A New Theoretical Perspective on Data Heterogeneity in Federated Optimization

Jiayi Wang[1] **Shiqiang Wang**[2] **Rong-Rong Chen**[1] **Mingyue Ji**[1]

Abstract

In federated learning (FL), data heterogeneity is the main reason that existing theoretical analyses are pessimistic about the convergence rate. In particular, for many FL algorithms, the convergence rate grows dramatically when the number of local updates becomes large, especially when the product of the gradient divergence and local Lipschitz constant is large. However, empirical studies can show that more local updates can improve the convergence rate even when these two parameters are large, which is inconsistent with the theoretical findings. This paper aims to bridge this gap between theoretical understanding and practical performance by providing a theoretical analysis from a new perspective on data heterogeneity. In particular, we propose a new and weaker assumption compared to the local Lipschitz gradient assumption, named the heterogeneity-driven pseudo-Lipschitz assumption. We show that this and the gradient divergence assumptions can jointly characterize the effect of data heterogeneity. By deriving a convergence upper bound for FedAvg and its extensions, we show that, compared to the existing works, local Lipschitz constant is replaced by the much smaller heterogeneity-driven pseudo-Lipschitz constant and the corresponding convergence upper bound can be significantly reduced for the same number of local updates, although its order stays the same. In addition, when the local objective function is quadratic, more insights on the impact of data heterogeneity can be obtained using the heterogeneity-driven pseudo-Lipschitz constant. For example, we can identify a

region where FedAvg can outperform mini-batch SGD even when the gradient divergence can be arbitrarily large. Our findings are validated using experiments.

1. Introduction

Federated learning (FL) has emerged as an important technique for locally training machine learning models over geographically distributed workers. It has advantages in improving training efficiency and preserving data privacy. In this paper, we consider the following optimization problem in FL:

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} F_i(\mathbf{x}) \right\}, \qquad (1)$$

where N is the number of workers; $F_i(\mathbf{x})$ is the expected loss function of worker i given by[1]

$$F_i(\mathbf{x}) := \mathbb{E}_{\mathbf{n}_i \sim \mathcal{D}_i}[\ell(\mathbf{x}; \mathbf{n}_i)], \qquad (2)$$

where $\ell(\cdot)$ is the loss function, $\mathbf{n}_i$ is the random data sample on worker i, and $\mathcal{D}_i$ is the data distribution on worker i. In addition, we use $\mathcal{D}$ to denote the global data distribution. In FL, each worker performs $I > 1$ local iterations using its local dataset to reduce the communication cost, which is called *local updates*. Federated averaging (FedAvg), also known as local stochastic gradient descent (local SGD), is one of the most popular algorithms to solve the above optimization problem (McMahan et al., 2017). In addition to FedAvg, a number of FL algorithms (Yu et al., 2019a; Karimireddy et al., 2020; Reddi et al., 2020; Li et al., 2020b; Wang et al., 2020a;b) have been proposed, whereas the core mechanism, local updates, is still the foundation of FL. Nevertheless, existing theoretical analyses are pessimistic on the convergence error caused by local updates. It is unclear if performing a large number of local updates can improve the convergence rate when the gradient divergence is large or data are highly non-IID/heterogeneous. This will be explained in detail as follows.

---

[1]Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, USA. [2]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Jiayi Wang <jiayi.wang@utah.edu>, Shiqiang Wang <wangshiq@us.ibm.com>, Rong-Rong Chen <rchen@utah.edu>, Mingyue Ji <mingyue.ji@utah.edu>.

---

[1]The objective function can be extended to weighted average by multiplying each local objective function by a possibly distinct constant.
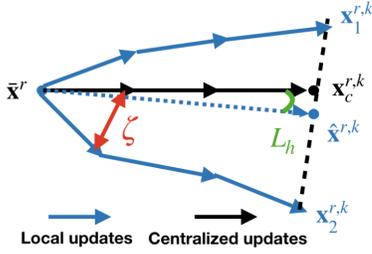
Figure 1: An illustrative comparison between local updates and centralized updates. $\bar{\mathbf{x}}^r$ is the global model at rth round. The local models after k local iterations at the rth round are denoted by $\mathbf{x}_1^{r,k}$ and $\mathbf{x}_2^{r,k}$. The average of $\mathbf{x}_1^{r,k}$ and $\mathbf{x}_2^{r,k}$ is $\hat{\mathbf{x}}^{r,k}$. The centralized model after k centralized iterations is denoted by $\mathbf{x}_c^{r,k}$. It can be seen that ζ shows the difference between $\mathbf{x}_c^{r,k}$ and $\mathbf{x}_i^{r,k}$, $i = 1, 2$ and L_h shows the difference between $\mathbf{x}_c^{r,k}$ and $\hat{\mathbf{x}}^{r,k}$.

There is a gap between the theoretical understanding and the experimental results. Unlike the centralized SGD running on a single machine, where the gradients are directly sampled from the global data distribution $\mathcal{D}$, the local gradients in FedAvg are sampled from the local data distributions $\{\mathcal{D}_i\}$, which are often highly heterogeneous (Kairouz et al., 2021). This can deteriorate FL's performance since the local models could drift to different directions during local updates (Zhao et al., 2018; Karimireddy et al., 2020). Therefore, a common understanding is that local SGD can have a larger convergence error than that of centralized SGD due to local updates. Existing theoretical analyses for non-convex objective functions (Yu et al., 2019a;b; Wang & Joshi, 2019; Yang et al., 2020) confirmed this intuition and showed that the convergence error caused by local updates grows fast when the number of local updates I becomes large. This limits the usefulness of local updates. However, in practice, a large number of local updates have been successfully applied (Li et al., 2020a; Niknam et al., 2020; Rieke et al., 2020) and showed superior experimental performance compared to mini-batch SGD with each worker performing $I = 1$ local iteration per round (McMahan et al., 2017; Lin et al., 2020). This means that, empirically, a large I can improve the convergence rate even when the data are highly non-IID. This inconsistency between the pessimistic theoretical results and the good experimental results for the local updates implies that the existing theoretical analyses may overestimate the error caused by local updates. In addition, it is indeed challenging to show theoretically when local SGD ($I > 1$) can outperform mini-batch SGD ($I = 1$) (Woodworth et al., 2020a;b).

Although local models could drift to different directions, the average of local models can still be close to the centralized model. To the best of our knowledge, the only metric of data heterogeneity in existing works (Yu et al., 2019b; Wang & Joshi, 2019; Woodworth et al., 2020b) is the gradient divergence (ζ), or its more general version, called gradient dissimilarity (Karimireddy et al., 2020), which characterizes the difference between the local gradient $\nabla F_i(\mathbf{x})$ of worker i and the global gradient $\nabla f(\mathbf{x})$. As shown in Figure 1, the intuition of the gradient divergence is that when ζ is large, the difference between local gradients and the global gradient is large. Then after multiple local iterations, the local models will drift to different directions. Previous theoretical results based on the gradient divergence show that when ζ is large, I has to be small to avoid the divergence of the FL algorithms. However, in FL, the final output is the global model on the server, which is the average of local models after local updates. As shown in Figure 1, although ζ is large, the averaged model $\hat{\mathbf{x}}^{r,k}$ can still be close to the centralized model $\mathbf{x}_c^{r,k}$ that can be obtained if we had used centralized SGD. This means that the convergence error caused by local updates can be close to zero. While ζ successfully characterizes the variance among local gradients, it cannot capture the difference between the averaged model and the centralized model. Consequently, relying solely on the gradient divergence in convergence analysis can lead to an overestimation of the convergence error caused by local updates. To obtain a better convergence upper bound, *it is necessary to introduce a new metric which can characterize the difference between the averaged model and the centralized model.*

To address the inconsistency between the theory and practice, we introduce a new metric L_h, referred to as the *heterogeneity-driven pseudo-Lipschitz constant*. As shown in Figure 1, the proposed metric L_h captures the difference between the averaged model and the centralized model, which cannot be characterized by ζ. In our analysis, we use the heterogeneity-driven pseudo-Lipschitz constant L_h and the global Lipschitz constant L_g to substitute the widely used local Lipschitz constant $\tilde{L}$. This is based on our important observation that $\tilde{L}$ is affected by the data heterogeneity, which has not been pointed out in previous theoretical studies. In the literature (Yu et al., 2019b; Yang et al., 2020; Khaled et al., 2020), $\tilde{L}$ is used to characterize the smoothness of the gradients for all local objective functions under any degree of data heterogeneity. However, as shown in Table 1 (Section 6), $\tilde{L}$ increases fast as the percentage of non-IID data increases. We use L_h to characterize the information on data heterogeneity and use L_g to characterize the smoothness of the global objective function. It can be proved that the new assumptions used in this paper are weaker than the local Lipschitz gradient commonly used in the literature.

Contribution of this paper. In this paper, we reveal the fundamental effect of data heterogeneity on FedAvg and its extensions by introducing a new metric L_h,

the *heterogeneity-driven pseudo-Lipschitz constant* in Assumption 4.2. In particular, our main contributions are as follows.

1. Using the new assumptions, which are proved to be weaker than those in the literature, we develop a novel analysis for FedAvg and its extensions, including FedAvg with momentum (Yu et al., 2019a) and FedAdam (Reddi et al., 2020), with general non-convex objective functions. We show that for the terms with the number of local updates (I), the local Lipschitz constant $\tilde{L}$ is replaced by the newly introduced heterogeneity-driven pseudo-Lipschitz constant L_h and the global Lipschitz constant L_g. Since L_h can be significantly smaller than $\tilde{L}$ in practice, a much larger number of local updates (I) can be used to achieve a small convergence upper bound even if the gradient divergence ζ is large. This bridges the gap between theory and practice.

2. Our analysis can incorporate partial participation where only a subset of workers are sampled to perform local updates in each round. We show that with partial participation, increasing I can still improve the convergence rate when the data are highly heterogeneous.

3. We discuss a number of insights seen from the proposed L_h metric. For example we identify a region where local SGD can outperform mini-batch SGD for some quadratic objective functions.

4. Our theoretical results are validated using experiments.

2. Related Work

FedAvg, also known as local SGD, was first proposed by McMahan et al. (2017). Since then, there has been considerable work analyzing the convergence rate of local SGD (Haddadpour, Farzin et al., 2019; Yu et al., 2019b; Kairouz et al., 2021) and its extensions such as FedAvg with momentum (Yu et al., 2019a), SCAFFOLD (Karimireddy et al., 2020) and adaptive methods (Reddi et al., 2020). There is also a line of work focusing on the partial participation (Yang et al., 2020), compression and quantization (Jiang & Agrawal, 2018; Richtárik et al., 2021) in local SGD. Despite the extensive analysis of local SGD and its extensions, it is hard to show that a large number of local updates can improve the convergence rate when data are highly heterogeneous (Woodworth et al., 2020a;b), while in practice, more local updates can improve the convergence. To address the gap between the theory and practice, there are two papers (Wang et al., 2022; Das et al., 2022) trying to find new assumptions that can better characterize the effect of data heterogeneity in local SGD. However, none of these works have noted that the local Lipschitz constant

increases with the data heterogeneity. Consequently, they still rely on the local Lipschitz assumption for convergence analysis, whereas our work introduces the heterogeneity-driven pseudo-Lipschitz constant, yielding an improved convergence bound. A detailed discussion on related work can be found in Appendix A.

3. Preliminaries

In FedAvg, each round is composed of the local update phase and the global update phase. The global model is initialized as $\bar{\mathbf{x}}^0$. At the start of round r, the server distributes the global model $\bar{\mathbf{x}}^r$ to all workers. During the local update phase, each worker updates its local model with the local learning rate γ and the stochastic gradients sampled from its own local data distribution $\mathcal{D}_i$,

$$\mathbf{x}_i^{r,k+1} = \mathbf{x}_i^{r,k} - \gamma \mathbf{g}(\mathbf{x}_i^{r,k}; \mathbf{n}_i), \qquad (3)$$

where $\mathbf{x}_i^{r,k}$ is the local model at the rth round and kth iteration at worker i. For simplicity, we use $\mathbf{g}_i(\cdot)$ to denote the stochastic gradient $\mathbf{g}(\cdot; \mathbf{n}_i)$. In addition, $\bar{\mathbf{g}}(\cdot)$ denotes the stochastic gradient sampled from the global dataset $\mathcal{D}$. We assume that the local stochastic gradient is an unbiased estimate of the full local gradient, i.e., $\mathbb{E}\big[\mathbf{g}_i(\mathbf{x}_i^{r,k})\big|\mathbf{x}_i^{r,k}\big] = \nabla F_i(\mathbf{x}_i^{r,k})$. After I local iterations at the rth round, worker i sends the local model update $\Delta_i^r := \bar{\mathbf{x}}^r - \mathbf{x}_i^{r,I}$ to the server. During the global update phase, the server updates the global model using the following equality:

$$\bar{\mathbf{x}}^{r+1} = \bar{\mathbf{x}}^r - \eta \cdot \frac{1}{N} \sum_{i=1}^{N} \Delta_i^r, \qquad (4)$$

where η is the global learning rate.

The following assumptions are widely used in the literature for the analysis of algorithms including FedAvg (Karimireddy et al., 2020; Yu et al., 2019b; Khaled et al., 2020; Wang et al., 2020a), FedAvg with momentum (Yu et al., 2019a) and adaptive methods (Reddi et al., 2020).

Assumption 3.1 (Local Lipschitz Gradient)**.**

$$\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \le \tilde{L} \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}, i. \qquad (5)$$

There are also some works (Khaled et al., 2020) assuming that Lipschitz gradient condition holds for each data sample $\|\nabla \ell(\mathbf{x}; \xi) - \nabla \ell(\mathbf{y}; \xi)\| \le L' \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \mathcal{D}$. Note that this is stronger and can imply local Lipschitz gradient condition.

Assumption 3.2 (Bounded Stochastic Gradient Variance)**.**

$$\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla F_i(\mathbf{x})\|^2\right] \le \sigma^2, \forall i, \mathbf{x}. \qquad (6)$$

Assumption 3.3 (Bounded Gradient Divergence)**.**

$$\|\nabla F_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \zeta^2, \forall i, \mathbf{x}. \qquad (7)$$

Assumption 3.3 is often the only metric of data heterogeneity in the literature (Yu et al., 2019a; Wang & Joshi, 2019), where it was shown that there is a term $O(\gamma^2 \tilde{L}^2 I^2 \zeta^2)$ in the convergence upper bound. This means that the gradient divergence (ζ) and the number of local updates (I) are coupled, and the error caused by ζ grows fast as I increases and the effect of $I^2 \zeta^2$ is amplified by $\tilde{L}^2$. In this paper, we find that this result can be pessimistic since it can be seen from Table 1 (in Section 6) that $\tilde{L}$ can be very large, which means that the error caused by $I^2 \zeta^2$ can become much larger due to the large $\tilde{L}^2$. In the next section, we will address this problem using Assumption 4.2 in the analysis.

4. Main Results

In this section, we present the convergence upper bound for non-convex objective functions using the proposed new assumption for both full participation and partial participation. We summarize the technical novelty and provide proofs for all theorems and propositions in Appendix B.

In the literature, three classes of assumptions on *stochastic gradient variance*, *gradient divergence* and *smoothness* are often made for theoretical analysis (Yu et al., 2019b; Wang et al., 2020a; Khaled et al., 2020). We keep Assumption 3.2 for stochastic gradient variance and Assumption 3.3 for gradient divergence. Assumptions 4.1 and 4.2 will replace Assumption 3.1. In Section 5, we will show that Assumptions 4.1 and 4.2 are weaker than Assumption 3.1.

Assumption 4.1 (Global Lipschitz Gradient). The global objective function $f(\mathbf{x})$ satisfies

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_g \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}. \quad (8)$$

In our analysis, the Lipschitz gradient condition is only needed for the global objective function instead of for each local objective function as in Assumption 3.1 or for each data sample as in (Khaled et al., 2020).

Assumption 4.2 (Heterogeneity-driven Pseudo-Lipschitz Condition on Averaged Gradients). There exists a constant $L_h \geq 0$ such that $\forall \mathbf{x}_i$,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}}) \right\|^2 \leq \frac{L_h^2}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2, \quad (9)$$

where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ and L_h is referred to as the *heterogeneity-driven pseudo-Lipschitz constant*.

We consider Assumption 4.2 as a new perspective on data heterogeneity for the following reasons. First, L_h can be used to characterize the convergence error caused by local updates. In particular, we will show that $\tilde{L}$ can be

replaced by L_h in the local-update related terms in existing convergence bounds in the literature. Second, unlike Assumption 3.3, L_h can characterize the difference between the *averaged* model and the *centralized* model. This difference captures the actual impact of data heterogeneity as discussed in Section 1 (see Figure 1). We will discuss these new perspectives of Assumption 4.2 and L_h in detail in this section and in Section 5.

Next, we present the convergence analysis for full participation. In the following, we define $\mathcal{F} := f(\mathbf{x}_0) - f^*$.

Theorem 4.3 (General Non-convex Objective Functions). *Assuming Assumptions 3.2, 3.3, 4.1, 4.2 hold, when $\gamma\eta \leq \frac{1}{2IL_g}$ and $\gamma \leq \min\left\{ \frac{1}{2\sqrt{30}IL_g}, \frac{1}{\sqrt{6(L_h^2+L_g^2)I}} \right\}$, after R rounds of FedAvg, we have*

$$\min_{r \in [R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 = \mathcal{O}\Bigg(\underbrace{\frac{\mathcal{F}}{\gamma\eta IR} + \frac{\gamma\eta L_g \sigma^2}{N}}_{\text{error unrelated to local updates}}$$

$$+ \underbrace{\gamma^2 \left(\frac{L_g^2}{N} + L_h^2 \right)(I-1)\sigma^2 + \gamma^2 L_h^2 (I-1)^2 \zeta^2}_{\text{error caused by local updates}} \Bigg),$$

$$(10)$$

where $[R] := \{0, 1, \ldots, R-1\}$ in this paper.

An improved bound by using Assumptions 4.1 and 4.2. In (10), the convergence error terms that are unrelated to local updates only depend on L_g, while in the error caused by local updates, σ^2 is coupled with both L_g and L_h, and ζ^2 is coupled only with L_h. In Yu et al. (2019a); Yang et al. (2020), the error caused by the stochastic gradient noise is $\mathcal{O}(\frac{\gamma\eta \tilde{L}\sigma^2}{N})$, and the error caused by local updates is $\mathcal{O}(\gamma^2 \tilde{L}^2(I-1)^2\zeta^2 + \gamma^2 \tilde{L}^2(I-1)\sigma^2)$, where we observe that $\tilde{L}$ is substituted by L_g and L_h, respectively, in (10). As shown by the experimental results in Table 1 (Section 6), L_g is smaller than $\tilde{L}$, and L_h can be much smaller than $\tilde{L}$. In addition, our experimental results show that $L_g + L_h$ is not larger than $\tilde{L}$, which intuitively implies that $\frac{L_g^2}{N} + L_h^2$ is not larger than $\tilde{L}^2$ (a formal analysis of this relation is left for future work). This means that the error caused by local updates can be significantly overestimated using the convergence results in existing works. Moreover, in Section 5, we show mathematically that L_h and L_g are smaller than $\tilde{L}$.

New insights about the effect of data heterogeneity. It can be observed that in the error caused by local updates, both ζ^2 and σ^2 are multiplied by L_h. *A key message is that when ζ^2 is large, as long as L_h^2 is small enough, the error caused by local updates can still be small.* Since L_h and ζ characterize the effect of data heterogeneity in different perspectives, we show that it is possible that $L_h = 0$ while

ζ can be arbitrarily large by providing an explicit example in Section 5. In that special case, no matter how large ζ is, the convergence error of local SGD is the same as that of centralized SGD, i.e., I can be arbitrarily large and only one aggregation is sufficient.

It is worth noting that although L_h increases with the percentage of heterogeneous data, it can still be small even if the percentage of heterogeneous data is large as shown by the experimental results in Table 1 (Section 6). The following corollary shows that more local updates can improve the convergence.

Corollary 4.4. *With $\gamma\eta = \min\left\{\sqrt{\frac{\mathcal{F}N}{RIL_g\sigma^2}}, \frac{1}{2IL_g}\right\}$ and $\gamma = \frac{1}{\sqrt{RI}}$, for FedAvg, we have*

$$\min_{r\in[R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2$$
$$= \mathcal{O}\left(\sqrt{\frac{\mathcal{F}L_g\sigma^2}{RIN}} + \frac{\mathcal{F}L_g + L_h^2\zeta^2 + (L_h^2 + L_g^2/N)\sigma^2/I}{R}\right). \tag{11}$$

It can be seen that the order of the dominant term is $\mathcal{O}(\frac{1}{\sqrt{RI}})$, which is consistent with the results in the literature (Yang et al., 2020; Karimireddy et al., 2020). Similar to Theorem 4.3, all $\tilde{L}$ in the existing works is replaced by L_h and L_g. Hence, the insights discussed after Theorem 4.3 still hold here. In Appendix A.5, we show that the new assumption can also be applied in the convergence analysis for strongly convex objective functions.

Analysis for Partial Participation. We also use the new assumption to derive the convergence upper bound for partial participation. At each round, M workers are uniformly sampled with replacement. The result provides insights into the relationship between local updates and partial participation. It is worth noting that the technique for partial participation in existing works cannot be directly applied in our analysis since the Lipschitz gradient (see Assumption 3.1) is often assumed for each local objective function in the literature. Therefore, we need to develop new techniques to incorporate the partial participation using L_h and L_g, which can be found in Appendix B.

Theorem 4.5 (Partial Participation). *Consider uniformly sampling M ($1 \leq M \leq N$) workers in each round of FedAvg with replacement. Assuming Assumptions 3.2, 3.3, 4.1, 4.2 hold, when $\gamma\eta \leq \frac{M}{16IL_g}, \gamma \leq \min\left\{\frac{1}{3\sqrt{10}L_gI}, \frac{1}{\sqrt{6(L_h^2+L_g^2)I}}\right\}$ after R rounds of FedAvg, we have*

$$\min_{r\in[R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2$$
$$= \mathcal{O}\left(\underbrace{\frac{\mathcal{F}}{\gamma\eta IR} + \frac{\gamma\eta L_g\sigma^2}{M}}_{\text{error unrelated to local updates}} + \underbrace{\frac{\gamma\eta L_g I\zeta^2}{M}}_{\text{error caused by p.p.}}\right.$$

$$\left. + \gamma^2\left(\frac{L_g^2}{N} + L_h^2\right)(I-1)\sigma^2 + \gamma^2 L_h^2(I-1)^2\zeta^2\right), \underbrace{}_{\text{error caused by local updates}}$$
$$\tag{12}$$

where "p.p." means partial participation.

Compared to Theorem 4.3, there are two differences in the convergence upper bound. First, the error caused by the stochastic noise $\mathcal{O}\left(\frac{\gamma\eta L_g\sigma^2}{M}\right)$ depends on M. This means that more workers sampled in each round can reduce the noise. Second, there is an additional term $\mathcal{O}\left(\frac{\gamma\eta L_g I\zeta^2}{M}\right)$ in the convergence upper bound, which denotes the error caused by partial participation. In the literature (Yang et al., 2020), this term is often multiplied by $\tilde{L}$. In (12), this term depends on L_g and not on L_h. This means that a small L_h cannot reduce the error caused by partial participation, which can be shown explicitly by the following corollary.

Corollary 4.6. *Consider uniformly sampling M workers at each round in FedAvg with replacement. With $\gamma\eta = \min\left\{\sqrt{\frac{M\mathcal{F}}{L_gIR(\sigma^2+I\zeta^2)}}, \frac{1}{15L_gI}\right\}$ and $\gamma = \frac{1}{\sqrt{RI}}$, we have*

$$\min_{r\in[R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2$$
$$= \mathcal{O}\left(\sqrt{\frac{\mathcal{F}L_g\zeta^2}{RM}} + \sqrt{\frac{\mathcal{F}L_g\sigma^2}{RIM}} \right.$$
$$\left. + \frac{\mathcal{F}L_g + L_h^2\zeta^2 + (L_h^2 + L_g^2/N)\sigma^2/I}{R}\right). \tag{13}$$

Compared to the existing results, where the dominant term is $\mathcal{O}\left(\frac{\mathcal{F}\tilde{L}\zeta^2}{RM}\right)$, $\tilde{L}$ is substituted by L_g in (13). Since the effect of partial participation is shown by the dominant term, this implies that a small L_h cannot reduce the error caused by partial participation. This is because L_h characterizes the difference between the averaged model over all workers and the centralized model (we will formally explain this property in Section 5). However, with partial participation, the global model on the server becomes a stochastic estimate of the average models over all workers since only a subset of workers are randomly sampled.

Applying Assumption 4.2 to other FL algorithms. Similar to Assumption 3.1, the proposed Assumption 4.2 can be used to analyze the performance of other FL algorithms using our methodology. In particular, we provide the convergence analyses for two examples including the FedAvg with momentum (Yu et al., 2019a) in Appendix A.3 and the FedAdam (Reddi et al., 2020) in Appendix A.4. In these two examples, the same conclusions on the effect of the heterogeneity-driven pseudo-Lipschitz constant as that for FedAvg can be made.

5. Discussions

In this section, we discuss the properties and advantages of our proposed L_h metric. First, we show that Assumption 4.1 (global Lipschitz gradient) and Assumption 4.2 (heterogeneity-driven pseudo-Lipschitz gradient) used in this paper are weaker than the commonly used Assumption 3.1 (local Lipschitz gradient). Then, we explain the significance of L_h by showing its ability to characterize the difference between the "virtual" averaged model (defined in (14)) and the centralized model. Afterwards, we illustrate some nice properties of L_h by considering an exemplar case of quadratic objective functions. By applying L_h in the convergence analysis of quadratic objective functions, we identify a region where local SGD can be better than mini-batch SGD.

5.1. Properties and Advantages of L_h

Additional definition. For the exposition of useful insights in our discussion, we define $\hat{\mathbf{x}}^{r,k}$ as the "virtual" averaged model during the local update phase and

$$\hat{\mathbf{x}}^{r,k+1} := \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i^{r,k+1} = \hat{\mathbf{x}}^{r,k} - \gamma \cdot \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{x}_i^{r,k}), \tag{14}$$

where $k \in \{0, 1, 2, \ldots, I-1\}$. Note that this virtual model $\hat{\mathbf{x}}^{r,k}$ may not be observed in the system, and is mainly used for the theoretical analysis. In addition, we define $\mathbf{x}_c^{r,k}$ as the model that would have been obtained by applying centralized updates at the kth iteration of the rth round given the averaged model $\hat{\mathbf{x}}^{r,k}$, which means that the gradient is sampled from the global data distribution $\mathcal{D}$.[2] Specifically,

$$\mathbf{x}_c^{r,k+1} := \hat{\mathbf{x}}^{r,k} - \gamma \bar{\mathbf{g}}(\hat{\mathbf{x}}^{r,k}), \tag{15}$$

where $\mathbb{E}\left[\bar{\mathbf{g}}(\hat{\mathbf{x}}^{r,k})\right] = \nabla f(\hat{\mathbf{x}}^{r,k})$.

Assumptions in this paper are weaker. In the following proposition, we show that Assumptions 4.1 and 4.2 are weaker than Assumption 3.1.

Proposition 5.1. *If Assumption 3.1 holds, then Assumption 4.1 holds by choosing $L_g = \tilde{L}$ and Assumption 4.2 holds by choosing $L_h = \tilde{L}$.*

Explanation of L_h. Assumption 4.2 captures the difference between the averaged model and centralized model, which can be seen from the following proposition. Recall that the virtual averaged model $\hat{\mathbf{x}}^{r,k}$ is defined in (14) and the centralized model $\mathbf{x}_c^{r,k}$ is defined in (15).

Proposition 5.2. *Given the virtual averaged model at the*

rth round and kth iteration $\hat{\mathbf{x}}^{r,k}$, we have

$$\left\| \mathbb{E}[\hat{\mathbf{x}}^{r,k+1} | \hat{\mathbf{x}}^{r,k}] - \mathbb{E}[\mathbf{x}_c^{r,k+1} | \hat{\mathbf{x}}^{r,k}] \right\|^2$$
$$\leq \gamma^2 \cdot \frac{L_h^2}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2. \tag{16}$$

Proposition 5.2 shows that although the difference among local models, captured by $\left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$ (which depends on both ζ and σ as shown in Lemma B.2), can be large after multiple local iterations, the difference between the averaged model and centralized model can still be small if L_h is small.

5.2. Analysis for Quadratic Objective Functions

In order to obtain an explicit relationship among L_h, $\tilde{L}$ and ζ, and demonstrate the benefit of using L_h, we consider the following quadratic objective function,[3]

$$F_i(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{\mathrm{T}}\mathbf{x} + c_i. \tag{17}$$

Using (1), the global objective function is given by $f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$, where $\mathbf{A} := \tfrac{1}{N}\sum_{i=1}^{N}\mathbf{A}_i$, $\mathbf{b} := \tfrac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i$ and $c := \tfrac{1}{N}\sum_{i=1}^{N}c_i$.

In Proposition 5.1, it is implied that $L_h \leq \tilde{L}$. Further, as shown in Table 1 (Section 6), L_h can be much smaller than $\tilde{L}$. In general, the explicit relationship between L_h and $\tilde{L}$ is challenging to derive. However, for quadratic objective functions, this relationship can be shown in the following proposition.

Proposition 5.3. *For quadratic objective functions defined in (17), Assumptions 3.1 and 4.2 hold with $\tilde{L} = \max_{i \in [N]} \|\mathbf{A}_i\|_2$ and $L_h = \max_{i \in [N]} \|\mathbf{A}_i - \mathbf{A}\|_2$, respectively, where $\|\cdot\|_2$ is the spectral norm.*

From Proposition 5.3, it can be seen that both L_h and $\tilde{L}$ capture the properties of Hessian matrices for quadratic objective functions. The heterogeneity-driven pseudo-Lipschitz constant L_h characterizes the largest absolute eigenvalue of the "deviation" of $\{\mathbf{A}_i\}$ from the global Hessian matrix $\mathbf{A}$, while $\tilde{L}$ characterizes the largest absolute eigenvalue of $\{\mathbf{A}_i\}$. We observe that when $\mathbf{A}_i = \mathbf{A}, \forall i$, which means that the difference of local Hessian matrices is zero, Assumption 4.2 holds with $L_h = 0$. Note that, at the same time, we can pick an $\mathbf{A}_i$ such that $\tilde{L} = \max_{i \in [N]} \|\mathbf{A}_i\|_2$ is much larger than zero. Hence, in this example, we explicitly show that L_h can be arbitrarily smaller than $\tilde{L}$.

In Proposition 5.2, it has been shown that even when ζ is large, the difference between the averaged model and the

---

[2]Note that the model $\mathbf{x}_c^{r,k}$ is different from the model obtained by applying the centralized updates from the beginning of the algorithm. We use this for the purpose of illustration only while not affecting the convergence bound results.

[3]Here we do not assume the Hessian matrix is positive definite so that the quadratic objective function can be non-convex.

centralized model can still be small as long as L_h is small. In the following proposition, we explicitly show that for quadratic functions, L_h can be zero when ζ is large.

Proposition 5.4. *For quadratic objective functions defined in (17), when $\zeta = 0$, Assumption 4.2 holds with $L_h = 0$, while when $L_h = 0$, ζ can be arbitrarily large.*

Proposition 5.4 shows that $L_h = 0$ is not a sufficient condition for $\zeta = 0$, which implies that only using ζ can overestimate the effect of the data heterogeneity. This is because, as we have seen in Proposition 5.3, for quadratic objective functions, the key effect of heterogeneity on the local updates is due to the difference between $\mathbf{A}$ and $\mathbf{A}_i$, while ζ depends not only on the difference between $\mathbf{A}$ and $\mathbf{A}_i$ but also on the difference between $\mathbf{b}$ and $\mathbf{b}_i$. In addition, we notice that in multi-label learning (Zhang & Zhou, 2014), when $\mathbf{A} = \mathbf{A}_i$, $\mathbf{b}$ can be very different from $\mathbf{b}_i$ since data examples sharing the same feature can have different labels. This means that $L_h = 0$ but $\zeta > 0$ is possible in practice.

On local SGD v.s. mini-batch SGD. In the following theorem, we consider the special case of $L_h = 0$, by which we show that local SGD can outperform mini-batch SGD even when ζ is arbitrarily large analytically. The extended discussion on the comparison between local SGD and mini-batch SGD for quadratic objective functions with $L_h > 0$ can be found in Appendix A.2, where similar conclusions still hold. Instead of directly applying $L_h = 0$ to Theorem 4.3, we develop a new proof technique for Theorem 5.5 below. The difference in the techniques can be shown by the requirement of the local learning rate γ, which no longer depends on I in Theorem 5.5 while Theorem 4.3 requires $\gamma \leq \min\left\{\frac{1}{2\sqrt{30}IL_g}, \frac{1}{\sqrt{6(L_h^2+L_g^2)I}}\right\}$.

In the following, we use t to denote the index of the total number of iterations, where $t \in [RI] := \{0, 1, \ldots, RI-1\}$. For some given r and k, we define $\hat{\mathbf{x}}^t$ as

$$\hat{\mathbf{x}}^t = \begin{cases} \hat{\mathbf{x}}^{r,k}, & \text{if } t = rI + k \text{ and } k \neq 0, \\ \bar{\mathbf{x}}^r, & \text{if } t = rI. \end{cases} \quad (18)$$

Theorem 5.5 (Special Case of $L_h = 0$)**.** *For quadratic objective functions defined in (17), with a common Hessian $\mathbf{A} = \mathbf{A}_i, \forall i$, when $\gamma \leq \frac{1}{L_g}$ and $\eta = 1$, for local SGD with I local iterations, we have*

$$\min_{t\in[RI]} \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 = \mathcal{O}\left(\frac{\mathcal{F}}{\gamma RI} + \frac{\gamma L_g}{N}\sigma^2\right); \quad (19)$$

and for mini-batch SGD with batch size I and learning rate $\gamma \leq \frac{1}{L_g}$, we have

$$\min_{t\in[RI]} \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 = \mathcal{O}\left(\frac{\mathcal{F}}{\gamma R} + \frac{\gamma L_g}{NI}\sigma^2\right). \quad (20)$$

In Theorem 5.5, the cost of communication and computation is the same for both local SGD and mini-batch SGD when R is fixed, since the number of aggregations is R and the total number of gradients sampled is NRI for both algorithms. The upper bound for γ is also the same. Comparing (19) with (20), we see that for local SGD, I is in the first term of (19), which means that local SGD uses more computation to reduce the error caused by initialization, since $\mathcal{F} = f(\bar{\mathbf{x}}^0) - f(\mathbf{x}^*)$. For mini-batch SGD, I is in the second term of (20), which means that mini-batch SGD uses more computation to reduce the error caused by the variance σ^2. Based on the above insights, we identify a region where local SGD can be better than mini-batch SGD in the following corollary.

Corollary 5.6. *Consider the quadratic objective function in Theorem 5.5. When $\sigma \leq \sqrt{\frac{\mathcal{F}NL_g}{RI}}$ and with appropriately chosen learning rates, for local SGD, we have*

$$\min_{t\in[RI]} \mathbb{E}\left[\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{F}L_g}{RI}\right); \quad (21)$$

for mini-batch SGD, we have

$$\min_{t\in[RI]} \mathbb{E}\left[\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{F}L_g}{R}\right). \quad (22)$$

First, it can be seen that the order of the dominant term for local SGD is $\mathcal{O}(\frac{1}{RI})$ while for mini-batch SGD, it is $\mathcal{O}(\frac{1}{R})$. This means that when I is large, local SGD can be much faster than mini-batch SGD. Second, the condition confirms the intuition that when the error caused by initialization is large ($\sigma \leq \sqrt{\frac{\mathcal{F}NL_g}{RI}}$), we should choose local SGD. Similar insights can also be shown in the results for $L_h > 0$ in Theorem A.2. It is worth noting that this result shows that the advantage of local SGD can be achieved even when ζ is arbitrarily large, while in the literature (Woodworth et al., 2020b), local SGD has been proved to be better than mini-batch SGD only when ζ is small.

A limitation of the result in Theorem 5.5 and Corollary 5.6 is that the left-hand side (LHS) of the convergence bound includes $\hat{\mathbf{x}}^t$, which can be either the virtual (non-observable) average model $\hat{\mathbf{x}}^{r,k}$, when $k \neq 0$ in (18), or the observable average model $\bar{\mathbf{x}}^r$, when $k = 0$ in (18). An extension to considering only the errors related to $\bar{\mathbf{x}}^r$ is left for future work.

6. Experiments

In this section, we present experimental results obtained from various datasets and models to validate our theoretical findings. In particular, we estimate $\tilde{L}$, L_h and L_g on MNIST (LeCun et al., 1998) with multilayer perceptron (MLP), CIFAR-10 (Krizhevsky & Hinton, 2009) with CNN and VGG-11, CIFAR-100 with VGG-16. Then we provide the
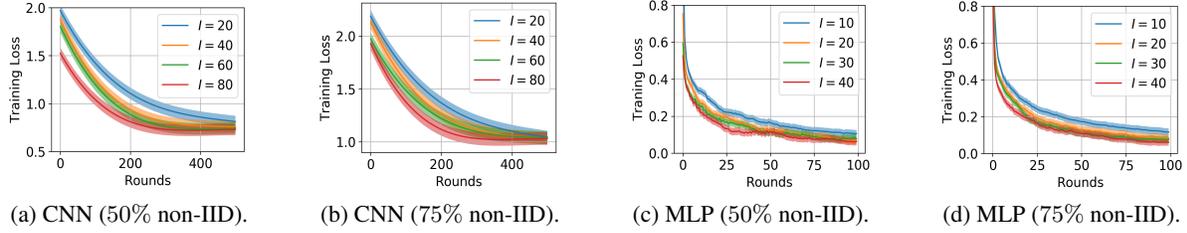
(a) CNN (50% non-IID).   (b) CNN (75% non-IID).   (c) MLP (50% non-IID).   (d) MLP (75% non-IID).

Figure 2: Results for CNN with CIFAR-10 and MLP with MNIST. For CNN, the learning rates are chosen as $\eta = 2$ and $\gamma = 0.05$. For MLP, the learning rates are chosen as $\eta = 2$ and $\gamma = 0.1$. Results for CNN are shown in (a) and (b). Results for 75% of MNIST are shown in (c) and (d).

Table 1: Estimated L_h, $\tilde{L}$, L_g for MLP with MNIST, CNN and VGG-11 with CIFAR-10 and VGG-16 with CIFAR-100. Since L_g only depends on the global dataset, L_g does not change with the percentage of non-IID (NIID) data.

Obj.	NIID	$\tilde{L}$	L_h	L_g
MLP	25%	130.97 ± 11.67	0.82 ± 0.11	
	50%	130.97 ± 11.67	0.82 ± 0.11	122.23
	75%	134.24 ± 12.23	1.66 ± 0.23	± 9.75
	100%	141.92 ± 12.78	2.36 ± 0.29	
CNN	25%	447.59 ± 22.27	0.96 ± 0.13	
	50%	898.49 ± 38.57	1.21 ± 0.19	323.35
	75%	1131.36 ± 47.82	1.63 ± 0.26	± 15.36
	100%	1662.24 ± 62.18	2.15 ± 0.34	
VGG-11	25%	161.07 ± 3.07	9.14 ± 0.32	
	50%	$246.68. \pm 7.67$	10.47 ± 0.35	151.95
	75%	338.07 ± 8.43	12.49 ± 0.41	± 5.42
	100%	536.92 ± 10.43	17.83 ± 1.12	
VGG-16	25%	363.11 ± 3.87	2.53 ± 0.12	
	50%	365.39 ± 1.1	3.26 ± 0.12	360.85
	75%	379.91 ± 8.82	3.96 ± 0.19	± 1.86
	100%	431.07 ± 25.10	4.67 ± 0.27	

Table 2: Special case of $L_h = 0$ with the quadratic objective functions. $I = 1$ is equivalent to mini-batch SGD. The number of rounds is the communication rounds needed to achieve the target of $f(\mathbf{x}) = 0.8$. For varying (η, γ), we fix $I = 10$ and for varying (I, s), we fix $\eta = 1$, $\gamma = 0.005$.

(η, γ)	$(1, 0.005)$	$(2, 0.0025)$	$(5, 0.001)$	$(10, 0.0005)$
# Rounds	86 ± 1.6	86 ± 1.6	86 ± 1.6	86 ± 1.6

(I, s)	$(1, 1) \& (1, 5)$	$(1, 10)$	$(5, 1)$	$(10, 1)$
# Rounds	927 ± 3.4	925 ± 1.7	187 ± 2.3	95 ± 2.4

Second, we observe that both L_g and L_h are smaller than $\tilde{L}$. This verifies the theoretical results in Proposition 5.1. Furthermore, the results in Table 1 show that L_h can be much smaller than $\tilde{L}$. This means that when characterizing the error caused by local updates, substituting $\tilde{L}$ by L_h can reduce the convergence upper bound.

The theoretical results in Theorem 4.3 and Theorem 4.5 are verified. In Figure 2, the convergence results for CNN with partial participation and MLP with full participation are provided. In Table 1, we see that L_h is relatively small in these cases. According to Theorem 4.3 and Theorem 4.5, for both full participation and partial participation, when L_h is small, the error caused by local updates is small, so a large I can still improve convergence. The experimental results in Figure 2 verify the theoretical results, because even when the percentage of heterogeneous data is more than 50%, the largest I ($I = 80$ for CNN and $I = 40$ for MLP) can still achieve the smallest training loss when R is fixed.

The insights gained from the analysis for quadratic objective functions are verified. We construct quadratic examples to verify Theorem 5.5. We consider $F_i(\mathbf{x}) = \frac{1}{2} \|\mathbf{U}\mathbf{x} - \mathbf{v}_i\|^2$, where $\mathbf{U} \in \mathbb{R}^{100 \times 100}$, $\mathbf{v}_i \in \mathbb{R}^{100}$. Each column of $\mathbf{U}$ and $\mathbf{v}_i$ is sampled from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In this case, the gradient divergence is $\|\mathbf{U}(\mathbf{v}_i - \mathbf{v})\|^2 > 0$. We set the stochastic gradient variance as $\sigma^2 = 0.01$. To distinguish the number of local updates from the mini-batch size in the experiments, we use a separate variable s to indicate the mini-batch size. Theorem 4.3 shows that when $L_h = 0$, using two-sided

results for FedAvg on MNIST and CIFAR-10 to verify the theoretical results in Theorem 4.3 and Theorem 4.5. Results with synthetic data for quadratic objective functions are also provided to verify the insights shown by Theorem 5.5.

The experimental setting is as follows. For training CNN with CIFAR-10, we partition the training dataset into 100 workers, and we uniformly sample 10 workers in each round. For other datasets and models, we partition the training dataset into 10 workers and use full participation. For the non-IID setting, the data on each worker is sampled in two steps. First, $X\%$ of the data on one worker is sampled from a single label, and we say that the percentage of heterogeneous data on this worker is $X\%$. Then, we uniformly partition the remaining data into all workers. Additional experimental details and results can be found in Appendix C.

The properties of $\tilde{L}$, L_h and L_g are verified. The experimental results for estimating $\tilde{L}$, L_h and L_g are shown in Table 1. First, it can be seen that $\tilde{L}$ grows fast as the percentage of the data heterogeneity increases, which implies that $\tilde{L}$ is related to the data heterogeneity in addition to the smoothness of local objective function.

learning rates does not have advantage over a single learning rate. This is validated by the experiments shown in Table 2, where there is no difference among the results with different learning rates when keeping the product of γ and η. When $s = 1$, comparing the results with $I = 1$, $I = 5$, and $I = 10$ in Table 2, we see that more local updates can reduce the number of rounds to achieve $f(\mathbf{x}) = 0.8$ (an arbitrarily chosen target value), which validates the results in Theorem 5.5. By the comparison between the results of $I = 1, s = 5$ and $I = 5, s = 1$ and the comparison between the results of $I = 1, s = 10$ and $I = 10, s = 1$, we can see that keeping the number of gradients sampled in one round the same, local SGD ($I > 1$) converges faster than mini-batch SGD ($I = 1$) when σ^2 is small (since $\sigma^2 = 0.01$ in this case), which validates the discussion for Theorem 5.5.

7. Conclusion

In this paper, we bridged the gap between the pessimistic theoretical results and the good experimental performance for FL by introducing a new theoretical perspective of the data heterogeneity, named the heterogeneity-driven pseudo-Lipschitz assumption, which can characterize the difference between the averaged model and the centralized model. This is the key to explain the benefit of local updates, especially when the gradient divergence is large. Using this assumption, we developed new analytical approaches to derive convergence upper bounds for FedAvg and its extensions, and for both non-convex and quadratic functions. These bounds can be much smaller than those in the literature and can better explain the effect of data heterogeneity using the heterogeneity-driven pseudo-Lipschitz constant. As a by-product, our approach can identify a region where local SGD can outperform mini-batch SGD without any constraint on the gradient divergence. All theoretical findings were also validated using experiments.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgment

References

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SAIM Review*, 60(2):223–311, 2018.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. CINIC-10 is not imagenet or CIFAR-10, 2018.

Das, R., Acharya, A., Hashemi, A., Sanghavi, S., Dhillon, I. S., and Topcu, U. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pp. 496–506. PMLR, 2022.

Haddadpour, Farzin et al. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems*, 2019.

Hubbard, J. H. and Hubbard, B. B. *Vector calculus, linear algebra, and differential forms: a unified approach*. Matrix Editions, 2015.

Jiang, P. and Agrawal, G. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *NeurIPS*, 2018.

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

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

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

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks, 2020b.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations*,

2020. URL https://openreview.net/forum?id=B1eyO1BFPr.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Niknam, S., Dhillon, H. S., and Reed, J. H. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6):46–51, 2020.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization, 2020.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. In *ICML*, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33:7611–7623, 2020a.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. Slowmo: Improving communication-efficient distributed SGD with slow momentum. In *International Conference on Learning Representations*, 2020b.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. On the unreasonable effectiveness of federated averaging with heterogeneous data, 2022.

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. Is local SGD better than minibatch SGD? In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10334–10343, 2020a.

Woodworth, B. E., Patel, K. K., and Srebro, N. Minibatch vs local SGD for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020b.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. In *International Conference on Learning Representations*, 2020.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *ICML*, pp. 7184–7193, Jun. 2019a.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *AAAI*, Jan.–Feb. 2019b.

Zhang, M.-L. and Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014. doi: 10.1109/TKDE.2013.39.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Appendix

This Appendix is composed of three sections. Appendix A provides more details of related work and additional theoretical results are presented. Appendix B provides all the proofs for theorems, corollaries and propositions in this paper. Appendix C provides additional details and results of experiments.

A. Additional Discussions

In this section, first we provide more discussions on related work. Second, we present the convergence analysis for the quadratic objective functions with $L_h > 0$, where we can also identify a parameter region where local SGD can be better than mini-batch SGD. Then we show that L_h and L_g can be successfully applied in the analysis for FedAvg with momentum (Yu et al., 2019a) and FedAdam (Reddi et al., 2020).

A.1. Additional Details of Related Work

There has been considerable work analyzing the convergence rate of federated learning algorithms (not limited to FedAvg), with non-convex objective functions (Haddadpour, Farzin et al., 2019; Yu et al., 2019b; Wang & Joshi, 2019; Karimireddy et al., 2020; Reddi et al., 2020). A key step shared by these analyses is to relate the difference of gradients,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}}) \right\|^2,$$

to the model divergence,

$$\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2, \tag{A.1}$$

which can be found, for example, in inequality (10) in the supplementary of Yu et al. (2019b), the inequality (6) in the supplementary of Reddi et al. (2020), and the proof of Lemma 19 in Karimireddy et al. (2020). In this step, the local Lipschitz gradient assumption (Assumption 3.1) is often applied, which amplifies the effect of data heterogeneity. In this paper, the pseudo-Lipschitz constant L_h is applied in this step so that the convergence error can be much smaller than that based on the local Lipshictz constant $\tilde{L}$, since it can be seen in Table 1 that L_h is often much smaller than $\tilde{L}$. Therefore, we believe that our techniques can also be applied to federated learning algorithms to improve the convergence analysis.

There are two papers (Wang et al., 2022; Das et al., 2022) closely related to our work. Both works assume the Lipschitz gradient for each local objective function while we only assume it for the global objective function. Wang et al. (2022) aim to re-characterize the data heterogeneity by extending the single gradient divergence assumption ((4) in Wang et al. (2022)) to the averaged gradient divergence assumption ((15) in Wang et al. (2022)). In addition, Wang et al. (2022) consider the convex objective function and their analysis cannot guarantee convergence to a stationary point while we consider general non-convex objective function and our results can guarantee convergence to a stationary point.

In the following, we provide more details about the difference between Wang et al. (2022) and our paper. In Wang et al. (2022), a new metric for data heterogeneity, ρ, the average drift at optimum, is proposed. The definition of ρ is

$$\rho = \left\| \frac{1}{\gamma I} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i^{r,I} - \bar{\mathbf{x}}^r \right) \right\|. \tag{A.2}$$

We discuss the difference between Wang et al. (2022) and our paper in the following three aspects.

First, the new metric ρ in Wang et al. (2022) focuses on the difference between models while in our paper, we still focus on the difference between the gradients. The key insight in Wang et al. (2022) is that if ρ is small, when the global model is $\mathbf{x}^*$, after multiple local updates, the averaged model does not change significantly. In our paper, the key insight is that since L_h can be small, the difference between the current global gradient and the current averaged local gradients can be small. In Wang et al. (2022), the gradient divergence (Assumption 3.3) is not used in the analysis. In our analysis, we still use the gradient divergence jointly with the proposed L_h to characterize the data heterogeneity.

Second, in Wang et al. (2022), it is only empirically shown that ρ can be small. In our paper, we not only empirically demonstrate that L_h can be small, but also mathematically proved that L_h is smaller than or equal to $\tilde{L}$ and provide an

analytical example to show the exact values of L_h and $\tilde{L}$. Our quadratic example can be non-convex, a case which ρ cannot cover.

Third, one weakness of using ρ is that in the convergence upper bound in Wang et al. (2022), the convergence error shown by ρ may not vanish. This means that by choosing $\gamma = \frac{1}{\sqrt{R}}$, when R goes to infinity, the convergence upper bound cannot guarantee that FedAvg can converge to the local minima of the global objective function. On the contrary, the convergence upper bound proved in this paper can guarantee the convergence to the local minima of the global objective function, which is shown by Corollary 4.4.

Das et al. (2022) introduce a parameter α to characterize the relationship between the difference of gradients to the model divergence shown in (A.1). This can also be done using the Assumption 4.2 and L_h. However, α cannot characterize the impact of L_h and Das et al. (2022) still assume Lipschitz gradient for each local objective function. They only use α as an intermediate step instead of theoretically analyzing the effect of data heterogeneity. In their theoretical results, the convergence error increases with I even when $\alpha = 0$.

A.2. Additional Results for Quadratic Objective Functions

In this section, we provide a comparison between mini-batch SGD and local SGD for quadratic objective functions with $L_h > 0$, i.e., $\boldsymbol{A}_i \neq \boldsymbol{A}, \forall i$.

Recall that we use t to denote the index of the total number of iterations, where $0 \leq t \leq RI - 1$, and the averaged model $\hat{\mathbf{x}}^t$ is defined in (18). In the appendix, we also define the local model on worker i at the tth iteration as

$$\mathbf{x}_i^t = \mathbf{x}_i^{r,k}, t = rI + k. \tag{A.3}$$

In the following, we introduce κ to characterize the difference between eigenvalues of the Hessian matrices $\{\boldsymbol{A}_i\}$. That is,

$$\kappa := \max_{i,j} 1 - \frac{\lambda_j(\boldsymbol{A}_i)}{\|\mathbf{A}_i\|_2}, \tag{A.4}$$

where $\lambda_j(\boldsymbol{A}_i)$ is the jth eigenvalue of $\boldsymbol{A}_i$ and $0 < \kappa \leq 2$.

It can be seen that κ is determined by the smallest eigenvalues of $\{\boldsymbol{A}_i\}$. Only when $\lambda_j(\boldsymbol{A}_i) < 0$ and $|\lambda_j(\boldsymbol{A}_i)| = \|\mathbf{A}_i\|_2$, κ is maximized and then we have $\kappa = 2$.

When analyzing the quadratic objective functions, it is worth noting that the gradient divergence is given by

$$\|\nabla F_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 = \|(\boldsymbol{A} - \boldsymbol{A}_i)\mathbf{x} + \mathbf{b} - \mathbf{b}_i\|^2, \forall i. \tag{A.5}$$

It can be seen that in this case, the gradient divergence cannot be bounded for all $\mathbf{x} \in \mathbb{R}^d$. Therefore, we apply the following assumption in the analysis for the quadratic objective functions.

Assumption A.1 (Weak Gradient Divergence). For FedAvg, with quadratic objective functions, for the global model $\bar{\mathbf{x}} \in \{\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^R\}$, we have

$$\|\nabla F_i(\bar{\mathbf{x}}) - \nabla f(\bar{\mathbf{x}})\|^2 \leq \zeta_q^2. \tag{A.6}$$

Theorem A.2 (Quadratic Objective Functions with $L_h > 0$). *With* $\gamma \leq \min\left\{\frac{1}{\lambda_{\max}}, \frac{1}{2L_h} \cdot \min\left\{\frac{1}{I}, \frac{([\varphi(\kappa)]^2 - 1)^3}{[\varphi(\kappa)]^{2(I+2)}}\right\}\right\}$, *for local SGD with quadratic objective functions that satisfy Assumptions 3.2 and A.1, we have*

$$\min_{t \in [T]} \mathbb{E}\|\nabla f(\hat{\mathbf{x}}^t)\|^2 \leq \frac{4\mathcal{F}}{\gamma T} + \frac{2\gamma L_g \sigma^2}{N} + 16\gamma^2 L_h^2 I \cdot \phi(\kappa, I) \cdot \zeta_q^2 + 4\gamma^2 L_h^2 \cdot \phi(\kappa, I) \cdot \sigma^2, \tag{A.7}$$

where $\lambda_{\max} = \max_i \|\mathbf{A}_i\|_2$,

$$\phi(\kappa, k) = \begin{cases} k & 0 \leq \kappa < 1 \\ \frac{\kappa^{2k} - 1}{\kappa^2 - 1} & 1 \leq \kappa \leq 2. \end{cases} \tag{A.8}$$

and

$$\varphi(\kappa) = \begin{cases} 1 & 0 \leq \kappa < 1 \\ \kappa & 1 \leq \kappa \leq 2. \end{cases} \tag{A.9}$$

The proof can be found in Appendix B.14. Compared to the theoretical results for general non-convex objective functions, the main improvement is on the choice of learning rate. We develop new techniques in the proof to achieve the improvement on the learning rate, which takes the advantage of the properties of quadratic objective functions. For the ease of comparison, the convergence bound for mini-batch SGD is as follows. With learning rate $\gamma \leq \frac{1}{L_g}$, for mini-batch SGD (Bottou et al., 2018), we have

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^t) \right\|^2 = \mathcal{O} \left(\frac{\mathcal{F}}{\gamma R} + \frac{\gamma L_g \sigma^2}{2NI} \right), \tag{A.10}$$

where $L_g = \|\mathbf{A}\|_2$.

Next, we will present the convergence rate for different values of σ. The following corollaries can be obtained directly from Theorem A.2 by plugging in the corresponding learning rate γ.

First, we consider the simplest case, that is $\sigma = 0$.

Corollary A.3 ($\sigma = 0$ for Quadratic Objective Functions)**.** *When $\sigma = 0$, with $\gamma = \frac{1}{(RI)^{\frac{1}{3}}}$, for local SGD, we have*

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^t) \right\|^2 = \mathcal{O} \left(\frac{\mathcal{F} + L_h^2 I \cdot \phi(\kappa, I) \cdot \zeta_q^2}{(RI)^{\frac{2}{3}}} \right), \tag{A.11}$$

while for mini-batch SGD, we have

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^t) \right\|^2 = \mathcal{O} \left(\frac{\mathcal{F} L_g}{R} \right). \tag{A.12}$$

In this case, when $\frac{R^{\frac{1}{3}}}{I^{\frac{2}{3}}} < \frac{\mathcal{F} L_g}{\mathcal{F} + L_h^2 I \phi(\kappa, I) \zeta_q^2}$, the convergence rate of local SGD is better than that of mini-batch SGD.

Second, we consider the case when $\sigma^2 \leq \frac{2N\mathcal{F}}{L_g R I}$.

Corollary A.4. *When $\sigma^2 \leq \frac{N\mathcal{F}}{\gamma^2 L_g R I}$, with $\gamma = \frac{1}{(RI)^{\frac{1}{3}}}$, for local SGD, we have*

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^t) \right\|^2 = \mathcal{O} \left(\frac{\mathcal{F} + L_h^2 I \cdot \phi(\kappa, I) \cdot \zeta_q^2}{(RI)^{\frac{2}{3}}} + \frac{\mathcal{F} N L_h^2 \phi(\kappa, I)}{L_g R^{\frac{5}{3}} I^{\frac{5}{3}}} \right), \tag{A.13}$$

while for mini-batch SGD, we have

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^t) \right\|^2 = \mathcal{O} \left(\frac{\mathcal{F} L_g}{R} + \frac{\mathcal{F} L_g}{R I^2} \right). \tag{A.14}$$

Similarly, when $\frac{R^{\frac{1}{3}}}{I^{\frac{2}{3}}} < \frac{\mathcal{F} L_g}{\mathcal{F} + L_h^2 I \phi(\kappa, I) \zeta_q^2}$, dominant term of local SGD is better than that of mini-batch SGD.

A.3. Applying L_h and L_g in the Analysis for FedAvg with Momentum

In this section, we apply Assumptions 4.1 and 4.2 in the analysis for FedAvg with Momentum in Yu et al. (2019a). First, we introduce the notations and the algorithm for clarification. We summarize the FedAvg with momentum in Algorithm 1.

The momentum of worker i at tth iteration is denoted by $\mathbf{u}_i^t \in \mathbb{R}^d$, where t is the index of the total number of iterations. That is, $t = aI + b, a, b \in \mathbb{N}$ and $0 \leq a \leq R - 1, 0 \leq b \leq I - 1$. At the start of the algorithm, the momentum is initialized as zero. That is, $\mathbf{u}_i^0 = \mathbf{0}, \forall i$. During local updates, we have

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \gamma \mathbf{u}_i^{t+1}, \tag{A.15}$$

and

$$\mathbf{u}_i^{t+1} = \beta \mathbf{u}_i^t + \mathbf{g}_i(\mathbf{x}_i^t), \tag{A.16}$$

Algorithm 1 FedAvg with momentum (Algorithm 1 in Yu et al. (2019a))

Input: $\gamma, \hat{\mathbf{x}}^0, I, \beta, \hat{\mathbf{u}}^0 = \mathbf{0}, \forall i$
Output: Global averaged model $\hat{\mathbf{x}}^{RI}$
for $t = 0$ *to* $RI - 1$ **do**
  for *Each worker i, in parallel* **do**
    if $t = aI, 0 \leq a \leq R - 1$ **then**
      $\mathbf{u}_i^t \leftarrow \hat{\mathbf{u}}^t$;
      $\mathbf{x}_i^t \leftarrow \hat{\mathbf{x}}^t$;
    end
    Sample the stochastic gradient $\mathbf{g}_i(\mathbf{x}_i^t)$;
    $\mathbf{u}_i^{t+1} \leftarrow \beta \mathbf{u}_i^t + \mathbf{g}_i(\mathbf{x}_i^t)$;
    $\mathbf{x}_i^{t+1} \leftarrow \mathbf{x}_i^t - \gamma \mathbf{u}_i^{t+1}$;
  end
  if $t = aI + I - 1, 0 \leq a \leq R - 1$ **then**
    $\hat{\mathbf{u}}^{t+1} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{u}_i^{t+1}$;
    $\hat{\mathbf{x}}^{t+1} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i^{t+1}$;
  end
end

---

where $\beta \in [0, 1)$. After local updates, the momentum is reset as

$$\mathbf{u}_i^t = \frac{1}{N} \sum_{j=1}^{N} \mathbf{u}_j^t, \forall t = (a+1)I. \tag{A.17}$$

Also, the local models are aggregated then updated as

$$\mathbf{x}_i^t = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^t, \forall t = (a+1)I. \tag{A.18}$$

Now we provide the theoretical results for the new analysis of FedAvg with momentum.

Theorem A.5 (FedAvg with Momentum). *By* $\gamma \leq \min\{\frac{(1-\beta)^2}{L_g(1+\beta)}, \frac{1-\beta}{\sqrt{18(L_g^2+L_h^2)I}}\}$ *, for FedAvg with momentum in Algorithm 1, with Assumptions 3.2, 3.3, 4.1 and 4.2, we have*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{2(1-\beta)(f_0 - f_*)}{\gamma T} + \frac{\gamma L_g \sigma^2}{N(1-\beta)^2} + \frac{3\gamma^2 L_h^2 I \sigma^2}{(1-\beta)^2} + \frac{9\gamma^2 L_h^2 I^2 \zeta^2}{(1-\beta)^2}. \tag{A.19}$$

The proof can be found in Section B.15. For the ease of comparison, we present the convergence bound in Theorem 1 by Yu et al. (2019a) as follows.

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{2(1-\beta)(f_0 - f_*)}{\gamma T} + \frac{\gamma \tilde{L} \sigma^2}{N(1-\beta)^2} + \frac{3\gamma^2 \tilde{L}^2 I \sigma^2}{(1-\beta)^2} + \frac{9\gamma^2 \tilde{L}^2 I^2 \zeta^2}{(1-\beta)^2}. \tag{A.20}$$

It can be seen that the difference is that in Theorem A.5, $\tilde{L}$ is substituted by L_g and L_h. It has been shown in Proposition 5.1 that $L_g \leq \tilde{L}$ and $L_h \leq \tilde{L}$. Therefore, by applying L_h and L_g in the analysis for FedAvg with momentum, we obtain a tighter convergence upper bound. Similar to Theorem 4.3, the insights of L_h can also be applied to FedAvg with momentum. That is, when L_h is small, the error caused by local updates can still be small.

A.4. Applying L_h and L_g in the Analysis for FedAdam

In this section, we present the theoretical results for the convergence analysis of FedAdam (Reddi et al., 2020). We use $x_{i,j}$ to denote the jth element of the model from the ith worker, and $[\mathbf{g}_i(\mathbf{x}_i)]_j$ to denote the jth element of the gradient. In addition to Assumptions 4.1 and 4.2, we use the following assumption in Reddi et al. (2020) for the analysis. The algorithm can be found in Algorithm 2.

Algorithm 2 FedAdam in (Reddi et al., 2020)

Input: $\gamma, \eta, \bar{\mathbf{x}}^0, I, \tau, \beta_1, \beta_2$
Output: Global averaged model $\bar{\mathbf{x}}^R$
for $r = 0$ *to* $R - 1$ **do**
    for *Each worker i in parallel* **do**
        $\mathbf{x}_i^{r,0} \leftarrow \bar{\mathbf{x}}^r$;
        for $k = 0$ *to* $I - 1$ **do**
            Sample a gradient $\mathbf{g}_i(\mathbf{x}_i^{r,k})$;
            $\mathbf{x}_i^{r,k+1} \leftarrow \mathbf{x}_i^{r,k} - \gamma \mathbf{g}_i(\mathbf{x}_i^{r,k})$;
        end
        $\Delta_r^i \leftarrow \mathbf{x}_i^{r,0} - \mathbf{x}_i^{r,I}$;
    end
    $\Delta_r = \frac{1}{N} \sum_{i=1}^N \Delta_r^i$;
    $\mathbf{m}^r \leftarrow \beta_1 \mathbf{m}^{r-1} + (1 - \beta_1)\Delta_r$;
    $\mathbf{v}^r \leftarrow \beta_2 \mathbf{v}^{r-1} + (1 - \beta_2)\Delta_r^2$;
    $\bar{\mathbf{x}}^{r+1} \leftarrow \bar{\mathbf{x}}^r - \eta \frac{\mathbf{m}^r}{\sqrt{\mathbf{v}^r} + \tau}$;
end

Assumption A.6 (Bounded Gradients (Assumption 3 in Reddi et al. (2020))).

$$|[\mathbf{g}_i(\mathbf{x})]_j| \leq G, \forall \mathbf{x} \in \mathbb{R}^d, \forall i, j, \tag{A.21}$$

where G is a positive constant.

Now we present the theoretical results. The proof can be found in Appendix B.16.

Theorem A.7. *Assuming Assumptions 3.2,3.3,4.1, 4.2 and A.6 hold, for FedAdam, with*

$$\gamma \leq \min \left\{ \frac{1}{16 L_g I}, \frac{1}{\sqrt{6(L_h^2 + L_g^2)}I}, \frac{\tau^{\frac{1}{3}}}{16 K (120 L_g^2 G)^{\frac{1}{3}}}, \frac{\tau}{6(2G + \eta L_g)} \right\},$$

we have

$$
\begin{aligned}
\min_r \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^r)\|^2 \leq & \left(\sqrt{\beta_2} \gamma I G + \tau \right) \left(\frac{8(f_0 - f_*)}{\gamma \eta I R} + \frac{\gamma L_g \sigma^2}{\tau N} + \frac{96 \gamma^2 I^2 L_h^2 \zeta^2}{\tau} + \frac{32 \gamma^2 L_h I \sigma^2}{\tau} \right) \\
& + \left(\sqrt{\beta_2} \gamma I G + \tau \right) \left(\sqrt{1 - \beta_2} G + \frac{\eta L_g}{2} \right) \left(\frac{32 \gamma}{N \tau^2} \sigma^2 + \frac{768 \gamma^3 L_h^2 I^3 \zeta^2}{\tau^2} + \frac{256 \gamma^3 L_h^2 I^2 \sigma^2}{\tau^2} \right).
\end{aligned}
\tag{A.22}
$$

From Theorem A.7, it can be seen that after applying L_h and L_g in the analysis for FedAdam, the error caused by local updates such as $\mathcal{O}(\gamma^2 L_h^2 I^2 \zeta^2)$ and $\mathcal{O}(\gamma^2 L_h^2 I \sigma^2)$ are related to L_h. In Theorem 4.3, it has been shown that when L_h is small, the error caused by local updates can be small. We see here that the insights shown by L_h can also be applied to FedAdam.

A.5. Applying L_h and L_g in the Analysis for Strongly Convex Objective Functions

In this section, we provide the theoretical results for the convergence analysis of FedAvg with strongly convex objective functions. For strongly convex objective functions, we have the following assumption.

Assumption A.8. The local objective function $F_i(\mathbf{x})$ is μ-convex for $\mu > 0$ and satisfies

$$\langle \nabla F_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq - \left(F_i(\mathbf{x}) - F_i(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right), \forall i, \mathbf{x}, \mathbf{y}. \tag{A.23}$$

Assumption A.8 implies that the global objective function $f(\mathbf{x})$ is also μ-convex, since we have

$$
\begin{aligned}
&\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\
&= \frac{1}{N} \sum_{i=1}^{N} \langle \nabla F_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\
&\leq -\frac{1}{N} \sum_{i=1}^{N} \left(F_i(\mathbf{x}) - F_i(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right) \\
&= -\left(f(\mathbf{x}) - f(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right).
\end{aligned}
\tag{A.24}
$$

The theoretical results for strongly convex objective functions are as follows.

Theorem A.9 (μ-convex). *For μ-strongly convex objective functions, which satisfy Assumption A.8, with Assumptions 3.2, 3.3, 4.1, 4.2, $\gamma\eta \leq \min\{\frac{1}{16L_g I}, \frac{1}{4L_h I}\}$ and $\gamma \leq \min\{\frac{1}{24L_g I}\sqrt{\frac{\mu}{L_g}}, \frac{1}{\sqrt{6(L_h^2+L_g^2)I}}\}$, we have the following convergence upper bound.*

$$
\mathbb{E}[f(\bar{\mathbf{x}}^R)] - f^* = \mathcal{O}\left(\mu\|\bar{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \exp(-\mu\gamma\eta IR) + \frac{\gamma\eta\sigma^2}{N} + \frac{\gamma^2(L_g^2/N + L_h^2)I\sigma^2}{\mu} + \frac{\gamma^2 L_h^2 I^2 \zeta^2}{\mu} \right).
\tag{A.25}
$$

From this convergence upper bound, it can be seen that the error caused by local updates is $\mathcal{O}\left(\frac{\gamma^2(L_g^2/N+L_h^2)I\sigma^2}{\mu} + \frac{\gamma^2 L_h^2 I^2 \zeta^2}{\mu} \right)$, where the impact of data heterogeneity can be characterized by $L_h^2\zeta^2$. This characterization is the same as that shown in Theorem 4.3 in the main paper. Using Lemma 1 in Karimireddy et al. (2020), by carefully choosing the learning rates, we have the following corollary.

Corollary A.10. *By choosing $\gamma = \frac{1}{L_g IR}\sqrt{\frac{\mu}{L_g}}$ and $\gamma\eta = \min\{\frac{\log(\max(1,\mu^2 RIN\|\bar{\mathbf{x}}^0-\mathbf{x}^*\|^2/\sigma^2))}{\mu RI}, \frac{1}{\max\{16L_g, L_h\}I}\}$, we have*

$$
\mathbb{E}[f(\bar{\mathbf{x}}^R)] - f^* = \tilde{\mathcal{O}}\left(\mu\|\bar{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \exp\left(-\frac{\mu R}{\max\{16L_g, L_h\}} \right) + \frac{\sigma^2}{\mu NIR} + \frac{(L_g^2/N + L_h^2)\sigma^2}{L_g^3 IR} + \frac{L_h^2\zeta^2}{L_g^3 R^2} \right),
\tag{A.26}
$$

where $\tilde{\mathcal{O}}(\cdot)$ means $\mathcal{O}(\cdot)$ ignoring logarithmic terms.

It can be seen that the dominant term $\mathcal{O}\left(\frac{\sigma^2}{\mu NIR} \right)$ is the same as that in Karimireddy et al. (2020) and the insight of the error caused by local updates still holds.

B. Proofs

The description of FedAvg with two-sided learning rates can be found in Algorithm 3. For full participation, we have $\mathcal{S}_r = \{1, 2, \ldots, N\}, \forall r$ and $M = N$. For partial participation, we have $M < N$.

B.1. Technical Novelty

Before proceeding to the proof of our theoretical results, we summarize the technical novelty as follows.

(1) We need to develop new techniques to incorporate Assumptions 4.1 and 4.2. In the proof of Theorem 4.3 shown in Section B.7, we need to characterize the difference between local gradients. In the literature, this is done by applying the local Lipschitz constant as shown in Assumption 3.1 in the main paper. In our paper, since Assumption 3.1 is replaced by our newly introduced Assumption 4.2, the proof techniques in the literature cannot be applied. It requires to develop new proof techniques to use Assumption 4.2 as shown in the proof of Lemma B.1-B.3. For example, in Lemma B.1, due to the application of Assumptions 4.1 and 4.2, we have to cope with a new term, the local gradient deviation $\|\frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i) - \nabla F_j(\mathbf{x}_j)\|^2$, which cannot be computed using existing techniques. Another example is that in the proof of Theorem 4.5. Due to that we only use the global Lipschitz gradient assumption, we have to derive a new method to bound and incorporate the sampling related term, which can be seen from (B.47) to (B.59) in Section B.8.

Algorithm 3 FedAvg with two-sided learning rates

Input: $\gamma, \eta, \bar{\mathbf{x}}^0, I$
Output: Global averaged model $\bar{\mathbf{x}}^R$
for $r = 0$ *to* $R - 1$ **do**
    Sample a subset of workers $\mathcal{S}_r$, $|\mathcal{S}_r| = M$;
    Distribute the current global model $\bar{\mathbf{x}}^r$ to workers in $\mathcal{S}_r$;
    for *Each worker i in $\mathcal{S}_r$, in parallel* **do**
        /* Local Update Phase */
        $k = 0$;
        while $k < I$ **do**
            Sample the stochastic gradient $\mathbf{g}_i(\mathbf{x}_i^{r,k})$;
            Update the local model
            $\mathbf{x}_i^{r,k+1} \leftarrow \mathbf{x}_i^{r,k} - \gamma \mathbf{g}_i(\mathbf{x}_i^{r,k})$;
            $k \leftarrow k + 1$;
        end
        Send $\Delta_i^r \leftarrow \bar{\mathbf{x}}^r - \mathbf{x}_i^{r,I}$ to the server;
    end
    /* Global Update Phase */
    Update the global model
    $\bar{\mathbf{x}}^{r+1} \leftarrow \bar{\mathbf{x}}^r - \eta \cdot \frac{1}{M} \sum_{i \in \mathcal{S}_r} \Delta_i^r$;
end

---

(2) In addition to using Assumption 4.2 to characterize the new convergence rate of FedAvg, we also validate this assumption from the theoretical perspective. We develop the proof for Proposition 5.1-5.2 in Sections B.9 and B.10.

(3) Another novelty of our techniques is that in Theorem 5.5, since we use iteration-by-iteration analysis in the proof, the learning rate is not a function of I which is in contrast to that in the literature. For example, it can be seen that in the literature, such as Theorem IV in Karimireddy et al. (2020), for quadratic objective functions, the learning rate is upper bounded by $\frac{1}{I}$. The advantage of that γ is not a function of I can be explained as follows. In Theorem 5.5, in order to obtain the optimal learning rate, we choose $\gamma = \frac{1}{\sqrt{RI}}$. This requires that $\frac{1}{\sqrt{RI}} \leq \frac{1}{L_g}$, which means that I can be as large as possible. However, if $\gamma \leq \frac{1}{IL_g}$ as in Karimireddy et al. (2020), we will have $\frac{1}{\sqrt{RI}} \leq \frac{1}{IL_g}$ such that $I \leq \frac{R}{L_g^2}$, which means that to achieve the convergence rate of $\mathcal{O}(\frac{1}{\sqrt{RI}})$, I cannot be arbitrarily large. Therefore, the range of the learning rate in Theorem 5.5 can significantly improves the convergence rate.

B.2. Additional Lemmas

In the proof, we use $\mathbf{x}_i$ to denote the local model of worker i regardless of the number of iterations, and use $\bar{\mathbf{x}} := \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$ to denote the averaged model. Following lemmas are useful in the proof for main theorems.

Lemma B.1 (Local Gradient Deviation). *With Assumption 3.3, 4.1 and 4.2, we have*

$$\frac{1}{N} \sum_{j=1}^N \left\| \frac{1}{N} \sum_{i=1}^N \nabla F_i(\mathbf{x}_i) - \nabla F_j(\mathbf{x}_j) \right\|^2 \leq 3(L_h^2 + L_g^2) \cdot \frac{1}{N} \sum_{j=1}^N \|\bar{\mathbf{x}} - \mathbf{x}_j\|^2 + 3\zeta^2. \tag{B.1}$$

Lemma B.2 (Model Divergence). *With $\gamma \leq \frac{1}{\sqrt{6(L_h^2 + L_g^2)I}}$, we have*

$$\sum_{k=0}^{I-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 \leq 12(I-1)^3 \gamma^2 \zeta^2 + 4(I-1)^2 \gamma^2 \sigma^2, \tag{B.2}$$

where $\hat{\mathbf{x}}^{r,k} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{r,k}$.

Lemma B.3 (The Change of Averaged Models). *With $\gamma \leq \frac{1}{2\sqrt{3}IL_g}$, at the rth round, we have*

$$\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2 \leq 5(I-1) \cdot \frac{\gamma^2\sigma^2}{N} + 30I\gamma^2 \sum_{k=0}^{I-1} \frac{L_h^2}{N} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2$$

$$+ 30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2. \tag{B.3}$$

Lemma B.4 (Model Divergence for FedAvg with Momentum). *With $1 - \frac{6\gamma^2I^2(L_h^2+L_g^2)}{(1-\beta)^2} > 0$, for FedAvg with momentum in Yu et al. (2019a), we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 \leq \frac{1}{1 - \frac{6\gamma^2I^2(L_h^2+L_g^2)}{(1-\beta)^2}} \cdot \left(\frac{2\gamma^2I\sigma^2}{(1-\beta^2)} + \frac{6\gamma^2I^2\zeta^2}{(1-\beta)^2}\right). \tag{B.4}$$

B.3. Proof of Lemma B.1

We start with the LHS of the inequality in Lemma B.1.

$$\frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i) - \nabla F_j(\mathbf{x}_j)\right\|^2$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}_j) + \nabla f(\mathbf{x}_j) - \nabla F_j(\mathbf{x}_j)\right\|^2$$

$$\leq 3\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}})\right\|^2 + 3 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}_j)\right\|^2 + 3 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f(\mathbf{x}_j) - \nabla F_j(\mathbf{x}_j)\right\|^2$$

$$\overset{(a)}{\leq} 3\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}})\right\|^2 + 3L_g^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}} - \mathbf{x}_j\right\|^2 + 3\zeta^2$$

$$\overset{(b)}{\leq} 3 \cdot \frac{L_h^2}{N}\sum_{i=1}^{N}\left\|\bar{\mathbf{x}} - \mathbf{x}_i\right\|^2 + 3L_g^2 \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}} - \mathbf{x}_j\right\|^2 + 3\zeta^2$$

$$= 3(L_h^2 + L_g^2) \cdot \frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}} - \mathbf{x}_j\right\|^2 + 3\zeta^2, \tag{B.5}$$

where (a) is due to Assumptions 3.3 and 4.1 and (b) is due to Assumption 4.2.

B.4. Proof of Lemma B.2

At the rth round of FedAvg, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2$$

$$= \frac{\gamma^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j(\mathbf{x}_j^{r,m})\right)\right\|^2$$

$$= \frac{\gamma^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m}) - \nabla F_i(\mathbf{x}_i^{r,m}) + \nabla F_i(\mathbf{x}_i^{r,m})\right.\right.$$

$$\left.\left. - \frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m}) + \frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m}) - \frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j(\mathbf{x}_j^{r,m})\right)\right\|^2$$

$$\leq 2 \cdot \frac{\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{m=0}^{k-1} \left(\nabla F_i(\mathbf{x}_i^{r,m}) - \frac{1}{N} \sum_{j=1}^{N} \nabla F_j(\mathbf{x}_j^{r,m}) \right) \right\|^2$$

$$+ 2 \cdot \frac{\gamma^2}{N} \sum_{i=1}^{N} \left\| \sum_{m=0}^{k-1} \left(\mathbf{g}_i(\mathbf{x}_i^{r,m}) - \nabla F_i(\mathbf{x}_i^{r,m}) + \frac{1}{N} \sum_{j=1}^{N} \nabla F_j(\mathbf{x}_j^{r,m}) - \frac{1}{N} \sum_{j=1}^{N} \mathbf{g}_j(\mathbf{x}_j^{r,m}) \right) \right\|^2$$

$$\overset{(a)}{\leq} 2 \cdot \frac{\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{m=0}^{k-1} \left(\nabla F_i(\mathbf{x}_i^{r,m}) - \frac{1}{N} \sum_{j=1}^{N} \nabla F_j(\mathbf{x}_j^{r,m}) \right) \right\|^2$$

$$+ 2 \cdot \frac{\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{m=0}^{k-1} \left(\mathbf{g}_i(\mathbf{x}_i^{r,m}) - \nabla F_i(\mathbf{x}_i^{r,m}) \right) \right\|^2$$

$$\leq 2 \cdot \frac{\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{m=0}^{k-1} \left(\nabla F_i(\mathbf{x}_i^{r,m}) - \frac{1}{N} \sum_{j=1}^{N} \nabla F_j(\mathbf{x}_j^{r,m}) \right) \right\|^2 + 2\gamma^2 k \sigma^2$$

$$\leq 2k \cdot \frac{\gamma^2}{N} \cdot \sum_{i=1}^{N} \sum_{m=0}^{k-1} \mathbb{E} \left\| \nabla F_i(\mathbf{x}_i^{r,m}) - \frac{1}{N} \sum_{j=1}^{N} \nabla F_j(\mathbf{x}_j^{r,m}) \right\|^2 + 2\gamma^2 k \sigma^2$$

$$\overset{(b)}{\leq} 2k\gamma^2 \sum_{m=0}^{k-1} \left(3(L_h^2 + L_g^2) \frac{1}{N} \sum_{k=1}^{N} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,m} - \mathbf{x}_k^{r,m} \right\|^2 + 3\zeta^2 \right) + 2\gamma^2 k \sigma^2$$

$$= 6k\gamma^2(L_h^2 + L_g^2) \sum_{m=0}^{k-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,m} - \mathbf{x}_i^{r,m} \right\|^2 + 6k^2\gamma^2\zeta^2 + 2\gamma^2 k \sigma^2, \tag{B.6}$$

where (a) is due to $\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i\|^2 - \|\bar{\mathbf{y}}\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i\|^2$, where $\mathbf{y}_i \in \mathbb{R}^d, \forall i$ and $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i$, and we let $\mathbf{y}_i = \sum_{m=0}^{k-1} [\mathbf{g}_i(\mathbf{x}_i^{r,m}) - \nabla F_i(\mathbf{x}_i^{r,m})]$, and (b) is due to Lemma B.1.

Note that when $k = I$, we have $\mathbf{x}_i^{r,k} = \mathbf{x}_i^{r+1,0} = \bar{\mathbf{x}}^{r+1}$, and when $k = 0$, we have $\mathbf{x}_i^{r,k} = \bar{\mathbf{x}}^r$. So we have $\left\| \mathbf{x}_i^{r,I} - \hat{\mathbf{x}}^{r,I} \right\|^2 = 0$, for $k = 0, I$. Then sum over k for one round on both sides, we have

$$\sum_{k=0}^{I-1} \frac{L_h^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$$

$$= \sum_{k=1}^{I-1} \frac{L_h^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$$

$$\leq \sum_{k=1}^{I-1} \left(6k\gamma^2(L_h^2 + L_g^2) \sum_{m=0}^{k-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,m} - \mathbf{x}_i^{r,m} \right\|^2 + 6k^2\gamma^2\zeta^2 + 2\gamma^2 k \sigma^2 \right)$$

$$\overset{(a)}{\leq} 3\gamma^2(L_h^2 + L_g^2)I(I-1) \sum_{m=0}^{I-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,m} - \hat{\mathbf{x}}^{r,m} \right\|^2$$

$$+ 6(I-1)^3\gamma^2\zeta^2 + 2(I-1)^2\gamma^2\sigma^2, \tag{B.7}$$

where (a) is due to that $k \leq I$ and $\sum_{k=1}^{I-1} \sum_{m=0}^{k-1} D_m \leq \frac{I(I-1)}{2} \sum_{m=0}^{I-1} D_m$ and we let $D_m = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,m} - \hat{\mathbf{x}}^{r,m} \right\|^2$.

Moving the first term on RHS of (B.7) to LHS, we have

$$\left(1 - 3\gamma^2(L_h^2 + L_g^2)I(I-1) \right) \sum_{k=0}^{I-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$$

$$\leq 6(I-1)^3\gamma^2\zeta^2 + 2(I-1)^2\gamma^2\sigma^2. \tag{B.8}$$

With $\gamma < \frac{1}{\sqrt{3(L_h^2 + L_g^2)I}}$, we have

$$1 - 3\gamma^2(L_h^2 + L_g^2)I(I-1) > 0. \tag{B.9}$$

Then we have

$$\sum_{k=0}^{I-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 \leq \frac{1}{1 - 3\gamma^2(L_h^2 + L_g^2)I(I-1)} \cdot \left(6(I-1)^3\gamma^2\zeta^2 + 2(I-1)^2\gamma^2\sigma^2 \right). \tag{B.10}$$

With $\gamma \leq \frac{1}{\sqrt{6(L_h^2 + L_g^2)I}}$, we have $\frac{1}{1-3\gamma^2(L_h^2+L_g^2)I(I-1)} \leq 2$. Then we obtain

$$\sum_{k=0}^{I-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 \leq 12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2. \tag{B.11}$$

B.5. Proof of Lemma B.3

At rth round, for $k = 0$, we have

$$\mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2 = 0. \tag{B.12}$$

At rth round, for $1 \leq k \leq I - 1$, we have

$$\mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2$$

$$= \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \frac{\gamma}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \bar{\mathbf{x}}^r \right\|^2$$

$$= \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) - \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k-1}) + \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k-1}) \right. \right.$$

$$\left. \left. - \nabla f(\hat{\mathbf{x}}^{r,k-1}) + \nabla f(\hat{\mathbf{x}}^{r,k-1}) - \nabla f(\bar{\mathbf{x}}^r) + \nabla f(\bar{\mathbf{x}}^r) \right) \right\|^2$$

$$\leq \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k-1}) - \nabla f(\hat{\mathbf{x}}^{r,k-1}) + \nabla f(\hat{\mathbf{x}}^{r,k-1}) - \nabla f(\bar{\mathbf{x}}^r) + \nabla f(\bar{\mathbf{x}}^r) \right) \right\|^2$$

$$+ \frac{\gamma^2\sigma^2}{N}$$

$$\stackrel{(a)}{\leq} \left(1 + \frac{1}{2I-1} \right) \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r \right\|^2 + \frac{\gamma^2\sigma^2}{N}$$

$$+ \gamma^2(1 + 2I - 1) \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k-1}) - \nabla f(\hat{\mathbf{x}}^{r,k-1}) + \nabla f(\hat{\mathbf{x}}^{r,k-1}) - \nabla f(\bar{\mathbf{x}}^r) + \nabla f(\bar{\mathbf{x}}^r) \right\|^2$$

$$\leq \left(1 + \frac{1}{2I-1} \right) \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r \right\|^2 + \frac{\gamma^2\sigma^2}{N} + 6I\gamma^2 \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k-1}) - \nabla f(\hat{\mathbf{x}}^{r,k-1}) \right\|^2$$

$$+ 6I\gamma^2 \mathbb{E} \left\| \nabla f(\hat{\mathbf{x}}^{r,k-1}) - \nabla f(\bar{\mathbf{x}}^r) \right\|^2 + 6I\gamma^2 \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2$$

$$\tag{B.13}$$

$$\stackrel{(b)}{\leq} \left(1 + \frac{1}{2I-1} + 6I\gamma^2 L_g^2 \right) \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r \right\|^2 + \frac{\gamma^2\sigma^2}{N} + \frac{6I\gamma^2 L_h^2}{N} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k-1} - \hat{\mathbf{x}}^{r,k-1} \right\|^2$$

$$+ 6I\gamma^2 \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2$$

21

$$(B.14)$$

$$\overset{(c)}{\leq} \left(1 + \frac{1}{I-1}\right) \mathbb{E}\left\|\hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r\right\|^2 + \frac{\gamma^2\sigma^2}{N} + \frac{6I\gamma^2 L_h^2}{N}\sum_{i=1}^N \mathbb{E}\left\|\mathbf{x}_i^{r,k-1} - \hat{\mathbf{x}}^{r,k-1}\right\|^2 + 6I\gamma^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2$$

$$\overset{(d)}{\leq} 5(I-1)\cdot\frac{\gamma^2\sigma^2}{N} + 30I\gamma^2\sum_{k=0}^{I-1}\frac{L_h^2}{N}\sum_{i=1}^N\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2, \qquad (B.15)$$

where (a) is due to that $\|\mathbf{x} + \mathbf{y}\|^2 \leq (1+p)\|\mathbf{x}\|^2 + (1+\frac{1}{p})\|\mathbf{y}\|^2, \forall p > 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

(b) is due to Assumptions 4.1 and 4.2, (c) is due to that by choosing $\gamma \leq \frac{1}{2\sqrt{3}L_g I}$, we have

$$1 + \frac{1}{2I-1} + 6I\gamma^2 L_g^2 \leq 1 + \frac{1}{2(I-1)} + \frac{1}{2I} \leq 1 + \frac{1}{I-1}, \qquad (B.16)$$

and (d) is due to $(1 + \frac{1}{q})^q < e, \forall q > 0$, where e is the natural exponent.

B.6. Proof of Lemma B.4

By Lemma 5 in Yu et al. (2019a), for FedAvg with momentum, we have

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2$$

$$\leq \frac{2\gamma^2 I\sigma^2}{(1-\beta^2)} + 2\gamma^2\cdot\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla F_i(\mathbf{x}_i^\tau) - \frac{1}{N}\sum_{j=1}^N\nabla F_j(\mathbf{x}_j^\tau)\right]\frac{1-\beta^{t-\tau}}{1-\beta}\right\|^2, \qquad (B.17)$$

where $t = aI + b, 1 \leq b \leq I$ and $t_0 = aI$.

Note that $t - t_0 \leq I$. For the second term in the RHS of (B.17), we have

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla F_i(\mathbf{x}_i^\tau) - \frac{1}{N}\sum_{j=1}^N\nabla F_j(\mathbf{x}_j^\tau)\right]\frac{1-\beta^{t-\tau}}{1-\beta}\right\|^2$$

$$\leq (t-t_0)\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\left[\nabla F_i(\mathbf{x}_i^\tau) - \frac{1}{N}\sum_{j=1}^N\nabla F_j(\mathbf{x}_j^\tau)\right]\frac{1-\beta^{t-\tau}}{1-\beta}\right\|^2$$

$$\leq (t-t_0)\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\nabla F_i(\mathbf{x}_i^\tau) - \frac{1}{N}\sum_{j=1}^N\nabla F_j(\mathbf{x}_j^\tau)\right\|^2\left(\frac{1-\beta^{t-\tau}}{1-\beta}\right)^2$$

$$\leq \frac{I}{(1-\beta)^2}\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\nabla F_i(\mathbf{x}_i^\tau) - \frac{1}{N}\sum_{j=1}^N\nabla F_j(\mathbf{x}_j^\tau)\right\|^2$$

$$\overset{(a)}{\leq} \frac{I}{(1-\beta)^2}\sum_{\tau=t_0}^{t-1}\left(3(L_h^2 + L_g^2)\cdot\frac{1}{N}\sum_{j=1}^N\|\bar{\mathbf{x}} - \mathbf{x}_j\|^2 + 3\zeta^2\right)$$

$$\leq \frac{3I(L_h^2 + L_g^2)}{(1-\beta)^2}\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\|\mathbf{x}_i^\tau - \hat{\mathbf{x}}^\tau\|^2 + \frac{3I^2\zeta^2}{(1-\beta)^2}, \qquad (B.18)$$

where (a) is due to Lemma B.1. Substituting back to (B.17), we obtain

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 \leq \frac{2\gamma^2 I\sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I(L_h^2 + L_g^2)}{(1-\beta)^2}\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\|\mathbf{x}_i^\tau - \hat{\mathbf{x}}^\tau\|^2 + \frac{6\gamma^2 I^2\zeta^2}{(1-\beta)^2}. \qquad (B.19)$$

Taking the average over t on both sides, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 \leq \frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I (L_h^2+L_g^2)}{(1-\beta)^2}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=t_0}^{t-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^\tau - \hat{\mathbf{x}}^\tau\right\|^2 + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2}$$

$$\leq \frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I^2 (L_h^2+L_g^2)}{(1-\beta)^2}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^\tau - \hat{\mathbf{x}}^\tau\right\|^2 + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2} \qquad \text{(B.20)}$$

Rearranging the above inequality, with $1 - \frac{6\gamma^2 I^2 (L_h^2+L_g^2)}{(1-\beta)^2} > 0$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 \leq \frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2+L_g^2)}{(1-\beta)^2}} \cdot \left(\frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2}\right). \qquad \text{(B.21)}$$

B.7. Proof of Theorem 4.3

With Assumption 4.1, we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\rangle + \frac{\gamma^2\eta^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2$$

$$= \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{E}_{\mathbf{x}_i^{r,k}}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k})\right]\right\rangle + \frac{\gamma^2\eta^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2$$

$$= \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle + \frac{\gamma^2\eta^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2. \qquad \text{(B.22)}$$

The second term in the RHS of (B.22) can be computed as follows.

$$- \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle$$

$$= -\frac{\gamma\eta}{I}\mathbb{E}\left\langle I\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle$$

$$= \frac{\gamma\eta}{2I}\left\{\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left(\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right)\right\|^2 - I^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\right\}$$

$$= \frac{\gamma\eta}{2I}\left\{\mathbb{E}\left\|\sum_{k=0}^{I-1}\left(\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k})\right) + \sum_{k=0}^{I-1}\left(\nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right)\right\|^2\right.$$

$$\left. - I^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\right\}$$

$$\leq \frac{\gamma\eta}{2I}\left\{2I\sum_{k=0}^{I-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k})\right\|^2 + 2I\sum_{k=0}^{I-1}\mathbb{E}\|\nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\|^2\right.$$

$$\left. - I^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\right\}$$

$$\overset{(a)}{\leq} \frac{\gamma\eta}{2I}\left\{\frac{2IL_h^2}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 2IL_g^2\sum_{k=0}^{I-1}\mathbb{E}\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\|^2\right.$$

$$-I^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\Bigg\},\tag{B.23}$$

where (a) is due to Assumption 4.1 and Assumption 4.2.

The third term in the RHS of (B.22) can be computed as follows.

$$\frac{\gamma^2\eta^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2$$

$$=\frac{\gamma^2\eta^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k}) + \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2$$

$$\leq \gamma^2\eta^2 L_g\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 + \gamma^2\eta^2 L_g\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2$$

$$\overset{(a)}{\leq} \gamma^2\eta^2 L_g\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 + \frac{\gamma^2\eta^2 I L_g\sigma^2}{N}.\tag{B.24}$$

Now we explain (a) in (B.24). We have

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{k=0}^{I-1}\sum_{k'=0}^{I-1}\mathbb{E}\left\langle\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k}), \mathbf{g}_{i'}(\mathbf{x}_{i'}^{r,k'}) - \nabla F_{i'}(\mathbf{x}_{i'}^{r,k'})\right\rangle.\tag{B.25}$$

When $i \neq i'$, we have

$$\mathbb{E}\left\langle\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k}), \mathbf{g}_{i'}(\mathbf{x}_{i'}^{r,k'}) - \nabla F_{i'}(\mathbf{x}_{i'}^{r,k'})\right\rangle$$

$$=\mathbb{E}\left(\mathbb{E}\left[\left\langle\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k}), \mathbf{g}_{i'}(\mathbf{x}_{i'}^{r,k'}) - \nabla F_{i'}(\mathbf{x}_{i'}^{r,k'})\right\rangle|\mathbf{x}_i^{r,k}, \mathbf{x}_{i'}^{r,k'}\right]\right)$$

$$=0.\tag{B.26}$$

When $i = i'$ but $k \neq k'$, suppose that $k \leq k'$,

$$\mathbb{E}\left\langle\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k}), \mathbf{g}_i(\mathbf{x}_i^{r,k'}) - \nabla F_i(\mathbf{x}_i^{r,k'})\right\rangle$$

$$\mathbb{E}\left(\mathbb{E}\left[\left\langle\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k}), \mathbf{g}_i(\mathbf{x}_i^{r,k'}) - \nabla F_i(\mathbf{x}_i^{r,k'})\right\rangle|\mathbf{x}_i^{r,0}, \mathbf{x}_i^{r,1}, \mathbf{x}_i^{r,2}, \dots, \mathbf{x}_i^{r,k'}\right]\right)$$

$$=0.\tag{B.27}$$

Therefore, we have

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2$$

$$=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{E}\left(\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k})\right\|^2|\mathbf{x}_i^{r,k}\right]\right)$$

$$\leq \frac{I\sigma^2}{N}.\tag{B.28}$$

Substituting (B.23) and (B.24) to (B.22), we have

$$\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})]$$

$$\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta L_h^2}{N} \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + \gamma\eta L_g^2 \sum_{k=0}^{I-1} \mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2$$

$$- \frac{\gamma\eta I}{2} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \gamma\eta \left(\frac{1}{2I} - \gamma\eta L_g\right) \mathbb{E}\left\|\frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N}$$

$$\overset{(a)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta L_h^2}{N} \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + \gamma\eta L_g^2 \sum_{k=0}^{I-1} \mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2$$

$$- \frac{\gamma\eta I}{2} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N}$$

$$\overset{(b)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta L_h^2}{N} \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 - \frac{\gamma\eta I}{2} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N}$$

$$+ \gamma\eta L_g^2 I \left(5(I-1)\frac{\gamma^2\sigma^2}{N} + 30I\gamma^2 \sum_{k=0}^{I-1} \frac{L_h^2}{N} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 30I(I-1)\gamma^2 \left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\right)$$

$$\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \left(\frac{\gamma\eta I}{2} - 30\gamma^3\eta L_g^2 I^2(I-1)\right) \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N} + 5\gamma^3\eta L_g^2 I(I-1)\frac{\sigma^2}{N}$$

$$+ \left(\frac{\gamma\eta L_h^2}{N} + \frac{30\gamma^3\eta L_g^2 L_h^2 I^2}{N}\right) \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2. \tag{B.29}$$

where (a) is due to $\gamma\eta < \frac{1}{2IL_g}$, (b) is due to Lemma B.3. By $\gamma \leq \frac{1}{2\sqrt{30}L_g I}$, we have

$$\frac{\gamma\eta I}{2} - 30\gamma^3\eta L_g^2 I^2(I-1) \leq \frac{\gamma\eta I}{4}, \tag{B.30}$$

and

$$\frac{\gamma\eta L_h^2}{N} + \frac{30\gamma^3\eta L_g^2 L_h^2 I^2}{N} \leq \frac{3\gamma\eta L_h^2}{2N}. \tag{B.31}$$

Substituting back to (B.29), we obtain

$$\mathbb{E}[f(\bar{\mathbf{x}}^{r+1})]$$

$$\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \frac{\gamma\eta I}{4} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N} + 5\gamma^3\eta L_g^2 I(I-1)\frac{\sigma^2}{N} + \frac{3\gamma\eta L_h^2}{2N} \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2$$

$$\overset{(a)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \frac{\gamma\eta I}{4} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 I L_g \sigma^2}{N} + 5\gamma^3\eta L_g^2 I(I-1)\frac{\sigma^2}{N}$$

$$+ \frac{3\gamma\eta L_h^2}{2} \left[12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right]. \tag{B.32}$$

where (a) is due to using Lemma B.2. Moving $-\frac{\gamma\eta I}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2$ to left and taking the average over r, we obtain

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 \leq \frac{4(f(\bar{\mathbf{x}}^0) - f^*)}{\gamma\eta IR} + \frac{4\gamma\eta L_g\sigma^2}{N} + \frac{20\gamma^2 L_g^2(I-1)\sigma^2}{N} + 24\gamma^2 L_h^2(I-1)\sigma^2 + 72\gamma^2 L_h^2(I-1)^2\zeta^2. \tag{B.33}$$

25

Then we have

$$
\begin{aligned}
\min_{r \in [R]} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2 &\leq \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2 \\
&= \mathcal{O} \left(\frac{f(\mathbf{x}^0) - f^*}{\gamma \eta I R} + \frac{\gamma \eta L_g \sigma^2}{N} + \gamma^2 \left(\frac{L_g^2}{N} + L_h^2 \right) (I-1) \sigma^2 + \gamma^2 L_h^2 (I-1)^2 \zeta^2 \right).
\end{aligned}
\tag{B.34}
$$

B.8. Proof of Theorem 4.5

In this section, we define an identity random variable to indicate the participation of workers in the following. In each round, the server performs M times of sampling. Then $\forall j \in [M], i \in [N]$, we have

$$
\mathbb{1}_{j,i}^r = \begin{cases} 1, \text{worker } i \text{ is chosen at } j\text{th sampling of } r\text{th round}, \\ 0, \text{else}, \end{cases}
\tag{B.35}
$$

where $\sum_{i=1}^N \mathbb{1}_{j,i}^r = 1$. Since we consider uniform sampling with replacement, we have

$$
p(\mathbb{1}_{j,i}^r = 1) = \frac{1}{N}, \forall i, j, r,
\tag{B.36}
$$

and

$$
\mathbb{E}_{\mathcal{S}_r} [\mathbb{1}_{j,i}^r] = \frac{1}{N},
\tag{B.37}
$$

where $\mathbb{E}_{\mathcal{S}_r}[\cdot]$ means taking the expectation over sampling at rth round. In addition, we have

$$
\mathbb{E}_{\mathcal{S}_r} \left\| \mathbb{1}_{j,i}^r \mathbf{z} \right\|^2 = \mathbb{E}_{\mathcal{S}_r} \mathbb{1}_{j,i}^r \left\| \mathbf{z} \right\|^2 = \frac{1}{N} \left\| \mathbf{z} \right\|^2,
\tag{B.38}
$$

for any $\mathbf{z}$ that is independent of $\mathbb{1}_{j,i}^r$, where the first equality is because $(\mathbb{1}_{j,i}^r)^2 = \mathbb{1}_{j,i}^r$; and for $i \neq i'$,

$$
\mathbb{E}_{\mathcal{S}_r} \left[\mathbb{1}_{j,i}^r \mathbb{1}_{j,i'}^r \right] = 0.
\tag{B.39}
$$

Here we assume that the sampling workers and sampling gradients are independent.

With Assumption 4.1, after one round of FedAvg, we have

$$
\mathbb{E} \left[f(\bar{\mathbf{x}}^{r+1}) \right] \leq \mathbb{E} \left[f(\bar{\mathbf{x}}^r) \right] - \gamma \eta \mathbb{E} \left\langle \nabla f(\bar{\mathbf{x}}^r), \frac{1}{M} \sum_{j=1}^M \sum_{i=1}^N \sum_{k=0}^{I-1} \mathbb{1}_{j,i}^r \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\rangle + \frac{\gamma^2 \eta^2 L_g}{2} \mathbb{E} \left\| \frac{1}{M} \sum_{j=1}^M \sum_{i=1}^N \sum_{k=0}^{I-1} \mathbb{1}_{j,i}^r \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\|^2.
\tag{B.40}
$$

It can be seen that the inner-product term is the same as that in (B.23). So we have

$$
\begin{aligned}
&- \gamma \eta \mathbb{E} \left\langle \nabla f(\bar{\mathbf{x}}^r), \frac{1}{M} \sum_{j=1}^M \sum_{i=1}^N \sum_{k=0}^{I-1} \mathbb{1}_{j,i}^r \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\rangle \\
&= -\gamma \eta \mathbb{E} \left\langle \nabla f(\bar{\mathbf{x}}^r), \frac{1}{M} \sum_{j=1}^M \sum_{i=1}^N \sum_{k=0}^{I-1} \mathbb{E}_{\mathcal{S}^r} [\mathbb{1}_{j,i}^r] \mathbb{E}_{\mathbf{x}_i^{r,k}} [\mathbf{g}_i(\mathbf{x}_i^{r,k})] \right\rangle \\
&= -\gamma \eta \mathbb{E} \left\langle \nabla f(\bar{\mathbf{x}}^r), \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\rangle \\
&\leq \frac{\gamma \eta}{2I} \left\{ \frac{2I L_h^2}{N} \sum_{k=0}^{I-1} \sum_{i=1}^N \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 + 2I L_g^2 \sum_{k=0}^{I-1} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2 \right. \\
&\left. - I^2 \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2 - \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\|^2 \right\}.
\end{aligned}
\tag{B.41}
$$

26

In this case, we consider $\mathbf{x}_i^{r,k}, i \notin \mathcal{S}_r$ as a virtual local model on worker i, which is not computed in the system. The virtual local model is mainly used for analysis. Similar to (B.24), for the third term in the RHS of (B.40), we have

$$
\mathbb{E}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2
$$

$$
= \mathbb{E}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k})-\nabla F_i(\mathbf{x}_i^{r,k})+\nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2
$$

$$
\leq 2\mathbb{E}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k})-\nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2 + 2\mathbb{E}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2
$$

$$
\leq \frac{2I\sigma^2}{M} + 2\mathbb{E}\left[\mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\right]. \tag{B.42}
$$

Now we consider the expectation on sampling. Let $Q_i = \sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})$, then for the second term in the RHS of (B.42), we have

$$
\mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{1}_{j,i}^{r}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 = \mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i\right\|^2
$$

$$
= \mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i - \frac{1}{N}\sum_{i=1}^{N}Q_i + \frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2
$$

$$
\overset{(a)}{=} \mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i - \frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2 + \left\|\frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2, \tag{B.43}
$$

where (a) is due to

$$
\mathbb{E}_{\mathcal{S}^r}\left\langle\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i - \frac{1}{N}\sum_{i=1}^{N}Q_i, \frac{1}{N}\sum_{i=1}^{N}Q_i\right\rangle = 0. \tag{B.44}
$$

Further, we have

$$
\mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i - \frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2
$$

$$
= \frac{1}{M^2}\mathbb{E}_{\mathcal{S}_r}\left[\sum_{j=1}^{M}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(N\mathbb{1}_{j,i}^{r}-1\right)Q_i\right\|^2 + \sum_{i\neq j'}\left\langle\frac{1}{N}\sum_{i=1}^{N}\left(N\mathbb{1}_{j,i}^{r}-1\right)Q_i, \frac{1}{N}\sum_{i=1}^{N}\left(N\mathbb{1}_{j',i}^{r}-1\right)Q_i\right\rangle\right]
$$

$$
= \frac{1}{M^2}\sum_{j=1}^{M}\mathbb{E}_{\mathcal{S}_r}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(N\mathbb{1}_{j,i}^{r}-1\right)Q_i\right\|^2
$$

$$
= \frac{1}{M^2}\sum_{j=1}^{M}\mathbb{E}_{\mathcal{S}_r}\left\|\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i - \frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2
$$

$$
= \frac{1}{M^2}\sum_{j=1}^{M}\left[\mathbb{E}_{\mathcal{S}_r}\left\|\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i\right\|^2 - 2\mathbb{E}_{\mathcal{S}_r}\left\langle\sum_{i=1}^{N}\mathbb{1}_{j,i}^{r}Q_i, \frac{1}{N}\sum_{i=1}^{N}Q_i\right\rangle + \left\|\frac{1}{N}\sum_{i=1}^{N}Q_i\right\|^2\right]
$$

$$= \frac{1}{M^2} \sum_{j=1}^{M} \mathbb{E}_{\mathcal{S}_r} \left\| \sum_{i=1}^{N} \mathbb{1}_{j,i}^{r} Q_i \right\|^2 - \frac{1}{M} \left\| \frac{1}{N} \sum_{i=1}^{N} Q_i \right\|^2$$

$$= \frac{1}{M^2} \sum_{j=1}^{M} \left[\sum_{i=1}^{N} \mathbb{E}_{\mathcal{S}_r} \left\| \mathbb{1}_{j,i}^{r} \right\|^2 \|Q_i\|^2 + \sum_{i \neq i'} \mathbb{E}_{\mathcal{S}_r} \left\langle \mathbb{1}_{j,i}^{r} Q_i, \mathbb{1}_{j,i'}^{r} Q_{i'} \right\rangle \right] - \frac{1}{M} \left\| \frac{1}{N} \sum_{i=1}^{N} Q_i \right\|^2$$

$$= \frac{1}{MN} \sum_{i=1}^{N} \|Q_i\|^2 - \frac{1}{M} \left\| \frac{1}{N} \sum_{i=1}^{N} Q_i \right\|^2. \tag{B.45}$$

Substituting above results back to (B.42), we obtain

$$\mathbb{E} \left\| \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \mathbb{1}_{j,i}^{r} \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\|^2 \leq \frac{2I\sigma^2}{M} + \frac{2}{MN} \sum_{i=1}^{N} \mathbb{E} \left\| \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\|^2 + 2 \cdot \frac{M-1}{M} \mathbb{E} \left\| \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{I-1} \nabla F_j(\mathbf{x}_j^{r,k}) \right\|^2. \tag{B.46}$$

For the second term of (B.46), we have

$$\mathbb{E} \left\| \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\|^2$$

$$= \mathbb{E} \left\| \sum_{k=0}^{I-1} \left[\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\mathbf{x}_i^{r,k}) + \nabla f(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k}) + \nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r) + \nabla f(\bar{\mathbf{x}}^r) \right] \right\|^2$$

$$\overset{(a)}{\leq} 4I^2 \zeta^2 + 4L_g^2 I \sum_{k=0}^{I-1} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 + 4L_g^2 I \sum_{k=0}^{I-1} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2 + 4I^2 \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2, \tag{B.47}$$

where (a) is due to Assumption 3.3 and Assumption 4.1. Substituting back and rearranging, we have

$$\frac{\gamma^2 \eta^2 L_g}{2} \mathbb{E} \left\| \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \mathbb{1}_{j,i}^{r} \mathbf{g}_i(\mathbf{x}_i^{r,k}) \right\|^2 \leq \frac{\gamma^2 \eta^2 L_g I \sigma^2}{M} + \frac{\gamma^2 \eta^2 L_g (M-1)}{M} \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\|^2$$

$$+ \frac{4\gamma^2 \eta^2 L_g I^2 \zeta^2}{M} + \frac{4\gamma^2 \eta^2 L_g^3 I}{MN} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$$

$$+ \frac{4\gamma^2 \eta^2 L_g^3 I}{MN} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2 + \frac{4\gamma^2 \eta^2 L_g I^2}{M} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2. \tag{B.48}$$

Substituting (B.41) and (B.48) back to (B.40), we have

$$\mathbb{E} \left[f(\bar{\mathbf{x}}^{r+1}) \right] \leq \mathbb{E} \left[f(\bar{\mathbf{x}}^r) \right] - \left(\frac{\gamma \eta I}{2} - \frac{4\gamma^2 \eta^2 L_g I^2}{M} \right) \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}^r) \right\|^2$$

$$- \left(\frac{\gamma \eta}{2I} - \frac{\gamma^2 \eta^2 L_g (M-1)}{M} \right) \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{I-1} \nabla F_i(\mathbf{x}_i^{r,k}) \right\|^2$$

$$+ \frac{\gamma^2 \eta^2 L_g I \sigma^2}{M} + \frac{4\gamma^2 \eta^2 L_g I^2 \zeta^2}{M} + \left(\gamma \eta L_h^2 + \frac{4\gamma^2 \eta^2 L_g^3 I}{M} \right) \cdot \frac{1}{N} \sum_{k=0}^{I-1} \sum_{i=1}^{N} \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2$$

$$+ \left(\gamma \eta L_g^2 + \frac{4\gamma^2 \eta^2 L_g^3 I}{M} \right) \sum_{k=0}^{I-1} \mathbb{E} \left\| \hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r \right\|^2. \tag{B.49}$$

28

By Lemma B.3, we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right] - \left(\frac{\gamma\eta I}{2} - \frac{4\gamma^2\eta^2 L_g I^2}{M}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 + \frac{\gamma^2\eta^2 L_g I\sigma^2}{M} + \frac{4\gamma^2\eta^2 L_g I^2\zeta^2}{M}$$

$$+ \left(\gamma\eta L_h^2 + \frac{4\gamma^2\eta^2 L_g^3 I}{M}\right) \cdot \frac{1}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2$$

$$+ \left(\gamma\eta I L_g^2 + \frac{4\gamma^2\eta^2 L_g^3 I^2}{M}\right)$$

$$\cdot \left(5(I-1)\cdot\frac{\gamma^2\sigma^2}{N} + 30I\gamma^2\sum_{k=0}^{I-1}\frac{L_h^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2\right). \quad \text{(B.50)}$$

By $\gamma\eta \leq \frac{M}{16IL_g}$ and $\gamma \leq \frac{1}{10\sqrt{3}IL_g}$, we have

$$-\left(\frac{\gamma\eta I}{2} - \frac{4\gamma^2\eta^2 L_g I^2}{M}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 + \left(\gamma\eta I L_g^2 + \frac{4\gamma^2\eta^2 L_g^3 I^2}{M}\right)\cdot 30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2$$

$$\leq -\left(\frac{\gamma\eta I}{2} - \frac{\gamma\eta I}{4}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 + \left(\gamma\eta I + \frac{\gamma\eta I}{4}\right)\cdot\frac{1}{10}\cdot\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2$$

$$\leq -\frac{\gamma\eta I}{8}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2. \quad \text{(B.51)}$$

By $\gamma\eta \leq \frac{M}{4IL_g}$, we have

$$\left(\gamma\eta I L_g^2 + \frac{4\gamma^2\eta^2 L_g^3 I^2}{M}\right)\cdot 5(I-1)\cdot\frac{\gamma^2\sigma^2}{N} \leq \gamma\eta I\cdot\frac{10\gamma^2 L_g^2(I-1)\sigma^2}{N}. \quad \text{(B.52)}$$

Then we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right] - \frac{\gamma\eta I}{8}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 + \frac{\gamma^2\eta^2 L_g I\sigma^2}{M} + \frac{4\gamma^2\eta^2 L_g I^2\zeta^2}{M} + \gamma\eta I\cdot\frac{10\gamma^2 L_g^2(I-1)\sigma^2}{N}$$

$$+ \left(\gamma\eta L_h^2 + \frac{4\gamma^2\eta^2 L_g^3 I}{M} + 30\gamma^2 I L_h^2\left(\gamma\eta I L_g^2 + \frac{4\gamma^2\eta^2 L_g^3 I^2}{M}\right)\right)\cdot\frac{1}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2. \quad \text{(B.53)}$$

With Lemma B.2, we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right] - \frac{\gamma\eta I}{8}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 + \frac{\gamma^2\eta^2 L_g I\sigma^2}{M} + \frac{4\gamma^2\eta^2 L_g I^2\zeta^2}{M} + \gamma\eta I\cdot\frac{10\gamma^2 L_g^2(I-1)\sigma^2}{N}$$

$$+ \left(\gamma\eta L_h^2 + \frac{4\gamma^2\eta^2 L_g^3 I}{M} + 30\gamma^2 I L_h^2\left(\gamma\eta I L_g^2 + \frac{4\gamma^2\eta^2 L_g^3 I^2}{M}\right)\right)\cdot\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right). \quad \text{(B.54)}$$

Then we obtain

$$\min_{r\in[R]}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 \leq \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^2 \leq \frac{8(f^0 - f^*)}{\gamma\eta IR} + \frac{8\gamma\eta L_g\sigma^2}{M} + \frac{32\gamma\eta L_g I\zeta^2}{M} + \frac{80\gamma^2 L_g^2(I-1)\sigma^2}{N}$$

$$+ 8\left(L_h^2 + \frac{4\gamma\eta L_g^3 I}{M} + 30\gamma^2 I L_h^2\left(I L_g^2 + \frac{4\gamma\eta L_g^3 I^2}{M}\right)\right)\cdot\left(12(I-1)^2\gamma^2\zeta^2 + 4(I-1)\gamma^2\sigma^2\right). \quad \text{(B.55)}$$

29

By $\gamma\eta \leq \frac{M}{16IL_g}$ and $\gamma \leq \frac{1}{10\sqrt{3}IL_g}$, we have

$$\frac{32\gamma\eta L_g^3 I}{M} \cdot \left(12(I-1)^2\gamma^2\zeta^2 + 4(I-1)\gamma^2\sigma^2\right) \leq \frac{\gamma\eta L_g}{M} \cdot \left(6I\zeta^2 + 2\sigma^2\right), \tag{B.56}$$

and

$$30\gamma^2 I L_h^2 \left(I L_g^2 + \frac{4\gamma\eta L_g^3 I^2}{M}\right) \cdot \left(12(I-1)^2\gamma^2\zeta^2 + 4(I-1)\gamma^2\sigma^2\right)$$

$$\leq \left(\frac{1}{10} + \frac{1}{40}\right) \cdot L_h^2 \cdot \left(12(I-1)^2\gamma^2\zeta^2 + 4(I-1)\gamma^2\sigma^2\right)$$

$$\leq \gamma^2 L_h^2 (I-1)^2\zeta^2 + \gamma^2 L_h^2 (I-1)\sigma^2. \tag{B.57}$$

Finally, we obtain

$$\min_{r\in[R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 \leq \frac{8(f^0 - f^*)}{\gamma\eta IR} + \frac{10\gamma\eta L_g\sigma^2}{M} + \frac{38\gamma\eta L_g I\zeta^2}{M} + \frac{80\gamma^2 L_g^2(I-1)\sigma^2}{N}$$

$$+ 97\gamma^2 L_h^2(I-1)^2\zeta^2 + 33\gamma^2 L_h^2(I-1)\sigma^2. \tag{B.58}$$

Rearrange,

$$\min_{r\in[R]} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 = \mathcal{O}\left(\frac{(f^0 - f^*)}{\gamma\eta IR} + \frac{\gamma\eta L_g\sigma^2}{M} + \frac{\gamma\eta L_g I\zeta^2}{M} + \frac{\gamma^2 L_g^2(I-1)\sigma^2}{N} + \gamma^2 L_h^2(I-1)\sigma^2 + \gamma^2 L_h^2(I-1)^2\zeta^2\right). \tag{B.59}$$

B.9. Proof of Proposition 5.1

First, using $\nabla f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\mathbf{x})$, it is straightforward to show that Assumption 3.1 implies Assumption 4.1 holds by choosing $L_g = \tilde{L}$.

Second, we can see that

$$\left\|\frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i) - \nabla f(\bar{\mathbf{x}})\right\|^2 = \left\|\frac{1}{N}\sum_{i=1}^{N} [\nabla F_i(\mathbf{x}_i) - \nabla F_i(\bar{\mathbf{x}})]\right\|^2$$

$$\leq \frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(\mathbf{x}_i) - \nabla F_i(\bar{\mathbf{x}})\|^2$$

$$\stackrel{(a)}{\leq} \frac{\tilde{L}^2}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2, \tag{B.60}$$

where (a) is due to Assumption 3.1. By choosing $L_h = \tilde{L}$, Assumption 4.2 holds. □

B.10. Proof of Proposition 5.2

Recall that $\hat{\mathbf{x}}^{r,k}$ is the virtual averaged model defined in (14) in the main paper. During one local iteration, we have

$$\mathbb{E}[\hat{\mathbf{x}}^{r,k+1}|\hat{\mathbf{x}}^{r,k}] = \hat{\mathbf{x}}^{r,k} - \gamma \cdot \frac{1}{N}\sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k}). \tag{B.61}$$

Using (15), if we use centralized update at this iteration, we have

$$\mathbb{E}[\mathbf{x}_c^{r,k+1}|\hat{\mathbf{x}}^{r,k}] = \hat{\mathbf{x}}^{r,k} - \gamma\nabla f(\hat{\mathbf{x}}^{r,k}). \tag{B.62}$$

Using Assumption 4.2, we obtain

$$
\left\| \mathbb{E}[\hat{\mathbf{x}}^{r,k+1}|\hat{\mathbf{x}}^{r,k}] - \mathbb{E}[\mathbf{x}_c^{r,k+1}|\hat{\mathbf{x}}^{r,k}] \right\|^2 = \gamma^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f\left(\hat{\mathbf{x}}^{r,k}\right) \right\|^2
$$

$$
\leq \gamma^2 \cdot \frac{L_h^2}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2. \tag{B.63}
$$

B.11. Proof of Proposition 5.3

For quadratic functions, we have

$$
\nabla F_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x} + \mathbf{b}_i, \mathbf{x} \in \mathbb{R}^d. \tag{B.64}
$$

Recall that $\mathbf{A} := \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i$ and $\mathbf{b} := \frac{1}{N} \sum_{i=1}^{N} \mathbf{b}_i$. We have

$$
\left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F_i(\mathbf{x}_i) - \nabla f\left(\bar{\mathbf{x}}\right) \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{A}_i \mathbf{x}_i + \mathbf{b}_i) - (\mathbf{A}\bar{\mathbf{x}} + \mathbf{b}) \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i - \mathbf{A}\bar{\mathbf{x}} \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i - 2\mathbf{A}\bar{\mathbf{x}} + \mathbf{A}\bar{\mathbf{x}} \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i - \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i \bar{\mathbf{x}} - \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}\mathbf{x}_i + \mathbf{A}\bar{\mathbf{x}} \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}_i(\mathbf{x}_i - \bar{\mathbf{x}}) - \frac{1}{N} \sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i - \bar{\mathbf{x}}) \right\|^2
$$

$$
= \left\| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{A}_i - \mathbf{A})(\mathbf{x}_i - \bar{\mathbf{x}}) \right\|^2
$$

$$
\leq \frac{1}{N} \sum_{i=1}^{N} \|(\mathbf{A}_i - \mathbf{A})(\mathbf{x}_i - \bar{\mathbf{x}})\|^2
$$

$$
\leq \frac{|\lambda_{\mathrm{diff}}|_{\max}^2}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2. \tag{B.65}
$$

For local Lipschitz gradient, we have

$$
\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\|
$$
$$
= \|\mathbf{A}_i \mathbf{x} - \mathbf{A}_i \mathbf{y}\|
$$
$$
\leq \|\mathbf{A}_i\| \|\mathbf{x} - \mathbf{y}\|. \tag{B.66}
$$

Since local Lipschitz gradient holds for each worker $i \in [N]$, we can choose $\tilde{L}$ as

$$
\tilde{L} = \max_i \|\mathbf{A}_i\|. \tag{B.67}
$$

B.12. Proof of Theorem 5.5

It can be observed that for quadratic objective functions when $\mathbf{A}_i = \mathbf{A}, \forall i$, we have $L_h = 0$ and $L_g = |\lambda(\mathbf{A})|$.

With Assumption 4.1, after one local iteration, we have

$$
\mathbb{E}\left[f(\hat{\mathbf{x}}^{t+1})\right] \leq \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - \gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle + \frac{\gamma^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2
$$

$$
= \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - \gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle + \frac{\gamma^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2. \tag{B.68}
$$

For the second term in the RHS of (B.68), we have

$$
- \gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle
$$

$$
= \frac{\gamma}{2}\left(\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t) - \nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2\right)
$$

$$
\leq \frac{\gamma}{2}\left(\frac{L_h^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2\right). \tag{B.69}
$$

For the third term of (B.68), we have

$$
\frac{\gamma^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2
$$

$$
= \frac{\gamma^2 L_g}{2}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}_i^t}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t) + \nabla F_i(\mathbf{x}_i^t)\right)\right\|^2\right]
$$

$$
\overset{(a)}{=} \frac{\gamma^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L_g}{2}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}_i^t}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t)\right)\right\|^2\right]
$$

$$
\leq \frac{\gamma^2 L_g}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N}, \tag{B.70}
$$

where $\mathbb{E}_{\mathbf{x}_i^t}[\cdot] = \mathbb{E}[\cdot|\mathbf{x}_i^t]$ and (a) is due to that $\mathbb{E}_{\mathbf{x}_i^t}\langle\nabla F_i(\mathbf{x}_i^t), \mathbf{g}_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t)\rangle = 0$.

Substitute (B.69) and (B.70) back to (B.68), we obtain

$$
\mathbb{E}\left[f(\hat{\mathbf{x}}^{t+1})\right]
$$

$$
\leq \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] + \frac{\gamma L_h^2}{2N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2
$$

$$
- \left(\frac{\gamma}{2} - \frac{\gamma^2 L_g}{2}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N}
$$

$$
\overset{(a)}{\leq} \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] + \frac{\gamma L_h^2}{2N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N}, \tag{B.71}
$$

where (a) is due to $\gamma < \frac{1}{L_g}$. Rearrange the above inequality with $L_h = 0$, we have

$$
\begin{aligned}
\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 &\leq \frac{2\mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - 2\mathbb{E}f(\hat{\mathbf{x}}^{t+1})}{\gamma} + \frac{L_h^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 + \frac{\gamma L_g \sigma^2}{N} \\
&= \frac{2\mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - 2\mathbb{E}f(\hat{\mathbf{x}}^{t+1})}{\gamma} + \frac{\gamma L_g \sigma^2}{N}.
\end{aligned}
\tag{B.72}
$$

Take the average over t on both sides, we obtain

$$
\min_{t\in[T]}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{2f(\hat{\mathbf{x}}^t) - 2f^*}{\gamma T} + \frac{\gamma L_g \sigma^2}{N}.
\tag{B.73}
$$

B.13. Proof of Corollary 5.6

In Corollary 5.6, for both local SGD and mini-batch SGD, we choose the learning rate as $\gamma = \frac{1}{L_g}$. The proof (order-wise, ignoring the constants) is as follows.

Let $h(\gamma)$ denote the order of the convergence upper bound in Theorem 5.5. For mini-batch SGD, we have

$$
h(\gamma) = \frac{\mathcal{F}}{\gamma R} + \frac{\gamma L_g \sigma^2}{NI}.
\tag{B.74}
$$

By minimizing $h(\gamma)$, we obtain $\gamma^* = \sqrt{\frac{\mathcal{F}NI}{RL_g\sigma^2}}$. Because $\sigma \leq \sqrt{\frac{\mathcal{F}NL_g}{RI}}$ as specified in Corollary 5.6, we have

$$
\gamma^* \geq \sqrt{\frac{\mathcal{F}NI}{RL_g}\cdot\frac{RI}{\mathcal{F}NL_g}} = \frac{I}{L_g}.
$$

Therefore, when $\gamma \in \left(0, \frac{I}{L_g}\right]$, $h(\gamma)$ is monotonically decreasing. Now, note that Theorem 5.5 requires $\gamma \in \left(0, \frac{1}{L_g}\right]$. When choosing $\gamma = \frac{1}{L_g} \leq \frac{I}{L_g}$ (since $I \geq 1$), $h(\gamma)$ is minimized under the condition of $\gamma \in \left(0, \frac{1}{L_g}\right]$ and we obtain

$$
h(\gamma) = h\left(\frac{1}{L_g}\right) = \frac{\mathcal{F}L_g}{R} + \frac{1}{L_g}\frac{L_g\sigma^2}{NI} \overset{(a)}{\leq} \frac{\mathcal{F}L_g}{R} + \frac{\mathcal{F}L_g}{RI^2},
$$

where (a) is due to $\sigma \leq \sqrt{\frac{\mathcal{F}NL_g}{RI}}$. The case of local SGD can be proven similarly.

B.14. Proof of Theorem A.2

First, we introduce a useful lemma, which is used in this section.

Lemma B.5. *With $x > 1$ and $k \in \mathbb{N}^+$, we have*

$$
\sum_{l=0}^{k-1}l^2 x^l \leq \frac{x^{k-1}}{(x-1)^3}\cdot k^2 x^2 \cdot x = \frac{k^2 x^{k+2}}{(x-1)^3}.
\tag{B.75}
$$

Proof. For the geometric series, we have

$$
\sum_{l=0}^{k-1}x^l = \frac{x^k - 1}{x - 1}.
\tag{B.76}
$$

When $k \geq 2$, taking the derivative over x on both sides, we obtain

$$
\sum_{l=0}^{k-1}lx^{l-1} = \frac{kx^{k-1}}{x-1} - \frac{x^k - 1}{(x-1)^2}.
\tag{B.77}
$$

Multiplying x on both sides, we obtain

$$\sum_{l=0}^{k-1} l x^l = \frac{k x^k}{x-1} - \frac{x^{k+1} - x}{(x-1)^2}. \tag{B.78}$$

Taking the derivative over x on both sides again, we obtain

$$\begin{aligned}
\sum_{l=0}^{k-1} l^2 x^{l-1} &= \frac{k^2 x^{k-1}}{x-1} - \frac{k x^k}{(x-1)^2} - \frac{(k+1)x^k - 1}{(x-1)^2} + \frac{2x^{k+1} - 2x}{(x-1)^3} \\
&= \frac{k^2 x^{k-1}(x-1)^2 - (2k+1)x^k(x-1) + (x-1) + 2x^{k+1} - 2x}{(x-1)^3} \\
&\leq \frac{k^2 x^{k-1}(x-1)^2 - (2k+1)x^k(x-1) + 2x^{k+1}}{(x-1)^3} \\
&= \frac{x^{k-1}}{(x-1)^3}\left[(k^2 - 2k + 1)x^2 - 2k^2 x + (2k+1)x + k^2\right] \\
&\overset{(a)}{\leq} \frac{x^{k-1}}{(x-1)^3}\left[(k^2 - 2k + 1)x^2 - k^2 x + (2k+1)x\right] \\
&\overset{(b)}{\leq} \frac{x^{k-1}}{(x-1)^3} \cdot k^2 x^2
\end{aligned} \tag{B.79}$$

where (a) is due to $x > 1$ and (b) is due to $k \geq 2$.

When $k = 1$, we have

$$\sum_{l=0}^{k-1} l^2 x^{l-1} = 0 \leq \frac{x^{k-1}}{(x-1)^3} \cdot k^2 x^2. \tag{B.80}$$

Then we have

$$\sum_{l=0}^{k-1} l^2 x^l \leq \frac{x^{k-1}}{(x-1)^3} \cdot k^2 x^2 \cdot x = \frac{k^2 x^{k+2}}{(x-1)^3}. \tag{B.81}$$

$\square$

For quadratic objective functions, the global objective functions is

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c. \tag{B.82}$$

The local objective function of worker i is

$$F_i(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A}_i \mathbf{x} + \mathbf{b}_i^T \mathbf{x} + c_i, \tag{B.83}$$

where $\mathbf{A} = \frac{1}{N} \sum_{i=1}^N \mathbf{A}_i$, $\mathbf{b} = \frac{1}{N} \sum_{i=1}^N \mathbf{b}_i$ and $c = \frac{1}{N} \sum_{i=1}^N c_i$. The local stochastic gradient is

$$\mathbf{g}_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x} + \mathbf{b}_i + \mathbf{n}_i, \tag{B.84}$$

where $\mathbf{n}_i \in \mathbb{R}^d$ is the noise vector. Since we assume the stochastic gradient is unbiased, we have

$$\mathbb{E}[\mathbf{g}_i(\mathbf{x})] = \mathbf{A}_i \mathbf{x} + \mathbf{b}_i + \mathbb{E}[\mathbf{n}_i] = \nabla F_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x} + \mathbf{b}_i, \forall \mathbf{x}. \tag{B.85}$$

Therefore, we can get $\mathbb{E}[\mathbf{n}_i] = \mathbf{0}$. By Assumption 3.2, we have

$$\mathbb{E}\left\|\mathbf{g}_i(\mathbf{x}) - \nabla F_i(\mathbf{x})\right\|^2 = \mathbb{E}\left\|\mathbf{n}_i\right\|^2 \leq \sigma^2. \tag{B.86}$$

By Assumption A.1, we have

$$\|\nabla F_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 = \|(\mathbf{A}_i - \mathbf{A})\mathbf{x} + (\mathbf{b}_i - \mathbf{b})\|^2 \leq \zeta_q^2, \forall i. \tag{B.87}$$

By Proposition 5.3, $L_g = |\lambda(\mathbf{A})|$ and $L_h = 2\max_i |\lambda(\mathbf{A} - \mathbf{A}_i)|$. In the following, we define $\lambda_{\max} := \max_i \|\mathbf{A}_i\|_2$.

In the following, we $\mathbf{n}_i^{r,k}$ to denote the noise vector on worker i at kth iteration of rth round. During local updates, we have

$$
\begin{aligned}
\mathbf{x}_i^{r,k} &= \mathbf{x}_i^{r,k-1} - \gamma \mathbf{g}_i(\mathbf{x}_i^{r,k-1}) \\
&= \mathbf{x}_i^{r,k-1} - \gamma(\mathbf{A}_i\mathbf{x}_i^{r,k-1} + \mathbf{b}_i + \mathbf{n}_i^{r,k-1}) \\
&= (\boldsymbol{I} - \gamma\mathbf{A}_i)\mathbf{x}_i^{r,k-1} - \gamma\left(\mathbf{b}_i + \mathbf{n}_i^{r,k-1}\right) \\
&= (\boldsymbol{I} - \gamma\mathbf{A}_i)\left[(\boldsymbol{I} - \gamma\mathbf{A}_i)\mathbf{x}_i^{r,k-2} - \gamma\left(\mathbf{b}_i + \mathbf{n}_i^{r,k-2}\right)\right] - \gamma\left(\mathbf{b}_i + \mathbf{n}_i^{r,k-1}\right) \\
&= (\boldsymbol{I} - \gamma\mathbf{A}_i)^2\mathbf{x}_i^{r,k-2} - \gamma(\boldsymbol{I} - \gamma\mathbf{A}_i)\left(\mathbf{b}_i + \mathbf{n}_i^{r,k-2}\right) - \gamma\left(\mathbf{b}_i + \mathbf{n}_i^{r,k-1}\right) \\
&= \ldots \\
&= (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\bar{\mathbf{x}}^r - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l(\mathbf{b}_i + \mathbf{n}_i^{r,k-1-l}) \\
&= (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\bar{\mathbf{x}}^r - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{b}_i - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,k-1-l} \\
&\overset{(a)}{=} \left[\boldsymbol{I} - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{A}_i\right]\bar{\mathbf{x}}^r - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{b}_i - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,k-1-l} \\
&= \bar{\mathbf{x}}^r - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\left[\mathbf{A}_i\bar{\mathbf{x}}^r + \mathbf{b}_i\right] - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,l-1-l} \\
&= \bar{\mathbf{x}}^r - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\nabla F_i(\bar{\mathbf{x}}^r) - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,k-1-l}.
\end{aligned}
\tag{B.88}
$$

Now we explain (a). For the sum of geometric series (Hubbard & Hubbard, 2015) of matrix $\boldsymbol{I} - \gamma\mathbf{A}_i$, we have

$$\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l = [\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)]^{-1}\left[\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\right] = \frac{1}{\gamma}\mathbf{A}_i^{-1}\left[\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\right]. \tag{B.89}$$

Since $\mathbf{A}_i$ is symmetric, $\mathbf{A}_i^{-1}$ and $\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)^k$ are also symmetric. Thus, we have

$$\frac{1}{\gamma}\mathbf{A}_i^{-1}\left[\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\right] = \frac{1}{\gamma}\left[\boldsymbol{I} - (\boldsymbol{I} - \gamma\mathbf{A}_i)^k\right]\mathbf{A}_i^{-1} = \sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l. \tag{B.90}$$

Multiplying $\mathbf{A}_i$ on both sides and rearranging, we obtain

$$(\boldsymbol{I} - \gamma\mathbf{A}_i)^k = \boldsymbol{I} - \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{A}_i. \tag{B.91}$$

Then for the model divergence, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 \\
&= \mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\nabla F_i(\bar{\mathbf{x}}^r) + \gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,k-1-l} - \gamma\cdot\frac{1}{N}\sum_{j=1}^{N}\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_j)^l\nabla F_j(\bar{\mathbf{x}}^r)\right.
\end{aligned}
$$

35

$$
- \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \mathbf{n}_j^{r,k-1-l} \bigg\|^2
$$

$$
= \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
+ \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \mathbf{n}_i^{r,l} - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \mathbf{n}_j^{r,k-1-l} \right\|^2
$$

$$
+ \mathbb{E} \left[\mathbb{E}_{\bar{\mathbf{x}}^r} \left\langle \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r), \right.\right.
$$

$$
\left.\left. \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \mathbf{n}_i^{r,l} - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \mathbf{n}_j^{r,k-1-l} \right\rangle \right]
$$

$$
\overset{(a)}{=} \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
+ \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \mathbf{n}_i^{r,l} - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \mathbf{n}_j^{r,k-1-l} \right\|^2, \tag{B.92}
$$

where (a) is due to $\mathbb{E}_{\bar{\mathbf{x}}^r}[\mathbf{n}_i^{r,l}] = 0$.

For the first term in the RHS of (B.92), we have

$$
\mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
= \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla f(\bar{\mathbf{x}}^r) + \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla f(\bar{\mathbf{x}}^r) \right.
$$

$$
- \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A})^l \nabla f(\bar{\mathbf{x}}^r) + \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A})^l \nabla f(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla f(\bar{\mathbf{x}}^r)
$$

$$
\left. + \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla f(\bar{\mathbf{x}}^r) - \gamma \cdot \frac{1}{N} \sum_{j=1}^{N} \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
\leq 4 \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla F_i(\bar{\mathbf{x}}^r) - \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla f(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
+ 4 \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_i)^l \nabla f(\bar{\mathbf{x}}^r) - \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A})^l \nabla f(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
+ 4 \cdot \frac{1}{N} \sum_{j=1}^{N} \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A})^l \nabla f(\bar{\mathbf{x}}^r) - \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla f(\bar{\mathbf{x}}^r) \right\|^2
$$

$$
+ 4 \cdot \frac{1}{N} \sum_{j=1}^{N} \mathbb{E} \left\| \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla f(\bar{\mathbf{x}}^r) - \gamma \sum_{l=0}^{k-1} (\boldsymbol{I} - \gamma \mathbf{A}_j)^l \nabla F_j(\bar{\mathbf{x}}^r) \right\|^2. \tag{B.93}
$$

For the first term in RHS of (B.93), we have

$$\mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\nabla F_i(\bar{\mathbf{x}}^r)-\gamma\sum_{l=0}^{k-1}(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\leq \gamma^2 k\sum_{l=0}^{k-1}\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\right\|^2\mathbb{E}\left\|\nabla F_i(\bar{\mathbf{x}}^r)-\nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\overset{(a)}{\leq}\gamma^2 k\cdot\phi(\kappa,k)\cdot\zeta_q^2, \tag{B.94}$$

where

$$\phi(\kappa,k)=\begin{cases}k & 0\leq\kappa<1\\ \frac{\kappa^{2k}-1}{\kappa^2-1} & 1\leq\kappa\leq 2.\end{cases} \tag{B.95}$$

Now we explain (a). We can rewrite $\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\right\|$ as

$$\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\right\|=\left[\max_j 1-\gamma\lambda_j(\mathbf{A}_i)\right]^l. \tag{B.96}$$

When $\mathbf{A}_i$ is positive definite, which means $\lambda_j(\mathbf{A}_i)>0,\forall i,j$, since $0<\gamma\leq\frac{1}{\lambda_{\max}}$, we have

$$1>\max_j 1-\gamma\lambda_j(\mathbf{A}_i)>\max_j 1-\frac{\lambda_j(\mathbf{A}_i)}{\lambda_{\max}}>0. \tag{B.97}$$

Then we have

$$\sum_{l=0}^{k-1}\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\right\|^2\leq k. \tag{B.98}$$

In this case, we also have

$$\kappa=\max_{i,j}1-\frac{\lambda_j(\boldsymbol{A}_i)}{\|\mathbf{A}_i\|_2}<1. \tag{B.99}$$

When $\mathbf{A}_i$ is not positive definite, which means that $\exists j\in[d]$, such that $\lambda_j(\mathbf{A}_i)\leq 0$, we have

$$1\leq\max_j 1-\gamma\lambda_j(\mathbf{A}_i)\leq\max_j 1-\frac{\lambda_j(\mathbf{A}_i)}{\lambda_{\max}}\leq\max_{i,j}1-\frac{\lambda_j(\boldsymbol{A}_i)}{\|\mathbf{A}_i\|_2}=\kappa. \tag{B.100}$$

Then we have

$$\sum_{l=0}^{k-1}\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\right\|^2\leq\sum_{l=0}^{k-1}\kappa^{2l}=\frac{\kappa^{2k}-1}{\kappa^2-1}. \tag{B.101}$$

For the second term in RHS of (B.93), we have

$$\mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I}-\gamma\mathbf{A}_i)^l\nabla f(\bar{\mathbf{x}}^r)-\gamma\sum_{l=0}^{k-1}(\boldsymbol{I}-\gamma\mathbf{A})^l\nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\leq\gamma^2 k\sum_{l=0}^{k-1}\left\|(\boldsymbol{I}-\gamma\mathbf{A}_i)^l-(\boldsymbol{I}-\gamma\mathbf{A})^l\right\|^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\overset{(a)}{\leq}\gamma^4 L_h^2 k\sum_{l=0}^{k-1}l^2[\varphi(\kappa)]^{2l}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\overset{(b)}{\leq}\gamma^4 L_h^2\cdot\frac{k^3[\varphi(\kappa)]^{2(k+2)}}{([\varphi(\kappa)]^2-1)^3}\cdot\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2, \tag{B.102}$$

where

$$\varphi(\kappa) = \begin{cases} 1 & 0 \le \kappa < 1 \\ \kappa & 1 \le \kappa \le 2, \end{cases} \tag{B.103}$$

(a) is due to

$$\left\|(\boldsymbol{I} - \gamma\mathbf{A}_i)^l - (\boldsymbol{I} - \gamma\mathbf{A})^l\right\|^2 \overset{(c)}{\le} l^2[\varphi(\kappa)]^{2l} \left\|\boldsymbol{I} - \gamma\mathbf{A}_i - \boldsymbol{I} + \gamma\mathbf{A}\right\|^2 = \gamma^2 l^2[\varphi(\kappa)]^{2l} \left\|\mathbf{A}_i - \mathbf{A}\right\|^2 \le \gamma^2 L_h^2 l^2[\varphi(\kappa)]^{2l}, \tag{B.104}$$

and (b) is due to Lemma B.5 by letting $x = [\varphi(\kappa)]^2$.

Now we prove (c). Let $\mathbf{B}_i = \boldsymbol{I} - \gamma\mathbf{A}_i$ and $\mathbf{B} = \boldsymbol{I} - \gamma\mathbf{A}$. Then we have

$$\|\mathbf{B}_i^l\| \le \begin{cases} 1 & 0 \le \kappa < 1 \\ \kappa^l & 1 \le \kappa \le 2. \end{cases} \tag{B.105}$$

Thus, we get $\|\mathbf{B}_i^l\| \le [\varphi(\kappa)]^l$.

Then we have

$$\begin{aligned}
\|\mathbf{B}_i^l - \mathbf{B}^l\| &= \|(\mathbf{B}_i - \mathbf{B})(\mathbf{B}_i^{l-1} + \mathbf{B}_i^{l-2}\mathbf{B} + \mathbf{B}_i^{l-3}\mathbf{B}^2 + \ldots + \mathbf{B}^{l-1})\| \\
&\le \|\mathbf{B}_i - \mathbf{B}\|\|\mathbf{B}_i^{l-1} + \mathbf{B}_i^{l-2}\mathbf{B} + \mathbf{B}_i^{l-3}\mathbf{B}^2 + \ldots + \mathbf{B}^{l-1}\| \\
&\le \|\mathbf{B}_i - \mathbf{B}\|(\|\mathbf{B}_i^{l-1}\| + \|\mathbf{B}_i^{l-2}\|\|\mathbf{B}\| + \|\mathbf{B}_i^{l-3}\|\|\mathbf{B}^2\| + \ldots + \|\mathbf{B}^{l-1}\|) \\
&\le l[\varphi(\kappa)]^{l-1}\|\mathbf{B}_i - \mathbf{B}\|.
\end{aligned} \tag{B.106}$$

Taking (B.94) and (B.102) back to (B.93), we can obtain

$$\begin{aligned}
&\mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\nabla F_i(\bar{\mathbf{x}}^r) - \gamma\cdot\frac{1}{N}\sum_{j=1}^{N}\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_j)^l\nabla F_j(\bar{\mathbf{x}}^r)\right\|^2 \\
&\le 8\gamma^2 k\cdot\phi(\kappa, k)\cdot\zeta_q^2 + 8\gamma^4 L_h^2\cdot\frac{k^3[\varphi(\kappa)]^{2(k+2)}}{([\varphi(\kappa)]^2 - 1)^3}\cdot\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2.
\end{aligned} \tag{B.107}$$

For the second term in RHS of (B.92), we have

$$\begin{aligned}
&\mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,k-1-l} - \gamma\cdot\frac{1}{N}\sum_{j=1}^{N}\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_j)^l\xi_j^{r,k-1-l}\right\|^2 \\
&= \mathbb{E}\left\|\gamma\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\mathbf{n}_i^{r,l}\right\|^2 + \mathbb{E}\left\|\gamma\cdot\frac{1}{N}\sum_{j=1}^{N}\sum_{l=0}^{k-1}(\boldsymbol{I} - \gamma\mathbf{A}_j)^l\xi_j^{r,k-1-l}\right\|^2 \\
&\le \gamma^2\sum_{l=0}^{k-1}\left\|(\boldsymbol{I} - \gamma\mathbf{A}_i)^l\right\|^2\sigma^2 + \gamma^2\sum_{l=0}^{k-1}\left\|(\boldsymbol{I} - \gamma\mathbf{A}_j)^l\right\|^2\sigma^2 \\
&\overset{(a)}{\le} 2\gamma^2\cdot\phi(\kappa, k)\cdot\sigma^2,
\end{aligned} \tag{B.108}$$

where (a) is due to $\sum_{l=0}^{k-1}\|\boldsymbol{I} - \gamma\mathbf{A}_i\|^2 \le \phi(\kappa, k)$.

Substituting (B.92), (B.107) and (B.108) back to (B.71), we have

$$
\mathbb{E}\left[f(\hat{\mathbf{x}}^{t+1})\right]
$$

$$
\leq \mathbb{E}\left[f(\hat{\mathbf{x}}^{t})\right] + \frac{\gamma L_h^2}{2N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\gamma^2 L_g}{2}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N}
$$

$$
\leq \mathbb{E}\left[f(\hat{\mathbf{x}}^{t})\right] + \frac{\gamma L_h^2}{2}\left(8\gamma^2 I\cdot\phi(\kappa,I)\cdot\zeta_q^2 + 8\gamma^4 L_h^2\cdot\frac{I^3[\varphi(\kappa)]^{2(I+2)}}{([\varphi(\kappa)]^2-1)^3}\cdot\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + 2\gamma^2\cdot\phi(\kappa,I)\cdot\sigma^2\right)
$$

$$
- \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N}
$$

$$
= \mathbb{E}\left[f(\hat{\mathbf{x}}^{t})\right] + \frac{\gamma^2 L_g\sigma^2}{2N} + \frac{\gamma}{2}\left(8\gamma^2 L_h^2 I\cdot\phi(\kappa,I)\cdot\zeta_q^2 + 2\gamma^2 L_h^2\cdot\phi(\kappa,I)\cdot\sigma^2\right)
$$

$$
- \left(\frac{\gamma}{2} - 4\gamma^5 L_h^4\cdot\frac{I^3[\varphi(\kappa)]^{2(I+2)}}{([\varphi(\kappa)]^2-1)^3}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2. \tag{B.109}
$$

Let $\gamma \leq \frac{1}{2L_h}\cdot\min\{\frac{1}{I}, \frac{([\varphi(\kappa)]^2-1)^3}{[\varphi(\kappa)]^{2(I+2)}}\}$, we have

$$
\frac{\gamma}{2} - 4\gamma^5 L_h^4\cdot\frac{I^3[\varphi(\kappa)]^{2(I+2)}}{([\varphi(\kappa)]^2-1)^3} \geq \frac{\gamma}{4}. \tag{B.110}
$$

Rearranging (B.109), we obtain

$$
\min_{t\in[T]}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{4\mathcal{F}}{\gamma T} + \frac{2\gamma L_g\sigma^2}{N} + 16\gamma^2 L_h^2 I\cdot\phi(\kappa,I)\cdot\zeta_q^2 + 4\gamma^2 L_h^2\cdot\phi(\kappa,I)\cdot\sigma^2,
$$
$$
\tag{B.111}
$$

where $T = RI$.

B.15. Proof of Theorem A.5

In this section, we provide the proof for the convergence analysis of FedAvg with momentum. First, we apply the technique of auxiliary sequence used by Yu et al. (2019a) to construct the proof. Then, we apply Assumptions 4.1 and 4.2 at key steps.

Before proceeding to the proof, we introduce the auxiliary sequence $\{\hat{\mathbf{z}}^t\}$. That is,

$$
\hat{\mathbf{z}}^t := \begin{cases} \hat{\mathbf{x}}^t & t = 0, \\ \frac{1}{1-\beta}\hat{\mathbf{x}}^t - \frac{\beta}{1-\beta}\hat{\mathbf{x}}^{t-1} & t > 0. \end{cases} \tag{B.112}
$$

Then during each iteration, we have

$$
\hat{\mathbf{z}}^{t+1} = \hat{\mathbf{z}}^t - \frac{\gamma}{1-\beta}\cdot\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t). \tag{B.113}
$$

First, we incorporate Lemma 4 in (Yu et al., 2019a) as follows to support our proof.

Lemma B.6 (Lemma 4 in (Yu et al., 2019a))**.** *For FedAvg with momentum, we have*

$$
\sum_{t=0}^{T-1}\left\|\hat{\mathbf{z}}^t - \hat{\mathbf{x}}^t\right\|^2 \leq \frac{\gamma^2\beta^2}{(1-\beta)^4}\sum_{t=0}^{T-1}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2. \tag{B.114}
$$

Using Assumption 4.1, we have

$$
\mathbb{E}f(\hat{\mathbf{z}}^{t+1}) \leq \mathbb{E}f(\hat{\mathbf{z}}^t) - \frac{\gamma}{1-\beta}\mathbb{E}\left\langle\nabla f(\hat{\mathbf{z}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle + \frac{\gamma^2 L_g}{2(1-\beta)^2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2. \tag{B.115}
$$

For the inner product in the RHS of (B.115), we have

$$-\frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{z}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle$$

$$=-\frac{\gamma}{1-\beta}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}_i^t}\left\langle \nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle\right] - \frac{\gamma}{1-\beta}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}_i^t}\left\langle \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle\right]$$

$$=-\frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle - \frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle. \tag{B.116}$$

For the first inner-product term, we have

$$-\frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle$$

$$\overset{(a)}{\leq} \frac{(1-\beta)}{2\beta L_g}\mathbb{E}\left\|\nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2\beta L_g}{2(1-\beta)^3}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2, \tag{B.117}$$

where (a) is due to $\langle a,b\rangle \leq \frac{c}{2}\|a\|^2 + \frac{1}{2c}\|b\|^2, c > 0$. Here, we let $a = \nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t)$, $b = \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)$ and $c = \frac{(1-\beta)^2}{\gamma\beta L_g}$.

For the second inner-product term, we have

$$-\frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle$$

$$=-\frac{\gamma}{1-\beta}\mathbb{E}\left\langle \nabla f(\hat{\mathbf{x}}^t), \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle$$

$$=\frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 - \frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2$$

$$-\frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2. \tag{B.118}$$

For the norm square in the RHS of (B.115), we have

$$\frac{\gamma^2 L_g}{2(1-\beta)^2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2$$

$$=\frac{\gamma^2 L_g}{2(1-\beta)^2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t) + \nabla F_i(\mathbf{x}_i^t)\right)\right\|^2$$

$$\leq \frac{\gamma^2 L_g\sigma^2}{2N(1-\beta)^2} + \frac{\gamma^2 L_g}{2(1-\beta)^2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2. \tag{B.119}$$

Substituting back to (B.115), we get

$$\mathbb{E}f(\hat{\mathbf{z}}^{t+1}) \leq \mathbb{E}f(\hat{\mathbf{z}}^t) - \frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2 L_g\sigma^2}{2N(1-\beta)^2} + \frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2$$

$$+\frac{(1-\beta)}{2\beta L_g}\mathbb{E}\left\|\nabla f(\hat{\mathbf{z}}^t) - \nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \left(\frac{\gamma}{2(1-\beta)} - \frac{\gamma^2 L_g}{2(1-\beta)^2} - \frac{\gamma^2\beta L_g}{2(1-\beta)^3}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2$$

$$
\leq \mathbb{E}f(\hat{\mathbf{z}}^t) - \frac{\gamma}{2(1-\beta)}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2 L_g \sigma^2}{2N(1-\beta)^2} + \frac{\gamma L_h^2}{2(1-\beta)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2
$$

$$
+ \frac{(1-\beta)L_g}{2\beta}\mathbb{E}\left\|\hat{\mathbf{z}}^t - \hat{\mathbf{x}}^t\right\|^2 - \left(\frac{\gamma}{2(1-\beta)} - \frac{\gamma^2 L_g}{2(1-\beta)^2} - \frac{\gamma^2 \beta L_g}{2(1-\beta)^3}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2. \tag{B.120}
$$

By Lemma B.6, we obtain

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\hat{\mathbf{z}}^t - \hat{\mathbf{x}}^t\right\|^2 \leq \frac{\gamma^2 \beta^2}{(1-\beta)^4}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2
$$

$$
= \frac{\gamma^2 \beta^2}{(1-\beta)^4}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{g}_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t) + \nabla F_i(\mathbf{x}_i^t)\right]\right\|^2
$$

$$
\leq \frac{\gamma^2 \beta^2 \sigma^2}{N(1-\beta)^4} + \frac{\gamma^2 \beta^2}{(1-\beta)^4}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 \tag{B.121}
$$

By Lemma B.4, with $\frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2 + L_g^2)}{(1-\beta)^2}} > 0$, we get

$$
\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{\mathbf{x}}^t - \mathbf{x}_i^t\right\|^2 \leq \frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2 + L_g^2)}{(1-\beta)^2}}\cdot\left(\frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2}\right). \tag{B.122}
$$

Substituting (B.121) and (B.122) back to (B.120), we obtain

$$
\frac{\gamma}{2(1-\beta)}\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{(f_0 - f_*)}{T} + \frac{\gamma^2 L_g \sigma^2}{2N(1-\beta)^2} + \frac{(1-\beta)L_g}{2\beta}\cdot\frac{\gamma^2 \beta^2 \sigma^2}{N(1-\beta)^4}
$$

$$
+ \frac{\gamma L_h^2}{2(1-\beta)}\cdot\frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2 + L_g^2)}{(1-\beta)^2}}\cdot\left(\frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2}\right)
$$

$$
- \left(\frac{\gamma}{2(1-\beta)} - \frac{\gamma^2 L_g}{2(1-\beta)^2} - \frac{\gamma^2 \beta L_g}{2(1-\beta)^3} - \frac{(1-\beta)L_g}{2\beta}\cdot\frac{\gamma^2 \beta^2}{(1-\beta)^4}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2. \tag{B.123}
$$

Dividing both sides by $\frac{\gamma}{2(1-\beta)}$, we obtain

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2
$$

$$
\leq \frac{2(1-\beta)(f_0 - f_*)}{\gamma T} + \frac{\gamma L_g \sigma^2}{N(1-\beta)^2} + L_h^2\cdot\frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2 + L_g^2)}{(1-\beta)^2}}\cdot\left(\frac{2\gamma^2 I \sigma^2}{(1-\beta^2)} + \frac{6\gamma^2 I^2 \zeta^2}{(1-\beta)^2}\right)
$$

$$
- \left(1 - \frac{\gamma L_g}{1-\beta} - \frac{2\gamma \beta L_g}{(1-\beta)^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2. \tag{B.124}
$$

By $\gamma \leq \frac{1-\beta}{\sqrt{18(L_g^2 + L_h^2)I}}$, we have

$$
\frac{1}{1 - \frac{6\gamma^2 I^2 (L_h^2 + L_g^2)}{(1-\beta)^2}} \leq \frac{3}{2}. \tag{B.125}
$$

By $\gamma \leq \frac{(1-\beta)^2}{L_g(1+\beta)}$, we have

$$1 - \frac{\gamma L_g}{1-\beta} - \frac{2\gamma\beta L_g}{(1-\beta)^2} \geq 1 - \frac{1-\beta}{1+\beta} - \frac{2\beta}{1+\beta} = 0. \tag{B.126}$$

With $\gamma \leq \min\{\frac{(1-\beta)^2}{L_g(1+\beta)}, \frac{1-\beta}{\sqrt{18(L_g^2+L_h^2)I}}\}$ we obtain

$$\min_t \mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{2(1-\beta)(f_0-f_*)}{\gamma T} + \frac{\gamma L_g\sigma^2}{N(1-\beta)^2} + \frac{3\gamma^2 L_h^2 I\sigma^2}{(1-\beta)^2} + \frac{9\gamma^2 L_h^2 I^2\zeta^2}{(1-\beta)^2}. \tag{B.127}$$

B.16. Proof of Theorem A.7

In this section, we use the techniques of Reddi et al. (2020) in the proof. We define the update at rth round Δ_r as

$$\Delta_r := \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{r,I} - \bar{\mathbf{x}}^r. \tag{B.128}$$

In FedAdam, the global update is given by

$$\bar{\mathbf{x}}^{r+1} = \bar{\mathbf{x}}^r + \eta\frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau}, \tag{B.129}$$

where

$$\mathbf{v}^r = \beta_2\mathbf{v}^{r-1} + (1-\beta_2)\Delta_r^2. \tag{B.130}$$

We use $\Delta_{r,j}$ to denote the jth element of Δ_r. We use v_j^r to denote the jth element of $\mathbf{v}^r$. The division is element-wise. $\frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau}$ means that for each element $j \in [d]$, we perform $\frac{\Delta_{r,j}}{\sqrt{v_j^r}+\tau}$, and Δ_r^2 means that for each element $j \in [d]$, we perform $\Delta_{r,j}^2$.

By Assumption 4.1, we obtain

$$f(\bar{\mathbf{x}}^{r+1}) \leq f(\bar{\mathbf{x}}^r) + \eta\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau}\right\rangle + \frac{\eta^2 L_g}{2}\sum_{j=1}^{d}\frac{\Delta_{r,j}^2}{(\sqrt{v_j^r}+\tau)^2}. \tag{B.131}$$

For the inner-product term, we have

$$\eta\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau}\right\rangle$$
$$= \eta\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau} - \frac{\Delta_r}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}\right\rangle + \eta\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}\right\rangle. \tag{B.132}$$

By (14)–(15) in Reddi et al. (2020), we have

$$\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\mathbf{v}^r}+\tau} - \frac{\Delta_r}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}\right\rangle$$
$$= \eta\sqrt{1-\beta_2}\mathbb{E}\sum_{j=1}^{d}\frac{G}{\tau}\cdot\frac{\Delta_{r,j}^2}{\sqrt{v_j^r}+\tau}, \tag{B.133}$$

and

$$\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{\Delta_r}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}\right\rangle$$
$$= -\gamma\eta I\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \eta\mathbb{E}\left\langle\frac{\nabla f(\bar{\mathbf{x}}^r)}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}, \Delta_r + \gamma I\nabla f(\bar{\mathbf{x}}^r)\right\rangle. \tag{B.134}$$

For the second term in the RHS of above inequality, we have

$$\eta\mathbb{E}\left\langle \frac{\nabla f(\bar{\mathbf{x}}^r)}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}, \Delta_r + \gamma I\nabla f(\bar{\mathbf{x}}^r)\right\rangle$$

$$= -\eta\gamma\mathbb{E}\left[\mathbb{E}_{\bar{\mathbf{x}}^r}\left\langle \frac{\nabla f(\bar{\mathbf{x}}^r)}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}, \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k}) - I\nabla f(\bar{\mathbf{x}}^r)\right\rangle\right]$$

$$= -\eta\gamma\mathbb{E}\left\langle \frac{\nabla f(\bar{\mathbf{x}}^r)}{\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau}, \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left(\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right)\right\rangle$$

$$\leq \frac{\eta\gamma I}{2}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{\eta\gamma}{2(\sqrt{\beta_2\mathbf{v}^{r-1}}+\tau)}\sum_{k=0}^{I-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\leq \frac{\eta\gamma I}{2}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{\eta\gamma}{2\tau}\sum_{k=0}^{I-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k}) + \nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\leq \frac{\eta\gamma I}{2}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{\eta\gamma L_h^2}{2\tau}\sum_{k=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + \frac{\eta\gamma I L_g^2}{2\tau}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2. \tag{B.135}$$

Since the local updates of FedAdam are the same as that of FedAvg, we can apply Lemma B.2 and Lemma B.3 in the above inequality. Then, by (B.29), with $\gamma \leq \frac{1}{\sqrt{6(L_h^2+L_g^2)I}}$ and $\gamma \leq \frac{1}{2\sqrt{3}IL_g}$, we obtain

$$\frac{\eta\gamma L_h^2}{2\tau}\sum_{k=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^r\right\|^2 + \frac{\eta\gamma I L_g^2}{2\tau}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2$$

$$\leq \frac{\gamma\eta L_g^2 I}{2\tau}\left(5(I-1)\frac{\gamma^2\sigma^2}{N} + 30I\gamma^2\sum_{k=0}^{I-1}\frac{L_h^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\right)$$

$$+ \frac{\eta\gamma L_h^2}{2\tau}\sum_{k=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^r\right\|^2$$

$$\leq \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N} + \frac{15\gamma^3\eta L_g^2 I^3}{\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \left(\frac{15\gamma^3\eta L_g^2 L_h^2 I^2}{\tau} + \frac{\eta\gamma L_h^2}{2\tau}\right)\sum_{k=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^r\right\|^2$$

$$\leq \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N} + \frac{15\gamma^3\eta L_g^2 I^3}{\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\eta\gamma}{\tau}\left(12(I-1)^3\gamma^2 L_h^2\zeta^2 + 4(I-1)^2\gamma^2 L_h^2\sigma^2\right). \tag{B.136}$$

According to (16) and its proof in Reddi et al. (2020), with $\gamma \leq \min\left\{\frac{1}{16L_g I}, \frac{\tau^{\frac{1}{3}}}{16K(120L_g^2 G)^{\frac{1}{3}}}\right\}$, we have

$$\frac{15\gamma^2 L_g^2 I^2}{\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}})\right\|^2 \leq \frac{1}{4}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\beta_2\sqrt{v_j^r}+\tau}. \tag{B.137}$$

Substituting (B.132)–(B.136) back to (B.131), we can get

$$\mathbb{E}f(\bar{\mathbf{x}}^{r+1}) \leq \mathbb{E}f(\bar{\mathbf{x}}^r) + \eta\sqrt{1-\beta_2}\mathbb{E}\sum_{j=1}^{d}\frac{G}{\tau}\cdot\frac{\Delta_{r,j}^2}{\sqrt{v_j^r}+\tau} - \gamma\eta I\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{\eta\gamma I}{2}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau}$$

$$+ \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N} + \frac{15\gamma^3\eta L_g^2 I^3}{\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\eta\gamma}{\tau}\left(12(I-1)^3\gamma^2 L_h^2\zeta^2 + 4(I-1)^2\gamma^2 L_h^2\sigma^2\right) + \frac{\eta^2 L_g}{2}\mathbb{E}\sum_{j=1}^{d}\frac{\Delta_{r,j}^2}{(\sqrt{v_j^r}+\tau)^2}$$

$$\overset{(a)}{\leq} \mathbb{E}f(\bar{\mathbf{x}}^r) + \left(\frac{\eta\sqrt{1-\beta_2}G}{\tau} + \frac{\eta^2 L_g}{2}\right) \mathbb{E}\sum_{j=1}^{d} \frac{\Delta_{r,j}^2}{\sqrt{v_j^r}+\tau} - \frac{\gamma\eta I}{4}\mathbb{E}\sum_{j=1}^{d} \frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N}$$

$$+ \frac{\eta\gamma}{\tau}\left(12(I-1)^3\gamma^2 L_h^2\zeta^2 + 4(I-1)^2\gamma^2 L_h^2\sigma^2\right), \tag{B.138}$$

where (a) is due to (B.137).

Similar to the proof of Lemma 4 in Reddi et al. (2020), we obtain

$$\mathbb{E}\sum_{j=1}^{d} \frac{\Delta_{r,j}^2}{(\sqrt{v_j^r}+\tau)^2}$$

$$\leq \mathbb{E}\sum_{j=1}^{d} \frac{\Delta_{r,j}^2}{\tau^2}$$

$$\leq 2\mathbb{E}\left\|\frac{\Delta_r + \gamma I\nabla f(\bar{\mathbf{x}}^r)}{\tau}\right\|^2 + 2\gamma^2 I^2\mathbb{E}\left\|\frac{\nabla f(\bar{\mathbf{x}}^r)}{\tau}\right\|^2. \tag{B.139}$$

Furthermore, by Lemma B.1 and Lemma B.2, we have

$$2\mathbb{E}\left\|\frac{\Delta_r + \gamma I\nabla f(\bar{\mathbf{x}}^r)}{\tau}\right\|^2$$

$$\leq \frac{4\gamma^2 I}{N\tau^2}\sigma^2 + \frac{4\gamma^2 I}{\tau^2}\mathbb{E}\sum_{k=0}^{I-1}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k}) + \nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right\|^2$$

$$\leq \frac{4\gamma^2 I}{N\tau^2}\sigma^2 + \frac{4\gamma^2 I}{\tau^2}\left(\frac{5\gamma^2 L_g^2 I^2\sigma^2}{N} + 30\gamma^2 L_g^2 I^3\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 + 24(I-1)^3\gamma^2 L_h^2\zeta^2 + 8(I-1)^2\gamma^2 L_h^2\sigma^2\right). \tag{B.140}$$

According to the proof for Theorem 2 in Reddi et al. (2020), with $\gamma \leq \min\{\frac{1}{16IL_g}, \frac{\tau}{6(2G+\eta L_g)}\}$, we have

$$\left(\sqrt{1-\beta_2}G + \frac{\eta L_g}{2}\right)\frac{2\gamma^2 I^2 + 120\gamma^4 L_g^2 I^4}{\tau^2} \leq \frac{\gamma I}{8}\frac{1}{\sqrt{\beta_2}\gamma IG + \tau}. \tag{B.141}$$

Substituting (B.139) and (B.140) back to (B.138), we obtain

$$\mathbb{E}f(\bar{\mathbf{x}}^{r+1}) \leq \mathbb{E}f(\bar{\mathbf{x}}^r) - \frac{\gamma\eta I}{4}\mathbb{E}\sum_{j=1}^{d} \frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N}$$

$$+ \left(\frac{\eta\sqrt{1-\beta_2}G}{\tau} + \frac{\eta^2 L_g}{2}\right)\left(\frac{4\gamma^2 I}{N\tau^2}\sigma^2 + \frac{2\gamma^2 I^2 + 120\gamma^4 L_g^2 I^4}{\tau^2}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 + \frac{96\gamma^4 L_h^2 I^4\zeta^2}{\tau^2} + \frac{32\gamma^4 L_h^2 I^3\sigma^2}{\tau^2}\right)$$

$$+ \frac{\eta\gamma}{\tau}\left(12(I-1)^3\gamma^2 L_h^2\zeta^2 + 4(I-1)^2\gamma^2 L_h^2\sigma^2\right)$$

$$\overset{(a)}{\leq} \mathbb{E}f(\bar{\mathbf{x}}^r) - \frac{\gamma\eta I}{8}\mathbb{E}\sum_{j=1}^{d} \frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}}+\tau} + \frac{5\gamma^3\eta L_g^2 I^2\sigma^2}{2\tau N} + \frac{\eta\gamma}{\tau}\left(12(I-1)^3\gamma^2 L_h^2\zeta^2 + 4(I-1)^2\gamma^2 L_h^2\sigma^2\right)$$

$$+ \left(\eta\sqrt{1-\beta_2}G + \frac{\eta^2 L_g}{2}\right)\left(\frac{4\gamma^2 I}{N\tau^2}\sigma^2 + \frac{96\gamma^4 L_h^2 I^4\zeta^2}{\tau^2} + \frac{32\gamma^4 L_h^2 I^3\sigma^2}{\tau^2}\right), \tag{B.142}$$

where (a) is due to (B.141).

Rearranging the above inequality, we have

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}+\tau}} \leq \frac{8(f_0-f_*)}{\gamma\eta IR} + \frac{\gamma L_g I\sigma^2}{\tau N} + \frac{96\gamma^2 I^2 L_h^2\zeta^2}{\tau} + \frac{32\gamma^2 L_h I\sigma^2}{\tau}$$

$$+\left(\sqrt{1-\beta_2}G + \frac{\eta L_g}{2}\right)\left(\frac{32\gamma}{N\tau^2}\sigma^2 + \frac{768\gamma^3 L_h^2 I^3\zeta^2}{\tau^2} + \frac{256\gamma^3 L_h^2 I^2\sigma^2}{\tau^2}\right). \tag{B.143}$$

By the proof of Theorem 2 in Reddi et al. (2020), we have

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2 v_j^{r-1}+\tau}} \geq \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\sum_{j=1}^{d}\frac{[\nabla f(\bar{\mathbf{x}}^r)]_j^2}{\sqrt{\beta_2}\gamma IG+\tau} \geq \frac{1}{\sqrt{\beta_2}\gamma IG+\tau}\min_r\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2. \tag{B.144}$$

Substituting (B.144) back to (B.143) and rearranging, we have

$$\min_r\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^r)\|^2 \leq \left(\sqrt{\beta_2}\gamma IG+\tau\right)\left(\frac{8(f_0-f_*)}{\gamma\eta IR} + \frac{\gamma L_g\sigma^2}{\tau N} + \frac{96\gamma^2 I^2 L_h^2\zeta^2}{\tau} + \frac{32\gamma^2 L_h I\sigma^2}{\tau}\right)$$

$$+\left(\sqrt{\beta_2}\gamma IG+\tau\right)\left(\sqrt{1-\beta_2}G + \frac{\eta L_g}{2}\right)\left(\frac{32\gamma}{N\tau^2}\sigma^2 + \frac{768\gamma^3 L_h^2 I^3\zeta^2}{\tau^2} + \frac{256\gamma^3 L_h^2 I^2\sigma^2}{\tau^2}\right). \tag{B.145}$$

B.17. Proof of Theorem A.9

In this section, we apply Assumption 4.2 in the convergence analysis for strongly convex objective functions in Karimireddy et al. (2020). First, we bound the term A_1 in the proof of Lemma 7 in Karimireddy et al. (2020) using techniques in our paper.

$$A_1 = \frac{2\gamma\eta}{N}\sum_{k,i}\langle\nabla F_i(\mathbf{x}_i^{k-1}), \mathbf{x}^*-\mathbf{x}\rangle$$

$$= \frac{2\gamma\eta}{N}\sum_{k,i}\langle\nabla F_i(\mathbf{x}_i^{k-1}) - \nabla f(\mathbf{x}) + \nabla f(\mathbf{x}), \mathbf{x}^*-\mathbf{x}\rangle$$

$$= \underbrace{\frac{2\gamma\eta}{N}\sum_{k,i}\langle\nabla F_i(\mathbf{x}_i^{k-1}) - \nabla f(\mathbf{x}), \mathbf{x}^*-\mathbf{x}\rangle}_{T_1} + \underbrace{\frac{2\gamma\eta}{N}\sum_{k,i}\langle\nabla f(\mathbf{x}), \mathbf{x}^*-\mathbf{x}\rangle}_{T_2}. \tag{B.146}$$

For T_1, we have

$$T_1 = 2\gamma\eta\left\langle\frac{1}{N}\sum_{k,i}\nabla F_i(\mathbf{x}_i^{k-1}) - I\nabla f(\mathbf{x}), \mathbf{x}^*-\mathbf{x}\right\rangle$$

$$\leq \gamma\eta\left(a\underbrace{\left\|\frac{1}{N}\sum_{k,i}\nabla F_i(\mathbf{x}_i^{k-1}) - I\nabla f(\mathbf{x})\right\|^2}_{T_3} + \frac{1}{a}\|\mathbf{x}^*-\mathbf{x}\|^2\right), \tag{B.147}$$

where $a > 0$ is a constant. Furthermore, for T_3, we have

$$T_3 = \left\|\frac{1}{N}\sum_{k,i}\nabla F_i(\mathbf{x}_i^{k-1}) - \sum_k\nabla f(\hat{\mathbf{x}}^{k-1}) + \sum_k\nabla f(\hat{\mathbf{x}}^{k-1}) - I\nabla f(\mathbf{x})\right\|^2$$

$$\leq \frac{2L_h^2 I}{N}\sum_{i,k}\|\mathbf{x}_i^{k-1}-\hat{\mathbf{x}}^{k-1}\|^2 + 2L_g^2 I\sum_k\|\hat{\mathbf{x}}^{k-1}-\mathbf{x}\|^2. \tag{B.148}$$

For T_2, according to Lemma 5 in Karimireddy et al. (2020), we obtain

$$
\begin{aligned}
T_2 &\leq \frac{2\gamma\eta}{N} \sum_{k,i} \left(f(\mathbf{x}^*) - f(\mathbf{x}) - \frac{\mu}{4} \|\mathbf{x} - \mathbf{x}^*\|^2 \right) \\
&= 2\gamma\eta I \left(f(\mathbf{x}^*) - f(\mathbf{x}) - \frac{\mu}{4} \|\mathbf{x} - \mathbf{x}^*\|^2 \right).
\end{aligned}
\tag{B.149}
$$

Substituting T_1, T_2 and T_3 back to A_1, we obtain

$$
\begin{aligned}
A_1 \leq & a\gamma\eta \left(\frac{2L_h^2 I}{N} \sum_{i,k} \|\mathbf{x}_i^{k-1} - \hat{\mathbf{x}}^{k-1}\|^2 + 2L_g^2 I \sum_k \|\hat{\mathbf{x}}^{k-1} - \mathbf{x}\|^2 \right) + \frac{\gamma\eta}{a} \|\mathbf{x}^* - \mathbf{x}\|^2 \\
& + 2\gamma\eta I \left(f(\mathbf{x}^*) - f(\mathbf{x}) - \frac{\mu}{4} \|\mathbf{x} - \mathbf{x}^*\|^2 \right).
\end{aligned}
\tag{B.150}
$$

Now we bound the term A_2 in the proof of Lemma 7 in Karimireddy et al. (2020).

$$
\begin{aligned}
A_2 &= \gamma^2\eta^2 I^2 \left\| \frac{1}{NI} \sum_{k,i} \left(\nabla F_i(\mathbf{x}_i^{k-1}) - \nabla f(\hat{\mathbf{x}}^{k-1}) + \nabla f(\hat{\mathbf{x}}^{k-1}) - \nabla f(\mathbf{x}) + \nabla f(\mathbf{x}) \right) \right\|^2 \\
&\leq \frac{3\gamma^2\eta^2 L_h^2 I}{N} \sum_{i,k} \left\| \mathbf{x}_i^{k-1} - \hat{\mathbf{x}}^{k-1} \right\|^2 + 3\gamma^2\eta^2 L_g^2 I \sum_k \left\| \hat{\mathbf{x}}^{k-1} - \mathbf{x} \right\|^2 + 3\gamma^2\eta^2 I^2 \|\nabla f(\mathbf{x})\|^2 \\
&\leq \frac{3\gamma^2\eta^2 L_h^2 I}{N} \sum_{i,k} \left\| \mathbf{x}_i^{k-1} - \hat{\mathbf{x}}^{k-1} \right\|^2 + 3\gamma^2\eta^2 L_g^2 I \sum_k \left\| \hat{\mathbf{x}}^{k-1} - \mathbf{x} \right\|^2 + 3\gamma^2\eta^2 L_g I^2 (f(\mathbf{x}) - f(\mathbf{x}^*)).
\end{aligned}
\tag{B.151}
$$

By Lemma 7 in Karimireddy et al. (2020), we have

$$
\left\| \bar{\mathbf{x}}^{r+1} - \mathbf{x}^* \right\|^2 \leq \left\| \bar{\mathbf{x}}^r - \mathbf{x}^* \right\|^2 + A_1 + A_2 + \frac{\gamma^2\eta^2 I\sigma^2}{N}.
\tag{B.152}
$$

Substituting A_1 and A_2 to the above inequality, we obtain

$$
\begin{aligned}
\left\| \bar{\mathbf{x}}^{r+1} - \mathbf{x}^* \right\|^2 \leq & \left(1 + \frac{\gamma\eta}{a} - \frac{\gamma\eta\mu I}{2} \right) \|\bar{\mathbf{x}}^r - \mathbf{x}^*\|^2 - \left(2\gamma\eta I - 3\gamma^2\eta^2 L_g I^2 \right) (f(\bar{\mathbf{x}}^r) - f^*) \\
& + \left(\frac{2a\gamma\eta L_h^2 I}{N} + \frac{3\gamma^2\eta^2 L_h^2 I}{N} \right) \sum_{i,k} \|\mathbf{x}_i^{r,k-1} - \hat{\mathbf{x}}^{r,k-1}\|^2 \\
& + \left(2a\gamma\eta L_g^2 I + 3\gamma^2\eta^2 L_g^2 I \right) \sum_k \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r \right\|^2 + \frac{\gamma^2\eta^2 I\sigma^2}{N}.
\end{aligned}
\tag{B.153}
$$

By choosing $a = \frac{4}{\mu I}$, we obtain

$$
\begin{aligned}
\left\| \bar{\mathbf{x}}^{r+1} - \mathbf{x}^* \right\|^2 \leq & \left(1 + \frac{\gamma\eta\mu I}{4} - \frac{\gamma\eta\mu I}{2} \right) \|\bar{\mathbf{x}}^r - \mathbf{x}^*\|^2 - \left(2\gamma\eta I - 3\gamma^2\eta^2 L_g I^2 \right) (f(\bar{\mathbf{x}}^r) - f^*) \\
& + \left(\frac{8\gamma\eta L_h^2 I}{\mu I N} + \frac{3\gamma^2\eta^2 L_h^2 I}{N} \right) \sum_{i,k} \|\mathbf{x}_i^{r,k-1} - \hat{\mathbf{x}}^{r,k-1}\|^2 \\
& + \left(\frac{8\gamma\eta L_g^2 I}{\mu I} + 3\gamma^2\eta^2 L_g^2 I \right) \sum_k \left\| \hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r \right\|^2 + \frac{\gamma^2\eta^2 I\sigma^2}{N} \\
\leq & \left(1 - \frac{\gamma\eta\mu I}{4} \right) \|\bar{\mathbf{x}}^r - \mathbf{x}^*\|^2 - \left(2\gamma\eta I - 3\gamma^2\eta^2 L_g I^2 \right) (f(\bar{\mathbf{x}}^r) - f^*) + \frac{\gamma^2\eta^2 I\sigma^2}{N} \\
& + \underbrace{\left(8\gamma\eta L_h \frac{L_h}{\mu} + 3\gamma^2\eta^2 L_h^2 I \right) \cdot \frac{1}{N} \sum_{i,k} \|\mathbf{x}_i^{r,k-1} - \hat{\mathbf{x}}^{r,k-1}\|^2}_{T_4}
\end{aligned}
$$

46

$$+ \underbrace{\left(8\gamma\eta L_g \frac{L_g}{\mu} + 3\gamma^2\eta^2 L_g^2 I\right) \sum_k \left\|\hat{\mathbf{x}}^{r,k-1} - \bar{\mathbf{x}}^r\right\|^2}_{T_5}. \tag{B.154}$$

Applying Lemma B.2 and Lemma B.3 to (B.154), we obtain

$$T_4 \leq \left(8\gamma\eta L_h \frac{L_h}{\mu} + 3\gamma^2\eta^2 L_h^2 I\right) \cdot \left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right), \tag{B.155}$$

and

$$
\begin{aligned}
T_5 &\leq \left(8\gamma\eta L_g I \frac{L_g}{\mu} + 3\gamma^2\eta^2 L_g^2 I^2\right) \cdot \left(5(I-1) \cdot \frac{\gamma^2\sigma^2}{N} + 30I(I-1)\gamma^2 L_g\left(f(\bar{\mathbf{x}}^r) - f^*\right)\right. \\
&\quad + \left. 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right) \\
&\leq \left(8\gamma\eta L_g I \frac{L_g}{\mu} + 3\gamma^2\eta^2 L_g^2 I^2 \cdot \frac{L_g}{\mu}\right) \cdot \left(5(I-1) \cdot \frac{\gamma^2\sigma^2}{N} + 30I(I-1)\gamma^2 L_g\left(f(\bar{\mathbf{x}}^r) - f^*\right)\right. \\
&\quad + \left. 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right) \\
&\leq \left(\gamma\eta I + \gamma^2\eta^2 L_g I^2\right) \cdot 8L_g \frac{L_g}{\mu} \cdot \left(5(I-1) \cdot \frac{\gamma^2\sigma^2}{N} + 30I(I-1)\gamma^2 L_g\left(f(\bar{\mathbf{x}}^r) - f^*\right)\right. \\
&\quad + \left. 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right). \tag{B.156}
\end{aligned}
$$

In particular, we have

$$
\begin{aligned}
&\left(\gamma\eta I + \gamma^2\eta^2 L_g I^2\right) \cdot 8L_g \frac{L_g}{\mu} \cdot 30I(I-1)\gamma^2 L_g\left(f(\bar{\mathbf{x}}^r) - f^*\right) \\
&\overset{(a)}{\leq} \frac{1}{2}\left(\gamma\eta I + \gamma^2\eta^2 L_g I^2\right)\left(f(\bar{\mathbf{x}}^r) - f^*\right), \tag{B.157}
\end{aligned}
$$

where (a) is due to $\gamma \leq \frac{1}{24 L_g I}\sqrt{\frac{\mu}{L_g}}$. Substituting T_4 and T_5 back to (B.154), by $\gamma \leq \frac{1}{24 L_g I}\sqrt{\frac{\mu}{L_g}}$ and $\gamma\eta \leq \frac{1}{16 L_g I}$, we obtain

$$
\begin{aligned}
&\left\|\bar{\mathbf{x}}^{r+1} - \mathbf{x}^*\right\|^2 \\
&\leq \left(1 - \frac{\gamma\eta\mu I}{4}\right)\left\|\bar{\mathbf{x}}^r - \mathbf{x}^*\right\|^2 - \left(\frac{1}{2}\gamma\eta I - 4\gamma^2\eta^2 L_g I^2\right)\left(f(\bar{\mathbf{x}}^r) - f^*\right) + \frac{\gamma^2\eta^2 I\sigma^2}{N} \\
&\quad + \left(8\gamma\eta L_h \frac{L_h}{\mu} + 3\gamma^2\eta^2 L_h^2 I\right) \cdot \left(3c(I-1)^3\gamma^2\zeta^2 + c(I-1)^2\gamma^2\sigma^2\right) \\
&\quad + \left(\gamma\eta I + \gamma^2\eta^2 L_g I^2\right)\left(40\gamma^2 L_g I \cdot \frac{L_g\sigma^2}{\mu N} + 8\frac{L_g^2}{\mu} \cdot 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right) \\
&\leq \left(1 - \frac{\gamma\eta\mu I}{4}\right)\left\|\bar{\mathbf{x}}^r - \mathbf{x}^*\right\|^2 - \frac{\gamma\eta I}{4}\left(f(\bar{\mathbf{x}}^r) - f^*\right) + \frac{\gamma^2\eta^2 I\sigma^2}{N} \\
&\quad + \left(8\gamma\eta L_h \frac{L_h}{\mu} + 3\gamma^2\eta^2 L_h^2 I\right) \cdot \left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right) \\
&\quad + \left(\gamma\eta I + \gamma^2\eta^2 L_g I^2\right)\left(40\gamma^2 L_g I \frac{L_g\sigma^2}{\mu N} + 8\frac{L_g^2}{\mu} \cdot 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right). \tag{B.158}
\end{aligned}
$$

47

Rearranging the above inequality, we have

$$
\begin{aligned}
f(\bar{\mathbf{x}}^r) - f^* \leq & \frac{4}{\gamma\eta I}\left(1 - \frac{\gamma\eta\mu I}{4}\right)\|\bar{\mathbf{x}}^r - \mathbf{x}^*\|^2 - \frac{4}{\gamma\eta I}\|\bar{\mathbf{x}}^{r+1} - \mathbf{x}^*\|^2 + \frac{4\gamma\eta\sigma^2}{N} \\
& + \underbrace{\left(32L_h\frac{L_h}{\mu} + 12\gamma\eta L_h^2 I\right)\cdot\left(12(I-1)^2\gamma^2\zeta^2 + 4(I-1)\gamma^2\sigma^2\right)}_{T_6} \\
& + \underbrace{(4 + 4\gamma\eta L_g I)\left(40\gamma^2 L_g I\frac{L_g\sigma^2}{\mu N} + 8\frac{L_g^2}{\mu}\cdot 30I\gamma^2 L_h^2\left(12(I-1)^3\gamma^2\zeta^2 + 4(I-1)^2\gamma^2\sigma^2\right)\right)}_{T_7}.
\end{aligned}
\tag{B.159}
$$

By $\gamma\eta \leq \frac{1}{4L_h I}$, we have

$$
T_6 \leq 45\gamma^2\frac{L_h^2}{\mu}I^2\zeta^2 + 15\gamma^2\frac{L_h^2}{\mu}I\sigma^2.
\tag{B.160}
$$

By $\gamma\eta \leq \frac{1}{16L_g I}$ and $\gamma \leq \frac{1}{24L_g I}\sqrt{\frac{L_g}{\mu}}$, we have

$$
T_7 \leq 80\gamma^2 L_g I\frac{L_g\sigma^2}{\mu N} + 18\gamma^2\frac{L_h^2}{\mu}I^2\zeta^2 + 6\gamma^2\frac{L_h^2}{\mu}I\sigma^2.
\tag{B.161}
$$

Substituting T_6 and T_7 back, we have

$$
\begin{aligned}
f(\bar{\mathbf{x}}^r) - f^* \leq & \frac{4}{\gamma\eta I}\left(1 - \frac{\gamma\eta\mu I}{4}\right)\|\bar{\mathbf{x}}^r - \mathbf{x}^*\|^2 - \frac{4}{\gamma\eta I}\|\bar{\mathbf{x}}^{r+1} - \mathbf{x}^*\|^2 + \frac{4\gamma\eta\sigma^2}{N} \\
& + 80\gamma^2 L_g I\frac{L_g\sigma^2}{\mu N} + 63\gamma^2\frac{L_h^2}{\mu}I^2\zeta^2 + 21\gamma^2\frac{L_h^2}{\mu}I\sigma^2.
\end{aligned}
\tag{B.162}
$$

By Lemma 1 in [2], using $\frac{1}{\mu R} \leq \gamma\eta I \leq \frac{1}{16L_g}$, we obtain

$$
\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}^R)] - f^* \leq & 4\mu\|\mathbf{x} - \mathbf{x}^*\|^2\exp\left(-\frac{\mu\gamma\eta IR}{4}\right) + \frac{4\gamma\eta\sigma^2}{N} \\
& + 80\gamma^2\frac{L_g^2}{\mu}I\frac{\sigma^2}{N} + 63\gamma^2\frac{L_h^2}{\mu}I^2\zeta^2 + 21\gamma^2\frac{L_h^2}{\mu}I\sigma^2.
\end{aligned}
\tag{B.163}
$$

C. Additional Details and Results of Experiments

In this section, we provide additional details of our experiments. More experimental results are provided for full participation with the MNIST dataset, CINIC-10 dataset (Darlow et al., 2018) and CIFAR-100 dataset.

Environment. All of our experiments are implemented in PyTorch and run on a server with four NVIDIA 2080Ti GPUs. The mini-batch size of SGD for MNIST and CIFAR-10 is 20. The mini-batch size of SGD for CIFAR-100 is 32. We run each experiment 5 times then plot their average and the standard deviation.

Model. For experiments with MLP model, we use a two-layer fully connected neural network, where the width of the networks is 100. For experimental results with CIFAR-10 dataset in the main paper, we use a CNN model. The structure of the CNN is $5 \times 5 \times 32$ Convolutional $\rightarrow 2 \times 2$ MaxPool $\rightarrow 5 \times 5 \times 32$ Convolutional $\rightarrow 2 \times 2$ MaxPool $\rightarrow 4096 \times 512$ Dense $\rightarrow 512 \times 128$ Dense $\rightarrow 128 \times 10$ Dense $\rightarrow$ Softmax. For experimental results with MNIST dataset, we use a two-layer neural network with cross-entropy loss and a linear regression model with MSE loss. For experimental results with CINIC-10 dataset (Darlow et al., 2018), we use VGG-16 with the cross-entropy loss.

Further explanation of the percentage of heterogeneous data. For example, the percentage of heterogeneous data is 50% means that 50% of the data on each worker are with the same label, e.g., 50% of the data on worker 1 are with label 1. Another 50% of the data are sampled uniformly from the remaining dataset.
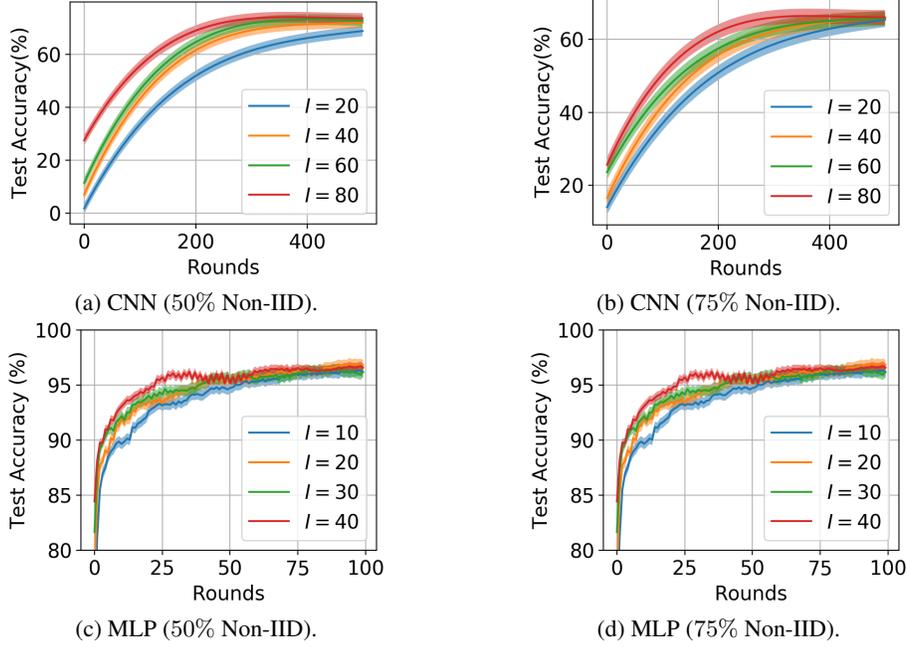
(a) CNN (50% Non-IID).

(b) CNN (75% Non-IID).

(c) MLP (50% Non-IID).

(d) MLP (75% Non-IID).

Figure C.1: The results of test accuracy for CNN with CIFAR-10 and MLP with MNIST, which are corresponding to the setting in Figure 2.
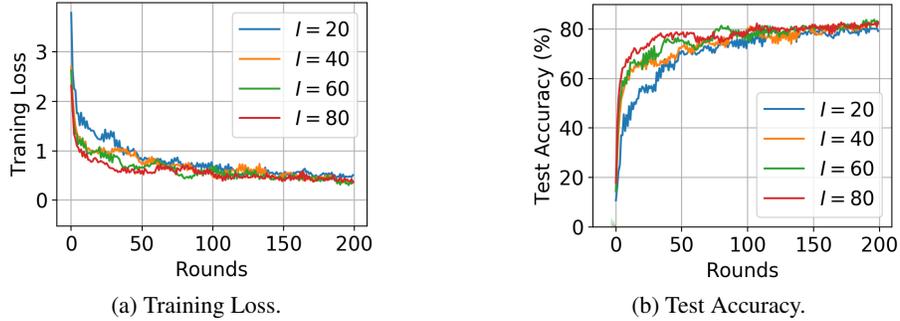

(a) Training Loss.

(b) Test Accuracy.

Figure C.2: Results with CIFAR-10 dataset. The model is VGG-11. The percentage of heterogeneous data is 50%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.

The estimate of L_h. Let the global model be $\bar{\mathbf{x}}$ and the local models be $\mathbf{x}_i, i = 1, 2, \ldots, N$, in the beginning of a round, then we estimate L_h using the following equations.

$$L_h^2 \approx \frac{\left\| \nabla f(\bar{\mathbf{x}}) - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\mathbf{x}_i) \right\|^2}{\frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i - \bar{\mathbf{x}} \right\|^2}.$$

Starting from a global model that is close to convergence, we perform FedAvg for 10 rounds and estimate L_h^2 in each round. Then we use the averaged L_h^2 over 10 rounds as the estimate for L_h^2. The reason for starting from a global model that is close to convergence is that this can make the variance of the estimate smaller. Similarly, the methods of estimating L_g and $\tilde{L}$ are given by

$$L_g \approx \frac{\| \nabla f(\bar{\mathbf{x}}) - \nabla f(\bar{\mathbf{y}}) \|}{\| \bar{\mathbf{x}} - \bar{\mathbf{y}} \|},$$

$$\tilde{L} \approx \max_i \frac{\| \nabla F_i(\bar{\mathbf{x}}) - \nabla F_i(\mathbf{x}_i) \|}{\| \bar{\mathbf{x}} - \mathbf{x}_i \|}.$$
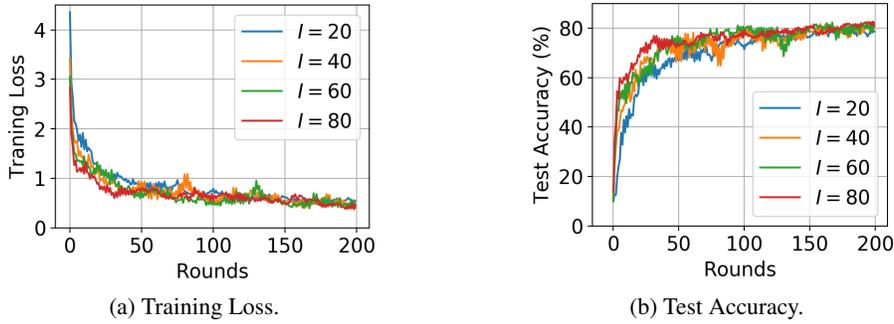
(a) Training Loss.

(b) Test Accuracy.

Figure C.3: Results with CIFAR-10 dataset. The model is VGG-11. The percentage of heterogeneous data is 75%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.
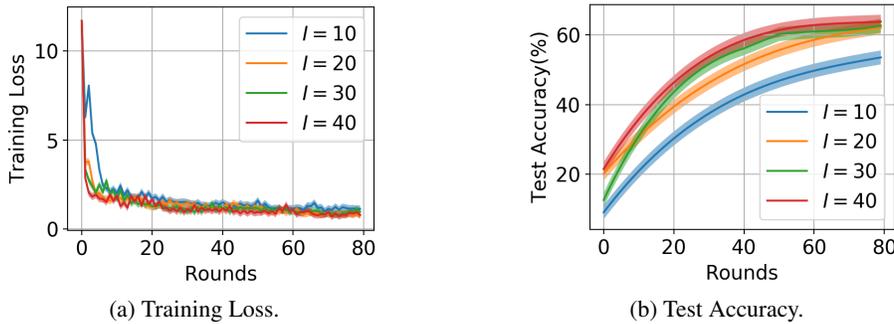

(a) Training Loss.

(b) Test Accuracy.

Figure C.4: Results with CINIC-10 dataset. The model is VGG-16. The percentage of heterogeneous data is 50%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.

Additional Experimental Results. We partition CIFAR-10, CINIC-10 and CIFAR-100 into 10 workers. During each round, all workers will perform the local updates. The results of test accuracy for CNN with CIFAR-10 and MLP with MNIST are provided in Figure C.1. The results of training loss can be found in Figure 2 of the main paper. As shown in Table 1 of the main paper, L_h is very small in this case. In Corollary 4.4, with full participation, it is shown that when L_h is small, increasing I can improve the convergence even when data are highly heterogeneous. As shown in both Figure C.1, the curve with the largest number of local iterations, converges the fastest and achieves the highest accuracy, which validates Theorem 4.3 and Corollary 4.4.

Results for CIFAR-10 with VGG-11 are shown in Figures C.2 and C.3. Results for CINIC-10 with VGG-16 are shown in Figures C.4 and C.5. Results for CIFAR-100 with VGG-16 are shown in Figure C.6 and C.7. It can be seen in both results, the curve with the largest number of local iterations converges the fastest and achieves the highest accuracy, which is

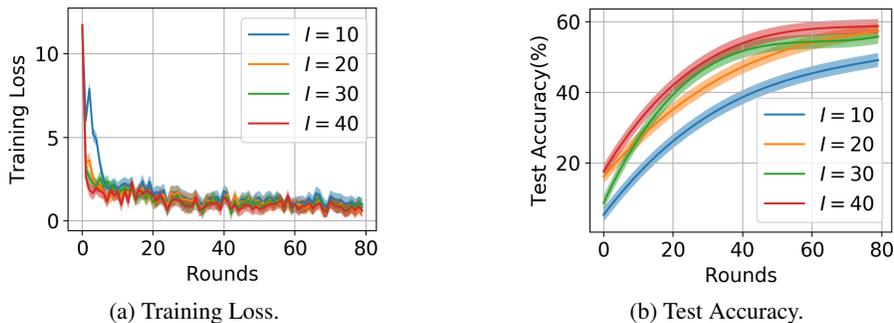(a) Training Loss.

(b) Test Accuracy.

Figure C.5: Results with CINIC-10 dataset. The model is VGG-16. The percentage of heterogeneous data is 75%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.
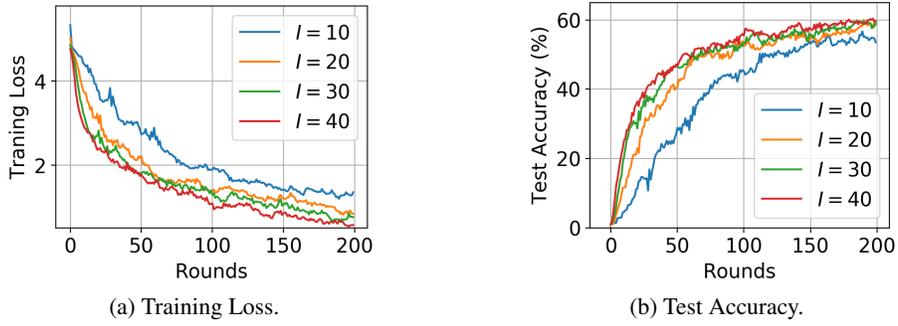
(a) Training Loss.

(b) Test Accuracy.

Figure C.6: Results with CIFAR-100 dataset. The model is VGG-16. The percentage of heterogeneous data is 50%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.02$.
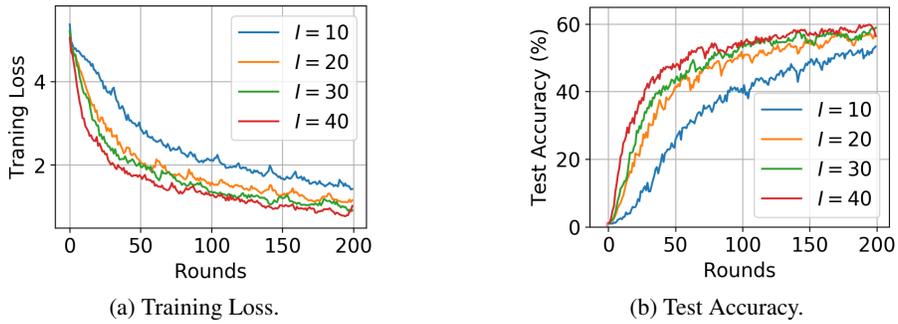

(a) Training Loss.

(b) Test Accuracy.

Figure C.7: Results with CIFAR-100 dataset. The model is VGG-16. The percentage of heterogeneous data is 75%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.02$.

consistent with our theoretical results in Theorem 4.3 and Corollary 4.4.