ and $\gamma\leq\frac{\sqrt{\bar{n}}E}{12HL}\leq\frac{1}{33EHL}$, we can further bound $\psi(\gamma,\sigma^2)$ as

$$\begin{aligned}\psi(\gamma,\sigma^2)=&\left(1044\gamma^4H^4L^2+72\gamma^2E^2H^2\right)E\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2+144\gamma^4EH^3L^2\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\&+42\gamma^2EH\sigma^2+18\gamma^2H^2\sigma^2+5\gamma^2E^3H\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sigma^2\\=&1044\gamma^4EH^4L^2\frac{1}{\bar{n}}\sigma^2+144\gamma^4L^2EH^3\frac{1}{\bar{n}}\sigma^2+\gamma^2\sigma^2\left\{72\frac{E^3H^2}{\bar{n}}+42EH+18H^2+5\frac{E^3H}{\bar{n}}\right\}\\\leq&146\gamma^2E^3H^2\sigma^2.\end{aligned}$$

To this end, we complete the proof of Lemma F.2.7.

G Recovering SCAFFOLD's Results

By comparing the communication complexity T required to achieve a ϵ-stationary solution, we can see that our result recovers that of SCAFFOLD when $N=1$ and $E=1$. Specifically, for our MTGC algorithm, to achieve an ϵ error bound, according to Corollary 4.1, we can find a T to satisfy

$$\sqrt{\frac{\mathcal{F}_0L\sigma^2}{\tilde{N}TEH}}\leq\frac{\epsilon}{3},\quad\left(\frac{\mathcal{F}_0L\sigma}{T}\right)^{\frac{2}{3}}\leq\frac{\epsilon}{3},\quad\frac{L\mathcal{F}_0}{T}\leq\frac{\epsilon}{3}.$$

Equivalently, $T \geq \frac{L\sigma^2\mathcal{F}_0}{\tilde{N}EH\epsilon^2}$, $T \geq \frac{L\sigma\mathcal{F}_0}{(\epsilon)^{\frac{3}{2}}}$, and $T \geq \frac{L\mathcal{F}_0}{\epsilon}$. In other words, the MTGC algorithm will have an expected error smaller than ϵ if T satisfies

$$T = \mathcal{O}\left(\frac{L\sigma^2\mathcal{F}_0}{\tilde{N}EH\epsilon^2} + \frac{L\sigma\mathcal{F}_0}{(\epsilon)^{\frac{3}{2}}} + \frac{L\mathcal{F}_0}{\epsilon}\right).$$

According to Theorem II of [18], to achieve the ϵ error bound, the number of global communication rounds SCAFFOLD needs to take can be expressed as

$$T = \mathcal{O}\left(\frac{L\sigma^2\mathcal{F}_0}{n_j H\epsilon^2} + \frac{L\mathcal{F}_0}{\epsilon}\right),$$

where we have converted the notation from [18] to our notation.

We see that the dominating term of the MTGC, $\mathcal{O}\left(\frac{L\sigma^2\mathcal{F}_0}{\tilde{N}EH\epsilon^2}\right)$, recovers that of SCAFFOLD when $N = 1$ (i.e., $\tilde{N} = n_j$) and $E = 1$, which corresponds to the case of a single group with a single (global) aggregator.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper. Additionally, the claims made in the abstract and introduction are supported by the theoretical analysis (see Sec. 4) and experiments (see Sec. 5).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation are included in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced. All assumptions are clearly stated in Sec. 4, and referenced in the statement of our theorem, i.e., Theorem 4.1. The proofs are provided in the supplemental material. Theorems and Lemmas that our proof relies upon are properly referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses all necessary information to reproduce the main experimental results in Sec. 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the datasets used in this work are open-sourced while the information about the models and parameters are clearly reported in Sec. 5. For the convenience of reproducibility, the code is attached to our submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details about the setting of our experiments are reported in Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the standard deviations which are computed based on 3 random trials (see Sec. 5).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information of the computer resources used for this work including the GPU type and memory information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the framework that our work is based on in Sec. 5.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.