On the Data Heterogeneity in Adaptive Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Adaptive federated learning, which benefits from the characteristic of both adaptive optimizer and federated training paradigm, has recently gained lots of attention. Despite achieving outstanding performances on tasks with heavy-tail stochastic gradient noise distributions, adaptive federated learning also suffers from the same data heterogeneity issue as standard federated learning: heterogeneous data distribution across the clients can largely deteriorate the convergence of adaptive federated learning. In this paper, we propose a novel adaptive federated learning framework with local gossip averaging to address this issue. Particularly, we introduce a client re-sampling mechanism and peer-to-peer gossip communications between local clients to mitigate the data heterogeneity without requiring additional gradient computation costs. We theoretically prove the fast convergence for our proposed method under non-convex stochastic settings and empirically demonstrate its superior performances over vanilla adaptive federated learning with client sampling. Moreover, we extend our framework to a communication-efficient variant, in which clients are divided into disjoint clusters determined by their connectivity or communication capabilities. We exclusively perform local gossip averaging within these clusters, leading to an enhancement in network communication efficiency for our proposed method.

1 Introduction

Federated learning (FL) (McMahan et al., 2017) has gained tons of attraction recently with the development of edge computing and edge devices such as IoT devices and smartphones. It enables clients to collaboratively learn a machine learning model by iteratively synchronizing with the central server without sharing their local private data. Standard SGD-based federated learning methods such as FedAvg (McMahan et al., 2017) work by aggregating the local-updated models via stochastic gradient descent. Recently, as the demand for training large-scale models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and ViT (Dosovitskiy et al., 2021), adaptive optimizations such as Adam (Kingma & Ba, 2014) and AMSGrad (Reddi et al., 2018) show their efficiency compared to stochastic gradient descent (SGD). This led to the development of adaptive federated learning methods such as FedAdam (Reddi et al., 2020) and FedAMS (Wang et al., 2022b) which take the advantage of efficient iterative synchronization and stable adaptive optimization methods.

Despite achieving superior model training performances on tasks with heavy-tail stochastic gradient noise distributions, adaptive federated learning still suffers from the same data heterogeneity issue as standard SGD-based federated learning. Specifically, the *statistical heterogeneity* of data distribution across clients can lead to overfitting issues of local models and degradation of global model convergence for adaptive federated learning. This data heterogeneity issue becomes noticeable, especially in practical cases where clients are not able to participate in each local training iteration due to system heterogeneity such as computational capabilities. Despite that various attempts have been made in solving the data heterogeneity issue for standard federated learning (Karimireddy et al., 2020b;a; Khanduri et al., 2021), few studies are tackling this issue in adaptive federated learning, especially for partial participation case where only selected clients have participated in each round of the training process.

In this work, we aim to develop a novel federated learning framework, **A**daptive **F**ederated learning with local **G**ossip **A**veraging (AFGA), that addresses the challenges of statistical heterogeneity in the adaptive federated learning setting. AFGA introduces a client re-sampling strategy and peer-to-peer gossip communication

between clients during local training steps to reduce the dissimilarity between local models, thus tackling the data heterogeneity issue. Note that AFGA does not incur extra communication between the central server and the sampled clients, and it also does not result in extra local gradient computations. Our contributions can be summarized as follows:

- We propose a novel adaptive federated optimization method, which benefits from both the client resampling strategy and decentralized gossip averaging, to mitigate the impact of data heterogeneity in adaptive federated optimization methods.
- We theoretically provide convergence guarantees of our proposed method in the stochastic nonconvex settings with data heterogeneity under partial client participation cases. Specifically, we prove that our proposed method can achieve a faster convergence rate than FedAMSGrad (Wang et al., 2022b) in partial participation settings.
- Moreover, we also extend our framework to a communication-efficient variant, CAFGA, where clients are divided into disjoint clusters and the local gossip communications are only performed within the clusters, thus leading to an overall efficient communication network. We demonstrate that CAFGA can achieve comparable performance and final accuracy to AFGA while simultaneously enhancing overall communication efficiency.
- Extensive experiments on several benchmark datasets demonstrate the proposed AFGA and CAFGA achieving outstanding performance with heterogeneous data in low client participation ratios.

2 Related Work

Federated learning. Federated learning (Konečný et al., 2016) play a critical role in collaboratively training models at edge devices with potential privacy protections. Basic optimization methods for federated learning include SGD-based global optimizer, e.g., FedAvg (McMahan et al., 2017) (a.k.a. Local SGD (Stich, 2018) and its variants (Li et al., 2019a; Yang et al., 2021), adaptive gradient optimization based global optimizer such as FedAdam, FedAdagrad, FedYogi (Reddi et al., 2020), FedAGM (Tong et al., 2020) and FedAMS (Wang et al., 2022b). While these optimization methods for federated learning show their ability on achieving stable results when data are heterogeneously distributed, they rarely study data heterogeneity itself. Recently, several works address the data heterogeneity issue through several aspects. For example, FedProx (Li et al., 2020a) adds a proximate term to align the local model with the global one, and FedDyn (Acar et al., 2021) involves dynamic regularization term for local and global model consistency. FedNova (Wang et al., 2020b) proposes a normalized averaging mechanism that reduces objective inconsistency with heterogeneous data. Moreover, several works study to eliminate the client drift caused by data heterogeneity from the aspect of variance reduction such as Karimireddy et al. (2020b;a); Khanduri et al. (2021); Cutkosky & Orabona (2019). They introduce additional control variables to track and correct the local model shift during local training, but they require extra communication costs for synchronizing these control variables. Besides, FedDC (Gao et al., 2022) involves both dynamic regularization terms and local drift variables for model correction.

Recent studies extend the decentralized training paradigm to federated learning with various adaption. For example, Guo et al. (2021) considered heterogeneous communications for modern communication networks that improve communication efficiency, and hierarchical federated learning algorithms (Liu et al., 2020; Abad et al., 2020; Castiglia et al., 2020) develop frameworks by aggregating client models to edge servers first before synchronizing them to the central server. [1]

3 Proposed Method

3.1 Preliminaries

In federated learning, we aim to minimize the following objective through N local clients:

[1] Due to space limitations, we leave the detailed related work for decentralized learning in Appendix.

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}_i}[f_i(\boldsymbol{x};\xi_i)], \tag{1}$$

where $\boldsymbol{x}$ denotes the model parameters, d denotes the dimension of model parameters $\boldsymbol{x}$, $f_i(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} f_i(\boldsymbol{x},\xi_i)$ is the local loss function corresponding to client i and let $\mathcal{D}_i$ denotes the local data distribution associated with client i. In this work, we focus on the non-convex optimization problem with heterogeneous data distributions, i.e., f_i are non-convex and the local data distribution $\mathcal{D}_i$ are non-i.i.d. distributed among clients. FedAvg (McMahan et al., 2017) is a basic optimization algorithm to solve Eq. 1, with the sequential implementation of local SGD updates and global averaging.

Adaptive federated optimizations. Adaptive federated learning is proposed to incorporate adaptive optimization methods (such as Adam (Kingma & Ba, 2014) and AMSGrad (Reddi et al., 2018)) to global optimizer by replacing the global averaging step in FedAvg. Reddi et al. (2020) summarizes several adaptive federated learning algorithms, and Jin et al. (2022) proposed FedDA, a momentum decoupling adaptive optimization method from the perspective of the dynamics of ODEs. FAFED (Wu et al., 2022) also studied adaptive federated learning but in the context of full participation. FedAMSGrad (Tong et al., 2020; Wang et al., 2022b) considers local SGD updates and global AMSGrad (Reddi et al., 2018) update on the central server. Specifically, at global round $r \in [R]$, the server broadcasts the model $\boldsymbol{x}_r$ to selected clients in the set $\mathcal{S}_r$. The selected client i conducts $\mathcal{I}$ steps of local SGD updates with local learning rate η_l and obtains the local model $\boldsymbol{x}^i_{r,\mathcal{I}}$. Then for the selected client i, it obtains a model difference $\Delta^i_r = \boldsymbol{x}^i_{r,\mathcal{I}} - \boldsymbol{x}_r$ and sends to the server. The server aggregates Δ^i_r then updates the global model $\boldsymbol{x}_{r+1}$ by taking Δ_r as a pseudo gradient for calculating momentum $\boldsymbol{m}_r$ and variance $\boldsymbol{v}_r$ for AMSGrad optimizer, and performs one step AMSGrad update with global learning rate η, i.e.,

$$\begin{aligned}
&\boldsymbol{m}_r = \beta_1\boldsymbol{m}_{r-1} + (1-\beta_1)\Delta_r, \boldsymbol{v}_r = \beta_2\boldsymbol{v}_{r-1} + (1-\beta_2)\Delta_r^2,\\
&\widehat{\boldsymbol{v}}_r = \max\{\widehat{\boldsymbol{v}}_{r-1}, \boldsymbol{v}_r\}, \boldsymbol{x}_{r+1} = \boldsymbol{x}_r + \eta\frac{\boldsymbol{m}_r}{\sqrt{\widehat{\boldsymbol{v}}_r + \epsilon}},
\end{aligned} \tag{2}$$

the server finally obtains model $\boldsymbol{x}_{r+1}$ by global round r. It's worth mentioning that if the set of selected clients $\mathcal{S}_r$ contains all clients, i.e., $|\mathcal{S}_r| = N$, it is known as full participation or without client sampling, and if $|\mathcal{S}_r| = M < N$, we denote it as partial participation or with client sampling.

Heterogeneous and inconsistency. Previous works show that FedAvg suffers from convergence degradation when data is non-i.i.d. distributed on local clients (Karimireddy et al., 2020b;a; Wang et al., 2020b). Several works on adaptive federated learning (Reddi et al., 2020; Wang et al., 2022b) empirically demonstrate that the unbalanced data distribution across clients may lead to worse performance, which implies these adaptive federated methods, unfortunately, do not escape from the convergence degradation as well. Theoretically, it has been shown that when the local data are heterogeneously distributed among clients, FedAMSGrad (Wang et al., 2022b) under partial participation setting is proved with convergence rate $\mathcal{O}(\sqrt{\mathcal{I}}/\sqrt{RM})$ w.r.t. global communication rounds R, local update iterations $\mathcal{I}$ and the number of participated clients M which has a certain gap between the desired rate $\mathcal{O}(1/\sqrt{R\mathcal{I}M})$. Although several works apply variance reduction techniques to show their ability to reduce the effect of data heterogeneity in federated learning (Karimireddy et al., 2020a;b), they are less compatible with the global adaptive optimizer. This motivates us to develop a new framework for mitigating the model inconsistency caused by data heterogeneity in adaptive federated learning.

3.2 AFGA: Adaptive Federated Learning with Local Gossip Averaging

To reduce the inconsistency between local models and achieve a better convergence rate under heterogeneous data in partial participation settings, we proposed a novel Adaptive Federated Learning with Local Gossip Averaging (AFGA) method with peer-to-peer gossip communication and client re-sampling framework. The peer-to-peer gossip communication implies that clients are able to communicate their local model without help from the server. Suppose there are in total N clients, we study the same objective function as other adaptive federated learning methods with a similar global framework but a different local updating process.

Algorithm 1 AFGA: Adaptive Federated Learning with Local Gossip Averaging

Input: initial point $\boldsymbol{x}_1$, local step size η_l and global step size η, optimizer hyperparameter $\beta_1, \beta_2, \epsilon$, doubly stochastic mixing matrix W

1: $\boldsymbol{m}_0 \leftarrow 0$, $\boldsymbol{v}_0 \leftarrow 0$
2: **for** $r = 1$ to R **do**
3: Randomly sample a subset of clients $\mathcal{S}_r$ for collecting local updates in round r
4: Clients Init: clients in $\mathcal{S}_r$ receive $\boldsymbol{x}_r$ from the server and broadcast to all clients with local communications
5: **for** $t = 0, ..., \mathcal{I} - 1$ **do**
6: Randomly **re-sample** a subset of clients $\mathcal{S}_{r,t}$ for gradient computation
7: **for** each client $i \in [N]$ in parallel **do**
8: **if** $i \in \mathcal{S}_{r,t}$ **then**
9: Compute $\boldsymbol{g}^i_{r,t} = \nabla F_i(\boldsymbol{x}^i_{r,t}; \xi^i_{r,t})$
10: $\boldsymbol{x}^i_{r,t'} = \boldsymbol{x}^i_{r,t} - \eta_l \boldsymbol{g}^i_{r,t}$
11: **else**
12: $\boldsymbol{x}^i_{r,t'} = \boldsymbol{x}^i_{r,t}$
13: **end if**
14: **Gossip:** $\boldsymbol{x}^i_{r,t+1} = \sum_{j \in \mathcal{N}^i} (W)_{i,j} \boldsymbol{x}^j_{r,t'}$
15: **end for**
16: **end for**
17: Clients $i \in \mathcal{S}_r$ send $\Delta^i_r = \boldsymbol{x}^i_{r,\mathcal{I}} - \boldsymbol{x}_r$ to the server
18: Aggregate model updates: $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{i \in \mathcal{S}_r} \Delta^i_r$
19: $\boldsymbol{m}_r = \beta_1 \boldsymbol{m}_{r-1} + (1 - \beta_1) \Delta_r$
20: $\boldsymbol{v}_r = \beta_2 \boldsymbol{v}_{r-1} + (1 - \beta_2) \Delta^2_r$
21: $\widehat{\boldsymbol{v}}_r = \max(\widehat{\boldsymbol{v}}_{r-1}, \boldsymbol{v}_r)$ and $\widehat{\boldsymbol{V}}_r = \mathrm{diag}(\widehat{\boldsymbol{v}}_r + \epsilon)$
22: Server update $\boldsymbol{x}_{r+1} = \boldsymbol{x}_r + \eta \frac{\boldsymbol{m}_r}{\sqrt{\widehat{\boldsymbol{v}}_r + \epsilon}}$
23: **end for**

If we take a deeper look at the local update steps (Lines 5-16), the major difference between AFGA and FedAMSGrad is the *re-sample* step (Line 5) and the *gossip communication* step (Line 14), which we will discuss in detail in the following.

Client re-sampling. Note that for each global training round, we have already sampled a subset of participating clients $\mathcal{S}_r$. Normally (e.g., in FedAMSGrad), this will be the fixed chosen subset of clients who actually performs local gradient computations throughout this global training round. Yet for AFGA, we perform client re-sampling at each local iteration to obtain a new subset $\mathcal{S}_{r,t}$ and only the selected clients in the subset $\mathcal{S}_{r,t}$ are active for gradient computation in that local iteration, while the other clients will stay idle. Note that such a design does not incur extra local gradient computations as the size of $\mathcal{S}_{r,t}$ is the same as $\mathcal{S}_r$.

Gossip communications. After finishing the local gradient computations and model update with respect to the selected subset of clients, AFGA introduces a gossip communication step that allow each client to communicate their model weights with their connected neighbors (the selected subset of clients are connected in a graph $\mathcal{G}$ with a corresponding mixing matrix W). This gossip communication step is conducted by all local clients despite being selected in $\mathcal{S}_{r,t}$ or not. While in practice, clients are not required to know the whole mixing matrix W. Instead, client i only needs to maintain the weights corresponding to those it receives from other clients.

It's important to note that AFGA does not necessitate the sampling of every client in specific rounds or iterations. In practical cases, if a client is unavailable, then it would not be sampled during local steps. If we remove the re-sampling step by setting $\mathcal{S}_{r,t} = \mathcal{S}_t$ and remove the gossip communication step by setting $\boldsymbol{x}^i_{r,t+1} = \boldsymbol{x}^j_{r,t'}$, AFGA will reduce to standard FedAMSGrad algorithm.

Remark 3.1. Our AFGA algorithm benefits from both gossip communications and client re-sampling while preserving the stable behavior of adaptive gradient methods. The steps of re-sampling in each local iteration help reduce the impact of data heterogeneity. It allows more clients to be included and participate in local training, which results in training a global model with a more balanced data distribution than without re-sampling. The peer-to-peer gossip communication is inspired by the recent advancement in decentralized optimization (Lian et al., 2017; Koloskova et al., 2020; Chen et al., 2021b), which has shown the ability to reduce the data heterogeneity issues between clients. By involving gossip communications in adaptive federated learning, it enables local models to average with their neighbors, thus preventing over-fitting on local data. The frequent re-sampling and gossip communications make AFGA effectively reduce the inconsistency between local clients, thus accelerating the overall convergence especially when the number of local steps increases. AFGA is also crucial to practical scenarios as it is compatible with low client participation ratios and limited local gradient computation capability while addressing the challenge of statistical heterogeneity in adaptive federated methods.

4 Convergence Analysis

In this section, we provide the theoretical convergence analysis of the proposed AFGA method. Before starting with the main theoretical results, let us first state the following assumptions based on stochastic optimization and adaptive gradient methods. For vector $\boldsymbol{x}$ and matrix A, we let $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$ and $\|A\| = \|A\|_F$. and $\|A\|_2$ represents the spectral norm of A. We denote $\mathbf{1}$ as vector with all elements equal to 1, and $\mathbf{I}$ as the identity matrix, with appropriate dimension.

Assumption 4.1. Each local objective function is L-smooth, i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\left| f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}) - \langle \nabla f_i(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \right| \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall i \in [N]$. This also implies the L-gradient Lipschitz condition, i.e., $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.

Assumption 4.2. The stochastic gradient on each client has a bounded local variance, i.e., $\forall \boldsymbol{x} \in \mathbb{R}^d, i \in [N]$, there is $\mathbb{E}\big[\|\nabla f_i(\boldsymbol{x}, \xi) - \nabla f_i(\boldsymbol{x})\|^2\big] \leq \sigma^2$.

Assumption 4.1 to 4.2 are standard assumptions in centralized and federated non-convex stochastic optimization problems (Kingma & Ba, 2014; Li et al., 2019a; Yang et al., 2021; Reddi et al., 2020).

Assumption 4.3. Each local objective function $f_i(\boldsymbol{x})$ has G-bounded stochastic gradient on ℓ_2, i.e., for all ξ, we have $\|\nabla f_i(\boldsymbol{x}, \xi)\| \leq G, \forall i \in [N]$.

Note that Assumption 4.3 is a standard assumption in non-convex *adaptive* optimization problems for under centralized and federated learning settings (Kingma & Ba, 2014; Reddi et al., 2018; Chen et al., 2020; Reddi et al., 2020; Wang et al., 2022a;b)

Assumption 4.4. The dissimilarity between client's objective function and the global objective function is bounded, i.e., $\forall \boldsymbol{x} \in \mathbb{R}^d$, there is $\frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma_g^2$.

Assumption 4.4 capture the objective dissimilarity in the cluster and between clusters. Similar data heterogeneity assumption, which considers the variance between local clients, is common in federated learning (Reddi et al., 2020; Yang et al., 2021) and decentralized learning (Lian et al., 2017; Li et al., 2019b; Koloskova et al., 2020).

Assumption 4.5 (Spectral gap)**.** For the gossip communications, clients are connected in the graph $\mathcal{G}$, and the corresponding weighting matrix W is a *doubly stochastic matrix*: $W \in [0,1]^{n \times n}$, $W\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W = \mathbf{1}^\top$ and $\text{null}(\mathbf{I} - W) = \text{span}(\mathbf{1})$. We assume the spectral gap ρ of matrix W satisfies: there exists $\rho \in [0, 1)$ such that $\|W - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq \rho$.

Assumption 4.5 is highly related to the gossip communication update process and is usually assumed for decentralized learning framework (Koloskova et al., 2020; Chen et al., 2021b; Guo et al., 2021). For a doubly stochastic matrix W, $\|W - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq 1$ naturally established [2], and the spectral gap $\rho \in [0, 1)$ describes the connectivity of the clients: a smaller spectral gap ρ indicates a denser connectivity between clients.

[2] Theoretical analysis is provided in the Appendix E.

Specifically, $\rho = 0$ indicates that all elements in the matrix W are $\frac{1}{n}$, and this means that each client would be connected and communicated with other clients in the network with a mixing weight of $\frac{1}{n}$. $\rho \to 1$ means the matrix W tends to be elements with either 0 or 1, corresponding to a graph that is nearly disconnected. We assume that there exists $\rho \in [0, 1)$ to satisfy $\|W - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq \rho$ since our proposed method is contributed by gossip communications between clients. Several works (Lian et al., 2017; Li et al., 2019b) alternatively assume the spectral gap ρ of a weighting matrix W as the second largest eigenvalue of a doubly stochastic matrix W, i.e., $\rho = |\lambda_2(W)|$, and this spectral gap holds the same role for revealing the connectivity of the graph.

Theorem 4.6 (Convergence analysis for Algorithm 1)**.** *Under Assumptions 4.1-4.5, if the local learning rate $\eta = \Theta(N\sqrt{\mathcal{I}/M})$ and $\eta_l = \Theta(1/\sqrt{R\mathcal{I}^2})$, and the network spectral gap satisfies $\rho \leq \frac{M}{M+N}$, then the iterates of Algorithm 1 in partial participation scenarios satisfy*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{R\mathcal{I}M}}\left[\Delta f + \frac{G\sqrt{M}}{N\sqrt{\mathcal{I}}}(\sigma_g^2 + \sigma^2)\right]\right) + \mathcal{O}\left(\frac{1}{R}\left[G^2 + L^2\rho^2\sigma_g^2 + \frac{\rho^2\sigma^2}{\mathcal{I}}\right]\right) + \widetilde{\mathcal{O}}\left(\frac{1}{R^{3/2}}\right), \tag{3}$$

where $\Delta f = f_0 - f_*$, $\widetilde{\mathcal{O}}(\cdot)$ hides all the absolute constants and problem dependent constants including $\rho, \sigma, \sigma_g^2, \mathcal{I}, M, N, G$.

Corollary 4.7. *Theorem 4.6 suggests that with sufficient amounts of global communication rounds R, i.e., $R \geq \mathcal{I}M$, our proposed method achieves a convergence rate of $\mathcal{O}\big(\frac{1}{\sqrt{R\mathcal{I}M}}\big)$. This improves the rate $\mathcal{O}\big(\frac{\sqrt{\mathcal{I}}}{\sqrt{RM}}\big)$ of adaptive federated optimization methods under partial client participation (Wang et al., 2022b), which suggests that AFGA can indeed bring accelerated convergence through local gossip communications.*

Remark 4.8. The sufficient amounts of global communication rounds that $R \geq \mathcal{I}M$ is a commonly-used condition to obtain the convergence rate in FL (Wang et al., 2022b; Yang et al., 2021). This condition is also practical as the algorithm usually converges after sufficient global rounds.

Remark 4.9. The impact of data heterogeneity is reflected in the variance σ_g within the convergence rate. From equation 3, the variance σ_g appears in $O\big(\frac{1}{\sqrt{RIM}}\big)$ and smaller order terms w.r.t. R, I, and M for AFGA. The partially participated FedAvg in Yang et al. (2021) and adaptive FL models like FedAMS (Wang et al., 2022b) obtain the rate of $O\big(\frac{\sqrt{I}}{\sqrt{RM}}\big)$, and the dominant $O\big(\frac{\sqrt{I}}{\sqrt{RM}}\big)$ term directly relates to the variance σ_g. This demonstrates improvements over partially participated adaptive FL models like FedAMS (Wang et al., 2022b).

Remark 4.10. The second and third terms in the convergence rate equation 3 contain terms with spectral gap ρ of the gossip communication network. A larger value of ρ corresponds to a sparser network, which results in a larger variance term in the convergence rate, indicating that the dissimilarity variance has not been completely eliminated. Specifically, when $\rho = 0$, i.e., all clients are fully-connected, we have the second term in equation 3 reduce to $\mathcal{O}(\frac{G^2}{R})$, which indicates that the non-dominant variance term of the convergence rate can be further reduced by sufficient local communications.

5 Communication-efficiency: Clustered-clients AFGA and Further Adaptation

CAFGA. While the frequent gossip communications enhance the overall performance of heterogeneous federated learning, it indeed involves extra peer-to-peer communication overhead, which makes our proposed AFGA less efficient in communication especially when clients are densely connected. Recent studies (Guo et al., 2021; Yuan et al., 2020) show that clients can be gathered into neighboring clusters based on locations or network capabilities, in which gossip communications are less expensive than communicating with the whole network. A similar idea of dividing clients into clusters has recently been studied in federated learning (Guo et al., 2021; Malinovsky et al., 2022; Long et al., 2022) and receives a lot of attention. Note that under a cluster-clients design, part of the network clients are grouped in a cluster, and clients within the cluster can be connected through high-bandwidth peer-to-peer communications, leading to an efficient gossip communication network and a relatively smaller spectral gap. This leverages the communication efficiency for applying gossip communications while maintaining comparable performance for our proposed AFGA.

Algorithm 2 CAFGA: Clustered-Client Adaptive Federated Learning with Local Gossip Averaging

Input: initial point $\boldsymbol{x}_1$, local step size η_l and global step size η, , optimizer hyperparameter $\beta_1, \beta_2, \epsilon$, doubly stochastic mixing matrix W

1: $\boldsymbol{m}_0 \leftarrow 0$, $\boldsymbol{v}_0 \leftarrow 0$
2: **for** $r = 1$ to R **do**
3: **for** each cluster $k \in [K]$ in parallel **do**
4: Randomly sample a subset $\mathcal{S}_r^k$ for collecting local updates in round r
5: Init: clients in $\mathcal{S}_r^k$ receive $\boldsymbol{x}_r$ from the server and broadcast $\boldsymbol{x}_r$ to all local neighbors
6: **for** $t = 0, ..., \mathcal{I} - 1$ **do**
7: Randomly **re-sample** a subset of clients $\mathcal{S}_{r,t}^k$ for gradient computation
8: **for** each client $i \in \mathcal{V}_k$ in parallel **do**
9: **if** $i \in \mathcal{S}_{r,t}^k$ **then**
10: Compute $\boldsymbol{g}_{r,t}^i = \nabla F_i(\boldsymbol{x}_{r,t}^i; \xi_{r,t}^i)$
11: $\boldsymbol{x}_{r,t'}^i = \boldsymbol{x}_{r,t}^i - \eta_l \boldsymbol{g}_{r,t}^i$
12: **else**
13: $\boldsymbol{x}_{r,t'}^i = \boldsymbol{x}_{r,t}^i$
14: **end if**
15: **Gossip** Comm: $\boldsymbol{x}_{r,t+1}^i = \sum_{j \in \mathcal{N}_k^i} (W)_{i,j} \boldsymbol{x}_{r,t'}^j$
16: **end for**
17: **end for**
18: Clients $i \in \mathcal{S}_r^k$ send $\Delta_r^i = \boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r$ to the server
19: **end for**
20: Aggregate: $\Delta_r = \frac{1}{K} \sum_{k \in [K]} \frac{1}{|\mathcal{S}_r^k|} \sum_{i \in S_r^k} \Delta_r^i$
21: Server update follows Lines 19-22, Algorithm 1
22: **end for**

Suppose there are still in total N clients, we study the same objective as Eq. equation 1 but we partition the clients into K disjoint clusters where each of them has n clients ($N = Kn$) [3]. We denote $\mathcal{V}_k$ as the set of local clients in cluster $k, k \in [K]$ and denote the neighbors for client $i \in \mathcal{V}_k$ as $\mathcal{N}_k^i$. Similar to Algorithm 1, we denote the weighted matrix of gossip averaging as W_k, and the corresponding spectral gap ρ_k. We then refer $\rho_{\max}$ as the maximum spectral gap among all clusters to represent the overall density in the network.

Algorithm 2 summarizes the proposed **C**lustered-clients paradigm **AFGA** (CAFGA). At the beginning of global round r, the server sample total M clients (for convenience, uniformly sampled m clients in each cluster) for global synchronization. The update rule *inside each cluster* follows the similar local update rule as Algorithm 1 with clients *re-sampling* and *gossip communications* in each local iteration, all clusters perform the training process parallelly. To be specific, at the t-th local iteration in cluster k, clients in the re-sampled subset $\mathcal{S}_{r,t}^k$ are active for gradient computation, while the unselcted clients stay idle. All clients in cluster k then perform a gossip communication step with mixing matrix W_k. The leftover global update process of the clustered-client framework is the same as Algorithm 1 and FedAMSGrad.

We also provide a complete theoretical convergence analysis for the Clustered-clients paradigm of AFGA (CAFGA); due to space limits, we referred interested readers to Appendix D for more details. In a nutshell, our theoretical analysis suggests that the convergence of CAFGA is related to $\rho_{\max}$ which aligns with the convergence rate of Algorithm 1, and implies that more densely connected gossip communications can help reduce the impact of data heterogeneity. Empirically we observe that under the same gossip communication structure (e.g., ring topology), the clustered-clients paradigm obtains performance improvement since grouping the whole network into disjoint clusters making the local clients more densely connected thus have

[3] We omit the clustering process in the algorithm for simplicity. The algorithm is compatible with various clustering methods, including clustering based on locations and network conditions, clustering based on client similarities, and random clustering.

smaller ρ values. The clustered-clients paradigm enables efficient and dense connections for adequate model averaging, which keeps the benefits of further mitigating the effects of data heterogeneity.

Communication-adapted: reduce the communication frequency. Despite AFGA achieves a faster convergence rate, one noticeable drawback is that it requires all clients to stay online to conduct frequent gossip averaging, even if some of the clients do not participate in local training (gradient computation). We want to emphasize that this design is mainly for the ease of theoretical analysis. In practice, we can avoid this issue by enforcing gossip communications only on selected clients in each round[4]. As shown in the next Section, such a communication-adapted AFGA actually enjoys similar model training performances compared to the original AFGA algorithm without requiring dense gossip communications for all clients. Also, note that this communication-adapted can also be applied here to CAFGA for further improving communication efficiency while achieving similar model performances.

6 Experiments

Datasets and models. We conduct experiments on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009) and Shakespeare (Caldas et al., 2018) dataset with various data sampling levels and client participation settings. We evaluate experiments on non-i.i.d. data distributions by a Dirichlet distribution partitioned strategy with parameter $\alpha = 0.6$ similar to Wang et al. (2020a;b). For image classification tasks on CIFAR-10 and CIFAR-100 datasets, we adopt a ConvMixer-256-8 network (Trockman & Kolter, 2022), which shares similar ideas to vision transformer (Dosovitskiy et al., 2021) to use patch embeddings to preserve locality and similarly, and is trained via adaptive gradient methods by default. For the next-character prediction task on Shakespeare, we adopt a 2-layer LSTM network, with 80-dimensional word embedding and 256 hidden units per layer, and follow a dropout layer with dropout rate 0.5.
Baselines and methods. We compare our method with several federated learning and adaptive federated learning baselines including FedAvg (McMahan et al., 2017), FedAdam (Reddi et al., 2020), FedAMSGrad(Wang et al., 2022b), SCAFFOLD (Karimireddy et al., 2020b) FedProx (Li et al., 2020b) and FedDyn (Acar et al., 2021).
Implementation overview. The number of local training iterations $\mathcal{I}$ on each client is set to 24 for experiments on CIFAR-10 and CIFAR-100 datasets, and $\mathcal{I} = 100$ for experiments on the Shakespeare dataset, and the batch size is set to 50 for all experiments by default. We report 500 rounds (refers to #R or # Rounds in tables and figures)for the CIFAR 10 and the Shakespeare datasets and 600 rounds for the CIFAR-100 dataset. For each dataset and setting, the number of local training iterations $\mathcal{I}$ and the total rounds of training #R are fixed across all baseline methods to ensure a fair comparison. For local update, we use the SGD optimizer with a learning rate from $\{0.1, 1\}$ for SGD-based global optimization methods (FedAvg, SCAFFOLD, FedProx, and FedDyn), and use SGD optimizer with a learning rate from $\{1,2,10\}$ for adaptive global optimization methods. For a fair comparison, the local SGD updates apply no momentum and no gradient clipping steps for all methods. We set the global learning rate as 1 for SGD-based global update, and set the global learning rate as 0.01 for global adaptive optimization, FedAdam, FedAMSGrad, and our proposed AFGA. See Appendix C for more details about the experimental setup including datasets, models, and hyperparameter details.

6.1 Main Results

We summarize the performance of our proposed methods and other federated learning baselines in Table 1, Table 2 and 3. Due to space limits, we leave the learning curves and most ablation studies in Appendix C. Our experiments based on two settings, `Setting 1:` 100 clients with 5% participation ratio and `Setting 2:` 50 clients with 10% participation ratio. For the Clustered-clients AGFA (CAFGA), we evenly partition clients into 5 clusters for both settings, i.e., for `Setting 1:` there are 20 clients in each cluster with participate

[4]For example, suppose we train AFGA with a ring topology and select M out of N clients to participate in each round. We can form a new ring topology over the M selected clients and only ask them to communicate over the new ring topology. In this way, the gossip communications only include these active clients while other unselected clients do not need to stay online and participate in the training process.

Table 1: The test accuracy of different methods on CIFAR-10 datasets. `Setting 1`: 100 clients, 5% participation. `Setting 2`: 50 clients, 10% participation. We report the average and the standard derivation over 3 runs with different random seeds.

Method	`Setting 1`		`Setting 2`	
	Acc. & std	R# (78%)	Acc.& std	R# (78%)
FedAvg	75.57 ± 1.10	313	76.85 ± 1.69	180
FedAdam	77.07 ± 0.05	425	78.46 ± 0.19	157
FedAMSGrad	77.53 ± 0.60	388	79.59 ± 0.76	154
SCAFFOLD	76.94 ± 1.17	273	76.46 ± 3.95	146
FedProx	75.63 ± 1.24	500+	76.91 ± 1.39	180
FedDyn	77.68 ± 0.06	297	78.55 ± 0.36	160
AFGA	78.45 ± 0.58	302	80.02 ± 2.00	152
CAFGA	**79.18** ± 1.02	**233**	**82.10** ± 0.67	**112**

Table 2: The test accuracy of different methods on CIFAR-100 datasets. `Setting 1`: 100 clients, 5% participation. `Setting 2`: 50 clients, 10% participation. We report the average and the standard derivation over 3 runs with different random seeds.

Method	`Setting 1`		`Setting 2`	
	Acc. & std	R# (52%)	Acc.& std	R# (52%)
FedAvg	49.88 ± 0.33	600+	51.21 ± 0.23	600+
FedAdam	50.40 ± 0.53	600+	51.99 ± 0.81	375
FedAMSGrad	50.58 ± 0.99	600+	52.15 ± 0.58	278
SCAFFOLD	51.71 ± 0.21	396	**55.12** ± 1.01	235
FedProx	49.75 ± 0.13	600+	51.28 ± 0.10	600+
FedDyn	49.93 ± 0.22	396	51.93 ± 0.52	600+
AFGA	52.08 ± 0.17	436	53.87 ± 0.68	369
CAFGA	**53.08** ± 0.72	**380**	54.41 ± 0.24	**230**

ratio 5%, and for `Setting 2`: there are 10 clients in each cluster with participate ratio 10%. We set ring topology as the default gossip communication topology.

Results on CIFAR-10 and CIFAR-100. Table 1 and Table 2 show the overall performance on training CIFAR-10 and CIFAR-100 datasets with ConvMixer-256-8 model. We observe that AFGA shows improvement upon other baselines, and the proposed CAFGA achieves better performance than AFGA. For `Setting 1` on CIFAR-10, based on the results on three random seeds, AFGA shows an average 1.5% improvement in accuracy compared to SCAFFOLD, nearly 0.8% improvement compared to FedDyn, and nearly 1% improvement compared to FedAMSGrad, The proposed CAFGA, our extension on clustered-clients setting, further improves 0.7% accuracy based over AFGA. For `Setting 2` on CIFAR-10, CAFGA demonstrated around 3.5% increase in accuracy over FedDyn and 2% increase over FedAMSGrad. Note that in both settings, AFGA and CAFGA show their superior performance in achieving desired test accuracy. This demonstrates our proposed AFGA and CAFGA achieve overall better performance than adaptive federated learning methods and other federated learning baselines in both settings.

Table 2 shows that in experiments on CIFAR-100, for `Setting 1`, AFGA and CAFGA obtain higher test accuracy among other baselines including SCAFFOLD and FedDyn. Specifically, CAFGA significantly outperforms all baselines with more than 1.3% increase over SCAFFOLD and more than 2% increase over FedAdam and FedAMSGrad. For `Setting 2`, our proposed AFGA and CAFGA still outperform other federated learning baselines expect for SCAFFOLD.

Table 3: The test accuracy of different methods on Shakespeare datasets. `Setting 1`: 100 clients, 5% participation. `Setting 2`: 50 clients, 10% participation. We report the average and the standard derivation over 3 runs with different random seeds.

Method	`Setting 1`		`Setting 2`	
	Acc.	R#(52%)	Acc.	R# (52%)
FedAvg	48.66 ± 0.01	500+	49.59 ± 0.56	500+
FedAdam	52.35 ± 0.10	391	52.23 ± 0.08	189
FedAMSGrad	52.09 ± 0.23	239	51.96 ± 0.06	220
SCAFFOLD	51.39 ± 0.02	500+	52.76 ± 0.06	166
FedProx	48.55 ± 0.01	252	49.33 ± 0.48	229
FedDyn	51.87 ± 0.03	500+	50.81 ± 0.08	500+
AFGA	53.08 ± 0.08	**121**	53.13 ± 0.04	108
CAFGA	**53.20** ± 0.05	152	**53.57** ± 0.02	**91**

Results on Shakespeare. Table 3 shows the overall performance of training the Shakespeare dataset with a 2-layer LSTM network. For `Setting 1` on Shakespeare, AFGA shows approximately 2.5% improvement in accuracy compared to SCAFFOLD, and 0.5% improvement compared to FedAMSGrad. The proposed CAFGA achieves even higher final accuracy than AFGA and significantly outperforms other baselines. For `Setting 2` on Shakespeare, AFGA addresses increasing in accuracy over FedAMSGrad, while CAFGA outperforms AFGA in final results. In both settings, AFGA or CAFGA show their superior performance in achieving desired test accuracy.

6.2 Communication run-time simulations

Table 4 presents a simulation study as a substitution of the real-world measurement similar to Guo et al. (2021). Consider a limited bandwidth setting where the average time of client-to-client communication cost is 1.8 seconds, and the average time of client-to-server communication cost is 18 seconds. Table 4 suggests that even when considering client-to-client communication costs, our proposed AFGA and CAFGA can still efficiently achieve high accuracy with less overall communication costs. This implies that though our proposed methods incur extra local gossip communications, it helps mitigate the impact of data heterogeneity thus improve the overall performance.

Table 4: The communication time under for CIFAR-10. Setting 1: 100 clients, 5% participation ratio.

Test Accuracy	70%	75%	78%	80%
FedAMSGrad time (h)	**76.0**	128.0	194.0	281.0
AFGA time (h)	91.76	130.98	223.48	250.86
CAFGA time (h)	76.22	**102.86**	**172.42**	**179.82**

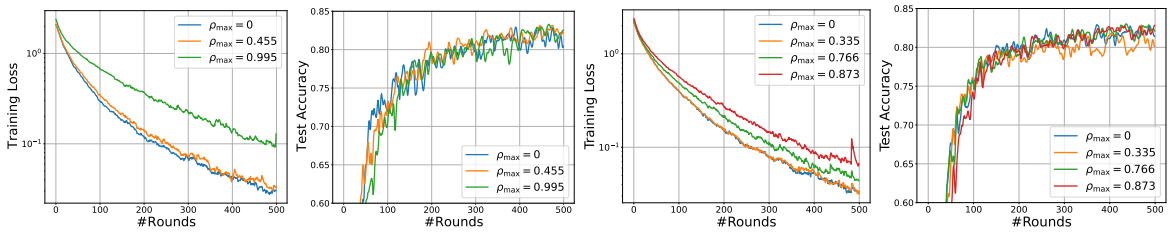
6.3 Ablation Studies

Sensitivity of gossip averaging and client re-sampling. We conduct experiments studying how the individual components, gossip averaging and re-sampling, and the clustered-clients framework contribute to the proposed AFGA and CAFGA. Table 6.3 presents the ablation study of the contribution of individual components, which indicates that the gossip averaging and client re-sampling simultaneously contribute to the accuracy improvements of AFGA. Furthermore, by the results from Table 6.3 (also with the observation of the learning curves in Appendix C), it shows that the clustered-clients paradigm further improves overall accuracy. These results show our intuition of utilizing gossip averaging and client re-sampling can effectively mitigate data heterogeneity, and also address the benefit of the clustered-client framework that consistently helps improve the performance. In addition to the aforementioned ablation studies, we have also conducted further ablation studies to investigate the effect of data heterogeneity, examine different gossip averaging topologies to understand the impact of the spectral gap on model performance, and explore the effects of varying the number of local iterations. Due to constraints on space, we provide detailed ablation studies and results in Appendix.

Table 5: Ablation of components at the last 5 rounds (in total 500 rounds) in training CIFAR-10 on ConvMixer-256-8 model.

Methods (FedAMSGrad)	Acc.
FedAMSGrad Only	79.59 ± 0.76
+Gossip	77.39 ± 0.01
+Gossip + Re-sampling (**AFGA**)	80.02 ± 2.00
+Gossip + Re-sampling + Clustered (**CAFGA**)	**82.10** ± 0.67

Ablation of gossip averaging topology. We also conduct ablation studies on how the gossip averaging topology affects the overall performance in (C)AFGA. Figure 1 shows the ablation study on (a) spectral gap in AFGA and (b) clusters' maximum spectral gap ρ for 5 clusters in CAFGA.

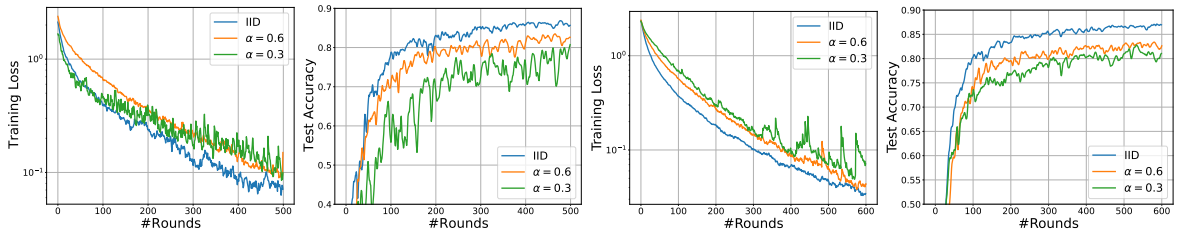


(a) AFGA (b) CAFGA

Figure 1: Ablation study with different heterogeneity degree of (C)AFGA in training CIFAR-10 on ConvMixer-256-8 model.

We follow the `Setting 2` in Table 1, i.e., 50 clients with a participation ratio of 0.1. For AFGA, we compare various of ρ from $\rho = \{0, 0.455, 0.995\}$ calculated by balanced fully-connected, random, and ring typologies correspondingly. We observe that the balanced fully-connected topology with $\rho = 0$ contributes to faster convergence, which aligns with the theoretical result that smaller ρ can help reduce the impact of data heterogeneity. Similar to CAFGA, we compare various of $\rho_{\max}$ from $\rho_{\max} = \{0, 0.335, 0.766, 0.873\}$ calculated by balanced fully-connected, unbalanced fully-connected [5], random, and ring typologies correspondingly. It shows that the fully-connected topology (relatively small $\rho_{\max}$ value) results in faster convergence as well.

Ablation on data heterogeneity. We further conduct experiments to investigate the impact of data



(a) AFGA (b) CAFGA

Figure 2: Ablation study with different heterogeneity degree of AFGA and CAFGA in training CIFAR-10 on ConvMixer-256-8 model.

heterogeneity as theoretically the proposed (C)AFGA show that the convergence rate is highly related to the model dissimilarity. We use Dirichlet(α) distribution for data partitioned in experiments, where α represents the degree of heterogeneity (smaller α implies more heterogeneous data distribution), and we choose α from

[5]This means that clients in the cluster are connected with all their neighbors but with random weighted averaging elements.

$\{0.3, 0.6\}$ together with the i.i.d. data partitioned setting for ablation study. The rest of the experimental setup is the same as `Setting 2` in Table 1. Figure 2 shows the learning curves for different non-i.i.d. degrees. We observe that the data heterogeneity across clients still significantly affects the convergence and generalization performance for our proposed (C)AFGA, as a more balanced data distribution attains faster convergence and higher accuracy.

6.4 Communication-Adapted AFGA and CAFGA

Table 6 shows the test accuracy and the total global round to reach target test accuracy of communication-adapted AFGA and communication-adapted CAFGA which we have discussed in the previous Section. We can observe that both communication-adapted methods achieve similar test accuracy compared with their original version. This suggests that in practice we can still solve the data heterogeneity issue without requiring all clients to participate in gossip communications. Due to space limitations, we left additional results of the CIFAR-100 dataset in Appendix C.

Table 6: The test accuracy (Acc.) and the total global rounds (R#) to reach 78% test accuracy of different methods when training ConvMixer-256-8 model on CIFAR-10 dataset, where (a) denotes the communication-adaptive version. `Setting 1`: 100 clients, 5% participation. `Setting 2`: 50 clients, 10% participation. `Setting 3`: 100 clients, 10% participation. `Setting 4`: 50 clients, 20% participation. To mitigate the effect of randomness and fluctuation of the accuracy, we take the average of the last 5 global rounds to represent final accuracy.

Method	`Setting 1`		`Setting 2`	
	Acc.	R#	Acc.	R#
AFGA	78.80	302	82.61	152
AFGA (a)	76.59	419	81.51	259
Method	`Setting 3`		`Setting 4`	
	Acc.	R#	Acc.	R#
CAFGA	81.56	142	83.15	69
CAFGA (a)	81.15	194	82.99	136

7 Conclusions and Future Works

In this paper, we propose a novel adaptive federated optimization algorithm, AFGA, that addresses data heterogeneity across clients and mitigates local model inconsistency by introducing gossip communications and client re-sampling during local training steps. We present a completed theoretical convergence analysis for the proposed AFGA. We prove that AFGA achieves a faster convergence rate than the previous adaptive federated optimization method for partial participation scenarios with heterogeneous data under non-convex stochastic settings. We extend AFGA to a more communication-efficient clustered-clients paradigm, where clients are divided into disjoint clusters and we only perform local gossip averaging within the clusters. The extended CAFGA algorithm is aimed at reducing the communication overhead introduced by gossip communications while maintaining the benefits of client re-sampling and gossip communications under heterogeneous data. Experiments on several benchmarks and ablation studies backup our theory.

Despite successfully tackling the data heterogeneity issue among clients by introducing gossip communications and client re-sampling, our current proposed methods also have certain limitations. First, extending the theoretical analysis to the communication-adapted versions is challenging and highly non-trivial. Moreover, gossip communications also incur extra challenges if attempting to further apply secure aggregation schemes to our method. Also if not all clients are trusted and there exist malicious clients, the frequent gossip communications between clients may increase the risks of model poisoning or privacy attacks. We leave those new challenges as future works.

References

Mehdi Salehi Heydar Abad, Emre Ozfatura, Deniz Gunduz, and Ozgur Ercetin. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8866–8870. IEEE, 2020.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=B7v4QMR6Z9w`.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Timothy Castiglia, Anirban Das, and Stacy Patterson. Multi-level local sgd: Distributed sgd for heterogeneous hierarchical networks. In *International Conference on Learning Representations*, 2020.

Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Xiangyi Chen, Belhal Karimi, Weijie Zhao, and Ping Li. On the convergence of decentralized adaptive gradient methods. *arXiv preprint arXiv:2109.03194*, 2021a.

Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip sgd with periodic global averaging. In *International Conference on Machine Learning*, pp. 1791–1802. PMLR, 2021b.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10112–10121, 2022.

Yuanxiong Guo, Ying Sun, Rui Hu, and Yanmin Gong. Hybrid local sgd for federated learning with heterogeneous communications. In *International Conference on Learning Representations*, 2021.

Jiayin Jin, Jiaxiang Ren, Yang Zhou, Lingjuan Lyu, Ji Liu, and Dejing Dou. Accelerated federated learning with decoupled adaptive optimization. In *International Conference on Machine Learning*, pp. 10298–10322. PMLR, 2022.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

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

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34: 6050–6061, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

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

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

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

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

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*, 2019b.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, pp. 1–20, 2022.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pp. 7111–7123. PMLR, 2021.

Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced proxskip: Algorithm, theory and application to federated learning. *arXiv preprint arXiv:2207.04338*, 2022.

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

Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *arXiv preprint arXiv:1901.09109*, 2019.

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

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

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

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Yunfei Teng, Wenbo Gao, Francois Chalus, Anna E Choromanska, Donald Goldfarb, and Adrian Weller. Leader stochastic gradient descent for distributed training of deep learning models. *Advances in Neural Information Processing Systems*, 32, 2019.

Qianqian Tong, Guannan Liang, and Jinbo Bi. Effective federated adaptive gradient methods with non-iid decentralized data. *arXiv preprint arXiv:2009.06557*, 2020.

Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

John Nikolas Tsitsiklis. Problems in decentralized decision making and computation. Technical report, Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=BkluqlSFDS`.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *Journal of Machine Learning Research*, 22(213):1–50, 2021. URL `http://jmlr.org/papers/v22/20-147.html`.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020b.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-compressed adaptive gradient method for distributed nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 6292–6320. PMLR, 2022a.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 22802–22838. PMLR, 2022b.

Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster adaptive federated learning. *arXiv preprint arXiv:2212.00974*, 2022.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Jinliang Yuan, Mengwei Xu, Xiao Ma, Ao Zhou, Xuanzhe Liu, and Shangguang Wang. Hierarchical federated learning through lan-wan orchestration. *arXiv preprint arXiv:2010.11612*, 2020.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

A Appendix

B Related Work

Federated learning. Federated learning Konečnỳ et al. (2016) play a critical role in collaboratively training models at edge devices with potential privacy protections. Basic optimization methods for federated learning include SGD-based global optimizer, e.g., FedAvg McMahan et al. (2017) (a.k.a. Local SGD Stich (2018) and its variants Li et al. (2019a); Yang et al. (2021), adaptive gradient optimization based global optimizer such as

FedAdam, FedAdagrad, FedYogi Reddi et al. (2020), FedAGM Tong et al. (2020) and FedAMS Wang et al. (2022b). While these optimization methods for federated learning show their ability on achieving stable results when data are heterogeneously distributed, they rarely study data heterogeneity itself. Recently, several works address the data heterogeneity issue through several aspects. For example, FedProx Li et al. (2020a) adds a proximate term to align the local model with the global one, and FedDyn Acar et al. (2021) involves dynamic regularization term for local and global model consistency. FedNova Wang et al. (2020b) proposes a normalized averaging mechanism that reduces objective inconsistency with heterogeneous data. Moreover, several works study to eliminate the client drift caused by data heterogeneity from the aspect of variance reduction such as Karimireddy et al. (2020b;a); Khanduri et al. (2021); Cutkosky & Orabona (2019). They introduce additional control variables to track and correct the local model shift during local training, but they require extra communication costs for synchronizing these control variables. Besides, FedDC Gao et al. (2022) involves both dynamic regularization terms and local drift variables for model correction.

Decentralized learning and beyond. Decentralized learning studies a distributed machine learning paradigm without a central server. It can been tracked back from gossip averaging techniques Tsitsiklis (1984); Boyd et al. (2006). Decentralized (gossip) SGD algorithms Lian et al. (2017); Li et al. (2019b); Boyd et al. (2006); Tang et al. (2018) are then proposed that consider client-to-client communications after each step of SGD update on the client for decentralized learning. Lu & De Sa (2021) proves a tight lower bound for decentralized training under the non-convex setting. Teng et al. (2019) proposes a leader-distributed SGD algorithm that pulls workers to the currently best-performing model among all models. There are recent studies generalized various distributed SGD algorithms under unified frameworks, where Wang & Joshi (2021) included reducing communication costs and decentralized training in i.i.d. settings, and Koloskova et al. (2020) studied a general network topology-changing gossip SGD methods that summarize several algorithms in distributed and federated learning.

Recent studies extend the decentralized training paradigm to federated learning with various adaption. For example, Guo et al. (2021) considered heterogeneous communications for modern communication networks that improve communication efficiency, and hierarchical federated learning algorithms Liu et al. (2020); Abad et al. (2020); Castiglia et al. (2020) develop frameworks by aggregating client models to edge servers first before synchronizing them to the central server.

C Additional Experiments

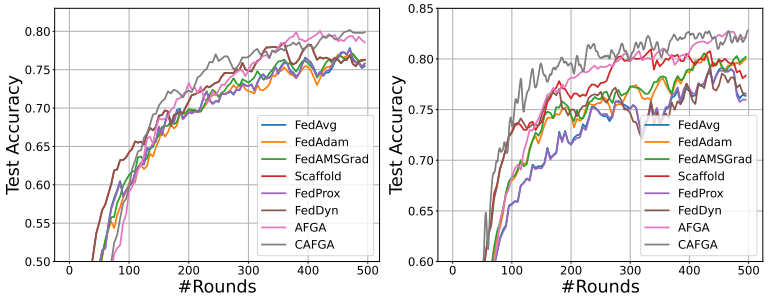
In this section, we present additional empirical results for our proposed algorithm AFGA and CAFGA in training ConvMixer-256-8 model Trockman & Kolter (2022) on CIFAR-10/100 Krizhevsky et al. (2009) datasets, and in training LSTM model on Shakespeare Caldas et al. (2018) dataset. All experiments in this paper are conducted on 4 NVIDIA RTX A6000 GPUs.

C.1 Additional Experimental Results

Additional Experimental Results on CIFAR-10. Figure 3 shows the overall test accuracy curves of experiments on CIFAR-10. It demonstrates that our proposed AFGA and CAFGA achieve overall better performance than adaptive federated learning methods and other federated learning baselines in both settings.

Additional Experimental Results on training ResNet-18 on CIFAR-10. Table 7 presents the empirical result for our proposed AFGA and CAFGA together with several federated learning baselines on training CIFAR-10 with ResNet-18 model. It shows that the proposed CAFGA outperforms other federated learning baselines, achieving a 0.4% improvement over FedAMSGrad and an enhancement of more than 1% compared to other baselines.

Additional Experimental Results on CIFAR-100. Figure 4 shows the empirical result for our proposed AFGA and CAFGA together with several federated learning baselines on training CIFAR-100 ConvMixer-256-8 model. They demonstrate that our proposed AFGA and CAFGA achieve overall better performance than adaptive federated learning methods and other federated learning baselines in both settings.



(a) CIFAR-10 100 clients 5% (b) CIFAR-10 50 clients 10%

Figure 3: The test accuracy for AFGA and CAFGA with several federated learning baselines in training CIFAR-10 data on ConvMixer-256-8 model.

Table 7: The test accuracy of training ResNet-18 model on CIFAR-10 datasets considering 50 clients and 10% participation ratio.

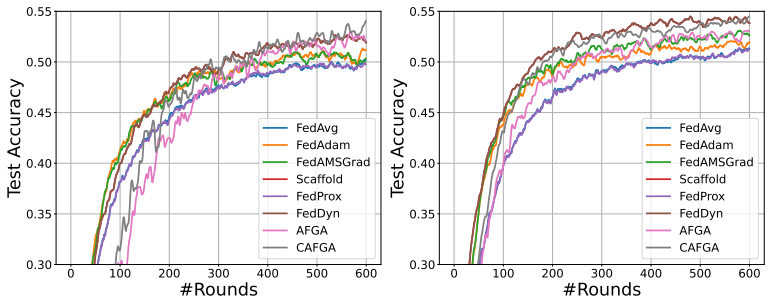
Method	Acc. & std.
FedAvg	70.32 ± 0.44
FedAdam	73.80 ± 0.58
FedAMSGrad	75.59 ± 0.73
SCAFFOLD	74.60 ± 0.67
FedProx	70.26 ± 0.45
FedDyn	74.15 ± 2.23
AFGA	74.72 ± 0.32
CAFGA	**75.96**± 0.67

C.1.1 Ablation Studies and Other Comparisons

Ablation of local iteration. We further study how the local iteration affects the convergence of our proposed CAFGA algorithm. Figure 6 shows the ablation study about local iterations, we compare the number of local iterations $\mathcal{I}$ from $\mathcal{I} = \{12, 24, 48\}$. We observe that larger $\mathcal{I}$ indeed helps accelerate convergence on training loss and helps to obtain a higher test accuracy. This result backs up our theory that the increasing number of local steps would help the overall performance.

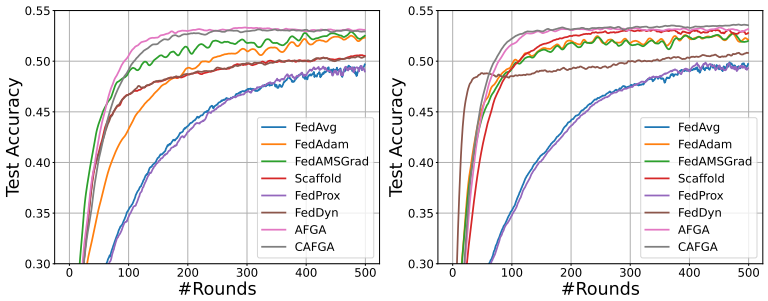
Comparisons to Decentralized Methods. We have briefly discussed decentralized learning in the related work in the main paper, here we provide more discussion about our proposed methods and the decentralized algorithms. Decentralized learning can certainly prevent single point failure without a central server, its performance is not on par with the conventional server-client FL setup, especially when data are heterogeneous distributed. In sharp contrast, the periodic synchronization between server and clients in our proposed method can help align local models for better convergence and ease the data heterogeneity issue in adaptive federated learning, which is mainly focus of this paper. Moreover, the central server setting allow us to easily apply adaptive optimizer for stable performances, while decentralized learning methods are mostly limited to SGD-based update as the adaptive optimizer needs the alignment of gradient update, otherwise suffers from some divergence issue Chen et al. (2021a).

We provide some experimental results comparing our proposed methods with decentralized algorithms including DSGD Lian et al. (2017), DAdam Nazari et al. (2019); Chen et al. (2021a) and PGA Chen et al. (2021b) under the same training settings. The following table shows the comparison result for several decentralized methods and our proposed AFGA and CAFGA. It shows that our proposed AFGA and CAFGA indeed attains better test accuracy results comparing to other decentralized baselines.



(a) CIFAR-100 100 clients 5% (b) CIFAR-100 50 clients 10%

Figure 4: The test accuracy for AFGA and CAFGA with several federated learning baselines in training CIFAR-100 data on ConvMixer-256-8 model.



(a) Shakespeare 100 clients 5% (b) Shakespeare 50 clients 10%

Figure 5: The test accuracy for AFGA and CAFGA with several federated learning baselines in training Shakespeare data on LSTM model.

Table 8: Comparison to decentralized algorithms.

Method	Test Accuracy(%)	Rounds (78%)
DSGD	80.23	**50**
DAdam	70.37	227
PGA	80.93	210
AFGA	82.16	152
CAFGA	**83.03**	112

C.2 Hyper-parameters Details

Hyper-parameter Settings. We conduct detailed hyper-parameter searches to find the best hyper-parameter for each baseline. We grid over the local learning rate $\eta_l \in \{0.001, 0.01, 0.1, 1.0\}$, and the global learning rate $\eta \in \{0.001, 0.01, 0.1, 1.0, 2.0, 5.0, 10.0\}$ for each method. For the global AMSGrad optimizer, we set $\beta_1 = 0.9$, $\beta_1 = 0.99$, and we search the best ϵ from $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}\}$. Table 9 summarizes the hyper-parameter details in our experiments.

Experiments are set up with `Setting 1:` 100 total clients, and `Setting 2`: 50 total clients in the network. For CAFGA, clients are equally divided into 5 clusters. The partial participation ratio is set to $p = 0.05$ for `Setting 1` and $p = 0.1$ for `Setting 2`, and the gossip communication topology is ring topology by default. For each method, we conduct $\mathcal{I} = 24$ iterations of local training with a batch size of 50 by default.

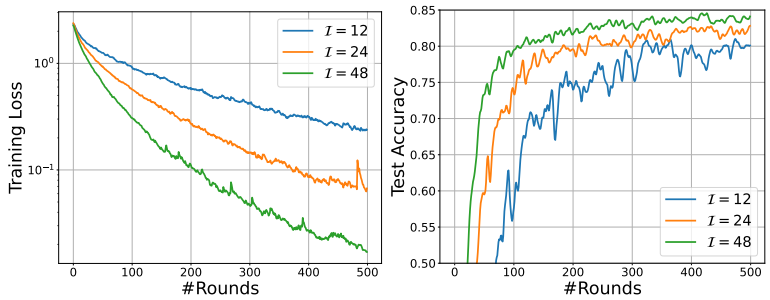


Figure 6: Ablation study with different heterogeneity degree of CAFGA in training CIFAR-10 on ConvMixer-256-8 model.

Table 9: Hyper-parameters details.

Setting 1 (100 clients 5% participation)																
	FedAvg		FedAdam		FedAMS		SCAFFOLD		FedProx		FedDyn		AFGA		CAFGA	
Data	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η
CIFAR-10	0.1	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	2.0	0.01	2.0
CIFAR-100	0.1	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	1.0	0.01	1.0
Shakespeare	1.0	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	10.0	0.01	10.0
Setting 2 (50 clients 10% participation)																
	FedAvg		FedAdam		FedAMS		SCAFFOLD		FedProx		FedDyn		AFGA		CAFGA	
Data&Model	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η	η_l	η
CIFAR-10	0.1	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	2.0	0.01	2.0
CIFAR-100	0.1	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	1.0	0.01	1.0
Shakespeare	1.0	1.0	1.0	0.01	1.0	1.0	0.1	1.0	0.1	1.0	0.1	1.0	0.01	10.0	0.01	10.0

D Convergence Analysis for Clustered-clients framework

We re-state two cluster related assumptions, the assumption of inter-client dissimilarity and the assumption of gossip mixing spectral gap, and the theorem in the following.

Assumption D.1. The dissimilarity between client's objective function and corresponding cluster's objective function is bounded, i.e., for all $\boldsymbol{x}, k \in [K]$, there is $\frac{1}{n}\sum_{i\in\mathcal{V}_k}\|\nabla f_i(\boldsymbol{x}) - \nabla \bar{f}_k(\boldsymbol{x})\|^2 \leq \sigma_k^2$. Similarly, clusters' objective function the global objective has a bounded dissimilarity variance: for $\alpha \geq 1$ and $\sigma_c \geq 0$, there is $\frac{1}{K}\sum_{k\in[K]}\|\nabla \bar{f}_k(\boldsymbol{x})\|^2 \leq \alpha^2\|\nabla f(\boldsymbol{x})\|^2 + \sigma_c^2$.

Assumption D.2 (Intra-cluster spectral gap)**.** Local clients in cluster k are connected in the graph $\mathcal{G}_k$ with weighting matrix W_k satisfies same characteristic as Assumption 4.5. We assume the spectral gap ρ_k satisfies: there exists $\rho_k \in [0, 1)$ such that $\|W_k - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq \rho_k$.

We further denote $\bar{\sigma}_l^2 = \frac{1}{K}\sum_{k=1}^K \sigma_k^2$ as the average dissimilarity between local clients in the same cluster, and denote $\rho_{\max} = \max_{k\in[K]} \rho_k$ as the maximum spectral gap among all K clusters.

Theorem D.3. *Under Assumptions 4.1-4.2, D.1 and D.2, if the local learning rate satisfies specific constraints, and the maximum spectral gap satisfies $\rho_{\max} \leq \frac{m}{m+n}$, then the iterates of Algorithm 1 in partial participation scenarios satisfy*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{R\mathcal{I}m}}\left[[f_0 - f_*] + \left(\widetilde{\sigma}^2 + \frac{\sigma^2}{K}\right)\left(1 + \frac{G\sqrt{m}}{n\sqrt{\mathcal{I}}}\right)\right]\right)$$
$$+ \mathcal{O}\left(\frac{1}{R}\left[G^2 + L^2(\sigma_c^2 + \rho_{\max}^2\bar{\sigma}_L^2) + \frac{\rho_{\max}^2\sigma^2}{\mathcal{I}}\right]\right) + \widetilde{\mathcal{O}}\left(\frac{1}{R^{3/2}}\right), \tag{4}$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides all the absolute constants and problem dependent constants including $\rho, \sigma, \sigma_c^2, \mathcal{I}, M, N$, and additionally we denote $\widetilde{\sigma}^2 = \sigma^2 + \bar{\sigma}_l^2 + \sigma_c^2$ as the variance summation.

E Preliminaries

About weighted matrix:

$$\text{null}(I_n - W_k) = \text{span}(\mathbf{x}|\mathbf{x} \in \mathbb{R}^n : (I_n - W_k)\mathbf{x} = \mathbf{0}) \tag{5}$$

If we have $\text{null}(I_n - W_k) = \text{span}(\mathbf{1})$, that means the following equation holds

$$\begin{pmatrix} 1-w_{11} & -w_{12} & ... & -w_{1n} \\ -w_{21} & 1-w_{22} & ... & -w_{2n} \\ \vdots & & & \\ -w_{n1} & ... & ... & 1-w_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{0} \tag{6}$$

if and only if $x_1 = x_2 = ... = x_n$. Since we assume $w_{ij} \in [0, 1]$, there is a counter-example if $W_k = (W_2, 0; 0, I_{n-2})$, then we have

$$\begin{pmatrix} 1-w_{11} & -w_{12} & ... & 0 \\ -w_{21} & 1-w_{22} & ... & 0 \\ \vdots & & & \\ 0 & & I_{n-2} & \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{0}, \tag{7}$$

then $(x_1, x_2, ..., x_n) = (c_1, c_2, 0, ..., 0)(c_1, c_2 \neq 0)$ can be a solution to the equation.

For the eigenvalues and eigenvectors of matrix $W_k - (1/n)\mathbf{1}\mathbf{1}^T$, we have

$$\begin{aligned} &W_k - (1/n)\mathbf{1}\mathbf{1}^T = W_k - J, \\ &(W_k - J)(W_k - J) = W_k - J, \\ &\lambda \boldsymbol{b} = (W_k - J)\boldsymbol{b} = (W_k - J)(W_k - J)\boldsymbol{b} = (W_k - J)\lambda \boldsymbol{b} = \lambda^2 \boldsymbol{b}, \\ &(\lambda^2 - \lambda)\boldsymbol{b} = 0, \quad \lambda = 1 \text{ or } \lambda = 0. \end{aligned} \tag{8}$$

The eigenvectors of $W_k - J$ are $(1, -1, 0, ..., 0)$, $(1, 0, -1, ..., 0)$,..., $(1, 1, 1, ..., 1)$.

The maximum of $\|W_k - (1/n)\mathbf{1}\mathbf{1}^T\|_2$ is obtained when W_k is equivalent to I_n, and we have $\max \|W_k - (1/n)\mathbf{1}\mathbf{1}^T\|_2 = 1$, which implies $\|W_k - (1/n)\mathbf{1}\mathbf{1}^T\|_2 \leq 1$

Preliminaries for the proof. We define the following auxiliary sequences, w.r.t. $\boldsymbol{x}_{r,t}$, $\boldsymbol{x}_{r,t}^k$. Firstly, we denote the average model on cluster k as

$$\bar{\boldsymbol{x}}_{r,t+1}^k = \bar{\boldsymbol{x}}_{r,t}^k - \eta_l \bar{\boldsymbol{g}}_{r,t}^k, \tag{9}$$

where $\bar{\boldsymbol{g}}_{r,t}^k = \frac{1}{n}\sum_{i \in \mathcal{V}_k} \boldsymbol{g}_{r,t}^i$. We also define the global average model

$$\bar{\boldsymbol{x}}_{r,t+1} = \bar{\boldsymbol{x}}_{r,t} - \eta_l \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}_{r,t}^i. \tag{10}$$

We next define sequences related to model differences, we denote the average model difference on cluster k as $\bar{\Delta}_r^k$, and the average global model difference $\bar{\Delta}_r$ without sampling consideration.

$$\begin{aligned} \bar{\Delta}_r^k &= \frac{1}{n}\sum_{i \in \mathcal{V}_k} \Delta_r^i = \frac{1}{n}\sum_{i \in \mathcal{V}_k} (\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r) = \bar{\boldsymbol{x}}_{r,\mathcal{I}}^k - \boldsymbol{x}_r = \bar{\boldsymbol{x}}_{r,0}^k - \eta_l \sum_{t=0}^{\mathcal{I}-1} \bar{\boldsymbol{g}}_{r,t}^k - \boldsymbol{x}_r = -\eta_l \sum_{t=0}^{\mathcal{I}-1} \bar{\boldsymbol{g}}_{r,t}^k \\ \bar{\Delta}_r &= \frac{1}{K}\sum_{k \in [K]} \frac{1}{n}\sum_{i \in \mathcal{V}_k} (\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r) = \frac{1}{K}\sum_{k \in [K]} \Delta_r^k = -\eta_l \frac{1}{K}\frac{1}{n}\sum_{t=0}^{\mathcal{I}-1}\sum_{k \in [K]}\sum_{i \in \mathcal{V}_k} \boldsymbol{g}_{r,t}^i, \end{aligned} \tag{11}$$

recall the definition of Δ_r

$$\Delta_r = \frac{1}{K}\sum_{k\in[K]}\frac{1}{m}\sum_{i\in S_r^k}\Delta_r^i = \frac{1}{K}\frac{1}{m}\sum_{k\in[K]}\sum_{i\in S_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r, \tag{12}$$

note that for $\bar{\Delta}_r$, we have the following result, which shows that Δ_r in the algorithm is the unbiased estimation of global average model difference $\bar{\Delta}_r$.

$$\mathbb{E}_{S_r}[\Delta_r] = \bar{\Delta}_r = \frac{1}{K}\sum_{k\in[K]}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r. \tag{13}$$

F Proof of Theorem 4.6

Proof of Theorem 4.6. From Section E, we have the following results,

$$\begin{aligned}
\Delta_r &= \frac{1}{Km}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r^k}\Delta_r^i = \frac{1}{Km}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r^k}(\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r) = \frac{1}{K}\sum_{k=1}^{K}\Delta_r^k,\\
\Delta_r^k &= \frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\Delta_r^i = \frac{1}{m}\sum_{i\in\mathcal{S}_r^k}(\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r) = \frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\sum_{t=0}^{\mathcal{I}-1}\sum_{i\in\mathcal{S}_{t,r}^k}\boldsymbol{g}_{r,t}^i.
\end{aligned} \tag{14}$$

Since we have two sampling process: sampling clients for global communication per global round r, and sampling selected clients for local gradient update per local iteration t, hence we state the following auxiliary equations

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_{t,r}^k}[\bar{\boldsymbol{g}}_{t,r}^k] &= \mathbb{E}\left[\frac{1}{n}\sum_{i\in\mathcal{S}_{t,r}^k}\boldsymbol{g}_{t,r}^i\right] = \frac{m}{n^2}\sum_{i=1}^{n}\boldsymbol{g}_{t,r}^i,\\
\mathbb{E}_{\mathcal{S}_r}[\Delta_r] &= \mathbb{E}_{\mathcal{S}_r}\left[\frac{1}{K}\sum_{i=1}^{K}\Delta_r^k\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{\mathcal{S}}[\Delta_r^k] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{\mathcal{S}}\left[\frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\sum_{t=0}^{\mathcal{I}-1}\sum_{i\in\mathcal{S}_{t,r}^k}\frac{1}{n}\boldsymbol{g}_{r,t}^i\right]\\
&= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{\mathcal{S}_r}\left[\frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\sum_{t=0}^{\mathcal{I}-1}\frac{m}{n^2}\sum_{i=1}^{n}\boldsymbol{g}_{t,r}^i\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{\mathcal{S}_r}\left[\frac{1}{n^2}\sum_{i\in\mathcal{S}_r^k}\sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{n}\boldsymbol{g}_{t,r}^i\right]\\
&= \frac{m}{n^2}\frac{1}{K}\sum_{i=1}^{K}\sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{n}\boldsymbol{g}_{t,r}^i = \frac{m}{n}\bar{\Delta}_r.
\end{aligned} \tag{15}$$

Thus, $\frac{n}{m}\Delta_r$ is the unbiased estimate of $\bar{\Delta}_r$. Similar to previous works (Zhou et al., 2018; Chen et al., 2020), we introduce a Lyapunov sequence $\boldsymbol{z}_r$: assume $\boldsymbol{x}_0 = \boldsymbol{x}_1$, for each $r \geq 1$, we have

$$\boldsymbol{z}_r = \boldsymbol{x}_r + \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}) = \frac{1}{1-\beta_1}\boldsymbol{x}_r - \frac{\beta_1}{1-\beta_1}\boldsymbol{x}_{r-1}. \tag{16}$$

For the difference of two adjacent element in sequence $\boldsymbol{z}_r$, we have

$$\begin{aligned}
\boldsymbol{z}_{r+1} - \boldsymbol{z}_r &= \frac{1}{\beta_1}(\boldsymbol{x}_{r+1} - \boldsymbol{x}_r) - \frac{\beta_1}{1-\beta_1}(\boldsymbol{x}_r - \boldsymbol{x}_{r-1})\\
&= \frac{1}{1-\beta_1}(\eta\widehat{\boldsymbol{V}}_r^{-1/2}\boldsymbol{m}_r) - \frac{\beta_1}{1-\beta_1}\eta\widehat{\boldsymbol{V}}_{r-1}^{-1/2}\boldsymbol{m}_{r-1}\\
&= \frac{1}{1-\beta_1}\eta\widehat{\boldsymbol{V}}_r^{-1/2}\Big[\beta_1\boldsymbol{m}_{r-1} + (1-\beta_1)\Delta_r\Big] - \frac{\beta_1}{1-\beta_1}\eta\widehat{\boldsymbol{V}}_{r-1}^{-1/2}\boldsymbol{m}_{r-1}\\
&= \eta\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r - \eta\frac{\beta_1}{1-\beta_1}\left(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\right)\boldsymbol{m}_{r-1}.
\end{aligned}$$

By Assumption 4.1, with the property of L-smoothness, for $r \in [R]$, taking conditional expectation at global round r, we have

$$
\begin{aligned}
\mathbb{E}[f(\boldsymbol{z}_{r+1})] - f(\boldsymbol{z}_r) &\leq \mathbb{E}[\langle \nabla f(\boldsymbol{z}_r), \boldsymbol{z}_{r+1} - \boldsymbol{z}_r \rangle] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{z}_{r+1} - \boldsymbol{z}_r\|^2] \\
&\leq \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{z}_r), \widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r \right\rangle\right] - \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{z}_r), \frac{\beta_1}{1-\beta_1}\left(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\right)\boldsymbol{m}_{r-1} \right\rangle\right] \\
&\quad + \frac{\eta^2 L}{2}\mathbb{E}\left[\left\| \widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r - \frac{\beta_1}{1-\beta_1}\left(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\right)\boldsymbol{m}_{r-1} \right\|^2\right] \\
&= \underbrace{\eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_r), \widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r \right\rangle\right]}_{I} \underbrace{- \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{z}_r), \frac{\beta_1}{1-\beta_1}\left(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\right)\boldsymbol{m}_{r-1} \right\rangle\right]}_{II} \\
&\quad + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}\left[\left\| \widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r - \frac{\beta_1}{1-\beta_1}\left(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\right)\boldsymbol{m}_{r-1} \right\|^2\right]}_{III} \\
&\quad + \underbrace{\eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{z}_r) - \nabla f(\boldsymbol{x}_r), \widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r \right\rangle\right]}_{IV}, \qquad (17)
\end{aligned}
$$

F.1 Bounding I

We have

$$
\begin{aligned}
I &= \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\widehat{\boldsymbol{v}}_r + \epsilon}} \right\rangle\right] \\
&= \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}} \right\rangle\right] + \eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\widehat{\boldsymbol{v}}_r + \epsilon}} - \frac{\Delta_r}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}} \right\rangle\right]. \qquad (18)
\end{aligned}
$$

For the first term in Eq. 18, recall that $\Delta_r = \frac{m}{n}\bar{\Delta}_r = -\frac{m}{n} \cdot \frac{\eta_l}{N}\sum_{i=1}^{N}\sum_{t=0}^{\mathcal{I}-1} \boldsymbol{g}_{r,t}^i$, we have

$$
\begin{aligned}
\eta \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}} \right\rangle\right] &= \frac{\eta m}{n}\mathbb{E}\left[\left\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}}, \bar{\Delta}_r \right\rangle\right] \\
&= -\frac{\eta \eta_l m}{n}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}_{r,t}^i \right\rangle\right] \\
&= -\frac{\eta \eta_l m}{n}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1} + \epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) \right\rangle\right], \qquad (19)
\end{aligned}
$$

where the last equation above holds by the unbiasedness of stochastic gradient, then we have

$$
\begin{aligned}
&-\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\rangle\Big] \\
&= -\frac{1}{2}\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\rangle\Big] \\
&\quad -\frac{1}{2}\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) \pm \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\rangle\Big] \\
&= -\frac{1}{2}\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\rangle\Big] \\
&\quad -\frac{1}{2}\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big(\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) \pm \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big)\Big\rangle\Big]. \qquad (20)
\end{aligned}
$$

Since we have inequalities, $\langle , \boldsymbol{b}\rangle = \|\|^2 + \|\boldsymbol{b}\|^2 - \|-\boldsymbol{b}\|^2$ and $\langle , \boldsymbol{b}\rangle \leq \frac{1}{2}\|\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$, then we have

$$
\begin{aligned}
&-\mathbb{E}\Big[\Big\langle \frac{\nabla f(\boldsymbol{x}_r)}{\sqrt{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}, \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\rangle\Big] \\
&\leq -\frac{1}{4}\mathbb{E}\Big[\Big\|\frac{\nabla f(\boldsymbol{x}_r)}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big\|^2 + \Big\|\frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2 \\
&\quad -\Big\|\frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big(\nabla f(\boldsymbol{x}_r) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big)\Big\|^2\Big] - \frac{1}{4}\mathbb{E}\Big[\Big\|\frac{\nabla f(\boldsymbol{x}_r)}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big\|^2 \\
&\quad +\Big\|\frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2 - \Big\|\frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big(\nabla f(\boldsymbol{x}_r) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big)\Big\|^2\Big] \\
&\quad +\frac{1}{4}\mathbb{E}\Big[\Big\|\frac{\nabla f(\boldsymbol{x}_r)}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big\|^2\Big] + \frac{1}{4}\mathbb{E}\Big[\Big\|\frac{1}{\sqrt[4]{\beta_2 \widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big(\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big)\Big\|^2\Big] \\
&\leq -\frac{1}{4C_0}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \frac{1}{4C_0}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big] - \frac{1}{4C_0}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big] \\
&\quad +\frac{1}{4\sqrt{\epsilon}}\mathbb{E}\Big[\Big\|\nabla f(\boldsymbol{x}_r) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big] + \frac{1}{4\sqrt{\epsilon}}\mathbb{E}\Big[\Big\|\nabla f(\boldsymbol{x}_r) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big] \\
&\quad +\frac{1}{4\sqrt{\epsilon}}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big], \qquad (21)
\end{aligned}
$$

where the second inequality holds by the property of variance $\widehat{\boldsymbol{v}}_r$: $\|\boldsymbol{x}\|^2 C_0^{-1} \leq \|\boldsymbol{x}\|^2(\beta_2\widehat{\boldsymbol{v}}_{r-1}+\epsilon)^{-1/2} \leq \|\boldsymbol{x}\|^2\epsilon^{1/2}$. The last three terms above are highly related to bound the inter-cluster consensus error $\|\boldsymbol{x}_r - \bar{\boldsymbol{x}}_{r,t}^k\|$ and intra-cluster consensus error $\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|$. For the second term in Eq. 18, we have

$$
\begin{aligned}
&\eta\mathbb{E}\Big[\Big\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\widehat{\boldsymbol{v}}_r+\epsilon}} - \frac{\Delta_r}{\sqrt{\beta_2\widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big\rangle\Big] \\
&\leq \eta\|\nabla f(\boldsymbol{x}_r)\|\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_r+\epsilon}} - \frac{1}{\sqrt{\beta_2\widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\Big\| \cdot \|\Delta_r\|\Big] \\
&\leq \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2], \qquad (22)
\end{aligned}
$$

where the second inequality holds by Assumption 4.3, Lemma G.8 and Lemma G.10.

Merging two parts together, we have the following result for bounding I

$$
\begin{aligned}
I &= \eta\mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\beta_2\widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\bigg\rangle\bigg] + \eta\mathbb{E}\bigg[\bigg\langle \nabla f(\boldsymbol{x}_r), \frac{\Delta_r}{\sqrt{\widehat{\boldsymbol{v}}_r+\epsilon}} - \frac{\Delta_r}{\sqrt{\beta_2\widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\bigg\rangle\bigg] \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg[-\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2 - \bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg\{\mathbb{E}\bigg[\bigg\|\nabla f(\boldsymbol{x}_r) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2\bigg] + \mathbb{E}\bigg[\bigg\|\nabla f(\boldsymbol{x}_r) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg]\bigg\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg[-\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2 - \bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg\{\mathbb{E}\bigg[\bigg\|\nabla f(\boldsymbol{x}_r) \pm \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2\bigg] \\
&\quad + \mathbb{E}\bigg[\bigg\|\nabla f(\boldsymbol{x}_r) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] + \mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg]\bigg\} \\
&\quad + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg[-\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2 - \bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg\{\frac{2L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\boldsymbol{x}_r - \bar{\boldsymbol{x}}_{r,t}^k\|^2] + \frac{2L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|^2] \\
&\quad + \frac{L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\boldsymbol{x}_r - \bar{\boldsymbol{x}}_{r,t}^k\|^2] + \frac{L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|^2]\bigg\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg[-\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2 - \bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg\{\frac{3L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\boldsymbol{x}_r - \bar{\boldsymbol{x}}_{r,t}^k\|^2] + \frac{3L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|^2]\bigg\} \\
&\quad + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\sum_{t=0}^{\mathcal{I}-1}\bigg[-\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2 - \bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\bigg[3L^2\mathcal{I}^2C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2+\sigma_c^2) + 3L^2\mathcal{I}^2C_1\rho_{\max}^2H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad + 3L^2\mathcal{I}^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\mathcal{I}^2+H_{\mathcal{I},\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\bigg(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg] + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_r\|^2],
\end{aligned}
\tag{23}
$$

where the last inequality holds by Lemma G.1 and G.2.

F.2 Bounding II

Bounding II mainly follows by the update rule and definition of virtual sequence $\boldsymbol{z}_r$,

$$
\begin{aligned}
II &= -\eta\mathbb{E}\Big[\Big\langle \nabla f(\boldsymbol{z}_r), \frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\rangle\Big] \\
&= -\eta\mathbb{E}\Big[\Big\langle \nabla f(\boldsymbol{z}_r) - \nabla f(\boldsymbol{x}_r) + \nabla f(\boldsymbol{x}_r), \frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\rangle\Big] \\
&\le \eta\mathbb{E}\Big[\|\nabla f(\boldsymbol{x}_r)\|\Big\|\frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|\Big] \\
&\quad + \eta L\mathbb{E}\Big[\|\boldsymbol{z}_r - \boldsymbol{x}_r\|\Big\|\frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|\Big] \\
&= \eta\mathbb{E}\Big[\|\nabla f(\boldsymbol{x}_r)\|\Big\|\frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|\Big] \\
&\quad + \eta^2 L\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{\widehat{\boldsymbol{v}}_{r-1}+\epsilon}}\frac{\beta_1}{1-\beta_1}\boldsymbol{m}_{r-1}\Big\|\Big\|\frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|\Big] \\
&\le \frac{\beta_1}{1-\beta_1}\frac{m}{n}\eta\eta_l\mathcal{I}G^2\mathbb{E}[\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\|_1] + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{m^2}{n^2}L\eta^2\eta_l^2\mathcal{I}^2G^2\epsilon^{-1/2}\mathbb{E}[\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\|_1], \quad (24)
\end{aligned}
$$

where the first iequality holds by Assumption 4.1, and the last one holds by Assumption 4.3 and Lemma G.8 about bounding $\nabla f(\boldsymbol{x}_r)$ and $\boldsymbol{m}_r$.

F.3 Bounding III

For bounding III, use the similar way for bounding II,

$$
\begin{aligned}
III &= \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r + \frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|^2\Big] \\
&\le \eta^2 L\mathbb{E}[\|\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r\|^2] + \eta^2 L\mathbb{E}\Big[\Big\|\frac{\beta_1}{1-\beta_1}(\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2})\boldsymbol{m}_{r-1}\Big\|^2\Big] \\
&\le \frac{\eta^2 L}{\epsilon}\mathbb{E}[\|\Delta_r\|^2] + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{m^2}{n^2}\eta^2\eta_l^2 L\mathcal{I}^2G^2\mathbb{E}[\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\|^2], \quad (25)
\end{aligned}
$$

where the first inequality holds by Cauchy-Schwarz inequality, and the second one follows by Assumption 4.3 and Lemma G.8 about bounding $\nabla f(\boldsymbol{x}_r)$ and $\boldsymbol{m}_r$.

F.4 Bounding IV

Similarly, we bound the last term in equation 17,

$$
\begin{aligned}
IV &= \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{z}_r) - \nabla f(\boldsymbol{x}_r), \eta\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r\Big\rangle\Big] \\
&\le \mathbb{E}\Big[\|\nabla f(\mathbf{z}_r) - \nabla f(\boldsymbol{x}_r)\|\Big\|\eta\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r\Big\|\Big] \\
&\le L\mathbb{E}\Big[\|\mathbf{z}_r - \boldsymbol{x}_r\|\Big\|\eta\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r\Big\|\Big] \\
&\le \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\frac{\beta_1}{1-\beta_1}\widehat{\boldsymbol{V}}_r^{-1/2}\boldsymbol{m}_r\Big\|^2\Big] + \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\widehat{\boldsymbol{V}}_r^{-1/2}\Delta_r\Big\|^2\Big] \\
&\le \frac{\eta^2 L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|\boldsymbol{m}_r\|^2] + \frac{\eta^2 L}{2\epsilon}\mathbb{E}[\|\Delta_r\|^2], \quad (26)
\end{aligned}
$$

where the first inequality holds due to Young's inequality, and the second one follows from Assumption 4.1 and the definition of virtual sequence $\mathbf{z}_r$. By Lemma G.7, we have

$$\sum_{r=1}^{R}\mathbb{E}[\|\boldsymbol{m}_r\|^2] \leq \sum_{r=1}^{R}\mathbb{E}[\|\Delta_r\|^2]. \tag{27}$$

Therefore, the summation of IV term is bounded by

$$\sum_{r=1}^{R} IV \leq \left(\frac{\eta^2 L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2 L}{2\epsilon}\right)\sum_{r=1}^{R}\mathbb{E}[\|\Delta_r\|^2]. \tag{28}$$

Merging I to IV together, we obtain the following result for bounding equation 17,

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{z}_{r+1})] - f(\boldsymbol{z}_1) \\
&\leq \frac{\eta\eta_l}{4C_0}\frac{m}{n}\Bigg\{ -\mathcal{I}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] - \sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Bigg[\Bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Bigg\|^2\Bigg] - \sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Bigg[\Bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Bigg\|^2\Bigg]\Bigg\} \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\frac{m}{n}\Bigg[3L^2\mathcal{I}^2C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2 H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2+\sigma_c^2) + 3L^2\mathcal{I}^2C_1\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad + 3L^2\mathcal{I}^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\mathcal{I}^2 + H_{\mathcal{I},\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad + \left(\eta\frac{\beta_1}{1-\beta_1}\frac{m}{n}\eta_l\mathcal{I}G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}\frac{m^2}{n^2}L\eta_l^2\mathcal{I}^2G^2\epsilon^{-1/2}\right)\mathbb{E}\big[\big\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\big\|_1\big] \\
&\quad + \eta^2 L\frac{\beta_1^2}{(1-\beta_1)^2}\frac{m^2}{n^2}\eta_l^2\mathcal{I}^2G^2\mathbb{E}\big[\big\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}\big\|^2\big] \\
&\quad + \left(\frac{\eta^2 L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2 L}{2\epsilon} + \frac{\eta^2 L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\mathbb{E}[\|\Delta_r\|^2],
\end{aligned}$$

then substituting the bound of $\|\Delta_r\|^2$ in Lemma G.5, and by applying Lemma G.4, then we have

$$
\begin{aligned}
&\frac{n}{m}\sum_{r=1}^{R}\big[\mathbb{E}[f(\boldsymbol{z}_{r+1})]-f(\boldsymbol{z}_1)\big]\\
&\leq -\frac{\eta\eta_l\mathcal{I}}{4C_0}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2]-\frac{\eta\eta_l}{4C_0}\sum_{r=1}^{R}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2\bigg]\\
&\quad-\frac{\eta\eta_l}{4C_0}\sum_{r=1}^{R}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\bigg[\bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg]\\
&\quad+\frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{r=1}^{R}\bigg[3L^2\mathcal{I}^2C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2+\sigma_c^2)+3L^2\mathcal{I}^2C_1\rho_{\max}^2H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2\\
&\quad+3L^2\mathcal{I}^2C_1\eta_l^2\sigma^2\rho_{\max}^2+3L^2C_1\big(\mathcal{I}^2+H_{\mathcal{I},\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\bigg(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg]\\
&\quad+\bigg(\eta\frac{\beta_1}{1-\beta_1}\eta_l\mathcal{I}G^2+\eta^2\frac{\beta_1^2}{(1-\beta_1)^2}\frac{m}{n}L\eta_l^2\mathcal{I}^2G^2\epsilon^{-1/2}\bigg)\sum_{r=1}^{R}\mathbb{E}\big[\big\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2}-\widehat{\boldsymbol{V}}_r^{-1/2}\big\|_1\big]\\
&\quad+\eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\frac{m}{n}\eta_l^2\mathcal{I}^2G^2\sum_{r=1}^{R}\mathbb{E}\big[\big\|\widehat{\boldsymbol{V}}_{r-1}^{-1/2}-\widehat{\boldsymbol{V}}_r^{-1/2}\big\|^2\big]\\
&\quad+\bigg(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{\eta^2L}{2\epsilon}+\frac{\eta^2L}{\epsilon}+\frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\bigg)\sum_{r=1}^{R}\bigg\{\frac{2\eta_l^2\mathcal{I}}{N}\sigma^2+\frac{64\eta_l^2\mathcal{I}(n-m)}{n(n-1)}[\bar{\sigma}_l^2+\sigma_c^2]\\
&\quad+\bigg(\frac{2n(n-m)}{m^2(n-1)}+\frac{64\eta_l^2\mathcal{I}L^2(n-m)}{n(n-1)}+4\eta_l^2(\mathcal{I}+1)^2L^2\bigg)\bigg[\mathcal{I}C_1\eta_l^2H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2)\\
&\quad+\mathcal{I}C_1\eta_l^2H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2+\mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2+\mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\bigg(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg]\\
&\quad+\frac{64\eta_l^2L^2(n-m)}{n(n-1)}\bigg[\mathcal{I}^3C_1\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2+\sigma_c^2)+\mathcal{I}^2C_1\eta_l^2\bigg(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg]\\
&\quad+\frac{16\eta_l^2(m-1)}{(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\bigg[\bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\bigg\|^2\bigg]+\frac{8\eta_l^2\mathcal{I}m}{n}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\bigg[\bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\bigg\|^2\bigg]\bigg\}. \qquad (29)
\end{aligned}
$$

we need the certain constraint on local learning rate

$$
\begin{aligned}
&C_{\beta,\eta}\eta\bigg[\frac{16\eta_l^2}{N^2}\frac{(m-1)}{(n-1)}+\frac{2(\mathcal{I}+2)\eta_l^2m}{N^2n}\bigg]\leq\frac{\eta\eta_l}{4C_0}\\
&\Rightarrow\eta_l\leq\frac{1}{4C_0C_{\beta,\eta}}\bigg[\frac{4}{N^2}\frac{(m-1)}{(n-1)}+\frac{2(\mathcal{I}+2)m}{N^2n}\bigg]^{-1}, \qquad (30)
\end{aligned}
$$

where $C_{\beta,\eta}=\Big(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\Big)=\mathcal{O}(\max\{\eta,1\})$, we further need the requirement of η_l, which is same as the requirement in full participation settings

$$
\begin{aligned}
&\frac{\eta\eta_l}{4\sqrt{\epsilon}}3L^2\mathcal{I}^2C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2H_{\mathcal{I},\rho})\alpha^2+\bigg(\frac{2n(n-m)}{m^2(n-1)}+\frac{32\eta_l^2L^2(n-m)}{n(n-1)}\bigg)C_{\beta,\eta}\eta\mathcal{I}^2C_1\eta_l^2H_{\mathcal{I},\rho}\rho_{\max}^2\alpha^2\\
&\quad+\frac{32\eta_l^2L^2(n-m)}{n(n-1)}C_{\beta,\eta}\eta\mathcal{I}^3C_1\eta_l^2\alpha^2\leq\frac{\eta\eta_l\mathcal{I}}{8C_0},\\
&\Rightarrow\frac{1}{4\sqrt{\epsilon}}3L^2\mathcal{I}C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2H_{\mathcal{I},\rho})\alpha^2+\bigg(\frac{2n(n-m)}{m^2(n-1)}+\frac{32\eta_l^2L^2(n-m)}{n(n-1)}\bigg)C_{\beta,\eta}\mathcal{I}C_1\eta_lH_{\mathcal{I},\rho}\rho_{\max}^2\alpha^2\\
&\quad+\frac{32\eta_l^2L^2(n-m)}{n(n-1)}C_{\beta,\eta}\mathcal{I}^2C_1\eta_l\alpha^2\leq\frac{1}{8C_0}, \qquad (31)
\end{aligned}
$$

$$\Rightarrow \eta_l \leq \frac{\sqrt[4]{\epsilon}}{\sqrt{18L^2C_0C_1\mathcal{I}(\mathcal{I}+\rho_{\max}^2 H_{\mathcal{I},\rho})\alpha^2}}, \tag{32}$$

thus we have

$$\begin{aligned}
&\frac{\eta\eta_l\mathcal{I}}{8C_0R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] \\
&\leq \frac{n}{Rm}\big[\mathbb{E}[f(\boldsymbol{z}_{r+1})]-f(\boldsymbol{z}_1)\big]+\frac{\eta\eta_l}{4R\sqrt{\epsilon}}\sum_{r=1}^{R}\Bigg[3L^2\mathcal{I}^2C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2H_{\mathcal{I},\rho})\sigma_c^2+3L^2\mathcal{I}^2C_1\rho_{\max}^2H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad+3L^2\mathcal{I}^2C_1\eta_l^2\sigma^2\rho_{\max}^2+3L^2C_1\big(\mathcal{I}^2+H_{\mathcal{I},\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg]+\frac{\beta_1}{1-\beta_1}\eta\eta_l\mathcal{I}G^2\frac{d}{R\sqrt{\epsilon}} \\
&\quad+2\frac{\beta_1^2}{(1-\beta_1)^2}\frac{m}{n}\eta^2\eta_l^2L\mathcal{I}^2G^2\frac{d}{\epsilon}+\left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{\eta^2L}{2\epsilon}+\frac{\eta^2L}{\epsilon}+\frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\Bigg\{\frac{2\eta_l^2\mathcal{I}}{N}\sigma^2 \\
&\quad+\left(\frac{2n(n-m)}{m^2(n-1)}+\frac{64\eta_l^2\mathcal{I}L^2(n-m)}{n(n-1)}+4\eta_l^2(\mathcal{I}+1)^2L^2\right)\Bigg[\mathcal{I}C_1\eta_l^2H_{\mathcal{I},\rho}\rho_{\max}^2\sigma_c^2 \\
&\quad+\mathcal{I}C_1\eta_l^2H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2+\mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2+\mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad+64\eta_l^2L^2\frac{(n-m)}{n(n-1)}\left[\mathcal{I}^3C_1\eta_l^2\sigma_c^2+\mathcal{I}^2C_1\eta_l^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\right]+\frac{64\eta_l^2\mathcal{I}(n-m)}{n(n-1)}[\bar{\sigma}_l^2+\sigma_c^2]\Bigg\}
\end{aligned} \tag{33}$$

since there is $H_{\mathcal{I},\rho}=\min\{\frac{1}{1-\rho_{\max}},\mathcal{I}\}\leq\mathcal{I}$ and $\rho_{\max}\leq 1$,

$$\begin{aligned}
&\frac{1}{8C_0R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] \\
&\leq \frac{n}{\eta\eta_lR\mathcal{I}m}\big[\mathbb{E}[f(\boldsymbol{z}_{r+1})]-f(\boldsymbol{z}_1)\big]+\frac{1}{4\sqrt{\epsilon}}\Bigg[C\cdot L^2\eta_l^2\mathcal{I}(\mathcal{I}+H_{\mathcal{I},\rho})\sigma_c^2 \\
&\quad+C\cdot L^2\eta_l^2\mathcal{I}\left(\rho_{\max}^2H_{\mathcal{I},\rho}\bar{\sigma}_l^2+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2+\sigma^2\rho_{\max}^2\left(1+\frac{1}{n}\right)\right)\Bigg]+C_\beta G^2\frac{d}{R\sqrt{\epsilon}}+2C_\beta^2\frac{m}{n}L\eta\eta_l\mathcal{I}G^2\frac{d}{\epsilon} \\
&\quad+\left(\frac{\eta L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{\eta L}{2\epsilon}+\frac{\eta L}{\epsilon}+\frac{\sqrt{1-\beta_2}G}{\epsilon}\right)\Bigg\{\frac{2\eta_l}{N}\sigma^2+\left(\frac{2n(n-m)}{m^2(n-1)}+\frac{64\eta_l^2\mathcal{I}L^2(n-m)}{n(n-1)}+4\eta_l^2(\mathcal{I}+1)^2L^2\right) \\
&\quad\cdot\Bigg[C_1\eta_lH_{\mathcal{I},\rho}\rho_{\max}^2\sigma_c^2+C_1\eta_lH_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2+C_1\eta_l\rho_{\max}^2\sigma^2+C_1\eta_l\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad+\frac{64(n-m)}{n(n-1)}\left[C_1L^2\eta_l^3\mathcal{I}^2\sigma_c^2+C_1L^2\eta_l^3\mathcal{I}\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)+\eta_l\bar{\sigma}_l^2+\eta_l\sigma_c^2\right]\Bigg\},
\end{aligned} \tag{34}$$

where C is a constant irrelevant to parameters and $\rho_{\max}=\max_{k\in[K]}\rho_k$, $H_{\mathcal{I},\rho}=\min\left\{\frac{1}{1-\rho_{\max}},\mathcal{I}\right\}$, $C_\beta=\frac{\beta_1}{1-\beta_1}$ and $\bar{\sigma}_L^2=\frac{1}{K}\sum_{k=1}^{K}\sigma_k^2$. This concludes the proof.

If we further apply the constraint of

$$\frac{n(n-m)}{m^2(n-1)}H_{\mathcal{I},\rho}\rho_{\max}^2\leq\frac{1}{n}, \tag{35}$$

where the condition Eq. 35 implies that the spectral gap $\rho_{\max}$ satisfies

$$\frac{\rho_{\max}^2}{1-\rho_{\max}}\leq\frac{m^2}{n^2}. \tag{36}$$

With the condition of Eq. 36, then we assume there is $\rho_{\max} \leq \frac{m}{m+n}$,

$$\begin{aligned}\frac{n(n-m)}{m^2(n-1)} H_{\mathcal{I},\rho}\rho_{\max}^2 &= \frac{n(n-m)}{m^2(n-1)} \frac{\rho_{\max}^2}{1-\rho_{\max}} \\ &\leq \frac{n(n-m)}{m^2(n-1)} \frac{m^2}{(m+n)n} \\ &\leq \frac{1}{n}. \end{aligned} \tag{37}$$

Also by choosing a constant $\widetilde{C}$, we have

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] \\
&\leq 8C_0\Bigg\{\frac{n[\mathbb{E}[f(\boldsymbol{z}_{r+1})]-f(\boldsymbol{z}_0)]}{\eta\eta_l R\mathcal{I}m} + \frac{1}{R}\left(\frac{C_\beta G^2 d}{\sqrt{\epsilon}} + \frac{2C_\beta^2\eta\eta_l\mathcal{I}LG^2dm}{\epsilon n}\right) \\
&\quad + \frac{CL^2\eta_l^2}{4\sqrt{\epsilon}}\left[\mathcal{I}(\mathcal{I}+H_{\mathcal{I},\rho})\sigma_c^2 + \left(\mathcal{I}\rho_{\max}^2 H_{\mathcal{I},\rho}\bar{\sigma}_l^2 + \frac{n+1}{n}\mathcal{I}\sigma^2\rho_{\max}^2 + \frac{m(n-m)}{n(n-1)}\mathcal{I}^2\bar{\sigma}_l^2\right)\right] \\
&\quad + \frac{\eta_l}{N}\left(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{3\eta L}{\epsilon} + \frac{2\sqrt{1-\beta_2}G}{\epsilon}\right)\sigma^2 \\
&\quad + \frac{C_3\eta_l}{n}\left(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{3\eta L}{\epsilon} + \frac{2\sqrt{1-\beta_2}G}{\epsilon}\right)[\bar{\sigma}_l^2+\sigma_c^2] \\
&\quad + \frac{C_2\eta_l}{n}\left(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{3\eta L}{\epsilon} + \frac{2\sqrt{1-\beta_2}G}{\epsilon}\right)\left[\sigma_c^2+\bar{\sigma}_l^2+\sigma^2+H_{\mathcal{I},\rho}\frac{\sigma^2}{\mathcal{I}^2 n} + \frac{m(n-m)}{n(n-1)}\frac{H_{\mathcal{I},\rho}}{\mathcal{I}}\bar{\sigma}_l^2\right] \\
&\quad + \frac{C_3L^2\eta_l^3\mathcal{I}}{n}\left(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{3\eta L}{\epsilon} + \frac{2\sqrt{1-\beta_2}G}{\epsilon}\right) \\
&\quad \cdot\left[(\mathcal{I}+H_{\mathcal{I},\rho}\rho_{\max}^2)\sigma_c^2 + H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2 + \frac{n+1}{n}\rho_{\max}^2\sigma^2 + \frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right] \\
&\quad + \frac{C_4(\mathcal{I}+1)^2\eta_l^3 m}{n}\left(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{3\eta L}{\epsilon} + \frac{2\sqrt{1-\beta_2}G}{\epsilon}\right)\left[\sigma_c^2+\bar{\sigma}_l^2+\sigma^2+H_{\mathcal{I},\rho}\frac{\sigma^2}{\mathcal{I}^2 n} + \frac{m(n-m)}{n(n-1)}\frac{H_{\mathcal{I},\rho}}{\mathcal{I}}\bar{\sigma}_l^2\right]
\end{aligned} \tag{38}$$

By adopting learning rates $\eta = \Theta(n\sqrt{\frac{\mathcal{I}}{m}})$ and $\eta_l = \Theta(\frac{1}{\sqrt{R\mathcal{I}}})$, then we have

$$\begin{aligned}\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] &= \mathcal{O}\left(\frac{1}{\sqrt{R\mathcal{I}m}}\left[[f_0-f_*] + \left([\sigma^2+\bar{\sigma}_L^2+\sigma_c^2] + \frac{\sigma^2}{K}\right)\left(1+\frac{G\sqrt{m}}{n\sqrt{\mathcal{I}}}\right)\right]\right) \\ &\quad + \mathcal{O}\left(\frac{1}{R}\left[G^2 + L^2(\sigma_c^2+\rho_{\max}^2\bar{\sigma}_L^2) + \frac{\rho_{\max}^2\sigma^2}{\mathcal{I}}\right]\right) + \widetilde{\mathcal{O}}\left(\frac{1}{R^{3/2}}\right).\end{aligned} \tag{39}$$

Based on the theoretical analysis above, we obtain the general convergence bound for (C)AFGA. For AFGA, let $m = M, n = N, \rho_{\max} = \rho$, $\sigma_c = 0$ and $\bar{\sigma}_L^2 = \sigma_g^2$, then we concludes the proof for Theorem 4.6.

□

G Supporting Lemmas

Lemma G.1 (Inter-cluster consensus error)**.** *For local learning rate which satisfying the condition $\eta_l \leq \frac{1}{8\mathcal{I}L}$, denote $C_{\mathcal{I}} = 1 + \frac{3}{2}\cdot\frac{1}{4\mathcal{I}-1}$, recall the definition for $\bar{\boldsymbol{x}}$ in Eq. 9, the inter-cluster model difference after s local*

steps satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t+1}^{k}-\boldsymbol{x}_r\|^2$$
$$\leq C_{\mathcal{I}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+8\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2]+\sigma_c^2)+\eta_l^2\frac{\sigma^2}{n}. \tag{40}$$

Proof. Note that the following proof is similar to Lemma 3 in (Reddi et al., 2020). By definition and auxiliary sequences, we have

$$\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t+1}^{k}-\boldsymbol{x}_r\|^2=\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r-\eta_l\bar{\boldsymbol{g}}_{r,t}^{k}\|^2$$
$$=\mathbb{E}\left\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r-\eta_l\left(\bar{\boldsymbol{g}}_{r,t}^{k}\mp\frac{1}{n}\sum_{i\in\mathcal{S}_{r,t}^{k}}\nabla f_i(\boldsymbol{x}_{r,t}^{i})\mp\frac{m}{n}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^{k})\mp\frac{m}{n}\nabla\bar{f}_k(\boldsymbol{x}_r)\right)\right\|^2$$
$$\leq(1+\gamma)\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+\eta_l^2\mathbb{E}\left\|\bar{\boldsymbol{g}}_{r,t}^{k}-\frac{1}{n}\sum_{i\in\mathcal{S}_{r,t}^{k}}\nabla f_i(\boldsymbol{x}_{r,t}^{i})\right\|^2+4(1+\gamma^{-1})\eta_l^2\mathbb{E}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}_{r,t}^{k}}[\nabla f_i(\boldsymbol{x}_{r,t}^{i})-\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^{k})]\right\|^2$$
$$+4(1+\gamma^{-1})\eta_l^2\mathbb{E}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}_{r,t}^{k}}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^{k})-\frac{m}{n}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^{k})\right\|^2+4(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^{k})-\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2]$$
$$+4(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2]$$
$$\leq(1+\gamma)\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+\eta_l^2\frac{\sigma^2}{n}+4(1+\gamma^{-1})\eta_l^2L^2\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+4(1+\gamma^{-1})\eta_l^2L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{x}_{r,t}^{i}-\bar{\boldsymbol{x}}_{r,t}^{k}\|^2$$
$$+4(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2]+\frac{m(n-m)}{n(n-1)}4(1+\gamma^{-1})\eta_l^2\sigma_k^2$$
$$\leq[(1+\gamma)+4(1+\gamma^{-1})\eta_l^2L^2]\cdot\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+4(1+\gamma^{-1})\eta_l^2L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{x}_{r,t}^{i}-\bar{\boldsymbol{x}}_{r,t}^{k}\|^2$$
$$+\eta_l^2\frac{\sigma^2}{n}+2(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2]+\frac{m(n-m)}{n(n-1)}4(1+\gamma^{-1})\eta_l^2\sigma_k^2, \tag{41}$$

where the first equality holds by Eq. 10. The first inequality holds due to $\boldsymbol{g}_{r,t}^{i}$ is an unbiased estimator of $\nabla f_i(\boldsymbol{x}_{r,t}^{i})$ and Young's inequality. The second inequality holds by Assumption 4.1 and 4.2, also the independency with $\boldsymbol{g}_{r,t}^{i}$ and $\boldsymbol{g}_{r,t}^{j}$ for $i\neq j$.

Averaging Eq. 41 over $k=1,...,K$ clusters, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t+1}^{k}-\boldsymbol{x}_r\|^2$$
$$\leq[(1+\gamma)+4(1+\gamma^{-1})\eta_l^2L^2]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+4(1+\gamma^{-1})\eta_l^2L^2\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{x}_{r,t}^{i}-\bar{\boldsymbol{x}}_{r,t}^{k}\|^2$$
$$+4(1+\gamma^{-1})\eta_l^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2]+\eta_l^2\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}4(1+\gamma^{-1})\eta_l^2\bar{\sigma}_l^2$$
$$\leq[(1+\gamma)+4(1+\gamma^{-1})\eta_l^2L^2]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^{k}-\boldsymbol{x}_r\|^2+4(1+\gamma^{-1})\eta_l^2L^2\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{x}_{r,t}^{i}-\bar{\boldsymbol{x}}_{r,t}^{k}\|^2$$
$$+4(1+\gamma^{-1})\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2]+\sigma_c^2)+\eta_l^2\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}4(1+\gamma^{-1})\eta_l^2\bar{\sigma}_l^2, \tag{42}$$

where the second inequality holds by Assumption 4.4. Choosing $\gamma = \frac{1}{4\mathcal{I}-1}$ with the condition of $\eta_l \leq \frac{1}{8\mathcal{I}L}$, we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t+1}^k - \boldsymbol{x}_r\|^2 \\
&\leq \left(1 + \frac{1}{4\mathcal{I}-1} + \frac{1}{2(4\mathcal{I}-1)}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + 16\mathcal{I}\eta_l^2 L^2 \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}\|\boldsymbol{x}_{r,t}^i - \bar{\boldsymbol{x}}_{r,t}^k\|^2 \\
&\quad + 16\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] + \sigma_c^2) + \eta_l^2\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}16\mathcal{I}\eta_l^2\bar{\sigma}_l^2 \\
&= C_{\mathcal{I}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + 16\mathcal{I}\eta_l^2 L^2 \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}\|\boldsymbol{x}_{r,t}^i - \bar{\boldsymbol{x}}_{r,t}^k\|^2 + 16\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] + \sigma_c^2) \\
&\quad + \eta_l^2\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}16\mathcal{I}\eta_l^2\bar{\sigma}_l^2,
\end{aligned} \tag{43}
$$

where $C_{\mathcal{I}} = 1 + \frac{3}{2}\cdot\frac{1}{4\mathcal{I}-1}$. This concludes the proof. □

G.1 Lemma for intra-cluster consensus error

Lemma G.2. *The intra-cluster consensus error $\sum_{i=1}^n \|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|^2$, also known as $\|X_{r,t}^{k,\perp}\|_F^2$, has the following upper bound,*

$$
\begin{aligned}
&\frac{1}{N}\mathbb{E}\sum_{k=1}^{K}\|X_{r,t+1}^{k,\perp}\|^2 \\
&\leq \left(\max_{k\in[K]}\rho_k^2(1+\zeta_k^{-1}) + \eta_l^2\cdot 4L^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\right)\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 \\
&\quad + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4L^2\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2,
\end{aligned} \tag{44}
$$

where ζ_k is some constant related to the Young's inequality, and it could be uniformly chosen for all $k = 1, ..., K$.

Proof. By definition we have $X_k = (\boldsymbol{x}_{r,t}^1, ..., \boldsymbol{x}_{r,t}^n)'$ and $X_{r,t}^{k,\perp} = X_{r,t}^k(I_n - J)$, where $J = \frac{1}{n}\mathbf{1}_n\cdot\mathbf{1}_n'$. Thus we have

$$
\begin{aligned}
\sum_{i=1}^{n}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_{r,t}^i\|^2 &= \|(\boldsymbol{x}_{r,t}^1, ..., \boldsymbol{x}_{r,t}^n)(I_n - J)\cdot I_n\cdot(I_n - J)(\boldsymbol{x}_{r,t}^1, ..., \boldsymbol{x}_{r,t}^n)'\|_F \\
&= \|X_{r,t}^k{}'(I_n - J)\cdot(I_n - J)X_{r,t}^k\|_F \\
&= \|X_{r,t}^{k,\perp\prime}\cdot X_{r,t}^{k,\perp}\|_F \\
&= \|X_{r,t}^{k,\perp}\|_F^2,
\end{aligned} \tag{45}
$$

Recall the update rule of AFGA, there is $X_{r,t+1}^{k,\perp} = (W_k - J)(X_{r,t}^{k,\perp} - \eta_l G_{r,t}^k)$, then we have

$$
\begin{aligned}
&\mathbb{E}\|X_{r,t+1}^{k,\perp}\|^2 = \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{r,t}^{k,\perp} - \eta_l G_{r,t}^k)\|^2 | \mathcal{F}_{r,t-1})) \\
&= \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{r,t}^{k,\perp} - \eta_l \nabla F(X_{r,t}^k) + \eta_l \nabla F(X_{r,t}^k) - \eta_l G_{r,t}^k)\|^2 | \mathcal{F}_{r,t-1})) \\
&= \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{r,t}^{k,\perp} - \eta_l \nabla F(X_{r,t}^k))\|^2 | \mathcal{F}_{r,t-1})) \\
&\quad + \eta_l^2 \mathbb{E}(\mathbb{E}(\|(W_k - J)(\nabla F(X_{r,t}^k) - G_{r,t}^k)\|^2 | \mathcal{F}_{r,t-1})) \\
&\leq \mathbb{E}(\|(W_k - J)(X_{r,t}^{k,\perp} - \eta_l \nabla F(X_{r,t}^k))\|^2) + \eta_l^2 \rho_k^2 n \sigma^2 \\
&\leq \rho_k^2 (1 + \zeta_k^{-1}) \cdot \mathbb{E}\|X_{r,t}^{k,\perp}\|^2 + \rho_k^2 (1 + \zeta_k) \eta_l^2 \mathbb{E}\|\nabla F(X_{r,t}^k))\|^2 + \eta_l^2 \rho_k^2 n \sigma^2,
\end{aligned} \tag{46}
$$

where the $\nabla F_k(X^k) \in \mathbb{R}^{n \times d}$ is associated to cluster k by stacking $\nabla f_i(\boldsymbol{x}^i)$ for $i \in \mathcal{V}_k$ row-wise. The third equality is due to the unbiasedness of stochastic gradient. The first inequality holds by Assumption 4.2 and $\|\nabla F(X_{r,t}^k) - G_{r,t}^k\|_F = \sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}_{r,t}^i) - \boldsymbol{g}_{r,t}^i\|^2$. For the Frobenius norm, there is $\|AB\|_F \leq \|A\|_2 \|B\|_F$. The second inequality holds by Young's inequality with some parameter $\zeta_k > 0$ and $\|AB\|_F \leq \|A\|_2 \|B\|_F$ as well. For $\nabla F_k(X_{r,t}^k)$, by definition, we have

$$
\begin{aligned}
&\|\nabla F_k(X_{r,t}^k)\|_F^2 = \sum_{i \in \mathcal{V}_k} \|\nabla f_i(\boldsymbol{x}_{r,t}^i)\|^2 \\
&= \sum_{i \in \mathcal{V}_k} \|\nabla f_i(\boldsymbol{x}_{r,t}^i) - \nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) + \nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) - \nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k) + \nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k) - \nabla \bar{f}_k(\boldsymbol{x}_r) + \nabla \bar{f}_k(\boldsymbol{x}_r)\|^2 \\
&\leq \sum_{i \in \mathcal{V}_k} \Big[4\|\nabla f_i(\boldsymbol{x}_{r,t}^i) - \nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\|^2 + 4\|\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) - \nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\|^2 + 4\|\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k) - \nabla \bar{f}_k(\boldsymbol{x}_r)\|^2 \\
&\quad + 4\|\nabla \bar{f}_k(\boldsymbol{x}_r)\|^2 \Big] \\
&\leq \sum_{i \in \mathcal{V}_k} \Big[4\|\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) - \nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\|^2 + 4L^2\|\boldsymbol{x}_{r,t}^i - \bar{\boldsymbol{x}}_{r,t}^k\|^2 + 4L^2\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + 4\|\nabla \bar{f}_k(\boldsymbol{x}_r)\|^2 \Big] \\
&\leq 4L^2\|X_{r,t}^{k,\perp}\|^2 + 4L^2 n\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + 4n\|\nabla \bar{f}_k(\boldsymbol{x}_r)\|^2 + 4n\sigma_k^2,
\end{aligned} \tag{47}
$$

where the first inequality holds by Cauchy inequality, and the last inequality holds by Assumption 4.4. Averaging Eq. 47 over $k = 1, ..., K$, we have the following iteration

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{r,t+1}^{k,\perp}\|^2 \\
&\leq \frac{1}{N}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k^{-1})\cdot\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 + \frac{1}{N}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)\eta_l^2\mathbb{E}\|\nabla F(X_{r,t}^k))\|^2 + \eta_l^2\sigma^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2 \\
&\leq \frac{1}{N}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k^{-1})\cdot\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 + \eta_l^2\frac{1}{N}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)\cdot 4L^2\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)\cdot 4L^2\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)\cdot 4\mathbb{E}\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2 \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)\cdot 4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2 \\
&\leq \left(\max_{k\in[K]}\rho_k^2(1+\zeta_k^{-1}) + \eta_l^2\cdot 4L^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\right)\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 \\
&\quad + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4L^2\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2. \qquad (48)
\end{aligned}
$$

This concludes the proof. □

G.2 Lemma for summation of intra-cluster and inter-cluster consensus errors

Lemma G.3. *If the local learning rate satisfies the condition: $\eta_l \leq \frac{1}{8\mathcal{I}L}$, the for all local round $s = 0, ..., \mathcal{I}-1$, there is*

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{r,t}^{k,\perp}\|^2 + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 \\
&\leq (t+1)C_1\eta_l^2(\mathcal{I} + \rho_{\max}^2 H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + (t+1)C_1\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad + (t+1)C_1\eta_l^2\sigma^2\rho_{\max}^2 + (t+1)C_1\left(1 + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\right)\eta_l^2\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right), \qquad (49)
\end{aligned}
$$

where C_1 is a constant independent to parameters.

Proof. Denote an auxiliary vector

$$
M_{r,t} = \left(\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{r,t}^{k,\perp}\|^2, \quad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2\right)^T. \qquad (50)
$$

From Lemma G.1 and G.2, we have the following inequality which is defined element-wise for $s = 0, ..., \mathcal{I}-1$

$$
M_{r,t+1} \leq G\cdot M_{r,t} + B_{r,t}, \qquad (51)
$$

where

$$
G = \begin{pmatrix} \max_{k\in[K]}\rho_k^2(1+\zeta_k^{-1}) + \eta_l^2\rho_L\cdot 4L^2 & \eta_l^2\rho_L\cdot 4L^2 \\ 16\mathcal{I}\eta_l^2 L^2 & C_{\mathcal{I}} \end{pmatrix} \qquad (52)
$$

$$
B_{r,t} = \begin{pmatrix} 4\rho_L\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + 4\rho_L\eta_l^2\bar{\sigma}_l^2 + \eta_l^2\sigma^2\rho_{\max}^2 \\ 16\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + \eta_l^2\frac{\sigma^2}{n} + \frac{16m(n-m)}{n(n-1)}\mathcal{I}\eta_l^2\sigma_k^2 \end{pmatrix} = \begin{pmatrix} b^{(1)} \\ b^{(2)} \end{pmatrix}. \qquad (53)
$$

Consider the eigen-decomposition of matrix G,

$$G = \frac{1}{16\mathcal{I}\eta_l^2 L^2(\lambda_1 - \lambda_2)} \begin{pmatrix} \lambda_1 - C_{\mathcal{I}} & \lambda_2 - C_{\mathcal{I}} \\ 16\mathcal{I}\eta_l^2 L^2 & 16\mathcal{I}\eta_l^2 L^2 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot \begin{pmatrix} 16\mathcal{I}\eta_l^2 L^2 & -\lambda_2 + C_{\mathcal{I}} \\ -16\mathcal{I}\eta_l^2 L^2 & \lambda_1 - C_{\mathcal{I}} \end{pmatrix}, \tag{54}$$

where we assume $\lambda_1 \leq \lambda_2$, thus we have

$$\begin{aligned} G^j B &= \frac{1}{16\mathcal{I}\eta_l^2 L^2(\lambda_1 - \lambda_2)} \begin{pmatrix} \lambda_1 - C_{\mathcal{I}} & \lambda_2 - C_{\mathcal{I}} \\ 16\mathcal{I}\eta_l^2 L^2 & 16\mathcal{I}\eta_l^2 L^2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 16\mathcal{I}\eta_l^2 L^2 & -\lambda_2 + C_{\mathcal{I}} \\ -16\mathcal{I}\eta_l^2 L^2 & \lambda_1 - C_{\mathcal{I}} \end{pmatrix} \begin{pmatrix} b^{(1)} \\ b^{(2)} \end{pmatrix} \\ &= \frac{1}{16\mathcal{I}\eta_l^2 L^2(\lambda_1 - \lambda_2)} \begin{pmatrix} (\lambda_1 - C_{\mathcal{I}})(\lambda_1^j 16\mathcal{I}\eta_l^2 L^2 b^{(1)} + \lambda_1^j(-\lambda_2 + C_{\mathcal{I}})b^{(2)}) + (\lambda_2 - C_{\mathcal{I}})(-\lambda_2^j 16\mathcal{I}\eta_l^2 L^2 b^{(1)} + \lambda_2^j(\lambda_1 - C_{\mathcal{I}})b^{(2)}) \\ 16\mathcal{I}\eta_l^2 L^2(\lambda_1^j 16\mathcal{I}\eta_l^2 L^2 b^{(1)} + \lambda_1^j(-\lambda_2 + C_{\mathcal{I}})b^{(2)}) + 16\mathcal{I}\eta_l^2 L^2(-\lambda_2^j 16\mathcal{I}\eta_l^2 L^2 b^{(1)} + \lambda_2^j(\lambda_1 - C_{\mathcal{I}})b^{(2)}) \end{pmatrix}. \end{aligned} \tag{55}$$

Therefore the sum of two elements has the following result

$$\begin{aligned} (1,1)G^j B_{r,t-j} &= \lambda_1^j b^{(1)} + \lambda_2^j b^{(2)} + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1}\Big(16\mathcal{I}\eta_l^2 L^2 b^{(1)} + 4\eta_l^2 \rho_L L^2 b^{(2)}\Big) \\ &\leq \lambda_2^j (b^{(1)} + b^{(2)}) + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1}\Big(16\mathcal{I}\eta_l^2 L^2 b^{(1)} + 4\eta_l^2 \rho_L L^2 b^{(2)}\Big). \end{aligned} \tag{56}$$

Therefore, we have the following result

$$\sum_{j=0}^{t} (1,1)G^j B_{r,t-j} \leq \sum_{j=0}^{t} \Bigg(\lambda_2^j (b_{r,t-j}^{(1)} + b_{r,t-j}^{(2)}) + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1}\big(16\mathcal{I}\eta_l^2 L^2 b^{(1)} + 4\eta_l^2 \rho_L L^2 b^{(2)}\big)\Bigg). \tag{57}$$

Since $\lambda_2 \geq C_{\mathcal{I}} > 1$, we have

$$\frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1} = \lambda_2^{l-1} \sum_{t=0}^{l-1} \left(\frac{\lambda_1}{\lambda_2}\right)^s \leq \lambda_2^{j-1} \min\left\{\frac{\lambda_2}{\lambda_2 - \lambda_1}, l\right\} \leq \lambda_2^j \min\left\{\frac{1}{\lambda_2 - \lambda_1}, l\right\}, \tag{58}$$

thus we have

$$\sum_{j=0}^{t} (1,1)G^j B_{r,t-j} \leq \sum_{j=0}^{t} \lambda_2^j (b_{r,t-j}^{(1)} + b_{r,t-j}^{(2)}) + \sum_{l=0}^{s} \Bigg(\lambda_2^j \min\left\{\frac{1}{\lambda_2 - \lambda_1}, l\right\}\big(16\mathcal{I}\eta_l^2 L^2 b^{(1)} + 4\eta_l^2 \rho_L L^2 b^{(2)}\big)\Bigg). \tag{59}$$

By the definition of $\rho_L = \max_{k\in[k]} \rho_k^2(1+\zeta_k)$ and by the Gershgorin's theorem, since $\eta_l > 0$, we have the upper bound for λ_2,

$$\begin{aligned} \lambda_2 &\leq \max\left\{\max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \eta_l^2 \rho_L \cdot 8L^2, C_{\mathcal{I}} + 16\mathcal{I}\eta_l^2 L^2\right\} \\ &< \max\left\{\max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \frac{\rho_L}{(4\mathcal{I}-1)2\mathcal{I}}, 1 + \frac{2}{4\mathcal{I}-1}\right\}, \end{aligned} \tag{60}$$

where the last inequality holds by the bound of $\eta_l^2 \leq \frac{1}{64\mathcal{I}^2 L^2} < \frac{1}{16\mathcal{I}(4\mathcal{I}-1)L^2}$. Define a distance constant $H_{\mathcal{I},\rho} = \min\left\{\mathcal{I}, \frac{1}{1-\rho_{\max}}\right\}$. Next we consider two cases: *small or dense* communication network with $\rho_{\max} \leq 1 - \frac{1}{\mathcal{I}}$ and *large and sparse* communication network with $\rho_{\max} > 1 - \frac{1}{\mathcal{I}}$.

Case 1: For $\rho_{\max} \leq 1 - \frac{1}{\mathcal{I}}$, i.e., $\frac{1}{1-\rho_{\max}} \leq \mathcal{I}$, thus we have $H_{\mathcal{I},\rho} = \frac{1}{1-\rho_{\max}}$. Let $\zeta_k = \frac{\rho_k}{1-\rho_k}$, then we have

$$\max_{k\in[k]} \rho_k^2(1+\zeta_k^{-1}) = \rho_{\max}, \quad \rho_L = \max_{k\in[k]}\left\{\frac{\rho_k^2}{1-\rho_k}\right\} = \frac{\rho_{\max}^2}{1-\rho_{\max}} = \rho_{\max}^2 H_{\mathcal{I},\rho}, \tag{61}$$

where the middle part of the second equality holds by the monotonically increasing of $\frac{x^2}{1-x}$. Then the bound for λ_2 is formalized as

$$\begin{aligned}\lambda_2 &\leq \max\left\{\rho_{\max} + \frac{\rho_{\max}^2}{(1-\rho_{\max})2\mathcal{I}(4\mathcal{I}-1)}, 1 + \frac{3}{2(4\mathcal{I}-1)}\right\} \\ &\leq \max\left\{1 - \frac{1}{\mathcal{I}} + \frac{(1-\frac{1}{\mathcal{I}})^2}{2(\mathcal{I}-1)}, 1 + \frac{3}{2(4\mathcal{I}-1)}\right\} \\ &< 1 + \frac{3}{4\mathcal{I}-1},\end{aligned} \tag{62}$$

where the second inequality holds by $\rho_{\max} \leq 1 - \frac{1}{\mathcal{I}}$. Then by $s \leq \mathcal{I}$ and $\lambda_2 \geq 1$ (just by the definition of matrix G can get this result), we can obtain the following bound

$$\sum_{j=0}^{t} \lambda_2^j b_{t-j}^1 \leq \left(\left(1 + \frac{2}{4\mathcal{I}-1}\right)^{\mathcal{I}}\right) \cdot \sum_{j=0}^{t} b_j^{(1)} \leq 3 \cdot \sum_{j=0}^{t} b_j^{(1)}. \tag{63}$$

We also have

$$\rho_{\max} + \eta_l^2 \rho_L 4L^2 \leq \rho_{\max} + \frac{\rho_{\max}}{(1-\rho_{\max})(4\mathcal{I}-1)4\mathcal{I}} \leq 1 - \frac{1}{\mathcal{I}} + \frac{(1-\frac{1}{\mathcal{I}})^2}{4(4\mathcal{I}-1)} \leq C_{\mathcal{I}}, \tag{64}$$

where the second inequality holds by the upper bound for $\rho_{\max}$. By the definition of matrix G, we bound the difference of $\lambda_2 - \lambda_1$,

$$\begin{aligned}\lambda_2 - \lambda_1 &= C_{\mathcal{I}} - \rho_{\max} - \eta_l^2 \rho_L 4L^2 \\ &\geq C_{\mathcal{I}} - \left(\rho_{\max} + \frac{\rho_{\max}}{(1-\rho_{\max})(4\mathcal{I}-1)4\mathcal{I}}\right) \\ &\geq C_{\mathcal{I}} - \left(\rho_{\max} + \rho_{\max} \cdot \frac{1-\frac{1}{\mathcal{I}}}{4(4\mathcal{I}-1)}\right) \\ &\geq 1 + \frac{1}{4\mathcal{I}-1} - \left(\rho_{\max} + \rho_{\max} \cdot \frac{1}{4\mathcal{I}-1}\right) \\ &= (1-\rho_{\max})\left(1 + \frac{1}{4\mathcal{I}-1}\right) \\ &\geq 1 - \rho_{\max}.\end{aligned} \tag{65}$$

where the first and second inequality hold by the defined notations. Then we have

$$\begin{aligned}\sum_{j=0}^{t}(1,1)G^j B_{r,t-j} &\leq \sum_{j=0}^{t} \lambda_2^j (b_{r,t-j}^{(1)} + b_{r,t-j}^{(2)}) + \sum_{j=0}^{t}\left(\lambda_2^j \min\left\{\frac{1}{\lambda_2 - \lambda_1}, j\right\}\left(16\mathcal{I}\eta_l^2 L^2 b_{r,t-j}^{(1)} + 4\eta_l^2 \rho_L L^2 b_{r,t-j}^{(2)}\right)\right) \\ &\leq \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} 3\eta_l^2\left(16\mathcal{I}L^2 b_{r,j}^{(1)} + 4\rho_L L^2 b_{r,j}^{(2)}\right)\left(\min\left\{\frac{1}{\lambda_2 - \lambda_1}, \mathcal{I}\right\}\right) \\ &\leq \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} 3\eta_l^2\left(16\mathcal{I}L^2 b_{r,j}^{(1)} + 4H_{\mathcal{I},\rho}^2 \rho_{\max}^2 L^2 b_{r,j}^{(2)}\right) \\ &\leq \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} \frac{H_{\mathcal{I},\rho}}{16\mathcal{I}(4\mathcal{I}-1)} \cdot \left(48\mathcal{I} b_{r,j}^{(1)} + 12 H_{\mathcal{I},\rho}\rho_{\max}^2 b_{r,j}^{(2)}\right) \\ &\leq \sum_{j=0}^{t} 4(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2} \cdot \rho_{\max}^2 b_{r,j}^{(2)}.\end{aligned} \tag{66}$$

Then by the definition of $b^{(1)}$ and $b^{(2)}$, we have

$$
\begin{aligned}
&\sum_{j=0}^{t}(1,1)G^j B_{r,t-j} \\
&\leq \sum_{j=0}^{t} 4\bigg((4\rho_L\eta_l^2 + 16\mathcal{I}\eta_l^2)(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + 4\rho_L\eta_l^2\bar{\sigma}_l^2 + \eta_l^2\sigma^2\rho_{\max}^2 + \eta_l^2\frac{\sigma^2}{n} + 16\frac{m(n-m)}{n(n-1)}\mathcal{I}\eta_l^2\bar{\sigma}_l^2\bigg) \\
&\quad + \sum_{j=0}^{t}\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\bigg(16\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + \eta_l^2\frac{\sigma^2}{n} + 16\frac{m(n-m)}{n(n-1)}\mathcal{I}\eta_l^2\bar{\sigma}_l^2\bigg) \\
&= (t+1)\bigg[\bigg(4(4\rho_L\eta_l^2 + 16\mathcal{I}\eta_l^2) + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2 16\mathcal{I}\eta_l^2\bigg)(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + 16\rho_L\eta_l^2\bar{\sigma}_l^2 \\
&\quad + 16\eta_l^2\sigma^2\rho_{\max}^2 + \bigg(4\eta_l^2 + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\eta_l^2\bigg)\bigg(\frac{\sigma^2}{n} + 16\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg] \\
&\leq (t+1)\bigg[\bigg(16\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2 + 64\mathcal{I}\eta_l^2 + 16\eta_l^2 H_{\mathcal{I},\rho}\cdot\rho_{\max}^2\bigg)(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + 16\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad + 16\eta_l^2\sigma^2\rho_{\max}^2 + \bigg(4\eta_l^2 + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\eta_l^2\bigg)\bigg(\frac{\sigma^2}{n} + 16\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg)\bigg] \\
&\leq (t+1)C_1\eta_l^2(\mathcal{I} + \rho_{\max}^2 H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + (t+1)C_1\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 \\
&\quad + (t+1)C_1\eta_l^2\sigma^2\rho_{\max}^2 + (t+1)C_1\bigg(1 + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\bigg)\eta_l^2\bigg(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\bigg),
\end{aligned} \tag{67}
$$

where C_1 is some universal constant. The inequality holds by $\rho_L = \rho_{\max}^2 H_{\mathcal{I},\rho}$ and $H_{\mathcal{I},\rho} \leq \mathcal{I}$.
Case 2: In this case we have $\rho_{\max} > 1 - \frac{1}{\mathcal{I}}$, which means $H_{\mathcal{I},\rho} = \mathcal{I}$. Let $\zeta_k = (4\mathcal{I} - 1)$, thus we have

$$
\max_{k\in[K]} \rho_k^2(1 + \zeta_k^{-1}) = \rho_{\max}^2(1 + (4\mathcal{I} - 1)^{-1}), \quad \rho_L = 4\mathcal{I}\rho_{\max}^2 H_{\mathcal{I},\rho}. \tag{68}
$$

The upper bound for λ_2 has the form of

$$
\begin{aligned}
\lambda_2 &\leq \max\bigg\{\max_{k\in[K]} \rho_k^2(1 + \zeta_k^{-1}) + \eta_l^2\rho_L\cdot 8L^2, C_{\mathcal{I}}\bigg\} \\
&\leq \max\bigg\{\rho_{\max}^2(1 + (4\mathcal{I} - 1)^{-1}) + \frac{2\rho_{\max}^2}{4\mathcal{I} - 1}, 1 + \frac{3}{2(4\mathcal{I} - 1)}\bigg\} \\
&\leq 1 + \frac{3}{4\mathcal{I} - 1}.
\end{aligned} \tag{69}
$$

By the fact of $\min\left\{\frac{1}{\lambda_2 - \lambda_1}, l\right\} \leq \mathcal{I} = H_{\mathcal{I},\rho}$, we have

$$
\begin{aligned}
\sum_{j=0}^{t}(1,1)G^j B_{r,t-j} &\leq \sum_{j=0}^{t}\lambda_2^j(b_{r,t-j}^{(1)} + b_{r,t-j}^{(2)}) + \sum_{j=0}^{t}\bigg(\lambda_2^j\min\bigg\{\frac{1}{\lambda_2 - \lambda_1}, l\bigg\}\eta_l^2\cdot 4\rho_L L^2 b_{r,t-j}^2\bigg) \\
&\leq \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t}\eta_l^2\cdot 16\rho_{\max}H_{\mathcal{I},\rho}L^2 b_l^{(2)}\cdot 3H_{\mathcal{I},\rho} \\
&\leq \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} 16\rho_{\max}H_{\mathcal{I},\rho}b_l^{(2)}\cdot\frac{3}{16}\frac{H_{\mathcal{I},\rho}}{\mathcal{I}^2} \\
&= \sum_{j=0}^{t} 3(b_{r,j}^{(1)} + b_{r,j}^{(2)}) + \sum_{j=0}^{t} 3\rho_{\max}b_l^{(2)}\cdot\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2},
\end{aligned} \tag{70}
$$

where the above inequalities hold by the fact that $\rho_L = 4\mathcal{I}\rho_{\max}^2 = 4H_{\mathcal{I},\rho}\rho_{\max}^2$ and the constraint on step size η_l. Thus we can get a similar upper bound as Eq. 67 in Case 1. This concludes the proof. □

Lemma G.4. *With the similar condition in Lemma G.3, we have the corresponding bound for the intra-cluster consensus error* $\|X_{r,t}^{k,\perp}\|_F^2$,

$$\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2] \leq (t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2) + (t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2 + (t+1)C_1\eta_l^2\rho_{\max}^2\sigma^2 + (t+1)C_1\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\eta_l^2\rho_{\max}^2\frac{\sigma^2}{n} + (t+1)C_1\frac{m(n-m)}{n(n-1)}\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}}\eta_l^2\rho_{\max}^2\bar{\sigma}_l^2. \tag{71}$$

Proof. With the same definition of the auxiliary vector $M_{r,t}$ and the matrix G and $B_{r,t}$ in the proof of Lemma G.3, there is

$$M_{r,t} = \left(\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2], \quad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2]\right)^\top,$$
$$M_{r,t} = G^{t+1}M_{r,0} + \sum_{j=0}^{t}G^j B_{r,t-j} = \sum_{j=0}^{t}G^j B_{r,t-j}, \tag{72}$$

hence we have

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2] &= (1,0)\cdot M_{r,t} = (1,0)\cdot\sum_{j=0}^{t}G^j B_{r,t-j}\\ &= \sum_{j=0}^{t}\left[\lambda_1^j b_{r,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{r,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{t}\left[\lambda_2^j b_{r,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{r,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{t}\left[\lambda_2^j b_{r,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{r,j}^{(2)}\right],\end{aligned} \tag{73}$$

with the similar proof techniques as in Lemma G.3, there is

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2] &\leq \sum_{j=0}^{t}\left[3b_{r,j}^{(1)} + \frac{1}{\mathcal{I}^2}H_{\mathcal{I},\rho}^2\rho_{\max}^2 b_{r,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{t}\Bigg[12\rho_L\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2) + 12\rho_L\eta_l^2\bar{\sigma}_l^2 + 3\eta_l^2\sigma^2\rho_{\max}^2\\ &\quad + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\rho_{\max}^2\left(16\mathcal{I}\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2) + 16\frac{m(n-m)}{n(n-1)}\mathcal{I}\eta_l^2\bar{\sigma}_l^2 + \eta_l^2\frac{\sigma^2}{n}\right)\Bigg]\\ &\leq (t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2) + (t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2\\ &\quad + (t+1)C_1\eta_l^2\rho_{\max}^2\sigma^2 + (t+1)C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\rho_{\max}^2\frac{\sigma^2}{n} + (t+1)C_1\frac{m(n-m)}{n(n-1)}\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}}\rho_{\max}^2\bar{\sigma}_l^2.\end{aligned} \tag{74}$$

This concludes the proof. □

G.3 Lemmas for model difference Δ_r

There is a corresponding Lemma about model difference Δ_r for the partial participation settings.

Lemma G.5. *The global model difference $\Delta_r = \sum_{k=1}^{K} \sum_{i \in cS_r} \Delta_r^i$ in partial participation settings satisfies*

$$\begin{aligned}
&\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{2\eta_l^2 \mathcal{I}}{N}\sigma^2 + 2\eta_l^2(\mathcal{I}-1)\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] + 4\eta_l^2\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\boldsymbol{x}}_{r,\mathcal{I}-1}^k)\right\|^2\right] \\
&+ 8\left(\frac{n-m}{m(n-1)} + \eta_l^2 L^2\right)\left(\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,\mathcal{I}-1}^{k,\perp}\|^2]\right) + \frac{2\eta_l^2\sigma^2}{N}\left(\frac{n-m}{m}\cdot\rho_{\max}^2\right). \qquad (75)
\end{aligned}$$

Proof. Recall the definition in E, there are $\bar{\boldsymbol{x}}_{r,t} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_{r,t}^i$ (without consideration of client sampling) and the intra-cluster average $\bar{\boldsymbol{x}}_{r,t+1}^k = \bar{\boldsymbol{x}}_{r,t}^k - \eta_l \bar{\boldsymbol{x}}_{r,t+1}^k$, where $\bar{\boldsymbol{x}}_{r,t+1}^k = \frac{1}{n}\sum_{i\in\mathcal{V}_k}\boldsymbol{g}_{r,t}^i$.

For the model difference Δ_r, we have

$$\begin{aligned}
\mathbb{E}[\|\Delta_r\|^2] &= \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i - \boldsymbol{x}_r\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in\mathcal{S}_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i \mp \bar{\boldsymbol{x}}_{r,\mathcal{I}} \mp \bar{\boldsymbol{x}}_{r,\mathcal{I}-1} \mp \cdots \mp \bar{\boldsymbol{x}}_{r,1} - \boldsymbol{x}_r\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i - \bar{\boldsymbol{x}}_{r,\mathcal{I}}\right\|^2\right] + 2\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{r,\mathcal{I}} \mp \bar{\boldsymbol{x}}_{r,\mathcal{I}-1} \mp \cdots \mp \bar{\boldsymbol{x}}_{r,1} - \boldsymbol{x}_r\right\|^2\right], \qquad (76)
\end{aligned}$$

where the inequality holds by Cauchy-Schwarz inequality. For the first term in Eq. 76, we have

$$\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i - \bar{\boldsymbol{x}}_{r,\mathcal{I}}\right\|^2\right] = \frac{K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\boldsymbol{x}_{r,\mathcal{I}}^i - \bar{\boldsymbol{x}}_{r,\mathcal{I}}\|^2] \qquad (77)$$

For the second part in Eq. 76, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{r,\mathcal{I}} \pm \cdots \pm \bar{\boldsymbol{x}}_{r,1} - \boldsymbol{x}_r\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\eta_l\sum_{t=0}^{\mathcal{I}-1}\bar{\boldsymbol{g}}_{r,t}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\boldsymbol{g}_{r,t}^i\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}(\boldsymbol{g}_{r,t}^i \pm \nabla f_i(\boldsymbol{x}_{r,t}^i))\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}(\boldsymbol{g}_{r,t}^i - \nabla f_i(\boldsymbol{x}_{r,t}^i))\right\|^2\right] + \mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&\leq \frac{\eta_l^2(\mathcal{I}-1)M}{N^2}\sigma^2 + \mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right], \qquad (78)
\end{aligned}$$

where

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\sum_{t=0}^{\mathcal{I}-1}\Big(\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\mp\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\mp\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big)\Big\|^2\Big]\\
&\le 2\mathbb{E}\Big[\Big\|\sum_{t=0}^{\mathcal{I}-1}\Big(\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)-\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big)\Big\|^2\Big]\\
&\quad+4\mathbb{E}\Big[\Big\|\sum_{t=0}^{\mathcal{I}-1}\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]+4\mathbb{E}\Big[\Big\|\sum_{t=0}^{\mathcal{I}-1}\Big(\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}[\nabla f_i(\boldsymbol{x}_{r,t}^i)-\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)]\Big)\Big\|^2\Big]\\
&\le 2\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)-\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]\\
&\quad+4(\mathcal{I}-1)\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]+4(\mathcal{I}-1)\sum_{t=0}^{\mathcal{I}-1}L^2\frac{\eta_l^2 M}{N^2}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]\\
&\le 4\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]+4\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]\\
&\quad+4(\mathcal{I}-1)\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]+4(\mathcal{I}-1)L^2\frac{\eta_l^2 M}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]\\
&=4\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]+4\mathcal{I}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]\\
&\quad+4(\mathcal{I}-1)L^2\frac{\eta_l^2 M}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]\\
&\le 8\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big]+4\mathcal{I}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]\\
&\quad+4(\mathcal{I}+1)L^2\frac{\eta_l^2 M}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2],
\end{aligned}\tag{79}
$$

where the last inequality follows

$$
\begin{aligned}
&4\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\Big\|^2\Big]\\
&=4\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}[\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)\mp\nabla f_i(\boldsymbol{x}_{r,t}^i)]\Big\|^2\Big]\\
&\le 8\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big]+8\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}[\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)-\nabla f_i(\boldsymbol{x}_{r,t}^i)]\Big\|^2\Big]\\
&\le 8\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\Big[\Big\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\Big\|^2\Big]+8\frac{\eta_l^2 M}{N^2}L^2\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]
\end{aligned}\tag{80}
$$

and we use the characteristic of conditional expectation, where we use the characteristic of conditional expectation, i.e., $\mathbb{E}[\mathbb{E}_{\mathcal{S}_s}[\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)]] = \mathbb{E}[\frac{\eta_l M}{N^2}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)]$, Update the expectation term, that is $\mathbb{E}[\mathbb{E}_{\mathcal{S}_s}[\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) - \frac{\eta_l M}{N^2}\sum_{i=1}^{N}\nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k)]] = 0$ and $\forall r \neq s, i \in \mathcal{S}_s^k$ is independent with $i \in \mathcal{S}_r^k$. Then we have

$$
\begin{aligned}
&8\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&= \frac{8\eta_l^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{P}\{i\in\mathcal{S}_s^k\}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&\leq \frac{8\eta_l^2 K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\nabla f_i(\boldsymbol{x}_{r,t}^i)\|^2] \\
&\quad + \frac{8\eta_l^2}{N^2}\frac{m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&= \frac{8\eta_l^2 K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(\boldsymbol{x}_{r,t}^i)\|^2] + \frac{8\eta_l^2}{N^2}\frac{m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right],
\end{aligned}
\tag{81}
$$

where the inequality holds by Lemma G.6.

$$
\begin{aligned}
&\sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(\boldsymbol{x}_{r,t}^i)\|^2] \\
&= \sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(\boldsymbol{x}_{r,t}^i) \mp \nabla f_i(\bar{\boldsymbol{x}}_{r,t}^k) \mp \bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k) \mp \bar{f}_k(\boldsymbol{x}_r)\|^2] \\
&\leq \sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\mathbb{E}[4L^2\|X_{r,t}^{k,\perp}\|^2 + 4L^2 n\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2 + 4n\|\nabla\bar{f}_k(\boldsymbol{x}_r)\|^2 + 4n\sigma_k^2] \\
&\leq \sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\left[4L^2\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2] + 4L^2 n\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k - \boldsymbol{x}_r\|^2] + 4n\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2] + 4n\sigma_k^2 + 4n\sigma_c^2\right].
\end{aligned}
\tag{82}
$$

Then combining the previous terms, we have

$$
\begin{aligned}
&\mathbb{E}[\|\Delta_r\|^2] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r^k}\boldsymbol{x}_{r,\mathcal{I}}^i-\bar{\boldsymbol{x}}_{r,\mathcal{I}}\right\|^2\right]+2\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{r,\mathcal{I}}\mp\bar{\boldsymbol{x}}_{r,\mathcal{I}-1}\mp\cdots\mp\bar{\boldsymbol{x}}_{r,1}-\boldsymbol{x}_r\right\|^2\right] \\
&\leq \frac{2K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}\|\boldsymbol{x}_{r,\mathcal{I}}^i-\bar{\boldsymbol{x}}_{r,\mathcal{I}}^k\|^2]+\frac{2\eta_l^2\mathcal{I}M}{N^2}\sigma^2 \\
&\quad+2\mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&\leq \frac{2K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\boldsymbol{x}_{r,\mathcal{I}}^i-\bar{\boldsymbol{x}}_{r,\mathcal{I}}^k\|^2]+\frac{2\eta_l^2\mathcal{I}M}{N^2}\sigma^2 \\
&\quad+16\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{\eta_l}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_s^k}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right]+8\mathcal{I}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\right\|^2\right] \\
&\quad+8(\mathcal{I}+1)\frac{\eta_l^2ML^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2] \\
&\leq \frac{2K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\boldsymbol{x}_{r,\mathcal{I}}^i-\bar{\boldsymbol{x}}_{r,\mathcal{I}}^k\|^2]+\frac{2\eta_l^2\mathcal{I}M}{N^2}\sigma^2 \\
&\quad+\frac{16\eta_l^2K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(\boldsymbol{x}_{r,t}^i)\|^2]+\frac{16\eta_l^2}{N^2}\frac{m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&\quad+8\mathcal{I}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\right\|^2\right]+8(\mathcal{I}+1)\frac{\eta_l^2ML^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2] \\
&\leq \frac{2K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,\mathcal{I}}^{k,\perp}\|^2]+\frac{2\eta_l^2\mathcal{I}M}{N^2}\sigma^2 \\
&\quad+\frac{16\eta_l^2K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\Big[4L^2\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]+4L^2n\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k-\boldsymbol{x}_r\|^2] \\
&\quad+4n\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)\|^2]+4n\sigma_k^2+4n\sigma_c^2\Big]+\frac{16\eta_l^2}{N^2}\frac{m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] \\
&\quad+8\mathcal{I}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{\eta_l m}{N}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\right\|^2\right]+8(\mathcal{I}+1)\frac{\eta_l^2ML^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2], \qquad (83)
\end{aligned}
$$

where we have

$$
\begin{aligned}
&\frac{16\eta_l^2 K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\left[4L^2\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]+4L^2 n\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k-\boldsymbol{x}_r\|^2]\right]\\
&=\frac{16\eta_l^2 K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\left[4L^2\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]+4L^2 n\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k-\boldsymbol{x}_r\|^2]\right]\\
&=\frac{64\eta_l^2 L^2}{N}\frac{m(n-m)}{n^2(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]+\frac{64\eta_l^2 L^2}{K}\frac{m(n-m)}{n^2(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k-\boldsymbol{x}_r\|^2]\\
&=64\eta_l^2 L^2\frac{m(n-m)}{n^2(n-1)}\sum_{t=0}^{\mathcal{I}-1}\left[\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\boldsymbol{x}}_{r,t}^k-\boldsymbol{x}_r\|^2]\right]\\
&\leq 64\eta_l^2 L^2\frac{m(n-m)}{n^2(n-1)}\Bigg[\mathcal{I}^2 C_1\eta_l^2(\mathcal{I}+\rho_{\max}^2 H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2+\sigma_c^2)+\mathcal{I}^2 C_1\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2\\
&\quad+\mathcal{I}^2 C_1\eta_l^2\sigma^2\rho_{\max}^2+\mathcal{I}^2 C_1\left(1+\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\right)\eta_l^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg],
\end{aligned}
\tag{84}
$$

we have

$$
\begin{aligned}
&\frac{2K}{(Np)^2}\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,\mathcal{I}}^{k,\perp}\|^2]\\
&=\frac{2}{Kn}\frac{(n-m)}{m(n-1)}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,\mathcal{I}}^{k,\perp}\|^2]\\
&\leq\frac{2(n-m)}{m(n-1)}\Bigg[\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2)+\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2\\
&\quad+\mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2+\mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg],
\end{aligned}
\tag{85}
$$

and

$$
\begin{aligned}
&8(\mathcal{I}+1)\frac{\eta_l^2 ML^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_r}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]\\
&\leq 8(\mathcal{I}+1)\frac{\eta_l^2 ML^2}{N^2}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\boldsymbol{x}_{r,t}^i-\bar{\boldsymbol{x}}_{r,t}^k\|^2]\\
&=8(\mathcal{I}+1)\frac{\eta_l^2 mL^2}{n}\sum_{t=0}^{\mathcal{I}-1}\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{r,t}^{k,\perp}\|^2]\\
&=8(\mathcal{I}+1)\frac{\eta_l^2 mL^2}{n}\sum_{t=0}^{\mathcal{I}-1}\Bigg[(t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2)+(t+1)C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2\\
&\quad+(t+1)C_1\eta_l^2\rho_{\max}^2\sigma^2+(t+1)C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\frac{\sigma^2}{n}\Bigg]\\
&=4(\mathcal{I}+1)^2\frac{\eta_l^2 mL^2}{n}\Bigg[\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2+\sigma_c^2)+\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2\\
&\quad+\mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2+\mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\left(\frac{\sigma^2}{n}+\frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg].
\end{aligned}
\tag{86}
$$

Then by merging pieces together, we have

$$
\begin{aligned}
&\mathbb{E}[\|\Delta_r\|^2] \\
&\leq \frac{2\eta_l^2 \mathcal{I} M}{N^2}\sigma^2 + \left(\frac{2(n-m)}{m(n-1)} + \frac{4(\mathcal{I}+1)^2\eta_l^2 m L^2}{n}\right)\Bigg[\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2 + \sigma_c^2) + \mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2 \\
&\quad + \mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2 + \mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\rho_{\max}^2\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad + 64\eta_l^2 L^2\frac{m(n-m)}{n^2(n-1)}\Bigg[\mathcal{I}^2 C_1\eta_l^2(\mathcal{I} + \rho_{\max}^2 H_{\mathcal{I},\rho})(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) \\
&\quad + \mathcal{I}^2 C_1\rho_{\max}^2 H_{\mathcal{I},\rho}\eta_l^2\bar{\sigma}_l^2 + \mathcal{I}^2 C_1\eta_l^2\sigma^2\rho_{\max}^2 + \mathcal{I}^2 C_1\eta_l^2\left(1 + \frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\cdot\rho_{\max}^2\right)\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad + \frac{16\eta_l^2 K}{N^2}\frac{m(n-m)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\sum_{k=1}^{K}[4n\sigma_k^2 + 4n\sigma_c^2] \\
&\quad + \frac{16\eta_l^2}{N^2}\frac{m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] + \frac{8\eta_l^2\mathcal{I}m^2}{n^2}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\right\|^2\right] \\
&\leq \frac{2\eta_l^2\mathcal{I}M}{N^2}\sigma^2 + \frac{64\eta_l^2\mathcal{I}m(n-m)}{n^2(n-1)}[\bar{\sigma}_l^2 + \sigma_c^2] \\
&\quad + \left(\frac{2(n-m)}{m(n-1)} + 64\eta_l^2 L^2\frac{m(n-m)}{n^2(n-1)}\mathcal{I} + \frac{4(\mathcal{I}+1)^2\eta_l^2 m L^2}{n}\right)\Bigg[\mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\boldsymbol{x}_r)]\|^2 + \sigma_c^2) \\
&\quad + \mathcal{I}C_1\eta_l^2 H_{\mathcal{I},\rho}\rho_{\max}^2\bar{\sigma}_l^2 + \mathcal{I}C_1\eta_l^2\rho_{\max}^2\sigma^2 + \mathcal{I}C_1\eta_l^2\frac{H_{\mathcal{I},\rho}^2}{\mathcal{I}^2}\rho_{\max}^2\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\Bigg] \\
&\quad + 64\eta_l^2 L^2\frac{m(n-m)}{n^2(n-1)}\left[\mathcal{I}^3 C_1\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\boldsymbol{x}_r)\|^2 + \sigma_c^2) + \mathcal{I}^2 C_1\eta_l^2\left(\frac{\sigma^2}{n} + \frac{m(n-m)}{n(n-1)}\mathcal{I}\bar{\sigma}_l^2\right)\right] \\
&\quad + \frac{16\eta_l^2 m(m-1)}{n(n-1)}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\boldsymbol{x}_{r,t}^i)\right\|^2\right] + \frac{8\eta_l^2\mathcal{I}m^2}{n^2}\sum_{t=0}^{\mathcal{I}-1}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\boldsymbol{x}}_{r,t}^k)\right\|^2\right]. \qquad (87)
\end{aligned}
$$

Thus it concludes the proof.

□

G.4 Additional Supporting Lemmas

Lemma G.6 (Cluster sampling)**.** *For model weights $\boldsymbol{y}_{r,t}^{k,i}, \forall k \in [K], i \in \mathcal{V}_k r \in [R], t \in [\mathcal{I}]$, there is*

$$
\mathbb{E}\left[\left\|\sum_{k\in[K]}\sum_{i=1}^{n}\mathbb{I}\{i \in \mathcal{S}_k^t\}\boldsymbol{y}_r^{k,i}\right\|^2\right] \leq \mathbb{E}\left[\frac{m(m-1)}{n(n-1)}\left\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\right\|^2 + \frac{Km(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\boldsymbol{y}_r^{k,i}\|^2\right]. \qquad (88)
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\sum_{i\in\mathcal{V}_k}\mathbb{I}\{i\in\mathcal{S}_k^t\}\boldsymbol{y}_r^{k,i}\Big\|^2\Big]=\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{P}\{i\in\mathcal{S}_k^t\}\boldsymbol{y}_r^{k,i}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\mathbb{P}\{i\in\mathcal{S}_k^t\}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2+\mathbb{P}\{i\neq j\in\mathcal{S}_k^t\}\sum_{k=1}^{K}\sum_{i\neq j\in\mathcal{V}_k}\big(\boldsymbol{y}_r^{k,i}\big)'\big(\boldsymbol{y}_r^{j}\big)\\
&\quad+\mathbb{P}\{i\in\mathcal{S}_k^t,j\in\mathcal{S}_l^t|k\neq l\in[K]\}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\boldsymbol{y}_r^{i,k}\big)'\big(\boldsymbol{y}_r^{j,l}\big)\Big]\\
&=\mathbb{E}\Big[\frac{m}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2+\frac{m(m-1)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\neq j\in\mathcal{V}_k}\big(\boldsymbol{y}_r^{k,i}\big)'\big(\boldsymbol{y}_r^{k,j}\big)\\
&\quad+\frac{m^2}{n^2}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\boldsymbol{y}_r^{i,k}\big)'\big(\boldsymbol{y}_r^{j,l}\big)\Big]\\
&=\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\Big\|^2+\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2\\
&\quad+\frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\boldsymbol{y}_r^{i,k}\big)'\big(\boldsymbol{y}_r^{j,l}\big)\Big]\\
&\leq\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\Big\|^2+\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2\\
&\quad+\frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\Big(\frac{1}{2}\big\|\boldsymbol{y}_r^{i,k}\big\|^2+\frac{1}{2}\big\|\boldsymbol{y}_r^{j,l}\big\|^2\Big)\Big],
\end{aligned}
\tag{89}
$$

where the third equation holds by the probability of random sampling with replacement, i.e., $\mathbb{P}\{i\in\mathcal{S}_k^t\}=\frac{m}{n},\mathbb{P}\{i\neq j\in\mathcal{S}_k^t\}=\frac{m(m-1)}{n(n-1)},\mathbb{P}\{i\in\mathcal{S}_k^t,j\in\mathcal{S}_l^t|k\neq l\in[K]\}=\frac{m^2}{n^2}$. The forth equation holds by $\langle\ ,\boldsymbol{b}\rangle=\frac{1}{2}[\|\|^2+\|\boldsymbol{b}\|^2-\|-\boldsymbol{b}\|^2]$, $\frac{1}{2}\sum_{i\neq j}\|_i-{}_j\|^2=\sum_{i=1}^{n}n\|_i\|^2-\|\sum_{i=1}^{n}{}_i\|^2$, and $\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\|^2=\sum_{k=1}^{K}\|\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\|^2+\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\langle\boldsymbol{y}_r^{k,i}\boldsymbol{y}_r^{l,j}\rangle$. The last inequality holds by $'\boldsymbol{b}\leq\frac{1}{2}\|\|^2+\frac{1}{2}\|\boldsymbol{b}\|^2$. Reorganize the last item,

$$
\frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\Big(\frac{1}{2}\big\|\boldsymbol{y}_r^{i,k}\big\|^2+\frac{1}{2}\big\|\boldsymbol{y}_r^{j,l}\big\|^2\Big)=\frac{m(n-m)}{n^2(n-1)}(K-1)n\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\boldsymbol{y}_r^{k,i}\|^2,
\tag{90}
$$

then we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{k\in[K]}\sum_{i=1}^{n}\mathbb{I}\{i\in\mathcal{S}_k^t\}\boldsymbol{y}_r^{k,i}\Big\|^2\Big]\\
&\leq\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\Big\|^2+\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2\\
&\quad+\frac{m(n-m)}{n^2(n-1)}(K-1)n\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{i}\big\|^2\Big]\\
&=\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{y}_r^{k,i}\Big\|^2+\frac{Km(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\big\|\boldsymbol{y}_r^{k,i}\big\|^2\Big].
\end{aligned}
\tag{91}
$$

This concludes the proof. □

Lemma G.7 (Lemma for momentum term in the update rule)**.** *The first order momentum terms $\boldsymbol{m}_r$ in Algorithm 1 hold the following relationship w.r.t. model difference Δ_r:*

$$\sum_{r=1}^{R}\mathbb{E}[\|\boldsymbol{m}_r\|^2] \leq \sum_{r=1}^{R}\mathbb{E}[\|\Delta_r\|^2]. \tag{92}$$

Proof. By the updating rule, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{m}_r\|^2] &= \mathbb{E}\left[\left\|(1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\Delta_u\right\|^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\Delta_{u,i}\right)^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\right)\left(\sum_{u=1}^{t}\beta_1^{t-u}\Delta_{u,i}^2\right)\right] \\
&\leq (1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2],
\end{aligned} \tag{93}$$

summing over $t = 1, ..., T$, we have

$$\begin{aligned}
\sum_{r=1}^{R}\mathbb{E}[\|\boldsymbol{m}_r\|^2] &= (1-\beta_1)\sum_{r=1}^{R}\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2] \\
&= (1-\beta_1)\sum_{u=1}^{T}\sum_{t=u}^{T}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2] \\
&\leq (1-\beta_1)\sum_{u=1}^{T}\frac{1}{1-\beta_1}\mathbb{E}[\|\Delta_u\|^2] \\
&= \sum_{u=1}^{T}\mathbb{E}[\|\Delta_u\|^2].
\end{aligned} \tag{94}$$

This concludes the proof. □

Lemma G.8. *Under Assumptions 4.3, for AFGA, we have $\|\nabla f(\boldsymbol{x})\| \leq G$, $\|\Delta_r\| \leq \eta_l\mathcal{I}G$, $\|\boldsymbol{m}_r\| \leq \eta_l\mathcal{I}G$ and $\|\boldsymbol{v}_r\| \leq \eta_l^2\mathcal{I}^2G^2$.*

Proof. Since f has G-bounded stochastic gradients, for any $\boldsymbol{x}$ and ξ, we have $\|\nabla f(\boldsymbol{x},\xi)\| \leq G$, we have

$$\|\nabla f(\boldsymbol{x})\| = \|\mathbb{E}_\xi\nabla f(\boldsymbol{x},\xi)\| \leq \mathbb{E}_\xi\|\nabla f(\boldsymbol{x},\xi)\| \leq G.$$

For AFGA, the model difference $\bar{\Delta}_r^k$ on cluster k satisfies,

$$\bar{\Delta}_r^k = \bar{\boldsymbol{x}}_{r,\mathcal{I}}^k - \boldsymbol{x}_r = -\eta_l\sum_{t=0}^{\mathcal{I}-1}\bar{\boldsymbol{g}}_{r,t}^k,$$

therefore,

$$\|\bar{\Delta}_r^k\| = \left\|\eta_l\sum_{t=0}^{\mathcal{I}-1}\bar{\boldsymbol{g}}_{r,t}^k\right\| = \left\|\eta_l\sum_{t=0}^{\mathcal{I}-1}\frac{1}{K}\sum_{i\in\mathcal{V}_k}\boldsymbol{g}_{r,t}^i\right\| \leq \eta_l\mathcal{I}G,$$

for the global model difference Δ_r,

$$\|\Delta_r\| = \left\| \frac{1}{K} \sum_{k\in[K]} \bar{\Delta}_r^k \right\| \leq \eta_l \mathcal{I} G.$$

Thus we can obtain the bound for momentum $\boldsymbol{m}_r$ and variance $\boldsymbol{v}_r$,

$$\|\boldsymbol{m}_r\| = \left\| (1-\beta_1) \sum_{\mathcal{I}=1}^{t} \beta_1^{t-\mathcal{I}} \Delta_r \right\| \leq \eta_l \mathcal{I} G, \quad \|\boldsymbol{v}_r\| = \left\| (1-\beta_2) \sum_{\mathcal{I}=1}^{t} \beta_2^{t-\mathcal{I}} \Delta_r^2 \right\| \leq \eta_l^2 \mathcal{I}^2 G^2.$$

This concludes the proof. □

Lemma G.9. *For the variance difference sequence* $\widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2}$*, we have*

$$\sum_{r=1}^{R} \left\| \widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2} \right\|_1 \leq \frac{d}{\sqrt{\epsilon}}, \quad \sum_{r=1}^{R} \left\| \widehat{\boldsymbol{V}}_{r-1}^{-1/2} - \widehat{\boldsymbol{V}}_r^{-1/2} \right\|^2 \leq \frac{d}{\epsilon} \tag{95}$$

Proof. The proof of Lemma G.9 is exactly the same as the proof of Lemma C.2 in (Wang et al., 2022b). □

Lemma G.10. *For the element-wise difference,* $W_r = \frac{1}{\sqrt{\boldsymbol{v}_r+\epsilon}} - \frac{1}{\sqrt{\beta_2 \boldsymbol{v}_{r-1}+\epsilon}}$*, we have* $\|W_r\| \leq \frac{\sqrt{1-\beta_2}}{\epsilon} \|\Delta_r\|$.

Proof. The proof of Lemma G.9 is exactly the same as the proof of Lemma C.1 in (Wang et al., 2022b). □