Accelerating Adaptive Federated Optimization with Local Gossip Communications

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, adaptive federated optimization methods, such as FedAdam and FedAMSGrad, have gained increasing attention for their fast convergence and stable performance especially in training models with heavy-tail stochastic gradient distributions. However, the implementation of such methods still faces several bottlenecks, such as the large client-to-server communication overhead and the intense sensitivity to heterogeneous data. More importantly, the two objectives may conflict with each other, i.e., the convergence rate gets worse as the number of local steps increases in the partial participation setting, making it challenging to further improve the efficiency of adaptive federated optimization. We refer this problem as the *dilemma of local steps*. In this paper, we propose a novel hybrid adaptive federated optimization method (HA-Fed) where the clients are partitioned into disjoint clusters inside which they can communicate by fast client-to-client links. We show that HA-Fed resolves the *dilemma of local steps* in prior adaptive federated optimization methods, i.e., achieves a faster convergence rate as the local steps increases, while reducing the client-to-server communication overhead under non-i.i.d. settings. Specifically, HA-Fed improves the convergence rate from $\mathcal{O}(\sqrt{\tau}/\sqrt{TM})$ in FedAMSGrad to $\mathcal{O}(1/\sqrt{T\tau M})$ in partial participation scenarios under nonconvex stochastic setting. Extensive experiments and ablation studies demonstrate the effectiveness and broad applicability of our proposed method.

1 Introduction

Federated Learning [24, 31] has become a crucial large-scale machine learning paradigm where multiple clients jointly train a machine learning model coordinated by a central server. Unlike the traditional centralized training, where data is stored in a single central server, in federated learning, training data are stored on edge devices such as smartphones, hospital databases, or IoT devices, and only the local trained models are iteratively exchanged and synchronized to the central server. FedAvg [31] (also known as Local SGD [35]) has become one of the most popular federated optimization methods, where each client locally performs multiple steps of SGD updates then aggregates together for the global model update. Aside from the advantage of data privacy protection, the design of multiple local update steps also intends to reduce the communication between the server and clients. Compared with distributed learning [31, 35] where each local update step is followed by server aggregation, federated learning can further reduce the communication rounds. Recently, as the booming interests in training large-scale models such as BERT [8], GPT-3 [3] and ViT[9], adaptive federated optimization methods such as FedAdam [32], FedAGM [38] and FedAMS [43] has also been proposed and attracted a lot of attention. Specifically, adaptive federated optimization retains the multiple steps of SGD update on local clients but changes the global update of FedAvg from one-step SGD to one-step adaptive gradient methods update. By introducing adaptivity into federated learning, it achieves fast convergence, especially for heavy-tail stochastic gradient noise distributions.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

While various adaptive federated optimization algorithms have been proposed, there still exist several key bottlenecks in applying adaptive federated optimization in practice, such as (1) *large client-to-server communication overhead* due to the limited bandwidth and repetitive transmission between the server and clients; (2) *intense sensitivity on data heterogeneity* [32, 43] since nonidentical data distribution on different clients introduce extra variance between clients and slow down the training process of federated learning. What's even worse, these two objectives may conflict with each other: while multiple steps of local updates and the use of a partial participation strategy can certainly reduce the communication rounds between the server and clients, it has been shown that the variance overhead term grows as the number of local steps increases in partial participation settings, which leads to worse convergence rate in adaptive federated optimization [32, 43]. Such worse convergence result is largely due to data heterogeneity as in the i.i.d setting, the increasing of local steps can indeed lead to a better convergence rate [32, 43]. In this work, we refer this problem as the *dilemma of local steps*. Similar issues have also been shown in FedAvg that multiple local SGD steps may cause over-fitting on local clients, also known as client-drift, which slows down the convergence or leads to an unstable result [21]. This motivates us to study the following question:

*Can we resolve the **dilemma of local steps** for adaptive federated optimization? i.e., achieving a faster convergence rate as the number of local steps increases while reducing the client-to-server communication overhead under non-i.i.d. settings?*

Note that previous studies have shown that traditional variance reduction techniques [19, 11] can help reduce the client-drift and improve the convergence rate in FedAvg by additionally computing and communicating a control variate or the full batch gradient [20, 21]. However, it still remains an open problem how to apply such variance reduction techniques to adaptive federated optimization as it requires precise characterization of each local SGD iteration, which is incompatible with adaptive federated optimization whose current analysis can only give the characterization of cumulative gradient estimators between two communication rounds. Therefore, we take a different route here to solve the *dilemma of local steps* in adaptive federated optimization: Since the core idea of variance reduction is to introduce large-batch gradients for gradient correction, we could mimic the effect of having large-batch gradients by enabling the local client-to-client communications similar to gossip averaging in decentralized training [7, 14] for reducing the dissimilarity variance between clients. Specifically, in this paper, we propose a novel hybrid adaptive federated optimization method, HA-Fed, which benefits from both adaptive federated optimization [32, 38, 43] and techniques in decentralized training [28, 23, 27]. HA-Fed is structured by partitioning a global network into disjoint network clusters, where clients in the same cluster are connected via intra-cluster client-to-client communication links. These intra-cluster communications are fast and frequent, which incurs neglectable extra communication overhead compared with client-to-server communication links. Our contributions can be summarized as follows:

1. We propose a new hybrid adaptive federated optimization method, HA-Fed, which resolves the *dilemma of local steps* in prior adaptive federated optimization methods, i.e., achieves faster convergence rate as the local steps increases, while reducing the client-to-server communication overhead under non-i.i.d. settings.

2. We provide theoretical analysis for our proposed HA-Fed in the stochastic nonconvex optimization settings. We prove that HA-Fed achieves a faster convergence rate than FedAMSGrad [1] on the non-dominant term in full participation scenarios. Moreover, we show that in the practical partial participation setting, HA-Fed improves the convergence rate (dominant term) from $\mathcal{O}(\sqrt{\tau}/\sqrt{TM})$ [43] to $\mathcal{O}(1/\sqrt{T\tau M})$ w.r.t. global iterations T, local update steps τ and the number of participation clients M.

3. Extensive experiments are conducted on several benchmarks dataset and show that our proposed HA-Fed effectively saves the client-to-server communication overhead while achieving faster convergence with heterogeneous data. Extensive ablation studies also show the broad applicability of our proposed method.

[1] The convergence rate of FedAMSGrad is obtained from the convergence analysis for FedAMS [43], where FedAMSGrad gets a similar convergence to FedAMS. FedAMSGrad is also included in [38].

2 Preliminaries on Adaptive Federated Optimization

Firstly, let's begin with the general federated learning problem under nonconvex stochastic optimization settings. Suppose we have N local clients, and our goal is to minimize the following objective:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}), \tag{2.1}$$

where $\mathbf{x}$ denotes the model parameters, $f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} f_i(\mathbf{x}, \xi_i)$ is the local nonconvex loss function corresponding to the i-th client, and $\mathcal{D}_i$ is the local data distribution associated with client i. FedAvg [31] is a popular optimization algorithm to solve equation (2.1), with the sequential implementation of local SGD updates and global averaging.

Adaptive federated optimization is then proposed to incorporate adaptivity in federated optimization methods by replacing the global averaging in FedAvg with one-step adaptive gradient optimization. For example, FedAMSGrad is designed with multi-steps of local SGD updates and followed by one step of global AMSGrad[33] update. Specifically, at global round t, the server broadcasts the model $\mathbf{x}_t$ to selected clients. Each client i conducts τ steps of local SGD updates with local learning rate η_l and obtains the local model $\mathbf{x}^i_{t,\tau}$. The model difference $\Delta^i_t = \mathbf{x}^i_{t,\tau} - \mathbf{x}_t$ is aggregated to the server, and the server averages them to Δ_t. The server updates the global model $\mathbf{x}_{t+1}$ by taking Δ_t as a pseudo gradient for calculating momentum $\mathbf{m}_t$ and variance $\mathbf{v}_t$ for AMSGrad optimizer, and performs one step AMSGrad update with global learning rate η, i.e.,

$$\begin{aligned}
&\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\Delta_t, \mathbf{v}_t = \beta_2\mathbf{v}_{t-1} + (1-\beta_2)\Delta_t^2,\\
&\widehat{\mathbf{v}}_t = \max\{\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t\}, \mathbf{x}_{t+1} = \mathbf{x} + \frac{\mathbf{m}_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}},
\end{aligned} \tag{2.2}$$

then it obtains model $\mathbf{x}_{t+1}$ after one global round.

The convergence of FedAMSGrad is affected by several factors such as the number of local steps τ, global rounds T, and the number of participating clients M. In full participation setting, where M equals the total number of clients N, FedAMSGrad enjoys a $\mathcal{O}(1/\sqrt{T\tau N})$ rate of convergence [43]. This suggests that even for heterogeneous data, a larger number of local steps can help save the client-to-server communication cost and lead to faster convergence. However, previous study shows that under more practical partial participation setting, FedAMSGrad can only achieve a convergence rate of $\mathcal{O}(\sqrt{\tau}/\sqrt{TM})$ regards to the dissimilarity between clients under non-i.i.d. settings [43]. This suggests that while larger τ can help save the communication overhead, it also scarifies the fast rate of convergence. We refer this problem as the *dilemma of local steps*.

The *dilemma of local steps* arises due to the heterogeneous data inducing global variance term. This global variance is proportional to the number of local steps τ and affects the convergence. For full participation settings, it is fortunate that this variance overhead only appears on the non-dominant term, thus it does not slow down the overall convergence. However, for partial participation settings, the larger τ amplifies the over-fitting issue on local clients as fewer clients participate in each round of global training and thus leads to a worse convergence rate. In the following, we will try to resolve the *dilemma of local steps* with our proposed new hybrid adaptive federated optimization method.

3 Proposed Method

In this paper, we propose a hybrid adaptive federated optimization method (HA-Fed) where the clients are partitioned into disjoint clusters inside which they can communicate by fast client-to-client links, and clusters communicate with the central server with client-to-server communication links. Specifically, assuming we have one central server and K disjoint clusters, each of which contains n local clients and there are connected by client-to-client links (denoted by the adjacency matrix W_k). Let's denote the total number of clients as $N = Kn$. Our goal is to solve the following optimization problem:

$$\min_{x\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} \bar{f}_k(\mathbf{x}), \tag{3.1}$$

where $f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} f_i(\mathbf{x}, \xi_i)$ is the nonconvex loss function for the i-th client, and $\bar{f}_k(\mathbf{x}) := \frac{1}{n}\sum_{i\in\mathcal{V}_k} f_i(\mathbf{x})$ is the average loss on cluster k. We consider $\mathcal{V}_k$ as the set of local clients in the cluster k, and clients in cluster k are linked by a connected graph $\mathcal{G}_k$[2].

In order to accelerate FedAMSGrad under heterogeneous data settings, our HA-Fed starts from FedAMSGrad and introduces intra-cluster gossip communications. Gossip communication is designed for clients in a network to communicate with their neighbors without a central server, and it has been a popular approach in decentralized learning [28, 23, 7]. Our proposed HA-Fed adds frequent client-to-client gossip communication inside each cluster to leverage the over-fitting issue within the cluster. These gossip communications rely on inexpensive client-to-client communications without incurring extra client-to-server communication rounds, but at the same time, prevent over-fitting on local clients since the model on each client sufficiently communicates with their neighbors.

Algorithm 1 summarizes the proposed HA-Fed in full participation scenarios. The major difference between HA-Fed and FedAMSGrad lies in the local update step within each cluster (Line 9 in Algorithm 1): at the s-th step of intra-cluster training for cluster k, after client i finishes their local update and obtains $\mathbf{x}^i_{t,s+\frac{1}{2}}$ by one step SGD, we conduct one gossip averaging step within the cluster, i.e., let each client communicate with its neighbors $\mathcal{N}^i_k$ and aggregate the nearby local models with a weighted matrix W_k. The rest part of the algorithm is similar to FedAMSGrad.

In order to further reduce client-to-server communication rounds, we also adapt HA-Fed in partial participation setting[3]. Generally, in partial participation settings, the server samples a subset of m clients in each cluster before each round starts and only broadcasts the current model to these m selected clients and the selected clients will broadcast the received model to other clients within the same cluster with client-to-client links. For global model updates, all selected clients send the model difference Δ^i_t to the central server, and the server aggregates them to Δ_t. The rest of the partial participation update is the same as full participation scenarios.

Algorithm 1 HA-Fed:full participation

Input: initial point $\mathbf{x}_1$, global step size η, local step size η_l, $\beta_1, \beta_2, \epsilon$, mixing matrix W_k for all clusters $k \in [K]$

1: $\mathbf{m}_0 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$
2: **for** $t = 1$ to T **do**
3: **for** each cluster $k \in [K]$ in parallel **do**
4: **for** each client $i \in \mathcal{V}_k$ in parallel **do**
5: Receive model from the server: $\mathbf{x}^i_{t,0} = \mathbf{x}_t$
6: **for** $s = 0, ..., \tau - 1$ **do**
7: Compute local stochastic gradient: $\mathbf{g}^i_{t,s} = \nabla F_i(\mathbf{x}^i_{t,s}; \xi^i_{t,s})$
8: Local update: $\mathbf{x}^i_{t,s+\frac{1}{2}} = \mathbf{x}^i_{t,s} - \eta_l \mathbf{g}^i_{t,s}$
9: Gossip communication: $\mathbf{x}^i_{t,s+1} = \sum_{j\in\mathcal{N}^i_k} (W_k)_{i,j} \mathbf{x}^j_{t,s+\frac{1}{2}}$
10: **end for**
11: Get the model difference: $\Delta^i_t = \mathbf{x}^i_{t,\tau} - \mathbf{x}_t$
12: **end for**
13: **end for**
14: Server gets model difference: $\Delta_t = \frac{1}{K}\sum_{k\in[K]} \frac{1}{n}\sum_{i\in\mathcal{V}_k} \Delta^i_t$
15: Update: $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\Delta_t$
16: Update: $\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\Delta^2_t$
17: $\widehat{\mathbf{v}}_t = \max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t)$ and $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t + \epsilon)$
18: Server updates $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{\mathbf{m}_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}}$
19: **end for**

In a nutshell, HA-Fed takes advantage of both adaptive federated optimization and decentralized training: utilizes the fast local client-to-client links to reduce the impact of heterogeneous data and resolve the *dilemma of local steps* for adaptive federated optimization. The intra-cluster communication steps avoid over-fitting in each client, hence alleviating the effect of heterogeneous data between

[2]The connected graph implies there is a path from any client to any other client in the graph.

[3]Due to the space limit, see details in Algorithm 2 in the Appendix.

clients, while the client-to-server communication enables joint training cross clusters, which reduces the inter-cluster heterogeneity. These communications allow the proposed HA-Fed to efficiently train models in adaptive federated optimization and reduce the variance overhead. In particular, for the partial participation setting, although HA-Fed has only M clients involved in the global model update at each round, all clients within each cluster still perform local training and frequently communicate through intra-cluster client-to-client links. While for traditional FedAMSGrad, only M clients are actively training the local models at each round in the partial participation setting. This justifies the faster convergence of HA-Fed compared to FedAMSGrad in the partial participation setting.

4 Convergence Analysis

In this section, we provide the theoretical convergence analysis of the proposed HA-Fed method. Due to space limits, we leave the Assumptions for the theoretical analysis in the Appendix C.1, and the detailed analysis of both full and partial participation settings in the Appendix C.2.

Full Participation. We study the convergence behaviour of HA-Fed under full participation scenarios.

Theorem 4.1 (HA-Fed full participation)**.** Under Assumptions C.1-C.6, if the local learning rate satisfies the following condition $\eta_l \leq \min\big\{\frac{\sqrt[4]{\epsilon}}{\alpha\sqrt{CC_0\tau(\tau+\rho_{\max}^2 D_{\tau,\rho})}}, \frac{\epsilon}{2\tau C_0 C_{\beta,\eta}}\big\}$, then the iterates of Algorithm 1 satisfy

$$\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq 8(\beta_2\eta_l^2\tau^2G^2+\epsilon)^{\frac{1}{2}}\left\{\frac{f_0-f_*}{\eta\eta_l\tau T}+\frac{\Psi}{T}+\Phi_1+\Phi_2\right\}, \tag{4.1}$$

where $\Psi = \frac{C_\beta G^2 d}{\sqrt{\epsilon}} + \frac{2C_\beta^2\eta\eta_l\tau L G^2 d}{\epsilon}$, $\Phi_1 = \frac{CL^2\eta_l^2}{4\sqrt{\epsilon}}\Big[\tau^2\sigma_g^2 + \tau\rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2 + \tau\sigma^2\big(\frac{1}{n}+\rho_{\max}^2\big)\Big]$, $\Phi_2 = C_{\beta,\eta}\frac{\eta_l}{2\epsilon N}\sigma^2$, $C_\beta = \frac{\beta_1}{1-\beta_1}$, $C_{\beta,\eta} = \big((C_\beta^2+3)\eta L + 2\sqrt{1-\beta_2}G\big)$, $C_0 = \sqrt{\beta_2\tau^2G^2+\epsilon}$, C is a constant irrelevant to parameters, $\rho_{\max} = \max_{k\in[K]}\rho_k$ is the maximum spectral gap of all K clusters, $D_{\tau,\rho} = \min\big\{\frac{1}{1-\rho_{\max}}, \tau\big\}$ describes the density and connectivity of clusters and $\bar{\sigma}_L^2 = \frac{1}{K}\sum_{k=1}^K \sigma_k^2$ is the average of intra-cluster variance.

Corollary 4.2. Suppose we pick the global learning rate $\eta = \Theta(\sqrt{\tau N})$ and local learning rate $\eta_l = \Theta\big(\frac{1}{\sqrt{T}\tau}\big)$, when T is sufficient large, i.e., $T \geq \tau N$, the convergence rate for HA-Fed (Algorithm 1) under full participation satisfies $\min_{t\in[T]}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] = \mathcal{O}\big(\frac{1}{\sqrt{T\tau N}}\big)$.

This shows that HA-Fed achieves the same convergence rate $\mathcal{O}\big(\frac{1}{\sqrt{T\tau N}}\big)$ as FedAMSGrad [43].

Partial Participation. In such settings, we assume that only selected clients participate in each round of global synchronization. Assume the sampling strategy is random sampling without replacement in each cluster. Generally, at the beginning of global iteration t, the server samples a subset $\mathcal{S}_t^k$ for cluster k that contains m clients, these $M = Km$ clients receive the model from the server and synchronize their model difference for the global update.

Theorem 4.3 (HA-Fed partial participation)**.** Under Assumptions C.1-C.6, if the local learning rate satisfies $\eta_l \leq \min\big\{\frac{1}{4C_0C_{\beta,\eta}(\tau-1)}, \frac{\sqrt[4]{\epsilon}}{2\alpha L\sqrt{\widetilde{C}C_0\tau(\tau+\rho_{\max}^2D_{\tau,\rho})}}, \frac{1}{128\alpha^2\widetilde{C}C_0C_{\beta,\eta}\rho_{\max}^2D_{\tau,\rho}}\big(\frac{n-m}{m(n-1)}+\frac{1}{\tau^2}\big)^{-1}\big\}$, then the iterates of Algorithm 1 in partial participation scenarios satisfy

$$\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq 8(\beta_2\eta_l^2\tau^2G^2+\epsilon)^{\frac{1}{2}}\left\{\frac{f_0-f_*}{\eta\eta_l\tau T}+\frac{\Psi}{T}+\Phi_1+\Phi_2+\Phi_3+\Phi_4\right\}, \tag{4.2}$$

where Ψ and Φ_1 are same as theorem 4.1, $\Phi_2 = C_{\beta,\eta}\big[1+\big(\frac{n-m}{m}\big)\rho_{\max}^2\big]\frac{\eta_l}{\epsilon N}\sigma^2$, $\Phi_3 = \widetilde{C}\cdot C_{\beta,\eta}\frac{n-m}{m(n-1)}D_{\tau,\rho}\rho_{\max}^2\eta_l\big[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\big]$, $\Phi_4 = \widetilde{C}L^2C_{\beta,\eta}D_{\tau,\rho}\rho_{\max}^2\eta_l^3\big[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\big]$ where C and $\widetilde{C}$ are constants irrelevant to parameters, $C_{\beta,\eta} = \big((C_\beta^2+3)\eta L+2\sqrt{1-\beta_2}G\big)$ and $\rho_{\max}, D_{\tau,\rho}, \bar{\sigma}_L^2, C_\beta$ are same defined as Theorem 4.1.

Corollary 4.4. If there is $K \leq n-1$, i.e., the amount of clients in each cluster is more than the amount of clusters in the network, and the maximum spectral gap of all the networks satisfies $\rho_{\max} \leq \sqrt{\frac{m}{4n(n-m)}}$, then by choosing the global learning rate $\eta = \Theta(\sqrt{\tau N})$ and local learning rate

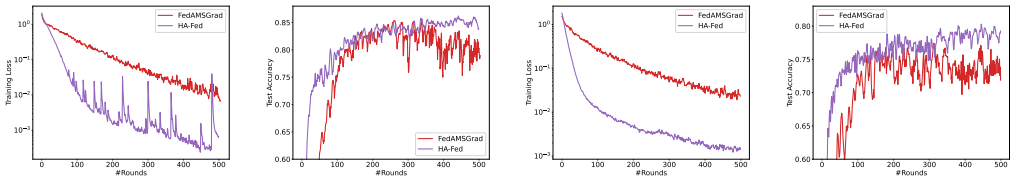
$\eta_l = \Theta(\frac{1}{\sqrt{T}\tau})$, when T is sufficient large, i.e., $T \geq \tau N$, the dominant term of the convergence rate for HA-Fed with partial participation scheme is $\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] = \mathcal{O}\big(\frac{f_0 - f_*}{\sqrt{T\tau N}} + \frac{\sigma^2+\sigma_l^2+\sigma_g^2}{\sqrt{T\tau N}}\big)$.

Corollary 4.4 shows that the HA-Fed achieves the same convergence rate as full participation settings with an additional condition on the spectral gap $\rho_{\max}$. This convergence rate significantly outperforms the rate of $\mathcal{O}\big(\frac{\sqrt{\tau}\tilde{\sigma}_g^2}{\sqrt{TM}}\big)$ in partial participation FedAMSGrad.

5 Experiments

Experimental Setup: We compare our proposed HA-Fed with FedAMSGrad, on CIFAR10 [25] dataset using (1) ResNet-18 [15] model, and (2) ConvMixer[4] model [39]. For HA-Fed, the global network topology is set up with 32 total clients, and they are equally divided into 4 clusters where each cluster contains 8 clients. We set the default partial participation ratio as $p = 0.25$, i.e., 2 clients participated per cluster per round. We adopt ring topology for all clusters by default with maximum spectral gap $\rho_{\max} = 0.805$. For FedAMSGrad, we set the number of clients and the partial participation ratio the same, i.e., 32 clients in total and 8 clients synchronize to the central server in each round. For both methods, we conduct $E = 2$ epochs of local training with a batch size of 50. We search for the best training hyper-parameter for both models. Due to the space limit, we leave the synthetic simulation studies, hyper-parameter and non-i.i.d. data sampling details, additional experiments on more datasets and ablation studies in Appendix H.

Figure 1 shows the convergence result of HA-Fed and FedAMSGrad on training CIFAR-10 with ResNet-18 and ConvMixer-256-8 model. We compare the training loss and test accuracy against global rounds for both models. For the ResNet-18 model, HA-Fed achieves faster convergence than FedAMSGrad in reducing training loss, and HA-Fed obtains an overall higher and stabler result in test accuracy. For the ConvMixer-256-8 model, HA-Fed again shows its faster convergence speed on training loss, in the meantime, HA-Fed still holds a higher test accuracy compared to FedAMSGrad under the same settings.



(a) ResNet-18 (b) ConvMixer-256-8

Figure 1: The learning curves for HA-Fed and FedAMSGrad in training CIFAR-10 data on (a) ResNet-18 model and (b) ConvMixer-256-8 model using ring topology for intra-cluster communications.

6 Conclusions

In this paper, we propose a novel hybrid adaptive federated optimization algorithm, HA-Fed, that overcomes the *dilemma of local steps* and achieves a faster convergence rate while reducing the client-to-server communication overhead under non-i.i.d. settings. HA-Fed mitigates the impact of data heterogeneity by adding inexpensive client-to-client communications hence resolving the *dilemma of local steps* without extra client-to-server communications. We present a completed theoretical convergence analysis for the proposed HA-Fed. We prove that HA-Fed achieves a faster convergence rate than the previous adaptive federated optimization method for both full and partial participation scenarios with heterogeneous data under nonconvex stochastic settings. Experiments on several benchmarks and ablation studies verify our theory.

[4] ConvMixer shares similar ideas to vision transformer [9] to use patch embeddings to preserve locality and similarly, and it is trained via adaptive gradient methods by default.

References

[1] 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)*, pages 8866–8870. IEEE, 2020.

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

[3] 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.

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

[5] 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.

[6] Mingzhe Chen, Nir Shlezinger, H Vincent Poor, Yonina C Eldar, and Shuguang Cui. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17), 2021.

[7] 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*, pages 1791–1802. PMLR, 2021.

[8] 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.

[9] 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.

[10] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[11] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

[13] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[14] 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.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[17] Divyansh Jhunjhunwala, Advait Gadhikar, Gauri Joshi, and Yonina C Eldar. Adaptive quantization of model updates for communication-efficient federated learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3110–3114. IEEE, 2021.

[18] Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

[19] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[20] 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*, 2020.

[21] 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*, pages 5132–5143. PMLR, 2020.

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

[23] 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*, pages 5381–5393. PMLR, 2020.

[24] 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.

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

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

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

[28] 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.

[29] 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)*, pages 1–6. IEEE, 2020.

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

[31] 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*, pages 1273–1282. PMLR, 2017.

[32] 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.

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

[34] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, 2020.

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

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

[37] 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.

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

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

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

[41] 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.

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

[43] Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. *arXiv preprint arXiv:2205.02719*, 2022.

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

[45] 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.

[46] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[47] 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 Related Work

Federated learning: Federated learning [24] has attracted growing interest recently due to the demand for training models locally at edge devices and the requirements of privacy protection. Federated optimization methods such as SGD-based optimization algorithm, FedAvg [31], also known as Local SGD [35], have been widely used in federated learning. Aside from FedAvg, since adaptive gradient methods such as Adam [22] and its variant AMSGrad [33] overcame the sensitivity to parameters and slow to convergence issue of SGD, adaptive federated optimizations such as FedAdam[32], FedAGM [38] and FedAMS [43] studied the corresponding adaptive optimization algorithms in federated learning. Moreover, several works [16, 12, 21, 26, 45] addressed and focused on the data heterogeneity issues of federated learning, where [21] proposed a federated learning variance reduction method that overcomes the data heterogeneity, but it requires extra communication costs for variance reduction operations. Hierarchical federated learning algorithms [29, 1, 4] are developed by aggregating client models to edge servers firstly before synchronizing them to the central server. However, this structure only considers the client-to-server communication, which is less helpful for solving the bottleneck of data heterogeneity.

Decentralized learning and other frameworks: Decentralized learning is a large-scale machine learning paradigm without a central server. It has been firstly studied from gossip averaging techniques [40, 2]. Decentralized (gossip) SGD algorithms [28, 27, 2, 36] are then proposed that consider client-to-client communications after each step of SGD update on the client. [30] proved a tight lower bound for decentralized training under the nonconvex setting. [37] proposes a leader-distributed SGD algorithm that pulls workers to the currently best-performing model among all models, which also utilizes inexpensive gossip communication. Moreover, recent studies generalized various distributed SGD algorithms under unified frameworks [41, 23], where [41] included reducing communication costs and decentralized training in i.i.d. settings, [7, 14] study decentralized training for local SGD, and [23] studied a general network topology-changing gossip SGD methods that summarize several algorithms in distributed and federated learning.

Communication-efficient federated learning: In terms of reducing the communication overhead in federated learning, one of the common approaches is to save the communication bits when synchronizing, such as the compressed and quantized FedAvg-based methods [34, 18, 17, 6]. Note that the bit compression strategy is orthogonal to our hybrid adaptive federated learning framework and can potentially be combined to further reduce communication overheads.

B HA-Fed Algorithm for Partial Participation

In the following, we summarize the proposed HA-Fed in partial participation settings.

Algorithm 2 HA-Fed:partial participation

Input: initial point $\mathbf{x}_1$, global step size η, local step size η_l, $\beta_1, \beta_2, \epsilon$, mixing matrix W_k for all clusters $k \in [K]$

1: $\mathbf{m}_0 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$
2: **for** $t = 1$ to T **do**
3: **for** each cluster $k \in [K]$ in parallel **do**
4: Random sample a subset $\mathcal{S}_t^k$ of clients with $|\mathcal{S}_t^k| = m$
5: **for** each client $i \in \mathcal{S}_t^k$ in parallel **do**
6: Receive model from the server $\mathbf{x}_{t,0}^i = \mathbf{x}_t$
7: Intra-cluster broadcast: $\mathbf{x}_{t,0}^j = \mathbf{x}_t, j \in \mathcal{V}_k$
8: **end for**
9: **for** each client $i \in \mathcal{V}_k$ in parallel **do**
10: **for** $s = 0, ..., \tau - 1$ **do**
11: Compute local stochastic gradient: $\mathbf{g}_{t,s}^i = \nabla F_i(\mathbf{x}_{t,s}^i; \xi_{t,s}^i)$
12: $\mathbf{x}_{t,s+\frac{1}{2}}^i = \mathbf{x}_{t,s}^i - \eta_l \mathbf{g}_{t,s}^i$
13: $\mathbf{x}_{t,s+1}^i = \sum_{j \in \mathcal{N}_k^i} (W_k)_{i,j} \mathbf{x}_{t,s+\frac{1}{2}}^j$
14: **end for**
15: Get the model difference: $\Delta_t^i = \mathbf{x}_{t,\tau}^i - \mathbf{x}_t$
16: **end for**
17: **end for**
18: Collect model differences from selected clients: $\Delta_t = \frac{1}{K} \sum_{k \in [K]} \frac{1}{m} \sum_{i \in S_t^k} \Delta_t^i$
19: $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \Delta_t$
20: $\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \Delta_t^2$
21: $\widehat{\mathbf{v}}_t = \max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t)$ and $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t + \epsilon)$
22: Server update $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \frac{\mathbf{m}_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}}$
23: **end for**

The proposed HA-Fed for partial participation settings is similar to full participation except for the broadcast and synchronization steps (Line 5-8 and Line 18 in Algorithm 2). For the partial participation setting, the server sends the current model $\mathbf{x}_t$ to m selected clients in each cluster for the broadcast step. Then these clients who received the model from the server will further broadcast the model $\mathbf{x}_t$ to their neighbors in the same cluster via fast client-to-client communication links. For example, if 8 clients within one cluster are grouped with ring topology, and two clients are sampled in each round, i.e., $m = 2$, then the partial participation setting needs 2 server-to-client broadcast

rounds and 6 much cheaper client-to-client broadcast rounds. In contrast, the full participation setting needs 8 server-to-client broadcast rounds for the same cluster grouping strategy which is much more expensive. Moreover, only the selected clients send the model difference Δ_t^i to the server (Line 18 in Algorithm 2), which also reduces the client-to-server communication overhead. The rest of the algorithm is the same as full participation scenarios.

C Convergence of HA-Fed

Notation: We consider column vectors throughout this paper except special explanations. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, denote $\sqrt{\mathbf{x}}, \mathbf{x}^2, \mathbf{x}/\mathbf{y}$ as the element-wise square root, square, and division of the vectors. For vector $\mathbf{x}$ and matrix $\mathbf{A}$, $\|\cdot\|$ abbreviates the ℓ_2 norm of the vector and Frobenius norm of the matrix, i.e., $\|\mathbf{x}\| = \|\mathbf{x}\|_2$ and $\|\mathbf{A}\| = \|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_2$ denotes the spectral norm of matrix $\mathbf{A}$. We denote $\mathbf{1}$ as vector with all elements equal to 1 with appropriate dimension, and $\mathbf{I}$ as the identity matrix with appropriate dimension.

C.1 Assumptions

Assumption C.1 (Smoothness)**.** Each loss function on the i-th client $f_i(\mathbf{x})$ is L-smooth, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\left|f_i(\mathbf{x}) - f_i(\mathbf{y}) - \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\right| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$.

Assumption C.2 (Bounded Gradient)**.** Each loss function on the i-th client $f_i(\mathbf{x})$ has G-bounded stochastic gradient on ℓ_2, i.e., for all ξ, we have $\|\nabla f_i(\mathbf{x}, \xi)\| \leq G$.

Assumption C.3 (Bounded Local Variance)**.** Each stochastic gradient on the i-th client has a bounded local variance, i.e., for all $\mathbf{x}, i \in [m]$,we have $\mathbb{E}\big[\|\nabla f_i(\mathbf{x}, \xi) - \nabla f_i(\mathbf{x})\|^2\big] \leq \sigma^2$.

Assumption C.1 also implies the L-gradient Lipschitz condition, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, it is a standard assumption in nonconvex optimization problems, and has been also adopted in [22, 33, 26, 45]. Assumption C.2 is usually adopted in studying adaptive gradient methods [22, 33, 47, 5]. Assumption C.3 is frequently adopted in distributed and federated learning optimization problem [32, 45, 7, 42].

Assumption C.4 (Bounded Intra-Cluster Variance)**.** The local objective function on each cluster has a bounded intra-cluster variance, i.e., for all $\mathbf{x}, k \in [K]$, we have $\frac{1}{n}\sum_{i\in\mathcal{V}_k}\|\nabla f_i(\mathbf{x}) - \nabla \bar{f}_k(\mathbf{x})\|^2 \leq \sigma_k^2$.

Assumption C.5 (Bounded Inter-Cluster Dissimilarity)**.** There is the following relationship between the loss function on each cluster and the global loss function: for $\alpha \geq 1$ and $\sigma_g \geq 0$, there is $\frac{1}{K}\sum_{k\in[K]}\|\nabla \bar{f}_k(\mathbf{x})\|^2 \leq \alpha^2\|\nabla f(\mathbf{x})\|^2 + \sigma_g^2$.

The assumption of bounded intra-cluster variance and inter-cluster dissimilarity represent the data heterogeneity in the cluster and between clusters. The similar data heterogeneity assumption, which only considers the variance between local clients, is common in federated learning [32, 45] and decentralized learning [28, 27, 23].

Assumption C.6 (Mixing Matrix)**.** The local clients in each cluster k are connected in the graph $\mathcal{G}_k$, and the corresponding mixing matrix W_k is a doubly stochastic matrix with the fact: $W_k \in [0,1]^{n\times n}$, $W_k\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W_k = \mathbf{1}^\top$ and $null(\mathbf{I} - W_k) = span(\mathbf{1})$. We further assume the spectral gap ρ_k: there exists $\rho_k \in [0, 1)$ such that $\|W_k - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq \rho_k$.

The assumption of the mixing matrix is usually assumed for decentralized learning framework [23, 7, 14]. Specifically, $\rho_k = 0$ means the matrix W_k with all elements $\frac{1}{n}$, corresponding to a better connected graph $\mathcal{G}_k$ and $\rho_k \to 1$ means the matrix W_k tends to be elements with either 0 or 1, corresponding to a less connective graph. Several works [28, 27] alternatively assume the spectral gap ρ of a mixing 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.

C.2 Technicalities

We examine the proposed Algorithm 1 and 2, Theorem 4.1 and 4.3 with introducing several remarks.

First let's rewrite Theorem 4.1.

Theorem C.7 (HA-Fed full participation)**.** Under Assumptions C.1-C.6, if the local learning rate satisfies the following condition $\eta_l \leq \min\left\{\frac{\sqrt[4]{\epsilon}}{\alpha\sqrt{CC_0\tau(\tau+\rho_{\max}^2 D_{\tau,\rho})}}, \frac{\epsilon}{2\tau C_0 C_{\beta,\eta}}\right\}$, then the iterates of Algorithm 1 satisfy

$$\min_{t\in[T]} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq 8(\beta_2\eta_l^2\tau^2G^2+\epsilon)^{\frac{1}{2}}\left\{\frac{f_0-f_*}{\eta\eta_l\tau T}+\frac{\Psi}{T}+\Phi_1+\Phi_2\right\}, \tag{C.1}$$

where $\Psi = \frac{C_\beta G^2 d}{\sqrt{\epsilon}} + \frac{2C_\beta^2\eta\eta_l\tau LG^2 d}{\epsilon}$, $\Phi_1 = \frac{CL^2\eta_l^2}{4\sqrt{\epsilon}}\left[\tau^2\sigma_g^2 + \tau\rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2 + \tau\sigma^2\left(\frac{1}{n}+\rho_{\max}^2\right)\right]$, $\Phi_2 = C_{\beta,\eta}\frac{\eta_l}{2\epsilon N}\sigma^2$, $C_\beta = \frac{\beta_1}{1-\beta_1}$, $C_{\beta,\eta} = \left((C_\beta^2+3)\eta L + 2\sqrt{1-\beta_2}G\right)$, $C_0 = \sqrt{\beta_2\tau^2G^2+\epsilon}$, C is a constant irrelevant to parameters, $\rho_{\max} = \max_{k\in[K]}\rho_k$ is the maximum spectral gap of all K clusters, $D_{\tau,\rho} = \min\left\{\frac{1}{1-\rho_{\max}}, \tau\right\}$ describes the density and connectivity of clusters and $\bar{\sigma}_L^2 = \frac{1}{K}\sum_{k=1}^K \sigma_k^2$ is the average of intra-cluster variance.

The convergence rate equation (C.1) is composed of four terms. The first and second terms are related to T and vanish as T increases. The third term Φ_1 represents the variance overhead introduced by both stochastic variance, intra-cluster and inter-cluster dissimilarity. The last term Φ_2 represents the stochastic variance from all N clients. Note that only the variance term Φ_1 is related to the network structure and intra-cluster topology, while the other three terms are identical to the corresponding term in the convergence rate of N-clients FedAMSGrad. In the following, we make comparisons between our proposed HA-Fed and FedAMSGrad. We first discuss the difference in the variance term Φ_1. Specifically, the dependency of Φ_1 for HA-Fed w.r.t. $\rho_{\max}, \tau, \sigma, \sigma_g, \bar{\sigma}_L$ and η_l is as follows

$$\Phi_1 = \mathcal{O}\left(\eta_l^2\tau^2\sigma_g^2 + \eta_l^2\rho_{\max}^2\tau^2\bar{\sigma}_L^2 + \eta_l^2\left(\frac{1}{n}+\rho_{\max}^2\right)\tau\sigma^2\right), \tag{C.2}$$

and the corresponding variance term $\widetilde{\Phi}_1$ for FedAMSGrad is has the order $\mathcal{O}(\eta_l^2\tau^2\widetilde{\sigma}_g^2 + \eta_l^2\tau\sigma^2)$, where $\widetilde{\sigma}_g^2$ is the global variance obtaining by a similar assumption on clients' loss function , i.e., the loss function on each client of FedAMSGard satisfies $\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \widetilde{\sigma}_g^2$.

Remark C.8. Comparing the convergence rate between FedAMSGrad and our proposed HA-Fed in full participation settings:

$$\text{HA-Fed: } \mathcal{O}\left(\frac{f_0-f_*}{\sqrt{T\tau N}} + \frac{\sigma^2}{\sqrt{T\tau N}} + \frac{G^2}{T} + \frac{\sigma_g^2+\rho_{\max}^2\bar{\sigma}_L^2}{T} + \frac{(\frac{1}{n}+\rho_{\max}^2)\sigma^2}{T\tau} + \frac{\sqrt{N\tau}}{T^{3/2}}\right),$$

$$\text{FedAMSGrad: } \mathcal{O}\left(\frac{f_0-f_*}{\sqrt{T\tau N}} + \frac{\sigma^2}{\sqrt{T\tau N}} + \frac{G^2}{T} + \frac{\widetilde{\sigma}_g^2}{T} + \frac{\sigma^2}{T\tau} + \frac{\sqrt{N\tau}}{T^{3/2}}\right).$$

This suggests that the intra-cluster network topology is only associated with the non-dominant term of the HA-Fed's convergence rate. For dense intra-cluster networks corresponding to small $\rho_{\max}$, HA-Fed outperforms FedAMSGrad thanks to the frequent intra-cluster client-to-client communication that weakens the impact of heterogeneous data. We also observe that the dominant term of HA-Fed's convergence rate is $\mathcal{O}(1/\sqrt{T\tau N})$. This matches the rate for general federated nonconvex optimization methods in full participation such as FedAvg [46, 45], SCAFFOLD [21], FedAdam [32] and FedAMS [43].

Remark C.9. When $\rho_{\max} = 0$, i.e., clients in each cluster are fully connected with their neighbors, HA-Fed with N clients works the same as FedAMSGrad with K clients, each of which has a n-times larger batch size[5]. In such case, the variance term Φ_1 in the convergence rate reduces to $\mathcal{O}\left(\eta_l^2\tau^2\sigma_g^2 + \eta_l^2\tau\frac{\sigma^2}{n}\right)$. This matches the result of FedAMSGrad with a smaller stochastic gradient variance $\frac{\sigma^2}{n}$, where this smaller stochastic gradient variance is obtained by the sufficient model averaging in each cluster.

Remark C.10. When $\rho_{\max} \to 1$, i.e., clients in each cluster are tending to be disconnected[6], HA-Fed tends will reduce to FedAMSGrad with N clients. In such case, the variance term Φ_1 becomes to $\mathcal{O}\left(\eta_l^2\tau^2(\sigma_g^2+\bar{\sigma}_L^2) + \eta_l^2\tau\sigma^2\right)$, which matches the variance term $\widetilde{\Phi}_1$ of N-clients FedAMSGrad with an inter-client variance $\sigma_g^2 + \bar{\sigma}_L^2$.

[5]For example, in HA-Fed, suppose clients in each cluster have the batch size b, when $\rho_{\max} = 0$, we can view each cluster as a single client in FedAMSGrad with a batch size of nb (merges all the data within the cluster).

[6]Since we assume clients are linked with a connected graph, it implies that $\rho_{\max} < 1$.

Theorem C.11 (HA-Fed partial participation)**.** Under Assumptions C.1-C.6, if the local learning rate satisfies $\eta_l \leq \min\Big\{\frac{1}{4C_0C_{\beta,\eta}(\tau-1)}, \frac{\sqrt[4]{\epsilon}}{2\alpha L\sqrt{\widetilde{C}C_0\tau(\tau+\rho_{\max}^2 D_{\tau,\rho})}}, \frac{1}{128\alpha^2\widetilde{C}C_0C_{\beta,\eta}\rho_{\max}^2 D_{\tau,\rho}}\big(\frac{n-m}{m(n-1)}+\frac{1}{\tau^2}\big)^{-1}\Big\}$, then the iterates of Algorithm 1 in partial participation scenarios satisfy

$$\min_{t\in[T]}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq 8(\beta_2\eta_l^2\tau^2G^2+\epsilon)^{\frac{1}{2}}\left\{\frac{f_0-f_*}{\eta\eta_l\tau T}+\frac{\Psi}{T}+\Phi_1+\Phi_2+\Phi_3+\Phi_4\right\}, \tag{C.3}$$

where Ψ and Φ_1 are same as theorem C.7, $\Phi_2 = C_{\beta,\eta}\big[1+\big(\frac{n-m}{m}\big)\rho_{\max}^2\big]\frac{\eta_l}{\epsilon N}\sigma^2$, $\Phi_3 = \widetilde{C}\cdot C_{\beta,\eta}\frac{n-m}{m(n-1)}D_{\tau,\rho}\rho_{\max}^2\eta_l\big[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\big]$, $\Phi_4 = \widetilde{C}L^2C_{\beta,\eta}D_{\tau,\rho}\rho_{\max}^2\eta_l^3\big[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\big]$ where C and $\widetilde{C}$ are constants irrelevant to parameters, $C_{\beta,\eta} = \big((C_\beta^2+3)\eta L+2\sqrt{1-\beta_2}G\big)$ and $\rho_{\max}, D_{\tau,\rho}, \bar{\sigma}_L^2, C_\beta$ are same defined as Theorem C.7.

Remark C.12. Comparing the convergence rate between FedAMSGrad and our proposed HA-Fed in partial participation settings with the two largest dominant terms,

$$\text{FedAMSGrad: } \mathcal{O}\bigg(\frac{\sqrt{\tau}\widetilde{\sigma}_g^2}{\sqrt{TM}}+\frac{f_0-f_*}{\sqrt{T\tau M}}+\frac{\sigma^2}{\sqrt{T\tau M}}\bigg),$$
$$\text{HA-Fed: } \mathcal{O}\bigg(\frac{f_0-f_*}{\sqrt{T\tau M}}+\frac{\sigma^2+\sigma_g^2+\bar{\sigma}_L^2}{\sqrt{T\tau M}}+\frac{\sigma_g^2+\rho_{\max}^2\bar{\sigma}_L^2+G^2}{T}\bigg).$$

It suggests that in the partial participation setting, the dominant term in equation (C.3) is $\mathcal{O}\big(\frac{1}{\sqrt{T\tau M}}\big)$. This implies that the intra-cluster variance σ_g^2 and inter-cluster variance $\bar{\sigma}_L^2$ in Φ_1 dominate the convergence rate in HA-Fed with the same order as stochastic variance σ^2, but they do not sacrifice on the convergence rate as local steps τ increases. Compared with the convergence rate of $\mathcal{O}\Big(\frac{\sqrt{\tau}}{\sqrt{TM}}\Big)$ [43] achieved by FedAMSGrad in partial participation setting, HA-Fed successfully resolves the *dilemma of local steps*: larger number of local steps τ can achieve faster convergence while reducing the client-to-server communication overhead.

Remark C.13. When $\rho_{\max}\to 0$, same as full participation, clients in each cluster are fully connected with their neighbors, HA-Fed with N clients works the same as FedAMSGrad with K clients, each of which has a n-times larger batch size. In such case, the variance term Φ_1 in the convergence rate reduces to $\mathcal{O}\big(\eta_l^2\tau^2\sigma_g^2+\eta_l^2\tau\frac{\sigma^2}{n}\big)$, $\Phi_2 = \mathcal{O}\big(\frac{\eta_l\sigma^2}{N}\max\{\eta,1\}\big)$ and $\Phi_3=\Phi_4=0$. This matches the result of FedAMSGrad with a smaller stochastic gradient variance $\frac{\sigma^2}{n}$, where this smaller stochastic gradient variance is obtained by the sufficient model averaging in each cluster.

Remark C.14. When $\rho_{\max}\to 1$ and $K=1$, i.e., all clients are tending to disconnected, HA-Fed will reduce to FedAMSGrad with N clients. There is $D_{\tau,\rho} = \min\{\frac{1}{1-\rho_{\max}},\tau\} = \tau$, hence Φ_3 term in theorem 3 turns to $\widetilde{C}\cdot\big((C_\beta^2+3)\eta L+2\sqrt{1-\beta_2}G\big)\big(\frac{N-M}{M(N-1)}+\frac{1}{\tau^2}\big)\tau\eta_l\big[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+\tau\frac{\sigma^2}{\tau^2 n}\big] = \mathcal{O}(\frac{\eta\eta_l\tau}{M}[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2])$, by appropriately choosing learning rates η and η_l, this Φ_3 term results to the rate of $\mathcal{O}(\frac{\sqrt{\tau}}{\sqrt{TM}}[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2])$, which is slightly larger than the dominant term in FedAMSGrad with an additional stochastic variance σ^2. [7]

D Preliminaries

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

$$\bar{\mathbf{x}}_{t,s+1}^k = \bar{\mathbf{x}}_{t,s}^k - \eta_l\bar{\mathbf{g}}_{t,s}^k, \tag{D.1}$$

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

$$\bar{\mathbf{x}}_{t,s+1} = \bar{\mathbf{x}}_{t,s} - \eta_l\frac{1}{N}\sum_{i=1}^N\mathbf{g}_{t,s}^i. \tag{D.2}$$

[7] This additional stochastic variance is because the techniques for bounding intra-cluster consensus error involves extra stochastic variance bound.

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

$$\bar{\Delta}_t^k = \frac{1}{n}\sum_{i\in\mathcal{V}_k}\Delta_t^i = \frac{1}{n}\sum_{i\in\mathcal{V}_k}(\mathbf{x}_{t,\tau}^i - \mathbf{x}_t) = \bar{\mathbf{x}}_{t,\tau}^k - \mathbf{x}_t = \bar{\mathbf{x}}_{t,0}^k - \eta_l\sum_{s=0}^{\tau-1}\bar{\mathbf{g}}_{t,s}^k - \mathbf{x}_t = -\eta_l\sum_{s=0}^{\tau-1}\bar{\mathbf{g}}_{t,s}^k$$

$$\bar{\Delta}_t = \frac{1}{K}\sum_{k\in[K]}\frac{1}{n}\sum_{i\in\mathcal{V}_k}(\mathbf{x}_{t,\tau}^i - \mathbf{x}_t) = \frac{1}{K}\sum_{k\in[K]}\Delta_t^k = -\eta_l\frac{1}{K}\frac{1}{n}\sum_{s=0}^{\tau-1}\sum_{k\in[K]}\sum_{i\in\mathcal{V}_k}\mathbf{g}_{t,s}^i, \tag{D.3}$$

recall the definition of Δ_t

$$\Delta_t = \frac{1}{K}\sum_{k\in[K]}\frac{1}{m}\sum_{i\in S_t^k}\Delta_t^i = \frac{1}{K}\frac{1}{m}\sum_{k\in[K]}\sum_{i\in S_t^k}\mathbf{x}_{t,\tau}^i - \mathbf{x}_t, \tag{D.4}$$

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

$$\mathbb{E}_{S_t}[\Delta_t] = \bar{\Delta}_t = \frac{1}{K}\sum_{k\in[K]}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{t,\tau}^i - \mathbf{x}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{t,\tau}^i - \mathbf{x}_t. \tag{D.5}$$

E Proof of Theorem 4.1: HA-Fed full participation

For full participation cases, we have $\bar{\Delta}_t = \Delta_t$. Similar to previous works about adaptive methods [47, 5], we introduce a Lyapunov sequence $\mathbf{z}_t$: assume $\mathbf{x}_0 = \mathbf{x}_1$, for each $t \geq 1$, we have

$$\mathbf{z}_t = \mathbf{x}_t + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1}) = \frac{1}{1-\beta_1}\mathbf{x}_t - \frac{\beta_1}{1-\beta_1}\mathbf{x}_{t-1}. \tag{E.1}$$

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

$$\begin{aligned}
\mathbf{z}_{t+1} - \mathbf{z}_t &= \frac{1}{\beta_1}(\mathbf{x}_{t+1} - \mathbf{x}_t) - \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1}) \\
&= \frac{1}{1-\beta_1}(\eta\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t) - \frac{\beta_1}{1-\beta_1}\eta\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{m}_{t-1} \\
&= \frac{1}{1-\beta_1}\eta\widehat{\mathbf{V}}_t^{-1/2}\Big[\beta_1\mathbf{m}_{t-1} + (1-\beta_1)\Delta_t\Big] - \frac{\beta_1}{1-\beta_1}\eta\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{m}_{t-1} \\
&= \eta\widehat{\mathbf{V}}_t^{-1/2}\Delta_t - \eta\frac{\beta_1}{1-\beta_1}\Big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big)\mathbf{m}_{t-1}.
\end{aligned}$$

By Assumption C.1, since f is L-smooth, taking conditional expectation at time t, we have

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{t+1})] - f(\mathbf{z}_t) \\
&\leq \mathbb{E}[\langle\nabla f(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t\rangle] + \frac{L}{2}\mathbb{E}[\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2] \\
&\leq \eta\mathbb{E}\Big[\Big\langle\nabla f(\mathbf{z}_t), \widehat{\mathbf{V}}_t^{-1/2}\Delta_t\Big\rangle\Big] - \eta\mathbb{E}\Big[\Big\langle\nabla f(\mathbf{z}_t), \frac{\beta_1}{1-\beta_1}\Big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big)\mathbf{m}_{t-1}\Big\rangle\Big] \\
&\quad + \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_t^{-1/2}\Delta_t - \frac{\beta_1}{1-\beta_1}\Big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big)\mathbf{m}_{t-1}\Big\|^2\Big] \\
&= \underbrace{\eta\mathbb{E}\Big[\Big\langle\nabla f(\mathbf{x}_t), \widehat{\mathbf{V}}_t^{-1/2}\Delta_t\Big\rangle\Big]}_{I_1}\underbrace{-\eta\mathbb{E}\Big[\Big\langle\nabla f(\mathbf{z}_t), \frac{\beta_1}{1-\beta_1}\Big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big)\mathbf{m}_{t-1}\Big\rangle\Big]}_{I_2} \\
&\quad \underbrace{+ \frac{\eta^2 L}{2}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_t^{-1/2}\Delta_t - \frac{\beta_1}{1-\beta_1}\Big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big)\mathbf{m}_{t-1}\Big\|^2\Big]}_{I_3} + \\
&\quad + \underbrace{\eta\mathbb{E}\Big[\Big\langle\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \widehat{\mathbf{V}}_t^{-1/2}\Delta_t\Big\rangle\Big]}_{I_4},
\end{aligned} \tag{E.2}$$

E.1 Bounding I_1

We have

$$
\begin{aligned}
I_1 &= \eta \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}} \Big\rangle\Big] \\
&= \eta \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big\rangle\Big] + \eta \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}} - \frac{\Delta_t}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big\rangle\Big]. \qquad \text{(E.3)}
\end{aligned}
$$

For the second term in equation (E.3), we have

$$
\begin{aligned}
&\eta \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}} - \frac{\Delta_t}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big\rangle\Big] \\
&\leq \eta \|\nabla f(\mathbf{x}_t)\| \mathbb{E}\Big[\Big\| \frac{1}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}} - \frac{1}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big\| \cdot \|\Delta_t\|\Big] \\
&\leq \frac{\eta \sqrt{1-\beta_2} G}{\epsilon} \mathbb{E}[\|\Delta_t\|^2], \qquad \text{(E.4)}
\end{aligned}
$$

where the second inequality holds by Assumption C.2, Lemma G.8 and Lemma G.10. For the first term in equation (E.3), recall that $\Delta_t = -\frac{\eta_l}{N} \sum_{i=1}^N \sum_{s=0}^{\tau-1} \mathbf{g}_{t,s}^i$, we have

$$
\begin{aligned}
&\eta \mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big\rangle\Big] = \eta \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \Delta_t \Big\rangle\Big] \\
&= -\eta \eta_l \sum_{s=0}^{\tau-1} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N} \sum_{i=1}^N \mathbf{g}_{t,s}^i \Big\rangle\Big] \\
&= -\eta \eta_l \sum_{s=0}^{\tau-1} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \Big\rangle\Big], \qquad \text{(E.5)}
\end{aligned}
$$

where we have

$$
\begin{aligned}
&- \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \Big\rangle\Big] \\
&= -\frac{1}{2} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \Big\rangle\Big] \\
&\quad - \frac{1}{2} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \pm \frac{1}{K} \sum_{k=1}^K \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \Big\rangle\Big] \\
&= -\frac{1}{2} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \Big\rangle\Big] \\
&\quad - \frac{1}{2} \mathbb{E}\Big[\Big\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \Big(\frac{1}{N} \sum_{i=1}^N \nabla f_i(\mathbf{x}_{t,s}^i) \pm \frac{1}{K} \sum_{k=1}^K \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \Big) \Big\rangle\Big]. \qquad \text{(E.6)}
\end{aligned}
$$

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

$$
\begin{aligned}
&- \mathbb{E}\left[\left\langle \frac{\nabla f(\mathbf{x}_t)}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}}, \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \right\rangle\right] \\
&\leq -\frac{1}{4}\mathbb{E}\left[\left\| \frac{\nabla f(\mathbf{x}_t)}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \right\|^2 + \left\| \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \right\|^2 \right. \\
&\quad \left. - \left\| \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \left(\nabla f(\mathbf{x}_t) - \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \right) \right\|^2 \right] - \frac{1}{4}\mathbb{E}\left[\left\| \frac{\nabla f(\mathbf{x}_t)}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \right\|^2 \right. \\
&\quad \left. + \left\| \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right\|^2 - \left\| \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \left(\nabla f(\mathbf{x}_t) - \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right) \right\|^2 \right] \\
&\quad + \frac{1}{4}\mathbb{E}\left[\left\| \frac{\nabla f(\mathbf{x}_t)}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \right\|^2 \right] + \frac{1}{4}\mathbb{E}\left[\left\| \frac{1}{\sqrt[4]{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \left(\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) - \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right) \right\|^2 \right] \\
&\leq -\frac{1}{4C_0}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \frac{1}{4C_0}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \right\|^2 \right] - \frac{1}{4C_0}\mathbb{E}\left[\left\| \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right\|^2 \right] \\
&\quad + \frac{1}{4\sqrt{\epsilon}}\mathbb{E}\left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \right\|^2 \right] + \frac{1}{4\sqrt{\epsilon}}\mathbb{E}\left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right\|^2 \right] \\
&\quad + \frac{1}{4\sqrt{\epsilon}}\mathbb{E}\left[\left\| \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) - \frac{1}{K}\sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) \right\|^2 \right], \qquad \text{(E.7)}
\end{aligned}
$$

where the second inequality holds by $\frac{\|\mathbf{x}\|^2}{C_0} \leq \frac{\|\mathbf{x}\|^2}{\sqrt{\beta_2 \widehat{\mathbf{v}}_{t-1} + \epsilon}} \leq \frac{\|\mathbf{x}\|^2}{\sqrt{\epsilon}}$. Bounding the last three terms above are equal to bound the inter-cluster consensus error $\|\mathbf{x}_t - \bar{\mathbf{x}}_{t,s}^k\|$ and intra-cluster consensus error $\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_{t,s}^i\|$.

Merging pieces together, we can finally bound I_1 here.

$$
\begin{aligned}
I_1 &= \eta\mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\beta_2\widehat{\mathbf{v}}_{t-1}+\epsilon}}\Big\rangle\Big] + \eta\mathbb{E}\Big[\Big\langle \nabla f(\mathbf{x}_t), \frac{\Delta_t}{\sqrt{\widehat{\mathbf{v}}_t+\epsilon}} - \frac{\Delta_t}{\sqrt{\beta_2\widehat{\mathbf{v}}_{t-1}+\epsilon}}\Big\rangle\Big] \\
&\le \frac{\eta\eta_l}{4C_0}\sum_{s=0}^{\tau-1}\Big[-\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{s=0}^{\tau-1}\Big\{\mathbb{E}\Big[\Big\|\nabla f(\mathbf{x}_t) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2\Big] + \mathbb{E}\Big[\Big\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big]\Big\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2] \\
&\le \frac{\eta\eta_l}{4C_0}\sum_{s=0}^{\tau-1}\Big[-\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{s=0}^{\tau-1}\Big\{\mathbb{E}\Big[\Big\|\nabla f(\mathbf{x}_t) \pm \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2\Big] + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2] \\
&\quad + \mathbb{E}\Big[\Big\|\nabla f(\mathbf{x}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] + \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i) - \frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big]\Big\} \\
&\le \frac{\eta\eta_l}{4C_0}\sum_{s=0}^{\tau-1}\Big[-\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{s=0}^{\tau-1}\Big\{\frac{2L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_t - \bar{\mathbf{x}}_{t,s}^k\|^2] + \frac{2L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_{t,s}^i\|^2] + \frac{L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_t - \bar{\mathbf{x}}_{t,s}^k\|^2] \\
&\quad + \frac{L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_{t,s}^i\|^2]\Big\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2] \\
&\le \frac{\eta\eta_l}{4C_0}\sum_{s=0}^{\tau-1}\Big[-\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{s=0}^{\tau-1}\Big\{\frac{3L^2}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_t - \bar{\mathbf{x}}_{t,s}^k\|^2] + \frac{3L^2}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_{t,s}^i\|^2]\Big\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2] \\
&\le \frac{\eta\eta_l}{4C_0}\sum_{s=0}^{\tau-1}\Big[-\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\Big[3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2+\sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2 + D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Big] + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2], \quad \text{(E.8)}
\end{aligned}
$$

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

E.2 Bounding I_2

The bound for I_2 mainly follows by the update rule and definition of virtual sequence $\mathbf{z}_t$,

$$
\begin{aligned}
I_2 &= -\eta\mathbb{E}\Big[\Big\langle \nabla f(\mathbf{z}_t), \frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1}\Big\rangle\Big] \\
&= -\eta\mathbb{E}\Big[\Big\langle \nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t), \frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1}\Big\rangle\Big], \quad \text{(E.9)}
\end{aligned}
$$

then by Assumption C.1, we have

$$
\begin{aligned}
I_2 &\leq \eta\mathbb{E}\bigg[\|\nabla f(\mathbf{x}_t)\|\bigg\|\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|\bigg] \\
&\quad+\eta L\mathbb{E}\bigg[\|\mathbf{z}_t-\mathbf{x}_t\|\bigg\|\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|\bigg] \\
&= \eta\mathbb{E}\bigg[\|\nabla f(\mathbf{x}_t)\|\bigg\|\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|\bigg] \\
&\quad+\eta^2 L\mathbb{E}\bigg[\bigg\|\frac{1}{\sqrt{\widehat{\mathbf{v}}_{t-1}+\epsilon}}\frac{\beta_1}{1-\beta_1}\mathbf{m}_{t-1}\bigg\|\bigg\|\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|\bigg] \\
&\leq \eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big\|_1\big]+\eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big\|_1\big],
\end{aligned}
\tag{E.10}
$$

where the last inequality holds by Assumption C.2 and Lemma G.8 about bounding $\nabla f(\mathbf{x}_t)$ and $\mathbf{m}_t$.

E.3 Bounding I_3

We have the following result for bounding I_3

$$
\begin{aligned}
I_3 &= \frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\widehat{\mathbf{V}}_t^{-1/2}\Delta_t+\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|^2\bigg] \\
&\leq \eta^2 L\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_t^{-1/2}\Delta_t\big\|^2\big]+\eta^2 L\mathbb{E}\bigg[\bigg\|\frac{\beta_1}{1-\beta_1}\big(\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big)\mathbf{m}_{t-1}\bigg\|^2\bigg] \\
&\leq \frac{\eta^2 L}{\epsilon}\mathbb{E}[\|\Delta_t\|^2]+\eta^2 L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_{t}^{-1/2}\big\|^2\big],
\end{aligned}
\tag{E.11}
$$

where the first inequality follows by Cauchy-Schwarz inequality, and the second one follows by Assumption C.2 and Lemma G.8 about bounding $\nabla f(\mathbf{x}_t)$ and $\mathbf{m}_t$.

E.4 Bounding I_4

$$
\begin{aligned}
I_4 &= \mathbb{E}\bigg[\bigg\langle\nabla f(\mathbf{z}_t)-\nabla f(\mathbf{x}_t),\eta\widehat{\mathbf{V}}_t^{-1/2}\Delta_t\bigg\rangle\bigg] \\
&\leq \mathbb{E}\bigg[\|\nabla f(\mathbf{z}_t)-\nabla f(\mathbf{x}_t)\|\bigg\|\eta\widehat{\mathbf{V}}_t^{-1/2}\Delta_t\bigg\|\bigg] \\
&\leq L\mathbb{E}\bigg[\|\mathbf{z}_t-\mathbf{x}_t\|\bigg\|\eta\widehat{\mathbf{V}}_t^{-1/2}\Delta_t\bigg\|\bigg] \\
&\leq \frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\frac{\beta_1}{1-\beta_1}\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\bigg\|^2\bigg]+\frac{\eta^2 L}{2}\mathbb{E}\bigg[\bigg\|\widehat{\mathbf{V}}_t^{-1/2}\Delta_t\bigg\|^2\bigg] \\
&\leq \frac{\eta^2 L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|\mathbf{m}_t\|^2]+\frac{\eta^2 L}{2\epsilon}\mathbb{E}[\|\Delta_t\|^2],
\end{aligned}
\tag{E.12}
$$

where the first inequality holds due to Young's inequality, and the second one follows from Assumption C.1 and the definition of virtual sequence $\mathbf{z}_t$. By Lemma G.7, we have

$$
\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{m}_t\|^2]\leq\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2].
\tag{E.13}
$$

Therefore, the summation of I_4 term is bounded by

$$
\sum_{t=1}^{T}I_4\leq\bigg(\frac{\eta^2 L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{\eta^2 L}{2\epsilon}\bigg)\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2].
\tag{E.14}
$$

E.5 Merging pieces together

Summing I_1 to I_4 from $t=1$ to T, we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) = \sum_{t=1}^{T}[I_1 + I_2 + I_3 + I_4] \\
&\leq \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\Big[-\tau\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \sum_{s=0}^{\tau-1}\mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \sum_{s=0}^{\tau-1}\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Big\{3L^2\tau^2C_1\eta_l^2(\tau + \rho_{\max}^2 D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2 D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2 + D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Big\} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2] \\
&\quad + \Big(\eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\Big)\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big\|_1\Big] \\
&\quad + \frac{\eta^2L}{\epsilon}\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2] + \eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big\|^2\Big] \\
&\quad + \Big(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon}\Big)\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2],
\end{aligned}
$$

Merge the similar pieces, then we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\Big[-\tau\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \sum_{s=0}^{\tau-1}\mathbb{E}\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2 - \sum_{s=0}^{\tau-1}\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Big\{3L^2\tau^2C_1\eta_l^2(\tau + \rho_{\max}^2 D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2 D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2 + D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Big\} \\
&\quad + \Big(\eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\Big)\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big\|_1\Big] \\
&\quad + \eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\sum_{t=1}^{T}\mathbb{E}\Big[\Big\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\Big\|^2\Big] \\
&\quad + \Big(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\Big)\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2],
\end{aligned}
$$

Further apply Lemma G.9 and Lemma G.5, we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\Bigg[-\tau\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \sum_{s=0}^{\tau-1}\mathbb{E}\Bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Bigg\|^2 - \sum_{s=0}^{\tau-1}\Bigg\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Bigg\|^2\Bigg] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Bigg\{3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2+\sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2+D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Bigg\} \\
&\quad + \left(\eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\right)\frac{d}{\sqrt{\epsilon}} + \eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\cdot\frac{d}{\epsilon} \\
&\quad + \left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\sum_{t=1}^{T}\Bigg\{\frac{\eta_l^2\tau}{N}\sigma^2 + \eta_l^2\tau\sum_{s=0}^{\tau-1}\mathbb{E}\Bigg[\Bigg\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Bigg\|^2\Bigg]\Bigg\},
\end{aligned}
\tag{E.15}
$$

drop extra terms in the second line with the following condition on learning rate,

$$
\eta\eta_l^2\tau\frac{C_{\beta,\eta}}{2\epsilon} \leq \frac{\eta\eta_l}{4C_0} \Rightarrow \eta \leq \frac{\epsilon}{2\tau C_0C_{\beta,\eta}},
\tag{E.16}
$$

where $C_{\beta,\eta} = \left(\eta L\frac{\beta_1^2}{(1-\beta_1)^2} + 3\eta L + 2\sqrt{1-\beta_2}G\right) = \mathcal{O}(\max\{\eta, 1\})$, merge the similar items,

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq -\left[\frac{\eta\eta_l\tau}{4C_0} - \frac{\eta\eta_l}{4\sqrt{\epsilon}}3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\alpha^2\right]\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Bigg\{3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\sigma_g^2 + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2+D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Bigg\} \\
&\quad + \eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2\frac{d}{\sqrt{\epsilon}} + 2\eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\cdot\frac{d}{\epsilon} \\
&\quad + \left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\frac{T\eta_l^2\tau}{N}\sigma^2,
\end{aligned}
\tag{E.17}
$$

note that we need the following requirement for local learning rate η_l:

$$
\eta_l \leq \frac{\sqrt[4]{\epsilon}}{\alpha\sqrt{6C_0C_1\tau(\tau+\rho_{\max}^2D_{\tau,\rho})}}.
\tag{E.18}
$$

Thus we have

$$
\begin{aligned}
&\frac{\eta\eta_l\tau}{8C_0T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\leq \frac{\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_T)}{T} + \eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2\frac{d}{T\sqrt{\epsilon}} + 2\eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\cdot\frac{d}{T\epsilon} \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\Bigg(3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\sigma_g^2 \\
&\quad + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2+D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Bigg) \\
&\quad + \left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\frac{\eta_l^2\tau}{N}\sigma^2,
\end{aligned}
\tag{E.19}
$$

since we have $D_{\tau,\rho} \leq \tau$ and the maximum spectral gap satisfies $\rho_{\max} \leq 1$,

$$
\begin{aligned}
&3L^2\tau^2 C_1 \rho_{\max}^2 D_{\tau,\rho}\eta_l^2 \bar{\sigma}_L^2 + 3L^2\tau^2 C_1 \eta_l^2 \sigma^2 \rho_{\max}^2 + 3L^2 C_1 \big(\tau^2 + D_{\tau,\rho}^2 \rho_{\max}^2\big)\eta_l^2 \frac{\sigma^2}{n} \\
&= 3C_1 L^2 \eta_l^2 \left(\tau^2 \rho_{\max}^2 (D_{\tau,\rho}\bar{\sigma}_L^2 + \sigma^2) + (\tau^2 + D_{\tau,\rho}^2 \rho_{\max}^2)\frac{\sigma^2}{n}\right) \\
&\leq 3C_1 L^2 \eta_l^2 \left(\tau^2 \rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2 + \tau^2 \rho_{\max}^2 \sigma^2 + (\tau^2 + \tau^2 \rho_{\max}^2)\frac{\sigma^2}{n}\right) \\
&\leq 6C_1 L^2 \eta_l^2 \left(\tau^2 \rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2 + \tau^2 \sigma^2 \left(\frac{1}{n} + \rho_{\max}^2\right)\right),
\end{aligned}
$$

and

$$
3L^2 \tau C_1 \eta_l^2 (\tau + \rho_{\max}^2 D_{\tau,\rho})\sigma_g^2 \leq 6L^2\tau^2 C_1 \eta_l^2 \sigma_g^2,
$$

hence with a universal constant C, we have the following derivation for iterations,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\leq 8C_0 \Bigg\{ \frac{\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_T)}{\eta\eta_l\tau T} + \frac{1}{T}\left(\frac{C_\beta G^2 d}{\sqrt{\epsilon}} + \frac{2C_\beta^2 \eta\eta_l \tau L G^2 d}{\epsilon}\right) \\
&\quad + \frac{CL^2\eta_l^2}{4\sqrt{\epsilon}}\left[\tau^2\sigma_g^2 + \left(\tau\rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2 + \tau\sigma^2\left(\frac{1}{n} + \rho_{\max}^2\right)\right)\right] + C_\beta \frac{\eta_l}{2\epsilon N}\sigma^2 \Bigg\},
\end{aligned} \tag{E.20}
$$

C is a constant irrelevant to parameters $\rho_{\max} = \max_{k\in[K]} \rho_k$, $D_{\tau,\rho} = \min\left\{\frac{1}{1-\rho_{\max}}, \tau\right\}$, $\bar{\sigma}_L^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2$ $C_\beta = \frac{\beta_1}{1-\beta_1}$ and $C_{\beta,\eta} = \left(\eta L \frac{\beta_1^2}{(1-\beta_1)^2} + 3\eta L + 2\sqrt{1-\beta_2}G\right) = \mathcal{O}(\max\{\eta, 1\})$.

F Proof of Theorem 4.3: HA-Fed partial participation

From Section C, we know that Δ_t is an unbiased estimation of $\bar{\Delta}_t$, thus the main difference between full participation and partial participation lies in the second order momentum estimation of model difference Δ_t, i.e., $\mathbb{E}[\|\Delta_t\|^2]$. Hence with a different bounded $\mathbb{E}[\|\Delta_t\|^2]$ in Section G.4 for partial participation, we have the following result starting from L-smooth expansion like equation (E.2),

$$
\mathbb{E}[f(\mathbf{z}_{t+1})] - f(\mathbf{z}_t) \leq I_1' + I_2' + I_3' + I_4', \tag{F.1}
$$

where I_i' corresponds to the similar term of I_i in equation (E.2).

By the result of Lemma G.1 and G.2, and the bound for model difference Δ_t in partial participation settings in Section G.4, we can obtain the bound for I_1'. The bound for I_2' I_3' and I_4' is obtained by the similar approach, and the derivation for them is similar to full participation settings, here we omit the derivation for bounding I_2' I_3' and I_4'.

Merging I_1' to I_4' together, we obtain the following result,

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\Big\{ -\tau\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \sum_{s=0}^{\tau-1}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2\Big] - \sum_{s=0}^{\tau-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big]\Big\} \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Big[3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2+\sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2+D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Big] \\
&\quad + \Big(\eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\Big)\sum_{t=1}^{T}\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_t^{-1/2}\big\|_1\big] \\
&\quad + \eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\sum_{t=1}^{T}\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_t^{-1/2}\big\|^2\big] \\
&\quad + \Big(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\Big)\sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2],
\end{aligned}
$$

then substituting the bound of Δ_t in Lemma G.6, we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq -\frac{\eta\eta_l\tau}{4C_0}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] - \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\sum_{s=0}^{\tau-1}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2\Big] - \frac{\eta\eta_l}{4C_0}\sum_{t=1}^{T}\sum_{s=0}^{\tau-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\Big\|^2\Big] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Big[3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2+\sigma_g^2) + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2+D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Big] \\
&\quad + \Big(\eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2 + \eta^2\frac{\beta_1^2}{(1-\beta_1)^2}L\eta_l^2\tau^2G^2\epsilon^{-1/2}\Big)\sum_{t=1}^{T}\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_t^{-1/2}\big\|_1\big] \\
&\quad + \eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\sum_{t=1}^{T}\mathbb{E}\big[\big\|\widehat{\mathbf{V}}_{t-1}^{-1/2}-\widehat{\mathbf{V}}_t^{-1/2}\big\|^2\big] \\
&\quad + \Big(\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}\Big)\sum_{t=1}^{T}\Big\{\frac{2\eta_l^2\tau}{N}\sigma^2 + 2\eta_l^2(\tau-1)\sum_{s=0}^{\tau-2}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\Big\|^2\Big] \\
&\quad + 4\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big\|^2\Big] + 8\Big(\frac{n-m}{m(n-1)}+\eta_l^2L^2\Big)\Big(\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2]\Big) \\
&\quad + \frac{2\eta_l^2\sigma^2}{N}\Big(\frac{n-m}{m}\cdot\rho_{\max}^2\Big)\Big\}, \qquad \text{(F.2)}
\end{aligned}
$$

we need the following constraint on local learning rate η_l

$$
\begin{aligned}
&2C_{\beta,\eta}\eta\eta_l^2(\tau-1) \leq \frac{\eta\eta_l}{4C_0}, \quad 4C_{\beta,\eta}\eta\eta_l^2 \leq \frac{\eta\eta_l}{4C_0} \\
&\Rightarrow \eta \leq \frac{1}{8C_0C_{\beta,\eta}(\tau-1)}, \quad \eta \leq \frac{1}{16C_0C_{\beta,\eta}}, \qquad \text{(F.3)}
\end{aligned}
$$

where $C_{\beta,\eta} = \left(\frac{\eta L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta L}{2\epsilon} + \frac{\eta L}{\epsilon} + \frac{\sqrt{1-\beta_2}G}{\epsilon}\right) = \mathcal{O}(\max\{\eta, 1\})$, and by applying Lemma G.4, then we have

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_1) \\
&\leq -\frac{\eta\eta_l\tau}{4C_0}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta\eta_l}{4\sqrt{\epsilon}}3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\alpha^2\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\sum_{t=1}^{T}\Bigg\{3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\sigma_g^2 \\
&\quad + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2 + D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Bigg\} \\
&\quad + \eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2\frac{d}{\sqrt{\epsilon}} + 2\eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\cdot\frac{d}{\epsilon} \\
&\quad + \left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\left[\frac{2T\eta_l^2\tau}{N}\sigma^2 + \frac{2T\eta_l^2\sigma^2}{N}\left(\frac{n-m}{m}\rho_{\max}^2\right)\right] \\
&\quad + 8T\left(\frac{n-m}{m(n-1)} + \eta_l^2L^2\right)\left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right) \\
&\quad \cdot\Bigg[\tau C_1\eta_l^2D_{\tau,\rho}\rho_{\max}^2(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) \\
&\quad + \tau C_1\eta_l^2D_{\tau,\rho}\rho_{\max}^2\bar{\sigma}_L^2 + \tau C_1\eta_l^2\rho_{\max}^2\sigma^2 + C_1D_{\tau,\rho}^2\tau^{-1}\rho_{\max}^2\eta_l^2\frac{\sigma^2}{n}\Bigg],
\end{aligned}
\tag{F.4}
$$

we further need the requirement of η_l, which is same as the requirement in full participation settings

$$
\begin{aligned}
&\frac{\eta\eta_l}{4\sqrt{\epsilon}}3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\alpha^2 + 8\left(\frac{n-m}{m(n-1)} + \eta_l^2L^2\right)C_{\beta,\eta}\eta\tau C_1\eta_l^2D_{\tau,\rho}\rho_{\max}^2\alpha^2 \leq \frac{\eta\eta_l\tau}{8C_0}, \\
&\Rightarrow \frac{1}{4\sqrt{\epsilon}}3L^2\tau C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\alpha^2 + 8\left(\frac{n-m}{m(n-1)} + \frac{1}{\tau^2}\right)C_{\beta,\eta}C_1\eta_lD_{\tau,\rho}\rho_{\max}^2\alpha^2 \leq \frac{1}{8C_0}, \\
&\Rightarrow \eta_l \leq \frac{\sqrt[4]{\epsilon}}{\sqrt{12L^2C_0C_1\tau(\tau+\rho_{\max}^2D_{\tau,\rho})\alpha^2}}, \quad \eta_l \leq \frac{1}{128C_0C_{\beta,\eta}C_1D_{\tau,\rho}\rho_{\max}^2\alpha^2}\left(\frac{n-m}{m(n-1)} + \frac{1}{\tau^2}\right)^{-1}
\end{aligned}
\tag{F.5}
$$

thus we have

$$
\begin{aligned}
&\frac{\eta\eta_l\tau}{8C_0T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\leq \frac{\mathbb{E}[f(\mathbf{z}_{T+1})] - f(\mathbf{z}_T)}{T} + \eta\frac{\beta_1}{1-\beta_1}\eta_l\tau G^2\frac{d}{T\sqrt{\epsilon}} + 2\eta^2L\frac{\beta_1^2}{(1-\beta_1)^2}\eta_l^2\tau^2G^2\cdot\frac{d}{T\epsilon} \\
&\quad + \frac{\eta\eta_l}{4\sqrt{\epsilon}}\cdot\sum_{t=1}^{T}\Bigg\{3L^2\tau^2C_1\eta_l^2(\tau+\rho_{\max}^2D_{\tau,\rho})\sigma_g^2 \\
&\quad + 3L^2\tau^2C_1\rho_{\max}^2D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 + 3L^2\tau^2C_1\eta_l^2\sigma^2\rho_{\max}^2 + 3L^2C_1\big(\tau^2 + D_{\tau,\rho}^2\cdot\rho_{\max}^2\big)\eta_l^2\frac{\sigma^2}{n}\Bigg) \\
&\quad + \left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right)\left[\frac{2\eta_l^2\tau}{N}\sigma^2 + \frac{2\eta_l^2\sigma^2}{N}\left(\frac{n-m}{m}\rho_{\max}^2\right)\right] \\
&\quad + 8\left(\frac{n-m}{m(n-1)} + \eta_l^2L^2\right)\left(\frac{\eta^2L}{2\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2} + \frac{\eta^2L}{2\epsilon} + \frac{\eta^2L}{\epsilon} + \frac{\eta\sqrt{1-\beta_2}G}{\epsilon}\right) \\
&\quad \cdot\Bigg[\tau C_1\eta_l^2D_{\tau,\rho}\rho_{\max}^2\sigma_g^2 + \tau C_1\eta_l^2D_{\tau,\rho}\rho_{\max}^2\bar{\sigma}_L^2 + \tau C_1\eta_l^2\rho_{\max}^2\sigma^2 + C_1D_{\tau,\rho}^2\tau^{-1}\rho_{\max}^2\eta_l^2\frac{\sigma^2}{n}\Bigg],
\end{aligned}
\tag{F.6}
$$

since there is $D_{\tau,\rho} \leq \tau$ and $\rho_{\max} \leq 1$, thus we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\leq 8C_0\Bigg\{\frac{\mathbb{E}[f(\mathbf{z}_{T+1})]-f(\mathbf{z}_T)}{\eta\eta_l\tau T}+\frac{1}{T}\bigg(\frac{C_\beta G^2 d}{\sqrt{\epsilon}}+\frac{2C_\beta^2\eta\eta_l\tau L G^2 d}{\epsilon}\bigg) \\
&\quad+\frac{1}{4\sqrt{\epsilon}}\bigg[C\cdot L^2\tau(\tau+D_{\tau,\rho})\eta_l^2\sigma_g^2+C\cdot L^2\eta_l^2\bigg(\tau\rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2+\tau\sigma^2\bigg(\frac{1}{n}+\rho_{\max}^2\bigg)\bigg)\bigg] \\
&\quad+\bigg(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\bigg)\bigg[\frac{\eta_l}{N}\sigma^2+\frac{\eta_l\sigma^2}{N}\bigg(\frac{n-m}{m}\rho_{\max}^2\bigg)\bigg] \\
&\quad+4\bigg(\frac{n-m}{m(n-1)}+\eta_l^2L^2\bigg)\bigg(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\bigg) \\
&\quad\cdot\bigg[C_1\eta_l D_{\tau,\rho}\rho_{\max}^2\sigma_g^2+C_1\eta_l D_{\tau,\rho}\rho_{\max}^2\bar{\sigma}_L^2+C_1\eta_l\rho_{\max}^2\sigma^2+C_1 D_{\tau,\rho}^2\tau^{-2}\rho_{\max}^2\eta_l\frac{\sigma^2}{n}\bigg]\Bigg\},
\end{aligned}
\tag{F.7}
$$

where C is a constant irrelevant to parameters and $\rho_{\max} = \max_{k\in[K]}\rho_k$, $D_{\tau,\rho} = \min\left\{\frac{1}{1-\rho_{\max}}, \tau\right\}$, $C_\beta = \frac{\beta_1}{1-\beta_1}$ and $\bar{\sigma}_L^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2$. Further apply the constraint of

$$
\frac{n-m}{m(n-1)}D_{\tau,\rho}\rho_{\max}^2 \leq \frac{1}{N},
\tag{F.8}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \\
&\leq 8C_0\Bigg\{\frac{\mathbb{E}[f(\mathbf{z}_{T+1})]-f(\mathbf{z}_T)}{\eta\eta_l\tau T}+\frac{1}{T}\bigg(\frac{C_\beta G^2 d}{\sqrt{\epsilon}}+\frac{2C_\beta^2\eta\eta_l\tau L G^2 d}{\epsilon}\bigg) \\
&\quad+\frac{CL^2\eta_l^2}{4\sqrt{\epsilon}}\bigg[\tau(\tau+D_{\tau,\rho})\sigma_g^2+\bigg(\tau\rho_{\max}^2 D_{\tau,\rho}\bar{\sigma}_L^2+\tau\sigma^2\bigg(\frac{1}{n}+\rho_{\max}^2\bigg)\bigg)\bigg] \\
&\quad+\bigg(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\bigg)\bigg[\frac{\eta_l}{N}\sigma^2+\frac{\eta_l\sigma^2}{N}\bigg(\frac{n-m}{m}\rho_{\max}^2\bigg)\bigg] \\
&\quad+\frac{1}{N}\bigg(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\bigg)\widetilde{C}\eta_l\bigg[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\bigg] \\
&\quad+\widetilde{C}L^2 D_{\tau,\rho}\rho_{\max}^2\bigg(\frac{\eta L}{\epsilon}\frac{\beta_1^2}{(1-\beta_1)^2}+\frac{3\eta L}{\epsilon}+\frac{2\sqrt{1-\beta_2}G}{\epsilon}\bigg)\eta_l^3\bigg[\sigma_g^2+\bar{\sigma}_L^2+\sigma^2+D_{\tau,\rho}\frac{\sigma^2}{\tau^2 n}\bigg]\Bigg\}
\end{aligned}
\tag{F.9}
$$

where the condition equation (F.8) implies that the spectral gap $\rho_{\max}$ satisfies

$$
\rho_{\max}^2 \leq \frac{m}{4n(n-m)}.
\tag{F.10}
$$

With the condition of equation (F.10), there is $\rho_{\max} \leq \frac{1}{2}$, hence when $\tau \geq 2$, there is $\frac{1}{1-\rho_{\max}} \leq \tau$, also assume $K \leq n-1$,

$$
\begin{aligned}
\frac{n-m}{m(n-1)}D_{\tau,\rho}\rho_{\max}^2 &= \frac{n-m}{m(n-1)}\frac{1}{1-\rho_{\max}}\rho_{\max}^2 \\
&\leq 2\frac{n-m}{m(n-1)}\frac{m}{4n(n-m)} \\
&\leq \frac{1}{n(n-1)} \\
&\leq \frac{1}{nK} = \frac{1}{N}.
\end{aligned}
\tag{F.11}
$$

Then we concludes the proof.

G Supporting Lemmas

G.1 Lemma for inter-cluster consensus error

Lemma G.1. For local learning rate which satisfying the condition $\eta_l \leq \frac{1}{8\tau L}$, denote $C_\tau = 1 + \frac{3}{2} \cdot \frac{1}{4\tau - 1}$, recall the definition for $\bar{\mathbf{x}}$ in equation (D.1), the inter-cluster model difference after s local steps satisfies

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s+1}^k - \mathbf{x}_t\|^2 \\
&\leq C_\tau \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 8\tau\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \sigma_g^2) + \eta_l^2\frac{\sigma^2}{n}.
\end{aligned} \tag{G.1}$$

Proof. Note that the following proof is similar to Lemma 3 in [32].

$$\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{x}}_{t,s+1}^k - \mathbf{x}_t\|^2 = \mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t - \eta_l\bar{\mathbf{g}}_{t,s}^k\|^2 \\
&= \mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t - \eta_l(\bar{\mathbf{g}}_{t,s}^k - \nabla\bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) + \nabla\bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) - \nabla\bar{f}_k(\mathbf{x}_t) + \nabla\bar{f}_k(\mathbf{x}_t))\|^2 \\
&\leq (1+\gamma)\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2\mathbb{E}\|\bar{\mathbf{g}}_{t,s}^k - \nabla\bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\|^2 \\
&\quad + 2(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) - \nabla\bar{f}_k(\mathbf{x}_t)\|^2] + 2(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\mathbf{x}_t)\|^2] \\
&\leq (1+\gamma)\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2\frac{\sigma^2}{n} + 2(1+\gamma^{-1})\eta_l^2L^2\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 2(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\mathbf{x}_t)\|^2] \\
&\leq [(1+\gamma) + 2(1+\gamma^{-1})\eta_l^2L^2]\cdot\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2\frac{\sigma^2}{n} + 2(1+\gamma^{-1})\eta_l^2\mathbb{E}[\|\nabla\bar{f}_k(\mathbf{x}_t)\|^2],
\end{aligned} \tag{G.2}$$

where the first equality holds by equation (D.2). The first inequality holds due to $\mathbf{g}_{t,s}^i$ is an unbiased estimator of $\nabla f_i(\mathbf{x}_{t,s}^i)$ and Young's inequality. The second inequality holds by Assumption C.1 and C.3, also the independency with $\mathbf{g}_{t,s}^i$ and $\mathbf{g}_{t,s}^j$ for $i \neq j$.

Averaging equation (G.2) over $k = 1, ..., K$ clusters, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s+1}^k - \mathbf{x}_t\|^2 \\
&\leq [(1+\gamma) + 2(1+\gamma^{-1})\eta_l^2L^2]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 2(1+\gamma^{-1})\eta_l^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\bar{f}_k(\mathbf{x}_t)\|^2] + \eta_l^2\frac{\sigma^2}{n} \\
&\leq [(1+\gamma) + 2(1+\gamma^{-1})\eta_l^2L^2]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 2(1+\gamma^{-1})\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \sigma_g^2) + \eta_l^2\frac{\sigma^2}{n},
\end{aligned} \tag{G.3}$$

where the second inequality holds by Assumption C.5. Choosing $\gamma = \frac{1}{4\tau-1}$ with the condition of $\eta_l \leq \frac{1}{8\tau L}$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s+1}^k - \mathbf{x}_t\|^2 \\
&\leq \left(1 + \frac{1}{4\tau-1} + \frac{1}{2(4\tau-1)}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 8\tau\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \sigma_g^2) + \eta_l^2\frac{\sigma^2}{n} \\
&= C_\tau\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 8\tau\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \sigma_g^2) + \eta_l^2\frac{\sigma^2}{n},
\end{aligned} \tag{G.4}$$

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

G.2 Lemma for intra-cluster consensus error

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

$$
\begin{aligned}
&\frac{1}{N}\mathbb{E}\sum_{k=1}^{K}\|X_{t,s+1}^{k,\perp}\|^2 \\
&\leq \left(\max_{k\in[K]}\rho_k^2(1+\zeta_k^{-1}) + \eta_l^2\cdot 4L^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\right)\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\|X_{t,s}^{k,\perp}\|^2 \\
&\quad + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4L^2\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2\max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^{K}\rho_k^2(1+\zeta_k)4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2,
\end{aligned}
\tag{G.5}
$$

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

$$
\begin{aligned}
\sum_{i=1}^{n}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_{t,s}^i\|^2 &= \|(\mathbf{x}_{t,s}^1, ..., \mathbf{x}_{t,s}^n)(I_n - J)\cdot I_n\cdot(I_n - J)(\mathbf{x}_{t,s}^1, ..., \mathbf{x}_{t,s}^n)'\|_F \\
&= \|X_{t,s}^k{}'(I_n - J)\cdot(I_n - J)X_{t,s}^k\|_F \\
&= \|X_{t,s}^{k,\perp\prime}\cdot X_{t,s}^{k,\perp}\|_F \\
&= \|X_{t,s}^{k,\perp}\|_F^2,
\end{aligned}
\tag{G.6}
$$

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

$$
\begin{aligned}
&\mathbb{E}\|X_{t,s+1}^{k,\perp}\|^2 = \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{t,s}^{k,\perp} - \eta_l G_{t,s}^k)\|^2|\mathcal{F}_{t,s-1})) \\
&= \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{t,s}^{k,\perp} - \eta_l\nabla F(X_{t,s}^k) + \eta_l\nabla F(X_{t,s}^k) - \eta_l G_{t,s}^k)\|^2|\mathcal{F}_{t,s-1})) \\
&= \mathbb{E}(\mathbb{E}(\|(W_k - J)(X_{t,s}^{k,\perp} - \eta_l\nabla F(X_{t,s}^k))\|^2|\mathcal{F}_{t,s-1})) \\
&\quad + \eta_l^2\mathbb{E}(\mathbb{E}(\|(W_k - J)(\nabla F(X_{t,s}^k) - G_{t,s}^k)\|^2|\mathcal{F}_{t,s-1})) \\
&\leq \mathbb{E}(\|(W_k - J)(X_{t,s}^{k,\perp} - \eta_l\nabla F(X_{t,s}^k))\|^2) + \eta_l^2\rho_k^2 n\sigma^2 \\
&\leq \rho_k^2(1+\zeta_k^{-1})\cdot\mathbb{E}\|X_{t,s}^{k,\perp}\|^2 + \rho_k^2(1+\zeta_k)\eta_l^2\mathbb{E}\|\nabla F(X_{t,s}^k))\|^2 + \eta_l^2\rho_k^2 n\sigma^2,
\end{aligned}
\tag{G.7}
$$

where the $\nabla F_k(X^k) \in \mathbb{R}^{n\times d}$ is associated to cluster k by stacking $\nabla f_i(\mathbf{x}^i)$ 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 C.3 and $\|\nabla F(X_{t,s}^k) - G_{t,s}^k\|_F = \sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_{t,s}^i) - \mathbf{g}_{t,s}^i\|^2$. 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 $\nabla F_k(X_{t,s}^k)$, by definition, we have

$$
\begin{aligned}
&\|\nabla F_k(X_{t,s}^k)\|_F^2 = \sum_{i\in\mathcal{V}_k} \|\nabla f_i(\mathbf{x}_{t,s}^i)\|^2 \\
&= \sum_{i\in\mathcal{V}_k} \|\nabla f_i(\mathbf{x}_{t,s}^i) - \nabla f_i(\bar{\mathbf{x}}_{t,s}^k) + \nabla f_i(\bar{\mathbf{x}}_{t,s}^k) - \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) + \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) - \nabla \bar{f}_k(\mathbf{x}_t) + \nabla \bar{f}_k(\mathbf{x}_t)\|^2 \\
&\leq \sum_{i\in\mathcal{V}_k} \Big[4\|\nabla f_i(\mathbf{x}_{t,s}^i) - \nabla f_i(\bar{\mathbf{x}}_{t,s}^k)\|^2 + 4\|\nabla f_i(\bar{\mathbf{x}}_{t,s}^k) - \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\|^2 + 4\|\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k) - \nabla \bar{f}_k(\mathbf{x}_t)\|^2 \\
&\quad + 4\|\nabla \bar{f}_k(\mathbf{x}_t)\|^2\Big] \\
&\leq \sum_{i\in\mathcal{V}_k} \Big[4\|\nabla f_i(\bar{\mathbf{x}}_{t,s}^k) - \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,s}^k)\|^2 + 4L^2\|\mathbf{x}_{t,s}^i - \bar{\mathbf{x}}_{t,s}^k\|^2 + 4L^2\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 4\|\nabla \bar{f}_k(\mathbf{x}_t)\|^2\Big] \\
&\leq 4L^2\|X_{t,s}^{k,\perp}\|^2 + 4L^2 n\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + 4n\|\nabla \bar{f}_k(\mathbf{x}_t)\|^2 + 4n\sigma_k^2,
\end{aligned} \tag{G.8}
$$

where the first inequality holds by Cauchy inequality, and the last inequality holds by Assumption C.4. Averaging equation (G.8) over $k = 1, ..., K$, we have the following iteration

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^K \mathbb{E}\|X_{t,s+1}^{k,\perp}\|^2 \\
&\leq \frac{1}{N}\sum_{k=1}^K \rho_k^2(1+\zeta_k^{-1})\cdot\mathbb{E}\|X_{t,s}^{k,\perp}\|^2 + \frac{1}{N}\sum_{k=1}^K \rho_k^2(1+\zeta_k)\eta_l^2\mathbb{E}\|\nabla F(X_{t,s}^k))\|^2 + \eta_l^2\sigma^2\frac{1}{K}\sum_{k=1}^K \rho_k^2 \\
&\leq \frac{1}{N}\sum_{k=1}^K \rho_k^2(1+\zeta_k^{-1})\cdot\mathbb{E}\|X_{t,s}^{k,\perp}\|^2 + \eta_l^2\frac{1}{N}\sum_{k=1}^K \rho_k^2(1+\zeta_k)\cdot 4L^2\mathbb{E}\|X_{t,s}^{k,\perp}\|^2 \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^K \rho_k^2(1+\zeta_k)\cdot 4L^2\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2\frac{1}{K}\sum_{k=1}^K \rho_k^2(1+\zeta_k)\cdot 4\mathbb{E}\|\nabla \bar{f}_k(\mathbf{x}_t)\|^2 \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^K \rho_k^2(1+\zeta_k)\cdot 4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2 \\
&\leq \left(\max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \eta_l^2\cdot 4L^2 \max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\right)\frac{1}{N}\sum_{k=1}^K \mathbb{E}\|X_{t,s}^{k,\perp}\|^2 \\
&\quad + \eta_l^2 \max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4L^2\mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 + \eta_l^2 \max_{k\in[K]}\{\rho_k^2(1+\zeta_k)\}\cdot 4(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) \\
&\quad + \eta_l^2\frac{1}{K}\sum_{k=1}^K \rho_k^2(1+\zeta_k)4\sigma_k^2 + \eta_l^2\sigma^2\rho_{\max}^2.
\end{aligned} \tag{G.9}
$$

This concludes the proof. □

G.3 Lemma for summation of intra-cluster and inter-cluster consensus errors

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

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^K \mathbb{E}\|X_{t,s}^{k,\perp}\|^2 + \frac{1}{K}\sum_{k=1}^K \mathbb{E}\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2 \\
&\leq (s+1)C_1\eta_l^2(\tau + \rho_{\max}^2 D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + (s+1)C_1\rho_{\max}^2 D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + (s+1)C_1\eta_l^2\sigma^2\rho_{\max}^2 + (s+1)C_1\left(1 + \frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\right)\eta_l^2\frac{\sigma^2}{n},
\end{aligned} \tag{G.10}
$$

where C_1 is a constant independent to parameters.

Proof. Denote an auxiliary vector

$$M_{t,s} = \left(\frac{1}{N} \sum_{k=1}^{K} \mathbb{E}\|X_{t,s}^{k,\perp}\|^2, \quad \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\|\bar{\mathbf{x}}_{t,s}^{k} - \mathbf{x}_t\|^2 \right)^T. \tag{G.11}$$

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

$$M_{t,s+1} \leq G \cdot M_{t,s} + B_{t,s}, \tag{G.12}$$

where

$$G = \begin{pmatrix} \max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \eta_l^2 \rho_L \cdot 4L^2 & \eta_l^2 \rho_L \cdot 4L^2 \\ 0 & C_\tau \end{pmatrix} \tag{G.13}$$

$$B_{t,s} = \begin{pmatrix} 4\rho_L \eta_l^2 (\alpha^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 4\rho_L \eta_l^2 \bar{\sigma}_L^2 + \eta_l^2 \sigma^2 \rho_{\max}^2 \\ 8\tau \eta_l^2 (\alpha^2 \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + \eta_l^2 \frac{\sigma^2}{n} \end{pmatrix} = \begin{pmatrix} b^{(1)} \\ b^{(2)} \end{pmatrix}. \tag{G.14}$$

Consider the eigen-decomposition of matrix G,

$$G = \begin{pmatrix} 1 & -\frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot \begin{pmatrix} 1 & \frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} \\ 0 & 1 \end{pmatrix}, \tag{G.15}$$

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

$$\begin{aligned} G^j B &= \begin{pmatrix} 1 & -\frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^j & 0 \\ 0 & \lambda_2^j \end{pmatrix} \begin{pmatrix} 1 & \frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b^{(1)} \\ b^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1^j (b^{(1)} + \frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} b^{(2)}) - \frac{4\eta_l^2 \rho_L L^2}{\lambda_1 - \lambda_2} \lambda_2^j b^{(2)} \\ \lambda_2^j b^{(2)} \end{pmatrix}. \end{aligned} \tag{G.16}$$

Therefore the sum of two elements has the following result

$$\begin{aligned} (1,1) G^j B_{t,s-j} &= \lambda_1^j b^{(1)} + \lambda_2^j b^{(2)} + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1} 4\eta_l^2 \rho_L L^2 b^{(2)} \\ &\leq \lambda_2^j (b^{(1)} + b^{(2)}) + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1} 4\eta_l^2 \rho_L L^2 b^{(2)}. \end{aligned} \tag{G.17}$$

Therefore, we have the following result

$$\sum_{j=0}^{s} (1,1) G^j B_{t,s-j} \leq \sum_{j=0}^{s} \left(\lambda_2^j (b_{t,s-j}^{(1)} + b_{t,s-j}^{(2)}) + \frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1} \eta_l^2 \cdot 4\rho_L L^2 b_{t,s-j}^{(2)} \right). \tag{G.18}$$

Since $\lambda_2 \geq C_\tau > 1$, we have

$$\frac{\lambda_2^j - \lambda_1^j}{\lambda_2 - \lambda_1} = \lambda_2^{l-1} \sum_{s=0}^{l-1} \left(\frac{\lambda_1}{\lambda_2} \right)^s \leq \lambda_2^{j-1} \min\left\{ \frac{\lambda_2}{\lambda_2 - \lambda_1}, l \right\} \leq \lambda_2^j \min\left\{ \frac{1}{\lambda_2 - \lambda_1}, l \right\}, \tag{G.19}$$

thus we have

$$\sum_{j=0}^{s} (1,1) G^j B_{t,s-j} \leq \sum_{j=0}^{s} \lambda_2^j (b_{t,s-j}^{(1)} + b_{t,s-j}^{(2)}) + \sum_{l=0}^{s} \left(\lambda_2^j \min\left\{ \frac{1}{\lambda_2 - \lambda_1}, l \right\} \eta_l^2 \cdot 4\rho_L L^2 b_{t,s-j}^{(2)} \right). \tag{G.20}$$

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,

$$\begin{aligned} \lambda_2 &\leq \max\left\{ \max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \eta_l^2 \rho_L \cdot 8L^2, C_\tau \right\} \\ &< \max\left\{ \max_{k\in[K]} \rho_k^2(1+\zeta_k^{-1}) + \frac{\rho_L}{(4\tau-1)2\tau}, 1 + \frac{3}{2(4\tau-1)} \right\}, \end{aligned} \tag{G.21}$$

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

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

$$\max_{k\in[k]} \rho_k^2(1+\zeta_k^{-1}) = \rho_{\max}, \quad \rho_L = \max_{k\in[k]}\left\{\frac{\rho_k^2}{1-\rho_k}\right\} = \frac{\rho_{\max}^2}{1-\rho_{\max}} = \rho_{\max}^2 D_{\tau,\rho}, \tag{G.22}$$

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

$$\begin{aligned}
\lambda_2 &\leq \max\left\{\rho_{\max} + \frac{\rho_{\max}^2}{(1-\rho_{\max})2\tau(4\tau-1)}, 1 + \frac{3}{2(4\tau-1)}\right\} \\
&\leq \max\left\{1 - \frac{1}{\tau} + \frac{(1-\frac{1}{\tau})^2}{2(\tau-1)}, 1 + \frac{3}{2(4\tau-1)}\right\} \\
&< 1 + \frac{2}{4\tau-1},
\end{aligned} \tag{G.23}$$

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

$$\sum_{j=0}^{s} \lambda_2^j b_{s-j}^1 \leq \left(\left(1 + \frac{2}{4\tau-1}\right)^{\tau}\right) \cdot \sum_{j=0}^{s} b_j^{(1)} \leq 3 \cdot \sum_{j=0}^{s} b_j^{(1)}. \tag{G.24}$$

We also have

$$\rho_{\max} + \eta_l^2 \rho_L 4L^2 \leq \rho_{\max} + \frac{\rho_{\max}}{(1-\rho_{\max})(4\tau-1)4\tau} \leq 1 - \frac{1}{\tau} + \frac{(1-\frac{1}{\tau})^2}{4(4\tau-1)} \leq C_\tau, \tag{G.25}$$

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

$$\begin{aligned}
\lambda_2 - \lambda_1 &= C_\tau - \rho_{\max} - \eta_l^2 \rho_L 4L^2 \\
&\geq C_\tau - \left(\rho_{\max} + \frac{\rho_{\max}}{(1-\rho_{\max})(4\tau-1)4\tau}\right) \\
&\geq C_\tau - \left(\rho_{\max} + \rho_{\max} \cdot \frac{1-\frac{1}{\tau}}{4(4\tau-1)}\right) \\
&\geq 1 + \frac{1}{4\tau-1} - \left(\rho_{\max} + \rho_{\max} \cdot \frac{1}{4\tau-1}\right) \\
&= (1-\rho_{\max})\left(1 + \frac{1}{4\tau-1}\right) \\
&\geq 1 - \rho_{\max}.
\end{aligned} \tag{G.26}$$

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

$$\begin{aligned}
\sum_{j=0}^{s}(1,1)G^j B_{t,s-j} &\leq \sum_{j=0}^{s} \lambda_2^j (b_{t,s-j}^{(1)} + b_{t,s-j}^{(2)}) + \sum_{j=0}^{s}\left(\lambda_2^j \min\left\{\frac{1}{\lambda_2-\lambda_1}, j\right\}\eta_l^2 \cdot 4\rho_L L^2 b_{t,s-j}^2\right) \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} 3\eta_l^2 \cdot 4\rho_L L^2 b_{t,j}^{(2)}\left(\min\left\{\frac{1}{\lambda_2-\lambda_1}, \tau\right\}\right) \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} 3\eta_l^2 \cdot D_{\tau,\rho}^2 \rho_{\max}^2 4L^2 b_{t,j}^{(2)} \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} \frac{12}{16\tau(4\tau-1)} \cdot D_{\tau,\rho}^2 \rho_{\max}^2 b_{t,j}^{(2)} \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} \frac{1}{\tau^2} \cdot D_{\tau,\rho}^2 \cdot \rho_{\max}^2 b_{t,j}^{(2)},
\end{aligned} \tag{G.27}$$

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

$$
\begin{aligned}
&\sum_{j=0}^{s}(1,1)G^j B_{t,s-j} \\
&\leq \sum_{j=0}^{s} 3\bigg((4\rho_L\eta_l^2 + 8\tau\eta_l^2)(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 4\rho_L\eta_l^2\bar{\sigma}_L^2 + \eta_l^2\sigma^2\rho_{\max}^2 + \eta_l^2\frac{\sigma^2}{n}\bigg) \\
&\quad + \sum_{j=0}^{s}\frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\bigg(8\tau\eta_l^2(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + \eta_l^2\frac{\sigma^2}{n}\bigg) \\
&= (s+1)\bigg[\bigg(3(4\rho_L\eta_l^2 + 8\tau\eta_l^2) + \frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2 8\tau\eta_l^2\bigg)(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 12\rho_L\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3\eta_l^2\sigma^2\rho_{\max}^2 + \bigg(3\eta_l^2 + \frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\eta_l^2\bigg)\frac{\sigma^2}{n}\bigg] \\
&\leq (s+1)\bigg[\bigg(12\rho_{\max}^2 D_{\tau,\rho}\eta_l^2 + 8\tau\eta_l^2 + 8\eta_l^2 D_{\tau,\rho}\cdot\rho_{\max}^2\bigg)(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + 12\rho_{\max}^2 D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + 3\eta_l^2\sigma^2\rho_{\max}^2 + \bigg(3\eta_l^2 + \frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\eta_l^2\bigg)\frac{\sigma^2}{n}\bigg] \\
&\leq (s+1)C_1\eta_l^2(\tau + \rho_{\max}^2 D_{\tau,\rho})(\alpha^2\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \sigma_g^2) + (s+1)C_1\rho_{\max}^2 D_{\tau,\rho}\eta_l^2\bar{\sigma}_L^2 \\
&\quad + (s+1)C_1\eta_l^2\sigma^2\rho_{\max}^2 + (s+1)C_1\bigg(1 + \frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\bigg)\eta_l^2\frac{\sigma^2}{n},
\end{aligned}
\tag{G.28}
$$

where C_1 is some universal constant. The inequality holds by $\rho_L = \rho_{\max}^2 D_{\tau,\rho}$ and $D_{\tau,\rho} \leq \tau$.

Case 2: In this case we have $\rho_{\max} > 1 - \frac{1}{\tau}$, which means $D_{\tau,\rho} = \tau$. Let $\zeta_k = (4\tau - 1)$, thus we have

$$
\max_{k\in[K]} \rho_k^2(1 + \zeta_k^{-1}) = \rho_{\max}^2(1 + (4\tau - 1)^{-1}), \quad \rho_L = 4\tau\rho_{\max}^2 D_{\tau,\rho}. \tag{G.29}
$$

The upper bound for λ_2 has the form of

$$
\begin{aligned}
\lambda_2 &\leq \max\bigg\{\max_{k\in[K]} \rho_k^2(1 + \zeta_k^{-1}) + \eta_l^2\rho_L\cdot 8L^2, C_\tau\bigg\} \\
&\leq \max\bigg\{\rho_{\max}^2(1 + (4\tau - 1)^{-1}) + \frac{2\rho_{\max}^2}{4\tau - 1}, 1 + \frac{3}{2(4\tau - 1)}\bigg\} \\
&\leq 1 + \frac{3}{4\tau - 1}.
\end{aligned}
\tag{G.30}
$$

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

$$
\begin{aligned}
\sum_{j=0}^{s}(1,1)G^j B_{t,s-j} &\leq \sum_{j=0}^{s}\lambda_2^j(b_{t,s-j}^{(1)} + b_{t,s-j}^{(2)}) + \sum_{j=0}^{s}\bigg(\lambda_2^j\min\bigg\{\frac{1}{\lambda_2 - \lambda_1}, l\bigg\}\eta_l^2\cdot 4\rho_L L^2 b_{t,s-j}^2\bigg) \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s}\eta_l^2\cdot 16\rho_{\max}D_{\tau,\rho}L^2 b_l^{(2)}\cdot 3D_{\tau,\rho} \\
&\leq \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} 16\rho_{\max}D_{\tau,\rho}b_l^{(2)}\cdot\frac{3}{16}\frac{D_{\tau,\rho}}{\tau^2} \\
&= \sum_{j=0}^{s} 3(b_{t,j}^{(1)} + b_{t,j}^{(2)}) + \sum_{j=0}^{s} 3\rho_{\max}b_l^{(2)}\cdot\frac{D_{\tau,\rho}^2}{\tau^2},
\end{aligned}
\tag{G.31}
$$

where the above inequalities hold by the fact that $\rho_L = 4\tau\rho_{\max}^2 = 4D_{\tau,\rho}\rho_{\max}^2$ and the constraint on step size η_l. Thus we can get a similar upper bound as equation (G.28) 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 $\|X_{t,s}^{k,\perp}\|_F^2$,

$$\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{t,s}^{k,\perp}\|^2] \leq (s+1)C_1\eta_l^2 D_{\tau,\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)]\|^2+\sigma_g^2) + (s+1)C_1\eta_l^2 D_{\tau,\rho}\rho_{\max}^2\bar{\sigma}_L^2 + (s+1)C_1\eta_l^2\rho_{\max}^2\sigma^2 + (s+1)C_1\eta_l^2\frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\frac{\sigma^2}{n}. \tag{G.32}$$

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

$$M_{t,s} = \left(\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{t,s}^{k,\perp}\|^2], \quad \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\mathbf{x}}_{t,s}^k - \mathbf{x}_t\|^2]\right)^\top,$$
$$M_{t,s} = G^{s+1}M_{t,0} + \sum_{j=0}^{s}G^j B_{t,s-j} = \sum_{j=0}^{s}G^j B_{t,s-j}, \tag{G.33}$$

hence we have

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{t,s}^{k,\perp}\|^2] &= (1,0)\cdot M_{t,s} = (1,0)\cdot\sum_{j=0}^{s}G^j B_{t,s-j}\\ &= \sum_{j=0}^{s}\left[\lambda_1^j b_{t,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{t,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{s}\left[\lambda_2^j b_{t,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{t,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{s}\left[\lambda_2^j b_{t,j}^{(1)} + \frac{\lambda_2^j-\lambda_1^j}{\lambda_2-\lambda_1}4\eta_l^2\rho_L L^2 b_{t,j}^{(2)}\right],\end{aligned} \tag{G.34}$$

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

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\mathbb{E}[\|X_{t,s}^{k,\perp}\|^2] &\leq \sum_{j=0}^{s}\left[3b_{t,j}^{(1)} + \frac{1}{\tau^2}D_{\tau,\rho}^2\rho_{\max}^2 b_{t,j}^{(2)}\right]\\ &\leq \sum_{j=0}^{s}\Bigg[12\rho_L\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)]\|^2+\sigma_g^2) + 12\rho_L\eta_l^2\bar{\sigma}_L^2 + 3\eta_l^2\sigma^2\rho_{\max}^2\\ &\quad + \frac{D_{\tau,\rho}^2}{\tau^2}\rho_{\max}^2\left(8\tau\eta_l^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)]\|^2+\sigma_g^2) + \eta_l^2\frac{\sigma^2}{n}\right)\Bigg]\\ &\leq (s+1)C_1\eta_l^2 D_{\tau,\rho}\rho_{\max}^2(\alpha^2\mathbb{E}[\|\nabla f(\mathbf{x}_t)]\|^2+\sigma_g^2) + (s+1)C_1\eta_l^2 D_{\tau,\rho}\rho_{\max}^2\bar{\sigma}_L^2\\ &\quad + (s+1)C_1\eta_l^2\rho_{\max}^2\sigma^2 + (s+1)C_1\eta_l^2\frac{D_{\tau,\rho}^2}{\tau^2}\cdot\rho_{\max}^2\frac{\sigma^2}{n}.\end{aligned} \tag{G.35}$$

□

G.4 Lemmas for model difference Δ_t

G.4.1 Full participation

Lemma G.5. The global model difference $\Delta_t = \sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\Delta_t^i$ in full participation cases satisfy

$$\mathbb{E}[\|\Delta_t\|^2] \leq \frac{\eta_l^2\tau}{N}\sigma^2 + \eta_l^2\tau\sum_{s=0}^{\tau-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_{t,s}^i)\right\|^2\right]. \tag{G.36}$$

Proof. Under full participation case, we have $\Delta_t = -\eta_l \sum_{s=0}^{\tau-1} \frac{1}{K} \sum_{k\in[K]} \frac{1}{n} \sum_{i\in\mathcal{V}_k} \mathbf{g}_{t,s}^i = -\eta_l \sum_{s=0}^{\tau-1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_{t,s}^i$

$$
\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\left[\left\|\eta_l \sum_{s=0}^{\tau-1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_{t,s}^i\right\|^2\right] \\
&= \eta_l^2 \mathbb{E}\left[\left\|\frac{1}{N} \sum_{s=0}^{\tau-1} \sum_{i=1}^{N} \mathbf{g}_{t,s}^i\right\|^2\right] \\
&= \eta_l^2 \mathbb{E}\left[\left\|\sum_{s=0}^{\tau-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_{t,s}^i - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i)\right) + \sum_{s=0}^{\tau-1} \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i)\right\|^2\right] \\
&\leq \frac{\eta_l^2 \tau}{N} \sigma^2 + \eta_l^2 \tau \sum_{s=0}^{\tau-1} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i)\right\|^2\right]. \qquad \text{(G.37)}
\end{aligned}
$$

where the inequalities holds by the fact of $\mathbf{g}_{t,s}^i$ is the unbiased estimator of $\nabla f_i(\mathbf{x}_{t,s}^i)$ and by Assumption C.3. This concludes the proof. □

G.4.2 Partial participation

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

Lemma G.6. The global model difference $\Delta_t = \sum_{k=1}^{K} \sum_{i\in cS_t} \Delta_t^i$ in partial participation settings satisfies

$$
\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] \leq \frac{2\eta_l^2 \tau}{N} \sigma^2 + 2\eta_l^2 (\tau - 1) \sum_{s=0}^{\tau-2} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i)\right\|^2\right] + 4\eta_l^2 \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\right\|^2\right] \\
+ 8\left(\frac{n-m}{m(n-1)} + \eta_l^2 L^2\right)\left(\frac{1}{N} \sum_{k=1}^{K} \mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2]\right) + \frac{2\eta_l^2 \sigma^2}{N}\left(\frac{n-m}{m} \cdot \rho_{\max}^2\right). \qquad \text{(G.38)}
\end{aligned}
$$

Proof. Recall the definition in C, there are $\bar{\mathbf{x}}_{t,s} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{t,s}^i$ (without consideration of client sampling) and the intra-cluster average $\bar{\mathbf{x}}_{t,s+1}^k = \bar{\mathbf{x}}_{t,s}^k - \eta_l \bar{\mathbf{x}}_{t,s+1}^k$, where $\bar{\mathbf{x}}_{t,s+1}^k = \frac{1}{n} \sum_{i\in\mathcal{V}_k} \mathbf{g}_{t,s}^i$. Consider the partial participation in the last step bfore the communication round, there is $\bar{\mathbf{x}}_{t+1} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{i\in\mathcal{S}_t^k} \mathbf{x}_{t,\tau}^i = \frac{1}{Np} \sum_{k=1}^{K} \sum_{i\in\mathcal{S}_t^k} \mathbf{x}_{t,\tau}^i$ and by Algorithm 2, there is $\bar{\mathbf{x}}_{t+1} - \mathbf{x}_t = \Delta_t$. For the model difference Δ_t, we have

$$
\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{i\in\mathcal{S}_t^k} \mathbf{x}_{t,\tau}^i - \mathbf{x}_t\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \frac{1}{m} \sum_{i\in\mathcal{S}_t^k} \mathbf{x}_{t,\tau}^i \mp \bar{\mathbf{x}}_{t,\tau-1} \mp \cdots \mp \bar{\mathbf{x}}_{t,1} - \mathbf{x}_t\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{1}{Np} \sum_{k=1}^{K} \sum_{i\in\mathcal{S}_t^k} \mathbf{x}_{t,\tau}^i - \bar{\mathbf{x}}_{t,\tau-1}\right\|^2\right] + 2\mathbb{E}\left[\left\|\bar{\mathbf{x}}_{t,\tau-1} \mp \cdots \mp \bar{\mathbf{x}}_{t,1} - \mathbf{x}_t\right\|^2\right], \qquad \text{(G.39)}
\end{aligned}
$$

where the inequality holds by Cauchy-Schwarz inequality. For the first term in equation (G.39), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_t^k}\mathbf{x}_{t,\tau}^i-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_t^k}\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\mathbf{g}_{t,\tau-1}^j)\Big)-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\mathbf{g}_{t,\tau-1}^j)\Big)-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\nabla f_i(\mathbf{x}_{t,\tau-1}^j))\Big)-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&\quad+\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)\Big)\Big\|^2\Big]. \qquad \text{(G.40)}
\end{aligned}
$$

For the first term in equation (G.40), we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j))\Big)-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j)-\bar{\mathbf{x}}_{t,\tau-1}^k)\Big)\Big\|^2\Big]\\
&\le 2\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\mathbf{x}_{t,\tau-1}^j-\bar{\mathbf{x}}_{t,\tau-1}^k-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j)+\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\big)\Big)\Big\|^2\Big]\\
&\quad+2\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big)\Big\|^2\Big]\\
&\le 2\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\mathbf{x}_{t,\tau-1}^j-\bar{\mathbf{x}}_{t,\tau-1}^k-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j)+\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\big)\Big)\Big\|^2\Big]\\
&\quad+2\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big\|^2\Big], \qquad \text{(G.41)}
\end{aligned}
$$

where the first equation holds because

$$
\begin{aligned}
\bar{\mathbf{x}}_{t,\tau-1}&=\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbf{x}_{t,\tau-1}^i=\frac{1}{K}\sum_{k=1}^{K}\bar{\mathbf{x}}_{t,\tau-1}^k=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\bar{\mathbf{x}}_{t,\tau-1}^k\\
&=\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\bar{\mathbf{x}}_{t,\tau-1}^k=\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\bar{\mathbf{x}}_{t,\tau-1}^k, \qquad \text{(G.42)}
\end{aligned}
$$

and the second inequality holds by the similar relationship

$$
\begin{aligned}
&\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big)\Big)\\
&=\frac{1}{Np}\sum_{k=1}^{K}np\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)=\frac{1}{K}\sum_{k=1}^{K}\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k). \qquad \text{(G.43)}
\end{aligned}
$$

Denotes $\mathbf{y}_t^{k,i} = \sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\mathbf{x}_{t,\tau-1}^j - \bar{\mathbf{x}}_{t,\tau-1}^k - \eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j) + \eta_l\nabla \bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\big)$

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{i=1}^n\sum_{i\in\mathcal{V}_k}\mathbb{I}\{i\in\mathcal{S}_k^t\}\mathbf{y}_t^{k,i}\Big\|^2\Big] = \mathbb{E}\Big[\Big\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbb{P}\{i\in\mathcal{S}_k^t\}\mathbf{y}_t^{k,i}\Big\|^2\Big]\\
&= \mathbb{E}\Big[\mathbb{P}\{i\in\mathcal{S}_k^t\}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2 + \mathbb{P}\{i\neq j\in\mathcal{S}_k^t\}\sum_{k=1}^K\sum_{i\neq j\in\mathcal{V}_k}\big(\mathbf{y}_t^{k,i}\big)'\big(\mathbf{y}_t^{j}\big)\\
&\quad + \mathbb{P}\{i\in\mathcal{S}_k^t, j\in\mathcal{S}_l^t|k\neq l\in[K]\}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\mathbf{y}_t^{i,k}\big)'\big(\mathbf{y}_t^{j,l}\big)\Big]\\
&= \mathbb{E}\Big[\frac{m}{n}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2 + \frac{m(m-1)}{n(n-1)}\sum_{k=1}^K\sum_{i\neq j\in\mathcal{V}_k}\big(\mathbf{y}_t^{k,i}\big)'\big(\mathbf{y}_t^{k,j}\big)\\
&\quad + \frac{m^2}{n^2}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\mathbf{y}_t^{i,k}\big)'\big(\mathbf{y}_t^{j,l}\big)\Big]\\
&= \mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\Big\|^2 + \frac{m(n-m)}{n(n-1)}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2\\
&\quad + \frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\big(\mathbf{y}_t^{i,k}\big)'\big(\mathbf{y}_t^{j,l}\big)\Big]\\
&\leq \mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\Big\|^2 + \frac{m(n-m)}{n(n-1)}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2\\
&\quad + \frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\Big(\frac{1}{2}\big\|\mathbf{y}_t^{i,k}\big\|^2 + \frac{1}{2}\big\|\mathbf{y}_t^{j,l}\big\|^2\Big)\Big],
\end{aligned}
\tag{G.44}
$$

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 $\langle\mathbf{a},\boldsymbol{b}\rangle = \frac{1}{2}[\|\mathbf{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\mathbf{a}-\boldsymbol{b}\|^2]$, $\frac{1}{2}\sum_{i\neq j}\|\mathbf{a}_i - \mathbf{a}_j\|^2 = \sum_{i=1}^n n\|\mathbf{a}_i\|^2 - \|\sum_{i=1}^n\mathbf{a}_i\|^2$, and $\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\|^2 = \sum_{k=1}^K\|\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\|^2 + \sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\langle\mathbf{y}_t^{k,i}\mathbf{y}_t^{l,j}\rangle$. The last inequality holds by $\mathbf{a}'\boldsymbol{b}\leq\frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$. Re-organize the last item,

$$
\frac{m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}\Big(\frac{1}{2}\big\|\mathbf{y}_t^{i,k}\big\|^2 + \frac{1}{2}\big\|\mathbf{y}_t^{j,l}\big\|^2\Big) = \frac{m(n-m)}{n^2(n-1)}(K-1)n\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\|\mathbf{y}_t^{k,i}\|^2,
\tag{G.45}
$$

then we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{k\in[K]}\sum_{i=1}^n\mathbb{I}\{i\in\mathcal{S}_k^t\}\mathbf{y}_t^{k,i}\Big\|^2\Big]\\
&\leq \mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\Big\|^2 + \frac{m(n-m)}{n(n-1)}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2\\
&\quad + \frac{m(n-m)}{n^2(n-1)}(K-1)n\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{i}\big\|^2\Big]\\
&= \mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\Big\|^2 + \frac{Km(n-m)}{n(n-1)}\sum_{k=1}^K\sum_{i\in\mathcal{V}_k}\big\|\mathbf{y}_t^{k,i}\big\|^2\Big].
\end{aligned}
\tag{G.46}
$$

By the definition of $\mathbf{y}_t^{k,i}$, we have

$$
\begin{aligned}
\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbf{y}_t^{k,i}\Big\|^2 &= \Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\mathbf{x}_{t,\tau-1}^{j}-\bar{\mathbf{x}}_{t,\tau-1}^{k}-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})+\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k})\big)\Big\|^2\\
&= \eta_l^2\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k})-\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})\big)\Big\|^2\\
&= \eta_l^2\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\nabla f_j(\bar{\mathbf{x}}_{t,\tau-1}^{k})-\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})\big)\Big\|^2\\
&\le \eta_l^2NL^2\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\|\bar{\mathbf{x}}_{t,\tau-1}^{k}-\mathbf{x}_{t,\tau-1}^{j}\|^2\\
&\le \eta_l^2NL^2\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\bar{\mathbf{x}}_{t,\tau-1}^{k}-\mathbf{x}_{t,\tau-1}^{i}\|^2\\
&= \eta_l^2NL^2\sum_{k=1}^{K}\|X_{t,\tau-1}^{k,\perp}\|^2,
\end{aligned}
\tag{G.47}
$$

where the second equation holds by $\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\bar{\mathbf{x}}_{t,\tau-1}^{k} = \sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\bar{\mathbf{x}}_{t,\tau-1}^{k}$, the third equation holds by $\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k}) = \sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\nabla f_j(\bar{\mathbf{x}}_{t,\tau-1}^{k})$, the first inequality holds by L-smoothness and Cauchy-Schwarz inequality, the second inequality holds by the assumption of doubly stochastic mixing matrix, and the last inequality is due to the previous definition of $X_{t,\tau-1}^{k,\perp}$. By the definition of $\mathbf{y}_t^{k,i}$, we also have

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\mathbf{y}_t^{k,i}\|^2 &= \sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\Big\|\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\big(\mathbf{x}_{t,\tau-1}^{j}-\bar{\mathbf{x}}_{t,\tau-1}^{k}-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})+\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k})\big)\Big\|^2\\
&\le \sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}\Big\|\mathbf{x}_{t,\tau-1}^{j}-\bar{\mathbf{x}}_{t,\tau-1}^{k}-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})+\eta_l\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k})\Big\|^2\\
&\le \sum_{k=1}^{K}\sum_{j\in\mathcal{V}_k}\Big(2\|\mathbf{x}_{t,\tau-1}^{j}-\bar{\mathbf{x}}_{t,\tau-1}^{k}\|^2+2\eta_l^2\|\nabla f_j(\mathbf{x}_{t,\tau-1}^{j})-\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^{k})\|^2\Big)\\
&\le \sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}2\|\mathbf{x}_{t,\tau-1}^{i}-\bar{\mathbf{x}}_{t,\tau-1}^{k}\|^2+\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}2\eta_l^2L^2\|\mathbf{x}_{t,\tau-1}^{i}-\bar{\mathbf{x}}_{t,\tau-1}^{k}\|^2\\
&= \sum_{k=1}^{K}2(1+\eta_l^2L^2)\|X_{t,\tau-1}^{k,\perp}\|^2,
\end{aligned}
\tag{G.48}
$$

where the first inequality holds by the fact of $0 \le (W_k))_{i,j} \le 1$ and$\|w_{11}\mathbf{a}_1+w_{12}\mathbf{a}_2+\cdots+w_{1n}\mathbf{a}_n\|^2 = w_{11}^2\|\mathbf{a}_1\|^2+w_{12}^2\|\mathbf{a}_2\|^2+\cdots+w_{1n}^2\|\mathbf{a}_n\|^2 \le w_{11}\|\mathbf{a}_1\|^2+w_{12}\|\mathbf{a}_2\|^2+\cdots+w_{1n}\|\mathbf{a}_n\|^2$. The second inequality holds by Cauchy-Schwarz inequality, the third one holds by L-smoothness (Assumption C.6) and the double-stochasticity of matrix W_k. Hence for (G.41), we have the following

result

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\mathbf{x}_{t,\tau-1}^j-\eta_l\nabla f_j(\mathbf{x}_{t,\tau-1}^j))\Big)-\bar{\mathbf{x}}_{t,\tau-1}\Big\|^2\Big]\\
&\le 2\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big\|^2\Big]+\frac{2}{(Np)^2}\Big(\frac{m(m-1)}{n(n-1)}\eta_l^2NL^2+\frac{Km(n-m)}{n(n-1)}2(1+\eta_l^2L^2)\Big)\sum_{k=1}^{K}\mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2]\\
&\le 2\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big\|^2\Big]+\Big(\frac{m-1}{Km(n-1)}2\eta_l^2L^2+\frac{n-m}{Kmn(n-1)}4(1+\eta_l^2L^2)\Big)\sum_{k=1}^{K}\mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2]\\
&\le 2\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla\bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k)\Big\|^2\Big]+\frac{4}{N}\Big(\frac{n-m}{m(n-1)}+\eta_l^2L^2\Big)\sum_{k=1}^{K}\mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2],
\end{aligned}
\tag{G.49}
$$

where the last inequality holds with the constraint on the learning rate $\eta_l \le \frac{1}{8\tau L}$.

For the second term in equation (G.40), we perform the similar strategies related to the client sampling and weighted gossip mixing (corresponding to equation (G.46))

$$
\begin{aligned}
&\eta_l^2\mathbb{E}\Big[\Big\|\frac{1}{Np}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{I}(i\in\mathcal{S}_t^k)\underbrace{\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)\Big)}_{\mathbf{e}_t^{k,i}}\Big\|^2\Big]\\
&=\frac{\eta_l^2}{(Np)^2}\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbf{e}_t^{k,i}\Big\|^2+\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\mathbf{e}_t^{k,i}\|^2\\
&\quad+\frac{2m(n-m)}{n^2(n-1)}\sum_{k\neq l}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_l}(\mathbf{e}_t^{i,k})'(\mathbf{e}_t^{j,l})\Big]\\
&=\frac{\eta_l^2}{(Np)^2}\mathbb{E}\Big[\frac{m(m-1)}{n(n-1)}\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbf{e}_t^{k,i}\Big\|^2+\frac{m(n-m)}{n(n-1)}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\|\mathbf{e}_t^{k,i}\|^2\Big],
\end{aligned}
\tag{G.50}
$$

where the second equation holds by the stochastic gradient noise is zero mean and independent distributed. By the independency of $(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)$ and the double-stochasticity of W_k, i.e., $W_k\mathbf{1}=\mathbf{1}$ and $\mathbf{1}^\top W_k=\mathbf{1}^\top$, we have

$$
\begin{aligned}
\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbf{e}_t^{k,i}\Big\|^2\Big]&=\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)\Big)\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)\Big\|^2\Big]\\
&=N\sigma^2,
\end{aligned}
\tag{G.51}
$$

and

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}[\|\mathbf{e}_t^{k,i}\|^2]&=\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\mathbb{E}\Big[\Big\|\Big(\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}(\nabla f_i(\mathbf{x}_{t,\tau-1}^j)-\mathbf{g}_{t,\tau-1}^j)\Big)\Big\|^2\Big]\\
&=\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_k}\sum_{j\in\mathcal{V}_k}(W_k)_{i,j}^2\sigma^2\\
&=\sum_{k=1}^{K}\|W_k\|^2\sigma^2.
\end{aligned}
\tag{G.52}
$$

For the Frobenius norm $\|W_k\|^2$, by the fact of the Frobenius norm of a matrix is equal to $L-2$ norm its singular values, denote the singular values of W_k as $d_{k,1} \leq d_{k,2} \leq \cdots \leq d_{k,n} = 1$, we have

$$\sum_{k=1}^{K} \|W_k\|^2 \sigma^2 = \sum_{k=1}^{K} \sigma^2 \sum_{i=1}^{n} (d_k^i)^2 \leq \sum_{k=1}^{K} \sigma^2 (1 + (n-1)\rho_k^2) \leq K\sigma^2 (1 + (n-1)\rho_{\max}^2). \quad \text{(G.53)}$$

Summarize the items, we have

$$\begin{aligned}
&\eta_l^2 \mathbb{E}\Bigg[\Bigg\| \frac{1}{Np} \sum_{k=1}^{K} \sum_{i \in \mathcal{V}_k} \mathbb{I}(i \in \mathcal{S}_t^k) \underbrace{\Bigg(\sum_{j \in \mathcal{V}_k} (W_k)_{i,j} (\nabla f_i(\mathbf{x}_{t,\tau-1}^j) - \mathbf{g}_{t,\tau-1}^j) \Bigg)}_{\mathbf{e}_t^{k,i}} \Bigg\|^2 \Bigg] \\
&\leq \frac{\eta_l^2}{(Np)^2} \frac{m(m-1)}{n(n-1)} N\sigma^2 + \frac{\eta_l^2}{(Np)^2} \frac{m(n-m)}{n(n-1)} K\sigma^2 (1 + (n-1)\rho_{\max}^2) \\
&= \frac{\eta_l^2}{Np} \frac{m-1}{n-1} \sigma^2 + \frac{\eta_l^2}{Np} \frac{n-m}{n(n-1)} \sigma^2 + \frac{\eta_l^2}{Np} \frac{n-m}{n} \sigma^2 \rho_{\max}^2 \\
&= \frac{\eta_l^2}{N} \sigma^2 + \frac{\eta_l^2}{N} \frac{n-m}{m} \sigma^2 \rho_{\max}^2. \qquad \text{(G.54)}
\end{aligned}$$

Therefore, we have the following result,

$$\begin{aligned}
&\mathbb{E}[\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_{t,\tau-1}\|^2] \\
&\leq 2\eta_l^2 \mathbb{E}\Bigg[\Bigg\| \frac{1}{K} \sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k) \Bigg\|^2 \Bigg] + 4\Bigg(\frac{n-m}{m(n-1)} + \eta_l^2 L^2 \Bigg) \Bigg(\frac{1}{N} \sum_{k=1}^{K} \mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2] \Bigg) \\
&\quad + \frac{\eta_l^2 \sigma^2}{N} \Bigg(1 + \frac{n-m}{m} \cdot \rho_{\max}^2 \Bigg). \qquad \text{(G.55)}
\end{aligned}$$

For the remaining part in equation (G.39), we have

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg\| \bar{\mathbf{x}}_{t,\tau-1} \pm \cdots \pm \bar{\mathbf{x}}_{t,1} - \mathbf{x}_t \Bigg\|^2 \Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\| \eta_l \sum_{s=0}^{\tau-2} \bar{\mathbf{g}}_{t,s} \Bigg\|^2 \Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\| \frac{\eta_l}{N} \sum_{s=0}^{\tau-2} \sum_{i=1}^{N} \mathbf{g}_{t,s}^i \Bigg\|^2 \Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\| \frac{\eta_l}{N} \sum_{s=0}^{\tau-2} \sum_{i=1}^{N} (\mathbf{g}_{t,s}^i \pm \nabla f_i(\mathbf{x}_{t,s}^i)) \Bigg\|^2 \Bigg] \\
&= \mathbb{E}\Bigg[\Bigg\| \frac{\eta_l}{N} \sum_{s=0}^{\tau-2} \sum_{i=1}^{N} (\mathbf{g}_{t,s}^i - \nabla f_i(\mathbf{x}_{t,s}^i)) \Bigg\|^2 \Bigg] + \mathbb{E}\Bigg[\Bigg\| \frac{\eta_l}{N} \sum_{s=0}^{\tau-2} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \Bigg\|^2 \Bigg] \\
&\leq \frac{\eta_l^2 (\tau-1)}{N} \sigma^2 + \eta_l^2 (\tau-1) \sum_{s=0}^{\tau-2} \mathbb{E}\Bigg[\Bigg\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \Bigg\|^2 \Bigg], \qquad \text{(G.56)}
\end{aligned}$$

where the forth equation holds since the stochastic noise $(\mathbf{g}_{t,s}^i - \nabla f_i(\mathbf{x}_{t,s}^i))$ is zero mean Hence for partial participation, the model difference Δ satisfies

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] \leq{}& \frac{2\eta_l^2 \tau}{N} \sigma^2 + 2\eta_l^2 (\tau-1) \sum_{s=0}^{\tau-2} \mathbb{E}\Bigg[\Bigg\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_{t,s}^i) \Bigg\|^2 \Bigg] + 4\eta_l^2 \mathbb{E}\Bigg[\Bigg\| \frac{1}{K} \sum_{k=1}^{K} \nabla \bar{f}_k(\bar{\mathbf{x}}_{t,\tau-1}^k) \Bigg\|^2 \Bigg] \\
&+ 8\Bigg(\frac{n-m}{m(n-1)} + \eta_l^2 L^2 \Bigg) \Bigg(\frac{1}{N} \sum_{k=1}^{K} \mathbb{E}[\|X_{t,\tau-1}^{k,\perp}\|^2] \Bigg) + \frac{2\eta_l^2 \sigma^2}{N} \Bigg(\frac{n-m}{m} \cdot \rho_{\max}^2 \Bigg). \qquad \text{(G.57)}
\end{aligned}$$

□

G.5 Additional Supporting Lemmas

Lemma G.7 (Lemma for momentum term in the update rule)**.** The first order momentum terms $\mathbf{m}_t$ in Algorithm 1 and 2 hold the following relationship w.r.t. model difference Δ_t:

$$\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{m}_t\|^2] \leq \sum_{t=1}^{T}\mathbb{E}[\|\Delta_t\|^2]. \tag{G.58}$$

Proof. By the updating rule, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{m}_t\|^2] &= \mathbb{E}\left[\left\|(1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\Delta_u\right\|^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\Delta_{u,i}\right)^2\right] \\
&\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{u=1}^{t}\beta_1^{t-u}\right)\left(\sum_{u=1}^{t}\beta_1^{t-u}\Delta_{u,i}^2\right)\right] \\
&\leq (1-\beta_1)\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2],
\end{aligned} \tag{G.59}$$

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

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{m}_t\|^2] &= (1-\beta_1)\sum_{t=1}^{T}\sum_{u=1}^{t}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2] \\
&= (1-\beta_1)\sum_{u=1}^{T}\sum_{t=u}^{T}\beta_1^{t-u}\mathbb{E}[\|\Delta_u\|^2] \\
&\leq (1-\beta_1)\sum_{u=1}^{T}\frac{1}{1-\beta_1}\mathbb{E}[\|\Delta_u\|^2] \\
&= \sum_{u=1}^{T}\mathbb{E}[\|\Delta_u\|^2].
\end{aligned} \tag{G.60}$$

This concludes the proof. □

Lemma G.8. Under Assumptions C.2, for HA-Fed, we have $\|\nabla f(\mathbf{x})\| \leq G$, $\|\Delta_t\| \leq \eta_l\tau G$, $\|\mathbf{m}_t\| \leq \eta_l\tau G$ and $\|\mathbf{v}_t\| \leq \eta_l^2\tau^2G^2$.

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

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

For HA-Fed, the model difference $\bar{\Delta}_t^k$ on cluster k satisfies,

$$\bar{\Delta}_t^k = \bar{\mathbf{x}}_{t,\tau}^k - \mathbf{x}_t = -\eta_l\sum_{s=0}^{\tau-1}\bar{\mathbf{g}}_{t,s}^k,$$

therefore,

$$\|\bar{\Delta}_t^k\| = \left\|\eta_l\sum_{s=0}^{\tau-1}\bar{\mathbf{g}}_{t,s}^k\right\| = \left\|\eta_l\sum_{s=0}^{\tau-1}\frac{1}{K}\sum_{i\in\mathcal{V}_k}\mathbf{g}_{t,s}^i\right\| \leq \eta_l\tau G,$$

for the global model difference Δ_t,

$$\|\Delta_t\| = \left\|\frac{1}{K}\sum_{k\in[K]}\bar{\Delta}_t^k\right\| \leq \eta_l\tau G.$$

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

$$\|\mathbf{m}_t\| = \left\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\Delta_t\right\| \leq \eta_l \tau G, \quad \|\mathbf{v}_t\| = \left\|(1-\beta_2)\sum_{\tau=1}^{t}\beta_2^{t-\tau}\Delta_t^2\right\| \leq \eta_l^2 \tau^2 G^2.$$

This concludes the proof. □

Lemma G.9. For the variance difference sequence $\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}$, we have

$$\sum_{t=1}^{T}\left\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\right\|_1 \leq \frac{d}{\sqrt{\epsilon}}, \quad \sum_{t=1}^{T}\left\|\widehat{\mathbf{V}}_{t-1}^{-1/2} - \widehat{\mathbf{V}}_t^{-1/2}\right\|^2 \leq \frac{d}{\epsilon} \tag{G.61}$$

Proof. The proof of Lemma G.9 is exactly the same as the proof of Lemma C.2 in [43]. □

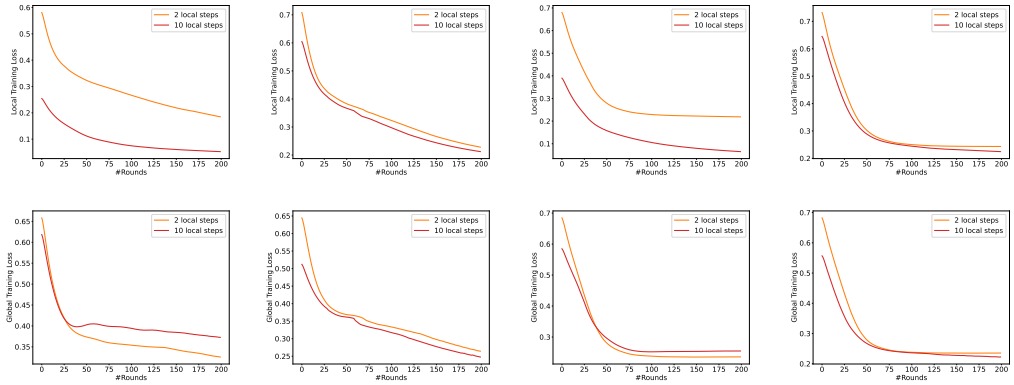
Lemma G.10. For the element-wise difference, $W_t = \frac{1}{\sqrt{\mathbf{v}_t+\epsilon}} - \frac{1}{\sqrt{\beta_2 \mathbf{v}_{t-1}+\epsilon}}$, we have $\|W_t\| \leq \frac{\sqrt{1-\beta_2}}{\epsilon}\|\Delta_t\|$.

Proof. The proof of Lemma G.9 is exactly the same as the proof of Lemma C.1 in [43]. □

H Additional Experiments

H.1 Simulation Study

We conduct the simulation study with synthetic data to verify the *dilemma of local steps*. The synthetic data is generated by Gaussian distribution and is heterogeneous among clients. We generate a simple set of data for 2 clients in the 2 dimensional space, in which there is assigned 10 data sample $(\mathbf{x}_i^{(k)}, y_i^{(k)})$ corresponding to client k. For each client k, $\mathbf{x}_i^{(k)}$ has mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}$, i.e., $x_i^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. The mean $\boldsymbol{\mu}_k$ varies from each client, thus data on each client are generating from different distributions. Specifically, we set $\boldsymbol{\mu}_1 = (1, 0)$ and $\boldsymbol{\mu}_2 = (0, 1)$, and labels of two clients are setup with $y_i^{(1)} = 0$ and $y_i^{(2)} = 1$. For both FedAMSGrad and HA-Fed, we simulate the full participation setting and use full batch gradients for better illustration. We perform two groups of comparison: (1) data sample $\mathbf{x}_i^{(k)}$ has covariance matrix $\mathbf{I}_2$, and we train the data on a simple multilayer-perceptron that has a hidden layer of 10 units; (2) $\mathbf{x}_i^{(k)}$ has a smaller covariance matrix $0.5 \cdot \mathbf{I}_2$, and we train the data on a simple multilayer-perceptron that has a hidden layer of 4 units; for both groups of comparison, we use ReLu activation function after the input layer.



(a) FedAMSGrad (b) HA-Fed (c) FedAMSGrad (d) HA-Fed

Figure 2: The comparison of local training loss and global training loss for FedAMSGrad and HA-Fed. (a)(b) is the comparison for group (1), (c)(d) is the comparison for group (2).

Figure 2 shows the comparison of *local training loss* and *global training loss* for FedAMSGrad and our proposed HA-Fed algorithms. The local training loss is calculated during the training process, which reflects the average training loss on each clients with the local updated model. The global training loss is calculated in each global round after the local updates, it merges the training data among clients and calculates the training loss with all data samples. For example, suppose we have N local client and each of which corresponds to a local loss function f_i, $i = 1, ..., N$. Denote $\mathbf{w}_i$ as the local model on client i, $\xi_i \sim \mathcal{D}_i$ is the data sample on client i with distribution $\mathcal{D}_i$, then the *local training loss* is calculated by $f_{\text{local}} = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{w}_i; \xi_i)$. Moreover, after each round of local updates, we aggregate the temporary local models $\mathbf{w}' = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_i$, together with the training data $\xi' = \bigcup_{i=1}^{N} \xi_i$ from all clients, the *global training loss* is calculated by $f_{\text{global}} = f(\mathbf{w}'; \xi')$, where $f = \frac{1}{N}\sum_{i=1}^{N} f_i$. The comparison between local and global training losses can reflect whether there exists over-fitting issues on local clients, if the local training loss decrease rapidly but the global training loss keeps a large value, it means the training with heterogeneous data causes over-fitting in clients.

From plot (a) and (c) in Figure 2, they show that FedAMSGrad with a larger number of local steps has higher global training loss compared to a smaller number of local steps. We observe that when local steps increase from $\tau = 2$ (orange) to $\tau = 10$ (red), FedAMSGrad gets worse in reducing global training loss though it improves the convergence to a locally optimal point. This shows that FedAMSGrad faces the *dilemma of local steps*, i.e., the convergence rate gets worse as the number of local steps increases. In contrast, plot (b) and (d) in Figure 2 show that our proposed HA-Fed effectively overcome this issue. For HA-Fed, the global training loss further decreases as the number of local steps increases from $\tau = 2$ to $\tau = 10$. This observation provides the empirical evidence that verifies our theory and shows that our proposed HA-fed can indeed overcome the *dilemma of local steps* in adaptive federated optimization methods.

H.2 Non-i.i.d. Sampling and Hyperparameter Settings

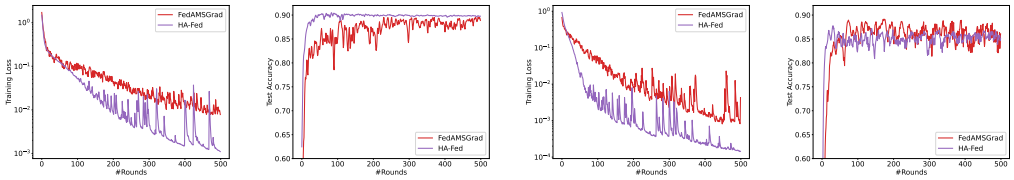
Non-i.i.d. data sampling For the non-i.i.d. data sampling, we sort the training data by labels, and divide the data by labels. For CIFAR-10 [25] dataset, for each label, we divide the data into 20 shards size of 250. Hence we get 200 shards of size 250 in total, and each client is randomly assigned six shards. For Fashion MNIST [44] dataset, we similarly divide the data into 20 shards size of 300 for each label, and each client is randomly assigned six shards as well. The similar non-i.i.d. data sampling strategy is adopted in [31, 45, 16].

Hyperparameter Settings We conduct detailed hyperparameter searches to find the best hyper-parameter for both FedAMSGrad and HA-Fed algorithms. We grid over the local learning rater $\eta_l \in \{0.001, 0.01, 0.1, 1.0\}$, and the global learning rate $\eta \in \{0.0001, 0.0005, 0.001, 0.01, 0.1\}$ for two methods. For the global AMSGrad optimizer, we set $\beta_1 = 0.9$, $\beta_1 = 0.99$, and we search the best ϵ from $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$.

Specifically, for the ResNet-18 experiments on CIFAR-10 dataset, we set the local learning rate $\eta_l = 0.1$ and the global learning rate $\eta = 0.0005$ for FedAMSGrad and $\eta_l = 0.1$, $\eta = 0.001$ for HA-Fed, set $\epsilon = 10^{-8}$ for both methods. For training ConvMixier-256-8 model on CIFAR-10 dataset, we set $\eta_l = 0.1, \eta = 0.001$ for FedAMSGrad, and $\eta_l = 0.1, \eta = 0.01$ for HA-Fed, set $\epsilon = 10^{-8}$ for both methods. For training ConvMixer-256-8 model on Fashion MNIST dataset, we set $\eta_l = 0.1, \eta = 0.0005$ for FedAMSGrad, $\eta_l = 0.1, \eta = 0.001$ for HA-Fed, and we set $\epsilon = 10^{-8}$ for both methods. For the CNN experiments on Fashion MNIST dataset, we set $\eta_l = 0.1, \eta = 0.01$ and $\epsilon = 10^{-8}$ for both two methods. All the experiments are set up with 32 total clients in the network, which are equally divided into 4 clusters. The partial participation ratio is set to $p = 0.25$, and the intra-cluster topology is ring topology by default. For both FedAMSGrad and HA-Fed, we conduct $E = 2$ epochs of local training with a batch size of 50.

H.3 Additional Experiments

In this section, we present additional empirical experiments for our proposed HA-Fed and FedAMSGrad algorithm in training CNN, and ConvMixer-256-8 model [39] on Fashion MNIST [44] dataset, where the CNN model [8] contains around 29 thousand of trainable parameters.



(a) CNN (b) ConvMixer-256-8

Figure 3: The learning curves for HA-Fed and FedAMSGrad in training Fashion MNIST data on (a) CNN model and (b) ConvMixer-256-8 model using ring topology for intra-cluster communications.

Figure 3 shows the empirical convergence result for HA-Fed and FedAMSGrad on training Fashion-MNIST with CNN and ConvMixer-256-8 model. We compare the training loss and test accuracy against global rounds for two algorithms. Plot (a) in Figure 3 shows that HA-Fed (purple) achieves faster convergence than FedAMSGrad when training the CNN model to reduce training losses and obtain high test accuracy. For the ConvMixer-256-8 model (Figure (b)), HA-Fed again shows its faster convergence in reducing training loss, and HA-Fed maintains a similar test accuracy as FedAMSGrad.

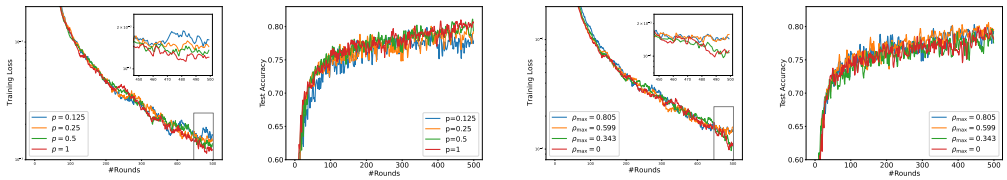
Model: CNN	Test Accuracy (%)	Model: ConvMixer-256-8	Test Accuracy (%)
FedAMSGrad	$88.79 \pm .16$	FedAMSGrad	83.54 ± 3.36
HA-Fed	$\mathbf{89.25} \pm \mathbf{.22}$	HA-Fed	$\mathbf{84.49} \pm \mathbf{1.57}$

Table 1: The test accuracy (with mean and standard error) results with three random seeds for training Fashion MNIST data on CNN and ConvMixer-256-8 model.

Moreover, Table 1, present the average test accuracy after three runs with different random seeds, and it shows that our proposed HA-Fed holds a higher accuracy than FedAMSGrad for all three datasets, with particular a stabler result for most of the experiments. These results empirically demonstrate the effectiveness and efficiency of our proposed HA-Fed method.

Now we study how the participation ratio and network topology would affect the convergence of our proposed HA-Fed algorithm. Figure 4 (a) illustrates the ablation study on the participation ratio p. Specifically, we test various values of p from 0.125 to 1. From Figure 4 (a), we can observe that a larger participation ratio slightly improves the performance, with faster convergence on training loss and higher test accuracy, in addition, we also observe that larger p also helps to stabilize the test accuracy. This is consistent with our theoretical results that increasing the number of participating clients improves the convergence rate, but the improvement is slight compared to a large number of global and local steps. Figure 4 (b) shows ablation study on on clusters' maximum spectral gap $\rho_{\max}$. Specifically, we compare various of $\rho_{\max}$ calculating from different network typologies, where $\rho_{\max} = 0.805$ corresponds to the default ring topology, $\rho_{\max} = 0$ corresponds to fully-connected networks in clusters, $\rho_{\max} = 0.599$ and $\rho_{\max} = 0.343$ are obtained from the Erdős-Rényi model [10, 13] for generating random graphs with edge probabilities 0.5 and 0.8. From plot (b) in Figure 4, we observe that smaller $\rho_{\max}$ roughly contributes to faster convergence on training loss. Moreover, we observe that smaller $\rho_{\max}$ helps achieve a stabler test accuracy curve but does not help obtain an obviously higher accuracy. This result matches the theoretical result that $\rho_{\max}$ holds the non-dominant term in the convergence of HA-Fed even for partial participation scenarios. This suggests that without a dense network topology, HA-Fed can still take the benefit of gossip communication to achieve the expected convergence result.

[8] The CNN model has two 5×5 convolution layers, where the first has 16 channels, the second has 32 channels, each followed with a 2×2 max pooling step.



(a) Ablation on the participation ratio p

(b) Ablation on $\rho_{\max}$

Figure 4: The learning curves with (a) different participating ratio p and (b) different maximum spectral gap $\rho_{\max}$ of clusters in training CIFAR-10 data on ConvMixer-256-8 model.

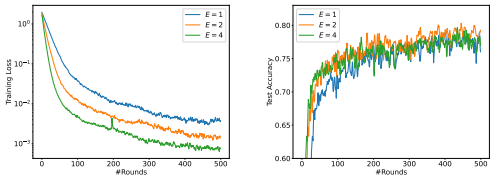


Figure 5: The learning curves with different E in training CIFAR-10 on ConvMixer-256-8 model.

We further study how the local epochs would affect the convergence of our proposed HA-Fed algorithm. Figure 5 shows the ablation study about local epochs, we compare the number of local epochs E from $E = \{1, 2, 4\}$. We observe that larger E indeed helps accelerate to convergence on training loss, as the green line in the left plot generally keeps a smaller training loss. For the test accuracy, in general, $E = 2$ has better performance for achieving a higher test accuracy though the curve for $E = 4$ grows rapidly at the beginning. This result backup our theory and show that HA-Fed indeed can resolve the *dilemma of local steps*.