Multimodal Federated Learning with Model Personalization

Ratun Rahman RR0110@UAH.EDU
Dinh C. Nguyen DINH.NGUYEN@UAH.EDU
Department of Electrical and Computer Engineering, The University of Alabama in Huntsville, Huntsville, Alabama, USA

Abstract

Federated learning (FL) has been widely studied to enable privacy-preserving machine learning (ML) model training. Most existing FL frameworks focus on unimodal data, where clients train on the same type of data, such as images or time series. However, many real-world applications naturally involve multimodal data from diverse sources. While multimodal FL has recently been proposed, it still faces challenges in managing data heterogeneity across diverse clients. This paper proposes a novel multimodal meta-FL framework termed mmFL that orchestrates multimodal learning and personalized learning. Our approach can enable the federated training of local ML models across data modality clusters while addressing the data heterogeneity across clients based on a meta-learning-based solution. Extensive simulation results show that our approach brings a significant improvement in the training performance (up to 7.18% in accuracy) compared with state-of-the-art algorithms.

1. Introduction

Federated learning (FL) has recently attracted significant attention for enabling privacy-aware machine learning by training models across distributed clients without sharing their data [9]. Traditional FL frameworks mostly consider unimodal data settings, where all clients train their machine learning (ML) models on the same type of data, e.g., image or time-series data. However, in real-world applications such as human activity monitoring and emotion recognition, systems often rely on multiple data modalities for a holistic understanding and reasoning. For example, in human activity recognition, a camera captures spatial features like body posture and movements, while wearable sensors record temporal features such as motion speed and acceleration. These complementary data sources enhance model training accuracy and robustness. Moreover, in multimodal FL systems, data across clients within each modality cluster is often non-independent and non-identically distributed (non-IID), creating new challenges for effective model training [3, 5, 10].

Multimodal learning and heterogeneous FL have been extensively studied in the literature. The first line of research focuses on *multimodal learning*, where data from different types are integrated into a dedicated server for ML model training [6, 12]. However, these works require centralized data collection and training, which raises data privacy concerns. The second line of research is *heterogeneous FL* with a focus on unimodal learning settings on a single data domain [11, 16]. To deal with data heterogeneity, personalization techniques have been proposed by allowing more diverse clients in FL and providing a performance-based impact on the global model [7, 8, 15].

Despite such research efforts, *a joint approach of multimodal learning and model personalization has been largely under-explored.* To fill this research gap, this paper proposes a novel multimodal meta-FL framework termed *mmFL* that orchestrates multimodal learning and personalized learning aimed at significantly improving the training performance of FL across heterogeneous clients. Extensive simulations performed on real-world datasets demonstrate that the proposed mmFL method improves training performance and overall accuracy up to 7.18% compared to existing methods.

2. Proposed MmFL Method

Fig. 1 illustrates our system's overall architecture, where there is a single server connected to a set of clients from different data modalities denoted as the set $\mathcal{M}$. It is assumed that each data modality $m \in \mathcal{M}$ has a set of clients $\mathcal{N}^m$. The raw data collected by the sources is denoted by $D^m_{n,k}$ for data modality m and client n. Every global communication round is denoted as $k \in \mathcal{K}$ where every client creates a local model from the local data and sends their local encoder model's weight to the decoder. The weights are denoted as $\boldsymbol{\theta}^{m,t}_{n,k}$ where $t \in \mathcal{T}$ is the local model iterations and $t = 1, 2, \dots, T$. For our meta-learning approach, we consider total j available learning rates denoted by η_i where $i \in j$, and local temporal round denoted by $t_{\text{temp}} \in \mathcal{T}_{\text{temp}}$ and $t_{\text{temp}} = 1, 2, \dots, T_{\text{temp}}$. The decoder receives the encoder models and proceeds with the decoder model for each data modality m and then does federated averaging afterward. As a result, for M types of datasets, the server has M decoder models and does M federated averaging. We can separate our system model into multiple steps as follows.

Step 1: For each global round k, the client n at data modality cluster m receives non-IID local dataset $D^m_{n,k}$ for local model raining, and the global model's weight $\boldsymbol{\theta}^m_{g,k}$. We denote the initial local model's weight as the global model before the local rounds, therefore, $\boldsymbol{\theta}^{m,0}_{n,k} = \boldsymbol{\theta}^m_{g,k}$.

Step 2: For meta-learning, in every data modality m, every client n creates a small testing dataset from $D^m_{n,k}$ denoted as $d^m_{n,k} \subset D^m_{n,k}$ and duplicate the local models as temporary local models $\boldsymbol{\theta}^m_{n,\text{temp}} = \boldsymbol{\theta}^{m,0}_{n,k}$.

Step 3: We denote the optimal learning rate as $\alpha^m_{n,k}$ for data modality m, and client n, and for calculation, we apply available learning rates η_i where $i \in j$ on $d^m_{n,k}$. The updated local model is calculated as follows.

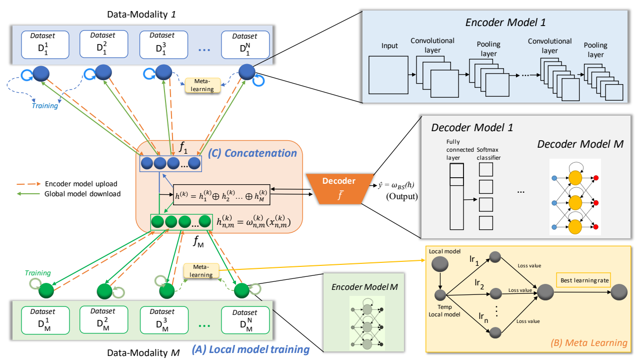


Fig. 1: The proposed mmFL framework with M modals where every client trains their encoder model and then calculates their best learning rates through meta-learning. Clients train the encoder model on the local client and the decoder model on the server and perform FL.

$$\boldsymbol{\theta}^{m,t_{\text{temp}}+1}_{n,\text{temp}} = \boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}} - \eta_i \nabla F(\boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}}, d^m_{n,k}). \tag{1}$$

Step 4: After T_{temp} rounds, we calculate the objective function (loss value) for learning rates as:

$$\min_{\boldsymbol{\theta} \in \mathrm{I}R^d} F(\boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}}) := \frac{1}{N} \sum_{n=1}^{N} f_n(\boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}}), \tag{2}$$

where $\mathrm{I}R^d$ denotes the d-dimensional real space in which the model parameters $\boldsymbol{\theta}$ reside local loss function, N is the total number of clients participating in FL.

Step 5: Here, $f_i : \mathrm{I}R^d \in \mathrm{I}R$ denotes the predicted loss value over the client's data distribution:

$$f^m_{n,k}(\boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}}) := \mathrm{I}E_{\xi_i}\left[f'^m_{n,k}(\boldsymbol{\theta}^{m,t_{\text{temp}}}_{n,\text{temp}}, \chi^m_{n,k})\right]. \tag{3}$$

Here, $\chi^m_{n,k}$ are non-IID data sample from $d^m_{n,k}$.

Step 6: We select the η_i as $\alpha^m_{n,k}$ for modality m and client n that produce the minimum $f^m_{n,k}$ value. So, we can explain it as:

$$\alpha^m_{n,k} := \underset{i\in j}{\operatorname{argmin}}(f^m_{n,k}(\eta_i)). \tag{4}$$

Step 7: We use $\alpha^m_{n,k}$ to do the encoder model training in k^{th} global round for m^{th} modality data n^{m-th} client, which can be expressed via model updating as

$$\boldsymbol{\theta}'^{m,t}_{n,k} = \boldsymbol{\theta}^{m,t}_{n,k} - \alpha^m_{n,k}\nabla F(\boldsymbol{\theta}^{m,t}_{n,k}). \tag{5}$$

Step 8: Then we send the encoder model's parameter from the client to the server for decoder model training using the same α and it is expressed as:

$$\boldsymbol{\theta}^{m,t+1}_{n,k} = \boldsymbol{\theta}'^{m,t}_{n,k} - \alpha^m_{n,k}\nabla F(\boldsymbol{\theta}'^{m,t}_{n,k}). \tag{6}$$

Step 9: After all the local rounds T, the server receives all completed models for every client n for aggregation. We also calculate the global round k^{th} loss and accuracy value from the weight $\boldsymbol{\theta}^m_{n,k}$ on the test data $D^m_{n,\text{test}}$.

$$f^m_{n,k}(\boldsymbol{\theta}^m_{n,k}) := \mathrm{I}E_{\xi_i}\left[f'^m_{n,k}(\boldsymbol{\theta}_{n,k}, D^m_{n,\text{test}})\right]. \tag{7}$$

Step 10: Once the server collects all the weights, it calculates the federated averaging separately for the modalities for the next global round $k+1$ as:

$$\boldsymbol{\theta}^m_{g,k+1} = \frac{1}{N}\sum_{n\in\mathcal{N}}\boldsymbol{\theta}^m_{n,k}. \tag{8}$$

Finally, the updated global model $\boldsymbol{\theta}^m_{g,k+1}$ is distributed across all $\mathcal{N}$ in $\mathcal{M}$ for the $K+1$ global round. After K global rounds, we finally get the optimal global model $\boldsymbol{\theta}^{*m}_n$.

Our proposed mmFL method enables clients with good datasets and better performance to use a different learning rate than those with poor performance and datasets. As a result, the global model is impacted separately for every client and the global model can have more accurate updates by using the personalized factor for enhancing decoder model training. The convergence analysis of the meta-learning-based FL within a data modaltiy cluster is given in the below appendix.

3. Experiments

3.1. Dataset and Data Processing

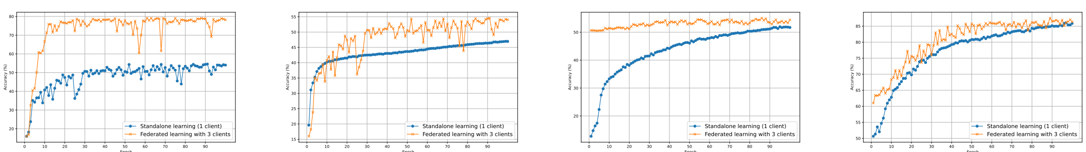
We use two datasets in this research: HAR (human activity recognition) [1] and CMU-MOSEI (Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity) [14].

HAR is a popular multimodal dataset that identifies human activities from smartphone sensor data consisting of accelerometers and gyroscopes. The activities include time series features for walking, walking upstairs, walking downstairs, sitting, standing, and laying. Accelerometer data captures linear acceleration along three axes (X, Y, and Z), allowing for the recognition of activities such as walking or running. Meanwhile, gyroscope data records angular velocity along the same axes, detecting rotational movements and changes in orientation, such as tilting or twisting, which can help identify activities like walking upstairs or sitting down. These sensors provide two types of data for effectively categorizing various human activities.

CMU-MOSEI dataset is a multimodal dataset for sentiment and emotion analysis with over 23,000 utterances from over 1,000 speakers from various YouTube videos. The emotions include happiness, sadness, anger, fear, disgust, and surprise with three perspectives: text (speech transcripts), audio (speech), and video (facial expressions and gestures). Text data is collected from the video's speech that helps in assessing words represent emotions. The audio data is made up of speech recordings, and prosodic elements such as tone, pitch, rhythm, and speech patterns are used to identify emotions.

3.2. Simulation Results

First, we compare FL results with the standalone method to show the effectiveness of collaborating learning over individual learning in Fig 2. From the figure, we can see that FL with 3 non-IID clients performs significantly better than running a model in one user in all the datasets.



(a) HAR Accelerometer (b) HAR Gyroscope (c) CMU-MOSEI Audio (d) CMU-MOSEI Text

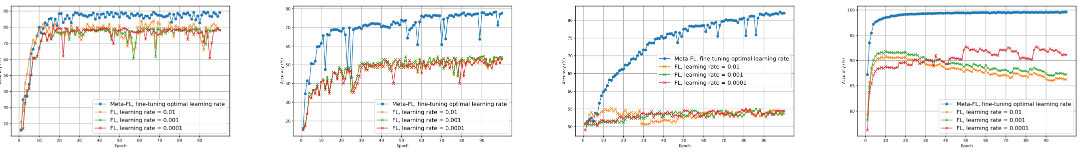
Fig. 2: Comparison between standalone learning (1 client) and FL (3 non-IID clients) in each data modality group.

Therefore, we proceed with our simulations for the collaborative approaches, particularly with FL and meta-FL. For calculating the performance, there are various loss functions available that directly affect the learning performance. In Table 1, we compare FL results on different datasets with different loss values: Cross-entropy loss, MSE loss, and BCE loss. From the table, we can see that the cross-entropy loss is more consistent than the other loss functions in handling both HAR and CMU-MOSEI data. Therefore, we will consider using cross-entropy loss for comparison.

	Features	CrossEntropy Loss		MSE Loss		BCE Loss	
		Loss	Accuracy	Loss	Accuracy	Loss	Accuracy
HAR	Multimodal FL Accelerometer	0.0672	87.93%	0.0196	95.73%	0.2885	75.13%
	Multimodal FL Gyroscope	0.2142	80.12%	0.0891	75.28%	0.3101	68.99%
	MmFL Accelerometer	0.1690	92.68%	0.0121	98.10%	0.1238	87.66%
	MmFL Gyroscope	0.1602	92.49%	0.0808	76.45%	0.0829	89.82%
CMU-MOSEI	Multimodal FL Text	0.0284	97.02%	0.0269	95.37%	0.2099	85.39%
	Multimodal FL Audio	0.2229	75.12%	0.0991	70.13%	0.1008	73.81%
	MmFL Text	0.0112	99.04%	0.0190	98.45%	0.1987	89.10%
	MmFL Audio	0.1877	87.51%	0.0880	75.55%	0.0772	87.82%

Table 1: Performance comparison (both loss and accuracy) between 3 different loss functions: CrossEntropy Loss, MSE Loss, and BCE Loss in both HAR and MOSEI multimodal data.

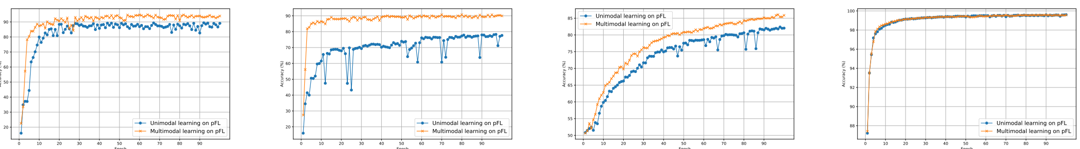
Then we compare our results with different learning rates and meta-learning that uses optimal learning rates using those learning rates in Fig. 3. The learning rates include 0.01, 0.001, and



(a) HAR Accelerometer (b) HAR Gyroscope (c) CMU-MOSEI Audio (d) CMU-MOSEI Text

Fig. 3: Comparison between different learning rates 0.01, 0.001, and 0.0001 for the FL and meta-FL with optimal learning (fine-tuning different learning rates).

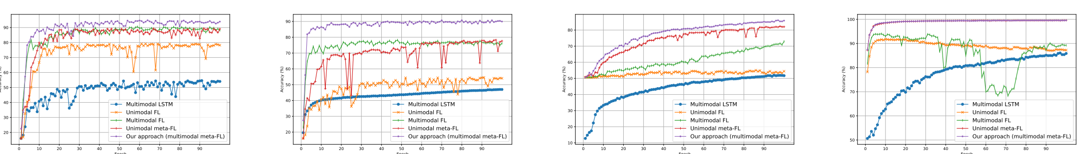
0.0001 used in comparison and meta-learning. The figure shows that meta-learning is performing significantly better than a fixed learning rate-based FL. Then we compare the unimodal meta-FL with mmFL for different datasets in Fig. 4.



(a) HAR Accelerometer (b) HAR Gyroscope (c) CMU-MOSEI Audio (d) CMU-MOSEI Text

Fig. 4: Comparison between unimodal meta-FL and mmFL in both HAR and CMU-MOSEI multimodal datasets.

From the graph, we can see that the accuracy of multimodal models is better than that of the unimodal models in all modalities of both datasets. Finally, we compare our model with other state-of-the-art methods in Fig. 5. For comparison, in the HAR model, we selected the multimodal LSTM [12], unimodal FL [4], multimodal FL [13], unimodal meta-FL [11], and our approach (mmFL). Similarly, for CMU-MOSEI, we also select multimodal LSTM [6], unimodal FL [2], multimodal FL [17], unimodal meta-FL [16], and our approach (mmFL). From the graphs, we can see that our method has outperformed all other approaches by around 7.18% across all datasets.



(a) HAR Accelerometer (b) HAR Gyroscope (c) CMU-MOSEI Audio (d) CMU-MOSEI Text

Fig. 5: Comparison between our approach and other state-of-the-art approaches in both HAR and CMU-MOSEI multimodal datasets.

4. Conclusion and Future Work

In this work, we proposed a novel meta-FL method called mmFL on two types of multimodal datasets, HAR and CMU-MOSEI. Then we compared different parameters in the multimodal to find the optimal settings. Simulation results also show that mmFL has improved the training performance than all other state-of-the-art methods by around 7.18%. However, for clients with large datasets, the mmFL method can be computationally demanding. Future works will be dedicated to resource-aware model training across data modalities, where split learning will be applied to resource-constrained clients.

References

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3, 2013.

[2] Timothy Castiglia, Shiqiang Wang, and Stacy Patterson. Flexible vertical federated learning with heterogeneous parties. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[3] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7314–7322, 2023.

[4] Gautham Krishna Gudur and Satheesh K Perepu. Federated learning with heterogeneous labels and models for mobile activity monitoring. In *Machine Learning for Mobile Health Workshop at NeurIPS 2020.*, 2020.

[5] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34:12876–12889, 2021.

[6] Louis-Philippe Morency, Paul Pu Liang, and Amir Zadeh. Tutorial on multimodal machine learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 33–38, 2022.

[7] Ratun Rahman and Dinh C Nguyen. Improved modulation recognition using personalized federated learning. *IEEE Transactions on Vehicular Technology*, 2024.

[8] Ratun Rahman, Neeraj Kumar, and Dinh C Nguyen. Electrical load forecasting in smart grid: A personalized federated learning approach. *arXiv preprint arXiv:2411.10619*, 2024.

[9] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022.

[10] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8432–8440, 2022.

[11] Jiaqi Wang, Xingyi Yang, Suhan Cui, Liwei Che, Lingjuan Lyu, Dongkuan DK Xu, and Fenglong Ma. Towards personalized federated learning via heterogeneous model reassembly. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5615–5623, 2024.

[13] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3063–3071, 2022.

[14] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[15] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–11244, 2023.

[16] Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

[17] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9100–9108, 2022.

5. Appendix

We here focus on conducting the analysis of theoretical convergence of the proposed meta-learning-based FL in a certain modality cluster, which is applicable to all modality clusters. To facilitate our theoretical convergence analysis, we summarize the key notations as follows.

- The set of global rounds: $\mathcal{K}$
- The set of local SGD iterations: $\mathcal{T}$
- The set of device: $\mathcal{N}$

To support our convergence analysis, we introduce two virtual sequences as:

$$\bar{\boldsymbol{\theta}}_k^t = \frac{1}{N}\sum_{n\in\mathcal{N}}\boldsymbol{\theta}_k^t, \qquad \bar{\boldsymbol{x}}_k^t = \frac{1}{N}\sum_{n\in\mathcal{N}}\boldsymbol{x}_k^t. \tag{9}$$

Subsequently, each client updates its personalized model as

$$\boldsymbol{x}_k^{t+1} = \boldsymbol{x}_k^t - \eta_t g_t, \qquad g_t = \nabla f(\boldsymbol{x}_k^t) + b_t + n_t \tag{10}$$

where for zero-mean noise $\mathbb{E}n_t = 0$ and bias b_k, g_t is a gradient oracle and η_t is the sequence of step sizes. If there is no bias, $b_t = 0$, it becomes the SGD setting and for no noise, $n_t = 0$, it becomes the classic gradient descent algorithm.

It is easy to observe that,

$$\bar{\boldsymbol{\theta}}_k^{t+1} = \boldsymbol{\theta}_k^t - \eta_k\nabla F(\boldsymbol{\theta}_k^t, \chi_k^t) + \eta_k B_k - \eta_k N_k, \tag{11}$$

To facilitate the analysis, we use the following common assumptions:

Assumption 1 *($L-$smoothness). Each local loss function F_n ($n \in \mathcal{N}$) is L-smooth ($L > 0$), i.e.*

$$F_n(\boldsymbol{\theta}') - F_n(\boldsymbol{\theta}) \leq \langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \nabla F(\boldsymbol{\theta})\rangle + \frac{L}{2}||\boldsymbol{\theta}' - \boldsymbol{\theta}||, \forall \boldsymbol{\theta}', \boldsymbol{\theta} \tag{12}$$

Assumption 2 *((M, σ^2)-bounded noise). There exists constant M, $\sigma^2 >= 0$ such that*

$$\mathbb{E}||n(\boldsymbol{\theta}, \xi)||^2 \leq M||\nabla F_n(\boldsymbol{\theta}) + b(\boldsymbol{\theta})||^2 + \sigma^2, \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{13}$$

Assumption 3 *((m, ζ)-bounded bias). There exists constants $0 \leq m < 1$ and $\zeta^2 \geq 0$ such that*

$$||b(\boldsymbol{\theta}||^2 \leq m||\nabla F_n(\boldsymbol{\theta})||^2 + \zeta^2, \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{14}$$

Assumption 4 *Finite parameter space: The parameter space Θ is finite, i.e, $\Theta = \chi_k^1, \chi_k^2, \chi_k^3, \ldots, \chi_k^t$. Also, $\chi_k^{'t} \in \Theta$*

There exist $0 < L_H < \infty$ such that

$$||\nabla F(\boldsymbol{\theta}, \chi_k^1) - \nabla F(\boldsymbol{\theta}, \chi_k^2)||_2 \leq L_H||\chi_k^1 - \chi_k^2||_2 \forall \chi_k^1, \chi_k^2 \in \Theta, \forall x \in X \tag{15}$$

Sampling variance is bounded by σ^2, such that

$$\mathbb{E}[||\nabla f(\boldsymbol{\theta}*, \xi) - \nabla F(\boldsymbol{\theta}_k^t, \chi_k^t)||_2^2|\chi_n] \leq \sigma^2, \forall \chi_k \in \Theta \tag{16}$$

Assumption 5 *The variance of stochastic gradients on local model training at each client is bounded:* $\mathbb{E}||\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t) - \nabla F(\boldsymbol{\theta}_{n,k}^t)||^2 \leq \sigma_g^2$

Lemma 1 *Under Assumption 4, there exist a constant* $C_1 > 0$ *such that for any* $\delta > 0$*, with probability* $1 - \delta$ *we have*

$$||\mathbb{E}_{\pi_t}\nabla F(\boldsymbol{\theta}_k, \chi_k^t) - \mathbb{E}_{\pi_1}\nabla F(\boldsymbol{\theta}_k, \chi_k^{'t})||_2^2 \leq C_1 \frac{logDt + log\frac{1}{\delta}}{Dt}, \tag{17}$$

$\forall x \in X, \forall t > 0$

Lemma 2 *Let F be L-smooth,* $\boldsymbol{x}_k^{t+1}$ *and* $\boldsymbol{x}_k^t$ *as in 11 with Assumption 2 and 3. Then for any stepsize* $\eta \leq \frac{1}{(M+1)L}$*, it holds*

$$\mathbb{E}_\xi[F(\boldsymbol{w_k^{t+1}}) - F(\boldsymbol{w_k^t})|\boldsymbol{w_k^t}] \leq \frac{\eta(m-1)}{2}||\nabla F(\boldsymbol{w_k^t})||^2 + \frac{\eta}{2}\zeta^2 + \frac{\eta^2 L}{2}\sigma^2 \tag{18}$$

when $M = m = \zeta^2 = c$ *for any constant c, we recover the standard descent lemma.*

Lemma 3 *Under Assumption 4, there exists a constant* $C_1 > 0$ *such that for any* $\delta > 0$*, with probability at least* $1 - \delta$ *we have*

$$||\mathbb{E}\nabla F(\boldsymbol{\theta}, \chi^t) - \mathbb{E}_{\pi_1}\nabla F(\boldsymbol{\theta}, \chi^{'t})||_2^2 \leq C_1 \frac{logDt + log\frac{1}{\delta}}{Dt}, \forall x \in X, \forall t > 0. \tag{19}$$

Lemma 4 *Under the assumption* $p = P_s$ *and series* $\sum_{l\geq 1} N_l e^{-l^2}$ *converges.*

$$\bar{v}[B(s, \frac{k}{\sqrt{n}})|X] \geq 1 - \delta, \forall s \in S, \epsilon, \delta \in (0,1) \tag{20}$$

with probability at least $1 - \epsilon$ *with respect to* P_s^n*, and where*
$k = k(\epsilon, \delta, v(s)) = inf\left[j \geq 1|\sum_{l\geq j} N_l e^{-l^2} \leq \epsilon\sqrt{\delta v(s)}\right]$
is independent of n *and nonincreasing with the positive parameters* $\epsilon, \delta, and v(s)$.

Lemma 5 *Let Assumption 5 hold, the expected upper bound of the variance of the stochastic gradient on local model training is given as*

$$\mathbb{E}||g_k^t - \bar{g}_k^t||^2 \leq \frac{\sigma_g^2}{N^2}. \tag{21}$$

Lemma 6 *The expected upper bound of the divergence of* $\boldsymbol{\theta}_{n,k}^t$ *is given as*

$$\left[\frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\left\|\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t\right\|^2\right] \leq 4\eta_k T B^2, \tag{22}$$

for some positive B.

Lemma 7 *The expected upper bound of* $\mathbb{E}[||\bar{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}^*||^2]$ *is given as*

$$\mathbb{E}||\bar{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}^*||^2$$
$$\leq ||(1-\mu\eta_k)||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \frac{1}{N}\sum_{n\in\mathcal{N}}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 + \frac{1}{4\eta_k}\frac{1}{N}\sum_{n\in\mathcal{N}}||\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t||^2 + \tag{23}$$
$$\frac{1}{2}\min(\Theta_1, \Theta_2, \Theta_3) + \eta_k^2||g_k^t - \bar{g}_k^t||^2$$

Theorem 1 *Under some Assumptions, for any* $\delta > 0$*, we have the probability at least* $1-\delta$*, for any* $T > 0$*, the following bound on the expected gradient of the final output under the true parameter* $\chi_k^{'t}$

(i) if the step size (η) *satisfies* $\eta_k = \frac{a}{\sqrt{K}}, \forall k \leq \mathcal{K}$*, for some constant* $a < \frac{\sqrt{\mathcal{K}}}{L_h}$*, then*

$$\mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{'t})||_2^2]$$
$$\leq [\tfrac{2(F(\boldsymbol{\theta}_1, \chi_k^{'t}) - \min_{x\in X} F(\boldsymbol{\theta}, \chi_k^{'t}))}{a\sqrt{\mathcal{K}}}] + [\tfrac{A_1}{\mathcal{K}} + \tfrac{A_2 log\mathcal{K}}{\mathcal{K}} + \tfrac{A_3 log^2\mathcal{K}}{\mathcal{K}}] + \tfrac{L_h a\sigma^2}{\sqrt{\mathcal{K}}}$$

where $A_1 = \frac{C_1(logD - log\delta)}{L_h D}, A_2 = \frac{C_1(logD - log\delta)}{L_h D} + \frac{C_1}{L_h D}, A_3 = \frac{C_1}{L_h D}$.

(ii) if the step size (η) *satisfies* $\eta_k = \frac{a}{k}, \forall k \leq \mathcal{K}$*, for some constant* $a < \frac{1}{L_h}$*, then*

$$\mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{'t})||_2^2]$$
$$\leq [\tfrac{2(F(\boldsymbol{\theta}_1, \chi_k^{'t}) - \min_{x\in X} F(\boldsymbol{\theta}, \chi_k^{'t}))}{a} + \tfrac{6C_1 + \pi^2 C_1(logD - log\delta)}{6D} + \tfrac{\pi^2 L_h a\sigma^2}{6}]\tfrac{1}{log\mathcal{K}}$$

(iii) if the step size (η) *satisfies* $\eta_k = \frac{a}{\sqrt{k}}, \forall k \leq \mathcal{K}$*, for some constant* $a < \frac{1}{L_h}$*, then*

$$\mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{t})||_2^2]$$
$$\leq [\tfrac{2(F(\boldsymbol{\theta}_1, \chi_k^{'t}) - \min_{x\in X} F(\boldsymbol{\theta}, \chi_k^{'t}))}{a\sqrt{\mathcal{K}}} + \tfrac{3C_1(logD - log\delta) + 4C_1}{D\sqrt{\mathcal{K}}} + \tfrac{L_h a\sigma^2}{\sqrt{\mathcal{K}}}] + \tfrac{L_h a\sigma^2 log\mathcal{K}}{\sqrt{\mathcal{K}}}$$

Theorem 2 *Given above Lemmas and Theorem 1, the convergence bound of our approach after* K *global communication rounds is given as*

$$\mathbb{E}\left[F(\boldsymbol{\theta}_K)\right] - F^* \leq \frac{L}{2(K + L/\mu)}\left[\frac{16\Phi_K}{15\mu^2} + \left(\frac{L}{\mu} + 1\right)\mathbb{E}||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||^2\right]. \tag{24}$$

6. Proofs of Convergence of Global Model Training

6.1. Proof of Lemma 2

By the quadratic upper bound in Assumption 1 and Assumption 2:

$$\begin{aligned}\mathbb{E}F(\boldsymbol{\theta}_k^{t+1}) &\leq F(\boldsymbol{\theta}_k^t) - \eta_k(\nabla F(\boldsymbol{\theta}_k^t), \mathbb{E}g_t) + \frac{\eta^2 L}{2}(\mathbb{E}||g_t - \mathbb{E}g_t||^2 + \mathbb{E}||\mathbb{E}g_t||^2) \\ &= F(\boldsymbol{\theta}_k^t) - \eta_k(\nabla F(\boldsymbol{\theta}_k^t), \nabla f(\boldsymbol{x}_k^t) + b_t + n_t) + \frac{\eta^2 L}{2}(\mathbb{E}||n_t||^2 + \mathbb{E}||\nabla f(\boldsymbol{x}_k^t) + b_t + n_t||^2) \\ &\leq F(\boldsymbol{\theta}_k^t) - \eta_k(\nabla F(\boldsymbol{\theta}_k^t), \nabla f(\boldsymbol{x}_k^t) + b_t + n_t) + \frac{\eta^2 L}{2}((M+1)\mathbb{E}||\nabla f(\boldsymbol{x}_k^t) + b_t + n_t||^2 + \sigma^2)\end{aligned}$$

By the choice of stepsize, $\eta \leq \frac{1}{(M+1)L}$, and Assumption 3:

$$\begin{aligned}\mathbb{E}F(\boldsymbol{\theta}_k^{t+1}) &\leq F(\boldsymbol{\theta}_k^t) + \frac{\eta_k}{2}(-2(\nabla F(\boldsymbol{\theta}_k^t), \nabla f(\boldsymbol{x}_k^t) + b_t + n_t) + ||\nabla f(\boldsymbol{x}_k^t) + b_t + n_t||^2) + \frac{\eta^2 L}{2}\sigma^2 \\ &= F(\boldsymbol{\theta}_k^t) + \frac{\eta_k}{2}(-||\nabla F(\boldsymbol{\theta}_k^t||^2 + ||b_t + n_t||^2) + \frac{\eta^2 L}{2}\sigma^2 \\ &= F(\boldsymbol{\theta}_k^t) + \frac{\eta_k}{2}(m-1)||\nabla F(\boldsymbol{\theta}_k^t)||^2 + \frac{\eta}{2}\zeta^2 + \frac{\eta^2 L}{2}\sigma^2\end{aligned} \tag{25}$$

This concludes the proof.

6.2. Proof of Lemma 3

The Hellignger distance between θ_1 and θ_2

$$d(\theta_1, \theta_2) = \sqrt{\frac{1}{2}\int_Y (\sqrt{f(y;\theta_1} - \sqrt{f(y;\theta_2)})^2} \tag{26}$$

There exists a constant A such that $||\theta_1 = \theta_2|| \leq Ad(\theta_1, \theta_2)$, where $||.||$ is the Euclidean norm. Let $B_k^t = B(\theta^c, \frac{k}{\sqrt{Dt}})$ be a ball centered at θ^c with radius $\frac{k}{\sqrt{Dt}}$ under distance d. Since Θ is finite, we can directly apply Lemma 4. So for $t \leq T, \epsilon, \delta \in (0,1)$ with probability at least $1 - \frac{6\delta}{\pi^2 t^2}$ with respect to $\mathbb{P}_{\theta^c}^t$, we have

$$\pi_t(B_{k(t)}^t) \geq 1 - \epsilon, \tag{27}$$

where $k(t) = inf\left[j \geq 1 | \sum_{i \geq j} |\Theta| e^{-i^2} \leq \frac{6\delta}{\pi^2 t^2}\sqrt{\epsilon \pi_0(\theta^c)}\right]$.

Note that $\sum_{i \geq j} e^{-i^2} \leq \frac{e}{e-1} e^{-j^2}$, we can set k(t) to be the solution of next equation.

$$\frac{e}{e-1}|\Theta| e^{-k(t)^2} = \frac{6\delta}{\pi^2 t^2}\sqrt{\epsilon \pi_0(\theta^c)} \tag{28}$$

we get $k(t) = \sqrt{log \frac{e|\Theta|\pi^2 t^2}{6\delta(e-1)\sqrt{\epsilon, \pi_0(\theta^c)}}}$. Now the bias in the gradient estimator can be bonded as follows.

$$
\begin{aligned}
&||\mathbb{E}_{\pi_t}\nabla_x F(\boldsymbol{\theta},\chi) - \mathbb{E}_{\pi_t}\nabla_x F(\boldsymbol{\theta},\chi^{'})||_2^2 \\
&\quad = ||\int (\nabla_x F(\boldsymbol{\theta},\chi) - \nabla_x F(\boldsymbol{\theta},\chi^{'}))\pi_t(\theta)d\theta||_2^2 \\
&\quad \leq \int ||(\nabla_x F(\boldsymbol{\theta},\chi)) - (\nabla_x F(\boldsymbol{\theta},\chi^{'}))_2^2||\pi_t(\theta)d\theta \\
&\quad \leq L_H^2||\chi-\chi^{'}||_2^2\pi_t(\theta)d\theta \\
&\quad = \int_{B^t_{k(t)}} L_H^2||\chi-\chi^{'}||_2^2\pi_t(\theta)\theta + \int_{(B^t_{k(t)})'} L_H^2||\chi-\chi^{'}||_2^2\pi_t(\theta)d\theta \\
&\quad \leq A^2L_H^2\frac{k(t)^2}{Dt}\int_{B^t_{k(t)}}\pi_t(\theta)d\theta + L_H^2\max_{\chi\in\Theta}||\chi-\chi^{'}||_2^2\int_{(B^t_{k(t)})'}\pi_t(\theta)d\theta \\
&\quad \leq A^2L_H^2\frac{k(t)^2}{Dt} + L_H^2\max_{\chi\in\Theta}||\chi-\chi^{'}||_2^2\epsilon
\end{aligned}
\tag{29}
$$

Here, D is the data batch size. Note that $\epsilon = \frac{1}{Dt}$ and $k(t) = \sqrt{log\frac{e|\Theta|\pi^2t^2\sqrt{Dt}}{6\delta(e-1)\sqrt{\pi_0(\chi')}}}$, we have

$$
\begin{aligned}
&||\mathbb{E}_{\pi_t}\nabla_x F(\boldsymbol{\theta},\chi) - \mathbb{E}_{\pi_t}\nabla_x F(\boldsymbol{\theta},\chi^{'})||_2^2 \\
&\quad \leq A^2L_H^2\frac{k(t)^2}{Dt} + L_H^2\max_{\chi\in\Theta}||\chi-\chi^{'}||_2^2\epsilon \\
&\quad \leq 2A^2L_H^2\max_{\theta\in\Theta}||\chi-\chi^{'}||_2^2\frac{log\frac{e|\Theta|\pi^2t^2\sqrt{Dt}}{6\delta(e-1)\sqrt{\pi_0(\chi')}}}{Dt} \\
&\quad = O\left(\frac{logDt + log\frac{1}{\delta}}{Dt}\right)
\end{aligned}
\tag{30}
$$

Let E_t denote the event that the above inequality holds, and E_t^c denote that the above inequality does not hold. Then

$$
\mathbb{P}(E_t^c) \leq \frac{6\delta}{\pi^2t^2} \tag{31}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}(\cap_{t=1}^{\infty}E_t) \\
&\quad = 1-\mathbb{P}(\cup_{t=1}^{\infty}E_t^c) \\
&\quad \geq 1-\sum_{t=1}^{\infty}\mathbb{P}(E_t^c)) \qquad \text{(Union bound)} \\
&\quad \geq 1-\sum_{t=1}^{\infty}\frac{6\delta}{\pi^2t^2} \\
&\quad = 1-\delta
\end{aligned}
\tag{32}
$$

6.3. Proof of Lemma 5

From Assumption 5, we have

$$\begin{aligned}\mathbb{E}||g_k^t - \bar{g}_k^t||^2 =& \mathbb{E}\Big\| \frac{1}{N}\sum_{n\in\mathcal{N}} \left(\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t) - \nabla F(\boldsymbol{\theta}_{n,k}^t)\right) \Big\|^2 \\ &= \frac{1}{N^2}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\| \left(\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t) - \nabla F(\boldsymbol{\theta}_{n,k}^t)\right) \Big\|^2 \le \frac{\sigma_g^2}{N^2}.\end{aligned} \tag{33}$$

6.4. Proof of Lemma 6

We know that in every global communication round, each client performs T rounds of local SGDs where there always exits $t' \le t$ such that $t - t' \le T$ and $\boldsymbol{\theta}_{n,k}^{t'} = \bar{\boldsymbol{\theta}}_k^{t'}, \forall n \in \mathcal{N}$. By using the fact that $\mathbb{E}||X - \mathbb{E}X||^2 = ||X||^2 - ||\mathbb{E}X||^2$ and $\bar{\boldsymbol{\theta}}_k^t = \mathbb{E}\boldsymbol{\theta}_{n,k}^t$, we have:

$$\begin{aligned}&\frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\|\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t\Big\|^2 = \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\|\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t\Big\|^2 = \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\|(\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^{t'}) - (\bar{\boldsymbol{\theta}}_k^t - \bar{\boldsymbol{\theta}}_k^{t'})\Big\|^2 \\ &\le \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\|\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^{t'}\Big\|^2 \le \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\| \left(\sum_{t=t'}^{t-1}(\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^{t'})\right) \Big\|^2 \\ &= \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\| \left(\sum_{t=t'}^{t-1} \eta_k \nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t)\right) \Big\|^2 \\ &\le \frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\| \left(\sum_{t=1}^{t-t'} \eta_k \nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t)\right) \Big\|^2,\end{aligned} \tag{34}$$

where the last inequality holds since the learning rate η_k is decreasing. Using the fact that $||\sum_{t=1}^{U} z^t||^2 \le U\sum_{t=1}^{U}||z^t||^2$, $t - t' \le T$ and assume that $\eta_k^{t'} \le 2\eta_k$ and $||\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t)||^2 \le B^2$ for positive constant B, we have

$$\begin{aligned}&\frac{1}{N}\sum_{n\in\mathcal{N}} \mathbb{E}\Big\|\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t\Big\|^2 \le \frac{1}{N}\sum_{n\in\mathcal{N}} \left(\mathbb{E}\sum_{t=1}^{t-t'} \eta_k^2 (t-t') \Big\|\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t)\Big\|^2\right) \\ &\le \frac{1}{N}\sum_{n\in\mathcal{N}} \left(\mathbb{E}\sum_{t=1}^{t-t'} \eta_k^2 T \Big\|\nabla F(\boldsymbol{\theta}_{n,k}^t, \chi_{n,k}^t)\Big\|^2\right) \\ &\le \frac{1}{N}\sum_{n\in\mathcal{N}} \left((\eta_k^{t'})^2 T \sum_{t=1}^{t-t'} B^2\right) \le \frac{1}{N}\sum_{n\in\mathcal{N}} (\eta_k^{t'})^2 T B^2 \le 4\eta_k T B^2.\end{aligned} \tag{35}$$

6.5. Proof of Theorem 1

The local SGD update at client n is followed as:

$$\bar{\boldsymbol{\theta}}_k^{t+1} = \boldsymbol{\theta}_k^t - \eta_k \nabla F(\boldsymbol{\theta}_k^t, \chi_k'^t) + \eta_k[\mathbb{E}_{\pi_1}\nabla F(\boldsymbol{\theta}_k^t, \chi_k^t) - \nabla F(\boldsymbol{\theta}_k^t, \bar{\chi}_k'^t)] - \eta_k[\nabla f(\boldsymbol{\theta}_k^t, \xi_k^t) - \mathbb{E}_{\pi_1}\nabla F(\boldsymbol{\theta}_k^t, \bar{\chi}_k^t)], \tag{36}$$

where F is a local loss function, $\eta > 0$ is the local learning rate, and χ is a sample uniformly chosen from the local dataset. Now if we will consider $\mathbb{E}_{\pi_1} \nabla F(\boldsymbol{\theta}_k^t, \chi_k^t) - \nabla F(\boldsymbol{\theta}_k^t, \bar{\chi}_k^{\prime t})$ as bias B_k and $\nabla f(\boldsymbol{\theta}_k^t, \xi_k^t) - \mathbb{E}_{\pi_1} \nabla F(\boldsymbol{\theta}_k^t, \bar{\chi}_k^t)$ as noise N_k, the equation becomes:

$$\bar{\boldsymbol{\theta}}_k^{t+1} = \boldsymbol{\theta}_k^t - \eta_k \nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t}) + \eta_k B_k - \eta_k N_k, \tag{37}$$

By Lemma 3, we know

$$\mathbb{E}[||B_t||_2^2] \leq C_1 \frac{logDt + log^1 \delta}{Dt} \tag{38}$$

By Assumption 4, we have

$$\mathbb{E}[||N_t||_2^2] \leq \sigma^2 \tag{39}$$

By proof of Lemma 2, we know that

$$\mathbb{E}[F(\boldsymbol{\theta}_k^{t+1}, \chi_k^{\prime t}) - F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})] \leq -\frac{\eta_k}{2} ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 + \frac{\eta_k}{2} C_1 \frac{logDt + log\frac{1}{\delta}}{Dt} + \frac{\eta_k^2}{2} L_h \sigma^2 \tag{40}$$

and after multiplying by 2,

$$2\mathbb{E}[F(\boldsymbol{\theta}_k^{t+1}, \chi_k^{\prime t}) - F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})] \leq -\eta_k ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 + \eta_k C_1 \frac{logDt + log\frac{1}{\delta}}{Dt} + \eta_k^2 L_h \sigma^2 \tag{41}$$

after rearranging,

$$\eta_k ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 \leq -2\mathbb{E}[F(\boldsymbol{\theta}_k^{t+1}, \chi_k^{\prime t}) - F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})] + \eta_k C_1 \frac{logDt + log\frac{1}{\delta}}{Dt} + \eta_k^2 L_h \sigma^2 \tag{42}$$

noting that $F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t}) \leq \min_{xinX} F(\boldsymbol{\theta}, \chi_k^{\prime t})$

$$\eta_k ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 \leq 2(F(\boldsymbol{\theta}_1, \chi_k^{\prime t}) - \min_{x \in X} F(\boldsymbol{\theta}, \chi_k^{\prime t})) + \eta_k C_1 \frac{logDt + log\frac{1}{\delta}}{Dt} + \eta_k^2 L_h \sigma^2 \tag{43}$$

summing over k from 1 to $\mathcal{K}$,

$$\begin{aligned} \sum_{k=1}^{\mathcal{K}} \eta_k ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 \leq & 2(F(\boldsymbol{\theta}_1, \chi_k^{\prime t}) - \min_{xinX} F(\boldsymbol{\theta}, \chi_k^{\prime t})) + C_1 \sum_{k=1}^{\mathcal{K}} \eta_k \frac{logDt + log\frac{1}{\delta}}{Dt} \\ & + L_h \sigma^2 \sum_{k=1}^{\mathcal{K}} \eta_k^2 \end{aligned} \tag{44}$$

Dividing both sides by $\sum_{k=1}^{\mathcal{K}} \eta_k$,

$$\begin{aligned} \frac{1}{\sum_{k=1}^{\mathcal{K}} \eta_k} \sum_{k=1}^{\mathcal{K}} \eta_k ||\nabla F(\boldsymbol{\theta}_k^t, \chi_k^{\prime t})||_2^2 \leq & \frac{1}{\sum_{k=1}^{\mathcal{K}} \eta_k} \Bigg[2(F(\boldsymbol{\theta}_1, \chi_k^{\prime t}) - \min_{xinX} F(\boldsymbol{\theta}, \chi_k^{\prime t})) + \\ & C_1 \sum_{k=1}^{\mathcal{K}} \eta_k \frac{logDt + log\frac{1}{\delta}}{Dt} + L_h \sigma^2 \sum_{k=1}^{\mathcal{K}} \eta_k^2 \end{aligned} \tag{45}$$

noting that $\frac{1}{\sum_{k=1}^{\mathcal{K}}\eta_k}\sum_{k=1}^{\mathcal{K}}\eta_k||\nabla F(\boldsymbol{\theta}_k^t,\chi_k'^t)||_2^2=\mathbb{E}||\nabla F(z_T,\chi_k'^t)||_2^2$,

$$\begin{aligned}\mathbb{E}||\nabla F(z_T,\chi_k'^t)||_2^2\leq&\frac{1}{\sum_{k=1}^{\mathcal{K}}\eta_k}\left[2(F(\boldsymbol{\theta}_1,\chi_k'^t)-\min_{xinX}F(\boldsymbol{\theta},\chi_k'^t))+C_1\sum_{k=1}^{\mathcal{K}}\eta_k\frac{logDt+log\frac{1}{\delta}}{Dt}\right.\\&\left.+L_h\sigma^2\sum_{k=1}^{\mathcal{K}}\eta_k^2\right]\end{aligned}\tag{46}$$

(i) if the step size (η) satisfies $\eta_k=\frac{a}{\sqrt{K}},\forall k\leq\mathcal{K}$, for some constant $a<\frac{\sqrt{\mathcal{K}}}{L_h}$.

Note that $\sum_{k=1}^{\mathcal{K}}\frac{1}{t}\leq log\mathcal{K}+1$ and $\sum_{k=1}^{\mathcal{K}}\frac{logk}{k}\leq log(log\mathcal{K}+1)$. Then

$$\begin{aligned}&\Theta_1\triangleq\mathbb{E}[||\nabla F(z_{\mathcal{K}},\chi_k'^t)||_2^2]\\&\leq\frac{2(F(\boldsymbol{\theta}_1,\chi_k'^t)-\min_{x\in X}F(\boldsymbol{\theta},\chi_k'^t))}{a\sqrt{\mathcal{K}}}+\frac{C_1(logD-log\delta)(log\mathcal{K}+1)}{L_hD\mathcal{K}}+\frac{C_1log\mathcal{K}(log\mathcal{K}+1)}{L_hD\mathcal{K}}\\&=\frac{2(F(\boldsymbol{\theta}_1,\chi_k'^t)-\min_{x\in X}F(\boldsymbol{\theta},\chi_k'^t))}{a\sqrt{\mathcal{K}}}+\frac{C_1(logD-log\delta)}{L_hD\mathcal{K}}+\frac{C_1(logD-log\delta)log\mathcal{K}}{L_hD\mathcal{K}}\\&+\frac{C_1log^2\mathcal{K}}{L_hD\mathcal{K}}+\frac{L_ha\sigma^2}{\sqrt{\mathcal{K}}}\end{aligned}\tag{47}$$

(ii) if the step size (η) satisfies $\eta_k=\frac{a}{k},\forall k\leq\mathcal{K}$, for some constant $a<\frac{1}{L_h}$. Let $M_{\mathcal{K}}=\sum_{k=1}^{\mathcal{K}}\frac{1}{k}$.

Note that

$$\sum_{k=1}^{\mathcal{K}}\frac{logk}{k^2}<\sum_{k=1}^{\infty}\frac{logk}{k^2}=\frac{\pi^2}{6}(12lnA-\gamma-ln2\pi)<1\tag{48}$$

where the Glaisher-Kinkelin constant $A\approx1.28$ and the Euler-Mascheroni constant $\gamma\approx0.58$.

Then we have

$$\begin{aligned}&\Theta_2\triangleq\mathbb{E}[||\nabla F(z_{\mathcal{K}},\chi_k'^t)||_2^2]\\&\leq\frac{2(F(\boldsymbol{\theta}_1,\chi_k'^t)-\min_{x\in X}F(\boldsymbol{\theta},\chi_k'^t))}{aM_{\mathcal{K}}}+\frac{C_1}{M_{\mathcal{K}}}\sum_{k=1}^{\mathcal{K}}\frac{logDk+log\frac{1}{\delta}}{Dk^2}+\sum_{k=1}^{\mathcal{K}}\frac{L_ha\sigma^2}{M_{\mathcal{K}}t^2}\\&\leq\left[\frac{2(F(\boldsymbol{\theta}_1,\chi_k'^t)-\min_{x\in X}F(\boldsymbol{\theta},\chi_k'^t))}{a}+\frac{6C_1+\pi^2C_1(logD-log\delta)+\frac{\pi^2L_ha\sigma^2}{6}}{6D}\right]\frac{1}{log\mathcal{K}}\end{aligned}\tag{49}$$

(iii) if the step size (η) satisfies $\eta_k = \frac{a}{\sqrt{k}}, \forall k \leq \mathcal{K}$, for some constant $a < \frac{1}{L_h}$. Let $Q_t = \sum_{k=1}^{\mathcal{K}} \frac{1}{\sqrt{k}}$. Note that $\sum_{k=1}^{\infty} \frac{1}{k\sqrt{k}} = \zeta(1.5) \approx 2.61 \leq 3$, $\sum_{t=1}^{\infty} \frac{logk}{k\sqrt{k}} < 4$, $\sum_{k=1}^{\mathcal{K}} \frac{1}{\sqrt{k}} \geq \sqrt{\mathcal{K}}$, where $\zeta(.)$ is the Riemann's zeta function. Then we have

$$\begin{aligned}
&\Theta_3 \triangleq \mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{'t})||_2^2] \\
&\leq \frac{2(F(\boldsymbol{\theta}_1, \chi_k^{'t}) - \min_{x\in X} F(\boldsymbol{\theta}, \chi_k^{'t}))}{aQ_{\mathcal{K}}} + \frac{C_1(logD - log\delta)}{DQ_{\mathcal{K}}} \sum_{k=1}^{\mathcal{K}} \frac{1}{k\sqrt{k}} + \frac{C_1}{DQ_{\mathcal{K}}} \sum_{k=1}^{\mathcal{K}} \frac{logk}{k\sqrt{k}} \\
&+ \frac{L_h a\sigma^2}{Q_{\mathcal{K}}} \sum_{k=1}^{\mathcal{K}} \frac{1}{k} \\
&\leq [\frac{2(F(\boldsymbol{\theta}_1, \chi_k^{'t}) - \min_{x\in X} F(\boldsymbol{\theta}, \chi_k^{'t}))}{a\sqrt{\mathcal{K}}} + \frac{3C_1(logD - log\delta) + 4C_1}{D\sqrt{\mathcal{K}}} + \frac{L_h a\sigma^2}{\sqrt{\mathcal{K}}}] \\
&+ \frac{L_h a\sigma^2 log\mathcal{K}}{\sqrt{\mathcal{K}}}
\end{aligned} \tag{50}$$

6.6. Proof of Theorem 2

From the SGD update rule $\bar{\boldsymbol{\theta}}_k^{t+1} = \bar{\boldsymbol{\theta}}_k^t - \eta_k g_k^t + \bar{\mathbf{v}}_k^t$ and $||a+b||^2 \leq 2||a||^2 + 2||b||^2$ for two real valued vectors a and b, we have

$$||\bar{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}^*||^2 = ||\bar{\boldsymbol{\theta}}_k^t - \eta_k g_k^t + \bar{\mathbf{v}}_k^t - \boldsymbol{\theta}^*||^2 \leq \underbrace{||\bar{\boldsymbol{\theta}}_k^t - \eta_k g_k^t - \boldsymbol{\theta}^*||^2}_{(A)} + ||\bar{\mathbf{v}}_k^t||^2 \tag{51}$$

We now focus on the bounding term (A) in 51. We have

$$\begin{aligned}
&||\bar{\boldsymbol{\theta}}_k^t - \eta_k g_k^t - \boldsymbol{\theta}^*||^2 = ||\bar{\boldsymbol{\theta}}_k^t - \eta_k g_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t + \eta_k \bar{g}_k^t||^2 \\
&= ||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2 + 2\eta_k \langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t, \bar{g}_k^t - g_k^t \rangle + \eta_k^2 ||g_k^t - \bar{g}_k^t||^2 \\
&= \underbrace{||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2}_{(B)} + \eta_k^2 ||g_k^t - \bar{g}_k^t||^2,
\end{aligned} \tag{52}$$

where $\langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t, \bar{g}_k^t - g_k^t \rangle = 0$. We now focus on bounding term (B). We have

$$\begin{aligned}
||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2 &= ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \eta_k^2 ||\bar{g}_k^t||^2 - 2\eta_k \frac{1}{N} \sum_{n\in\mathcal{N}} \langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle \\
&\le ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \eta_k^2 \frac{1}{N} \sum_{n\in\mathcal{N}} ||\nabla F(\boldsymbol{\theta}_{n,k}^t)||^2 - 2\eta_k \frac{1}{N} \sum_{n\in\mathcal{N}} \langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t + \boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle \\
&\le ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + 2\eta_k^2 \frac{L}{N} \sum_{n\in\mathcal{N}} (F(\boldsymbol{\theta}_{n,k}^t) - F^*) - 2\eta_k \frac{1}{N} \sum_{n\in\mathcal{N}} \langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle \\
&\quad - 2\eta_k \frac{1}{N} \sum_{n\in\mathcal{N}} \langle \boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle,
\end{aligned} \tag{53}$$

where in the first inequality we applied $||\sum_{n\in\mathcal{N}} z_n||^2 \le N \sum_{n\in\mathcal{N}} ||z_n||^2$, and in the second inequality we applied L-smoothness $||\nabla F(\boldsymbol{\theta}_{n,k}^t)||^2 \le 2L(F(\boldsymbol{\theta}_{n,k}^t) - F^*)$. For the third term in 53, by using the Cauchy–Schwarz inequality and arithmetic and geometric means (AM-GM) inequality: $2\langle a, b\rangle \le \frac{1}{\varepsilon}||a||^2 + \varepsilon ||b||^2$ for $\varepsilon > 0$, we have

$$-2\langle \bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle = 2\langle \boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle \le \frac{1}{\eta_k} ||\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t||^2 + \eta_k ||\nabla F(\boldsymbol{\theta}_{n,k}^t)||^2 \tag{54}$$

$$\le \frac{1}{\eta_k} ||\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t||^2 + 2\eta_k L(F(\boldsymbol{\theta}_{n,k}^t) - F^*). \tag{55}$$

For the last term in 53, by using μ-strong convexity, we have

$$-\langle \boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*, \nabla F(\boldsymbol{\theta}_{n,k}^t) \rangle \le -(F(\boldsymbol{\theta}_{n,k}^t) - F^*) - \frac{\mu}{2} ||\boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*||^2. \tag{56}$$

Therefore, 53 can be rewritten as

$$\begin{aligned}
&+ 2\eta_k L(F(\boldsymbol{\theta}_{n,k}^t) - F^*) \\
&- 2\eta_k \tfrac{1}{N} \textstyle\sum_{n\in\mathcal{N}} (F(\boldsymbol{\theta}_{n,k}^t) - F^*) - \mu\eta_k \tfrac{1}{N} \sum_{n\in\mathcal{N}} \tfrac{\mu}{2} ||\boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*||^2 \\
&\le ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + 2\eta_k(2\eta_k L - 1) \tfrac{1}{N} \textstyle\sum_{n\in\mathcal{N}} (F(\boldsymbol{\theta}_{n,k}^t) - F^*) + \tfrac{1}{N} \sum_{n\in\mathcal{N}} ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 \\
&- \mu\eta_k \tfrac{1}{N} \textstyle\sum_{n\in\mathcal{N}} ||\boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*||^2 \\
&= (1 - \mu\eta_k) ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + 2\eta_k(2\eta_k L - 1) \tfrac{1}{N} \textstyle\sum_{n\in\mathcal{N}} (F(\boldsymbol{\theta}_{n,k}^t) - F^*) + \\
&\tfrac{1}{N} \textstyle\sum_{n\in\mathcal{N}} ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2,
\end{aligned}$$

where we used the fact: $\frac{1}{N} \sum_{n\in\mathcal{N}} ||\boldsymbol{\theta}_{n,k}^t - \boldsymbol{\theta}^*||^2 = ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2$. We assume $\eta_k \le \frac{1}{4L}$, it holds $\eta_k L \le \frac{1}{4} \implies 2\eta_k L - 1 \le -\frac{1}{2}$. Thus

$$||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2 \le (1 - \mu\eta_k) ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \frac{1}{N} \sum_{n\in\mathcal{N}} ||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 - \frac{1}{2} \frac{1}{N} \sum_{n\in\mathcal{N}} (F(\boldsymbol{\theta}_{n,k}^t) - F^*) \tag{57}$$

$$||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2 \leq (1-\mu\eta_k)||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \frac{1}{N}\sum_{n\in\mathcal{N}}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 - \underbrace{\frac{1}{2}\mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{\prime t})||_2^2]}_{(C)} \tag{58}$$

We can bound (C) using equation 47, 49, and 50

$$\mathbb{E}[||\nabla F(z_{\mathcal{K}}, \chi_k^{\prime t})||_2^2] = \min(\Theta_1, \Theta_2, \Theta_3) \tag{59}$$

where the first inequality results from the convexity of $F_n(.)$, the second inequality is derived from the AM-GM inequality, and the third inequality results from the smoothness of $F_n(.)$. Therefore, 58 is further expressed as

$$||(\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^* - \eta_k \bar{g}_k^t||^2 \leq (1-\mu\eta_k)||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \frac{1}{N}\sum_{n\in\mathcal{N}}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 + \frac{1}{4\eta_k}\frac{1}{N}\sum_{n\in\mathcal{N}}||\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t||^2 + \frac{1}{2}\min(\Theta_1, \Theta_2, \Theta_3) \tag{60}$$

By plugging 60 into 51 and taking expectation we obtain

$$\begin{aligned}&\mathbb{E}||\bar{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}^*||^2 \\ &\leq ||(1-\mu\eta_k)||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + \frac{1}{N}\sum_{n\in\mathcal{N}}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}_{n,k}^t||^2 + \frac{1}{4\eta_k}\frac{1}{N}\sum_{n\in\mathcal{N}}||\boldsymbol{\theta}_{n,k}^t - \bar{\boldsymbol{\theta}}_k^t||^2 + \\ &\frac{1}{2}\min(\Theta_1, \Theta_2, \Theta_3) + \eta_k^2||g_k^t - \bar{g}_k^t||^2\end{aligned} \tag{61}$$

From Lemmas 5, 6, and 7, we have

$$\mathbb{E}||\bar{\boldsymbol{\theta}}_k^{t+1} - \boldsymbol{\theta}^*||^2 \leq (1-\mu\eta_k)\mathbb{E}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2 + 4\left(1 + \frac{1}{\eta_k}\right)\eta_k T B^2 + \frac{\eta_k^2\sigma_g^2}{N^2} \tag{62}$$

Let us define $Y_k^t = \mathbb{E}||\bar{\boldsymbol{\theta}}_k^t - \boldsymbol{\theta}^*||^2$ and $\Phi_k = 4\left(\frac{\eta_k+1}{\eta_k^2}\right)TB^2 + \frac{\sigma_g^2}{N^2}$, from 62 we have

$$\sum_{t=1}^{T} Y_k^{t+1} \leq \sum_{t=0}^{T-1}(1-\mu\eta_k)Y_k^t + \eta_k^2\Phi_k, \tag{63}$$

By $Y_k = \sum_{t=0}^{T-1} Y_k^t$, 63 is rewritten as

$$Y_k^{t+1} \leq (1-\mu\eta_k)Y_k^t + \eta_k^2\Phi_k, \tag{64}$$

We define a diminishing stepsize $\eta_k = \frac{4\theta}{k+\omega}$ for some $\theta > \frac{1}{4\mu}$ and $\omega > 0$. By defining $m = \max\{\frac{\theta^2\Phi_k}{4\theta\mu-1}, (\omega+1)Y_0\}$, we prove that $Y_k \leq \frac{m}{k+\omega}$ by induction. Due to $4\theta\mu > 1$, from 64 we have

$$\begin{aligned}Y_{k+1} &= \left(1 - \frac{4\theta\mu}{k+\omega}\right)\frac{m}{k+\omega} + \frac{16\theta^2}{(k+\omega)^2}\Phi_k \leq \frac{k+\omega-1}{(k+\omega)^2}m + \frac{16\theta^2}{(k+\omega)^2}\Phi_k \\ &\leq \frac{k+\omega-1}{(k+\omega)^2}m + \frac{16\theta^2}{(k+\omega)^2}\Phi_k - \frac{4\theta\mu-1}{(k+\omega)^2} \leq \frac{k+\omega-1}{(k+\omega)^2}m - \frac{4\theta\mu-1}{(k+\omega)^2} \\ &\leq \frac{k+\omega-4\theta\mu}{(k+\omega)^2}m \leq \frac{k+\omega-4\theta\mu}{(k+\omega)^2-(4\theta\mu)^2}m = \frac{1}{k+\omega+4\theta\mu}m \leq \frac{1}{k+\omega+1}m\end{aligned} \tag{65}$$

We choose $\theta = \frac{4}{\mu}$ and $\omega = \frac{L}{\mu}$, it follows that

$$m = \max\{\frac{\theta^2 \Phi_k}{4\theta\mu - 1}, (\omega+1)Y_0\} \leq \frac{\theta^2 \Phi_k}{4\theta\mu - 1} + (\omega+1)Y_0 = \frac{16\Phi_k}{15\mu^2} + \left(\frac{L}{\mu} + 1\right) Y_0 \tag{66}$$

By using the L-smoothness of $F(.)$, we have

$$\mathbb{E}\left[F(\bar{\boldsymbol{\theta}}_k)\right] - F^* \leq \frac{L}{2} Y_k \leq \frac{L}{2} \frac{m}{(k+\omega)} \leq \frac{L}{2(k+L/\mu)} \left[\frac{16\Phi_k}{15\mu^2} + \left(\frac{L}{\mu} + 1\right) \mathbb{E}||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||^2\right] \tag{67}$$

Finally, by applying 67 recursively, the convergence bound of our approach after K global communication rounds can be given as

$$\mathbb{E}\left[F(\boldsymbol{\theta}_K)\right] - F^* \leq \frac{L}{2(K+L/\mu)} \left[\frac{16\Phi_K}{15\mu^2} + \left(\frac{L}{\mu} + 1\right) \mathbb{E}||\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*||^2\right], \tag{68}$$

which completes the proof.