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 optimization, data heterogeneity is the main reason that existing theoretical analyses are pessimistic about the convergence error caused by local updates. However, experimental results have shown that more local updates can improve the convergence rate and reduce the communication cost when data are heterogeneous. This paper bridges this gap between theoretical understanding and the practical performance by providing a general theoretical analysis for federated averaging (FedAvg) with non-convex objective functions from a new perspective on data heterogeneity. Identifying the limitations in the commonly used assumption of bounded gradient divergence, we propose a new assumption, termed the heterogeneity-driven Lipschitz assumption, which characterizes the fundamental effect of data heterogeneity on local updates. We find the widely used local Lipschitz constant is affected by data heterogeneity, which is neglected in the literature. The proposed heterogeneity-driven Lipschitz constant can capture the information about data heterogeneity contained in local Lipschitz constant. At the same time, the information about the gradient smoothness is captured by the global Lipschitz assumption. Based on the new assumption, we derive novel convergence bounds for both full participation and partial participation, which are tighter and show that more local updates can improve the convergence rate even when data are highly heterogeneous. Furthermore, the assumptions used in this paper are weaker than those used in the literature.

[1]Department of Electrical & Computer Engineering, University of Utah, Salt Lake City, UT, USA [2]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Mingyue Ji <mingyue.ji@utah.edu>.

Workshop of Federated Learning and Analytics in Practice, colocated with 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

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 data privacy. We consider the following optimization problem in federated learning:

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

where N is the number of workers and $F_i(\mathbf{x})$ is the expected loss function of worker i,

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

where $\ell(\cdot)$ is the loss function, ξ_i is the random data sample on worker i, and $\mathcal{D}_i$ is the data distribution on worker i. In addition, let $\mathcal{D}$ be 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) (McMahan et al., 2017), also known as local SGD, is the most popular algorithm in FL.

There is a gap between the theoretical understanding and the experimental results. Unlike centralized stochastic gradient descent (SGD) 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 when using local updates. Existing theoretical analyses for non-convex objective functions (Yu et al., 2019a;b; Wang & Joshi, 2019; Yang et al., 2020) are pessimistic about the convergence error caused by local updates due to the data heterogeneity, since it is shown that the convergence error grows very fast when the number of local updates I increases. Even for convex objective functions, it is challenging to show theoretically when local SGD (with $I > 1$) can outperform mini-batch SGD ($I = 1$) (Woodworth et al., 2020a;b). However, in practice, 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 (McMahan et al., 2017; Lin et al., 2020). This means that increasing I can improve the convergence rate and reduce the communication cost when 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 analysis may overestimate the error caused by local updates.

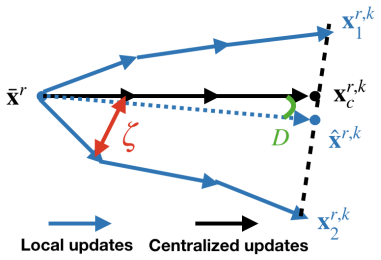


Figure 1: An informal and illustrative comparison between D and ζ in local updates and centralized updates. $\bar{\mathbf{x}}^r$ is the global model at rth round. The local models after k local iterations at the rth round are denoted by $\mathbf{x}_1^{r,k}$ and $\mathbf{x}_2^{r,k}$. The average of $\mathbf{x}_1^{r,k}$ and $\mathbf{x}_2^{r,k}$ is $\hat{\mathbf{x}}^{r,k}$. The centralized model after k 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 D shows the difference between $\mathbf{x}_c^{r,k}$ and $\hat{\mathbf{x}}^{r,k}$.

The existing metric cannot fully characterize the effect of data heterogeneity. The most common metric of data heterogeneity in existing works (Yu et al., 2019b; Wang & Joshi, 2019; Karimireddy et al., 2020; Woodworth et al., 2020b) is called gradient divergence (ζ), which characterizes the difference between the expected local gradient $\nabla F_i(\mathbf{x})$ of worker i and the expected global gradient $\nabla f(\mathbf{x})$. As shown in Figure 1, the intuition behind the gradient divergence is that when ζ is large, the difference between local models $\mathbf{x}_1^{r,k}$, $\mathbf{x}_2^{r,k}$ and the centralized model $\mathbf{x}_c^{r,k}$ becomes large since the centralized model is updated with the gradients sampled from $\mathcal{D}$ while the local model is updated with the gradients sampled from $\mathcal{D}_i$. 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 algorithm, which means that a large number of aggregations are needed to guarantee the convergence. However, as we show in Section C, there exists a case where ζ can be arbitrarily large while only one aggregation is sufficient. This mismatch between the large gradient divergence and the small number of aggregations is because that the gradient divergence cannot characterize the relationship between the averaged model $\hat{\mathbf{x}}^{r,k}$ and the centralized model $\mathbf{x}_c^{r,k}$. If the averaged model remains close to the centralized model after several local updates, it indicates that the effect of data heterogeneity on the disparity between local and centralized updates is small. In this case, performing more local updates is beneficial. However, as shown in Figure 1, when the difference between the averaged model $\hat{\mathbf{x}}^{r,k}$ and the centralized model $\mathbf{x}_c^{r,k}$ is small, ζ can be large.

Another observation is that the widely used local Lipschitz constant $\tilde{L}$ (in Assumption 5) is affected by data heterogeneity, which is neglected by previous theoretical results. 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 the data heterogeneity. However, as shown in Table 1, $\tilde{L}$ increases fast as the percentage of non-IID data increases, which means that the local Lipschitz constant contains the information about data heterogeneity. Neglecting the information about data heterogeneity contained in $\tilde{L}$ can lead to a loose convergence bound since the error term related to I is proportional to $\tilde{L}^2$ in the literature.

A deeper understanding of the behavior of local updates is needed. In addition to FedAvg, there have been 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). Nevertheless, the core mechanism, local updates, is still the foundation of all FL algorithms. Therefore, it is important to understand the behavior of local updates when data are highly heterogeneous so that more insights can be provided for designing FL algorithms. However, existing theoretical results overestimate convergence error caused by local updates. It is unclear how to fully take advantage of the local updates to reduce communication cost.

Contribution of this paper. In this paper, we reveal the fundamental effect of the data heterogeneity on local updates by introducing a new perspective shown by D, the *heterogeneity-driven Lipschitz constant* in Assumption 4. The proposed metric D captures previously overlooked information about the data heterogeneity contained in the local Lipschitz gradient assumption. In addition, we only assume the Lipschitz gradient for the global objective function instead of for each local objective function. In Section 2, using the new assumption, we develop a novel analysis for FedAvg with general non-convex objective functions, which shows that if D is small enough, even for a large ζ, the convergence error caused by local updates is small so that a large I can still be used to reduce communication costs. Our analysis can incorporate partial participation where only a subset of workers are sampled to perform the local updates in each round. In Section C, the insights behind the new assumption are discussed. We show that the assumptions used in this paper are weaker than those used in the literature and D can characterize the difference between the averaged model and the centralized model, which the gradient divergence cannot characterize. We further provide a (possibly non-convex) quadratic example with $D = 0$ to explicitly show that local SGD can be superior than mini-batch SGD even when $\zeta > 0$ is arbitrarily large.

2. Main Results

In this section, we present the theoretical results for non-convex objective functions using the proposed new assumption. All proofs can be found in appendix.

Table 1: Estimated D, $\tilde{L}$, L with the MNIST dataset. Heterogeneity is shown by the percentage of data on each worker that are not uniformly sampled from the global dataset.

Obj. Function	Two-layer Neural Network				Linear Regression			
Heterogeneity	25%	50%	75%	100%	25%	50%	75%	100%
$\tilde{L}$	127.62	130.97	134.24	141.92	2010.51	3577.35	20563.42	25402.19
D	0.35	0.82	1.66	2.36	226.15	916.20	3172.41	4610.54
L	122.23	122.23	122.23	122.23	869.07	869.07	869.07	869.07

In the literature, three classes of assumptions on *stochastic 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 1 for stochastic gradient variance and Assumption 2 for gradient divergence. Assumption 3 and 4 will replace Assumption 5 in appendix. In Section C, we will show that Assumptions 3 and 4 are weaker than Assumption 5.

Assumption 1 (Bounded Stochastic Gradient Variance)**.**

$$\mathbb{E}\left[\left\|\mathbf{g}_i(\mathbf{x}) - \nabla F_i(\mathbf{x})\right\|^2\right] \leq \sigma^2, \forall i, \mathbf{x}. \tag{3}$$

Assumption 2 (Bounded Gradient Divergence)**.**

$$\left\|\nabla F_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq \zeta^2, \forall i, \mathbf{x}. \tag{4}$$

Assumption 2 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 $O(\gamma^2\tilde{L}^2I^2\zeta^2)$ 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 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 solve this problem using our new assumption and analysis.

Assumption 3 (Global Lipschitz Gradient)**.** *The global objective function $f(\mathbf{x})$ satisfies*

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\| \leq L\left\|\mathbf{x} - \mathbf{y}\right\|, \forall \mathbf{x}, \mathbf{y}. \tag{5}$$

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 5 or for each data sample as in (Khaled et al., 2020).

Assumption 4 (Heterogeneity-driven Lipschitz Condition on Averaged Gradients)**.** *There exists a constant $D \geq 0$ such that $\forall \mathbf{x}_i$,*

$$\left\|\frac{1}{N}\sum_{i=1}^N \nabla F_i(\mathbf{x}_i) - \nabla f\left(\bar{\mathbf{x}}\right)\right\|^2 \leq \frac{D^2}{N}\sum_{i=1}^N \left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|^2, \tag{6}$$

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

Assumption 4 can be regarded as a new perspective on data heterogeneity. First, it has been shown in Table 1 that when the percentage of heterogeneous data increases, D becomes larger. Second, D shows how the difference between $\frac{1}{N}\sum_{i=1}\nabla F_i(\mathbf{x}_i)$ and $\nabla f\left(\bar{\mathbf{x}}\right)$ on the LHS of (6) depends on the difference between the local models $\{\mathbf{x}_i\}$ and the global model $\bar{\mathbf{x}}$ on the RHS of (6). When data are less heterogeneous, $\nabla F_i(\mathbf{x})$ is similar to $\nabla f(\mathbf{x})$. The LHS of (6) mainly depends on the difference between local models $\{\mathbf{x}_i\}$ and the global model $\bar{\mathbf{x}}$, which can be characterized by RHS of (6) so D is small. When data are highly heterogeneous, the LHS of (6) does not only depend on the difference on the models but also depend on the difference between the local gradients $\{\nabla F_i(\mathbf{x})\}$ and the global gradient $\nabla f(\mathbf{x})$ so D can be large. We will show in Section C that D can indeed characterize the difference between the averaged model and centralized model, which is not fully characterized by the gradient divergence ζ in Assumption 2. Next, we present the theoretical results for full participation.

Theorem 1 (General Non-convex Objective Functions)**.** *Assuming Assumptions 1, 2, 3, 4 hold, when $\gamma \leq \frac{1}{30(D+L)I}$ and $\gamma\eta \leq \frac{1}{4IL}$, then after R rounds of FedAvg,*

$$\min_{r\in[R]}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 = \mathcal{O}\Bigg(\underbrace{\frac{\mathcal{F}}{\gamma\eta RI} + \frac{\gamma\eta L\sigma^2}{N}}_{\text{error caused by SGD}} + \underbrace{\gamma^2D^2(I-1)^2\zeta^2 + \gamma^2D^2(I-1)\sigma^2}_{\text{error caused by local updates}}\Bigg), \tag{7}$$

where $\mathcal{F} := f(\mathbf{x}_0) - f^*$.

A tighter bound by using new assumption. In (7), the stochastic variance in the error caused by SGD depends L while the error caused by local updates depend D. In the literature, all the mentioned errors depend on $\tilde{L}$ instead of D and L. However, as shown by the experimental results in Table 1, both D and L are smaller than $\tilde{L}$. In Section C, we also prove theoretically that D and L are smaller than $\tilde{L}$. In particular, D can be far less than $\tilde{L}$. Therefore, existing theoretical results overestimate both the error caused by SGD and the error caused by local updates while the convergence upper bound using new assumption is tighter.

New insights about the effect of data heterogeneity. It can be observed that only one term in the error caused by

local updates depends on ζ while both terms in the error caused by local updates depend on D. *A key message is that when ζ^2 is large, as long as D^2 is small enough, the error caused by local updates can still be small.* Since D and ζ characterize the effect of the data heterogeneity in different perspectives, we show that it is possible that $D = 0$ while ζ can be arbitrarily large by providing an example in Section C. In this case, no matter how large ζ is, the convergence error of local SGD is the same as that of centralized SGD, which means that I can be arbitrarily large and only one aggregation is sufficient. Moreover, when $D = 0$, we can see that the impacts of γ and η on the convergence upper bound are the same since the error caused by local updates is zero and the error caused by SGD is a function of $\gamma\eta$. In this case, the two-sided learning rates may not help and only a single learning rate, e.g., let $\eta = 1$, suffices to achieve the desired convergence upper bound.

It is noteworthy that although the value of D increases with the percentage of heterogeneous data, it is possible for D to be small even when the percentage of heterogeneous data is large as shown by the experimental results for the two-layer neural network in Table 1 and the experimental results for CNN in Section D. The following corollary shows that when D is small, more local iterations can improve the convergence rate.

Corollary 1. *Given $c > 0$, when $D \leq \frac{c}{I}$, let $\gamma\eta = \frac{1}{c} \cdot \sqrt{\frac{4\mathcal{F}N}{RIL\sigma^2}}$ and $\gamma = \frac{1}{c} \cdot \frac{1}{\sqrt{RIN}}$, when R is sufficiently large so that $\gamma \leq \frac{1}{30(D+L)I}$ and $\gamma\eta \leq \frac{1}{4IL}$ are satisfied, we have*

$$\min_{r\in[R]} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 = \mathcal{O}\left(\sqrt{\frac{\mathcal{F}L\sigma^2}{RIN}} + \frac{\zeta^2 + \sigma^2/I}{RIN}\right). \tag{8}$$

From (8), it can be seen that it achieves *linear speedup* in the number of iterations RI with respect to the total number of workers N. The constant c controls the tradeoff between the learning rates η, γ and the number of local iterations I. Given D, if c is small, the learning rates are large, I will have to be small to satisfy $D \leq \frac{c}{I}$, and vice versa. The constant c is absorbed by the $\mathcal{O}(\cdot)$ in (8). Corollary 1 shows that as long as the condition of $D \leq \frac{c}{I}$ holds, the convergence error decreases if I increases. Although ζ can be large, if D is small enough, more local updates can still improve the convergence rate. In addition, to achieve the same accuracy, we can decrease R and increase I so that the communication cost can be reduced, as long as the condition $D \leq \frac{c}{I}$ still holds.

Analysis for Partial Participation. We also use the new assumption to develop the theoretical analysis for partial participation. Here we consider the sampling strategy where M workers are uniformly sampled with replacement at the start of each round. The result can show the insights into the relationship between local updates and partial participation. It is worth noting that the technique for partial participation in the literature cannot be directly applied in our analysis since the Lipschitz gradient (see Assumption 5) is assumed for each local objective function. Therefore, we develop new techniques to incorporate the partial participation using D and L, which can be found in the supplementary material.

Theorem 2 (Partial Participation)**.** *Consider uniformly sampling M ($1 \leq M \leq N$) workers in each round of FedAvg. Assuming Assumptions 1, 2, 3, 4 hold, when $\gamma \leq \frac{1}{30(D+L)I}$ and $\gamma\eta \leq \frac{1}{4IL}$, after R rounds of FedAvg,*

$$\min_{r\in[R]} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 = \mathcal{O}\Bigg(\underbrace{\frac{\mathcal{F}}{\gamma\eta RI} + \frac{\gamma\eta L\sigma^2}{M}}_{\text{error caused by SGD}} + \underbrace{\frac{\gamma\eta LI\zeta^2}{M}}_{\text{error caused by p.p.}} + \underbrace{\gamma^2D^2(I-1)^2\zeta^2 + \gamma^2D^2(I-1)\sigma^2}_{\text{error caused by local updates}}\Bigg), \tag{9}$$

where "p.p." means partial participation.

Compared with Theorem 1, there are two differences in the convergence bound. First, the stochastic variance term in the error caused by SGD depends on M. This means that more workers sampled in each round can reduce the stochastic variance. Second, there is an extra term $\frac{\gamma\eta LI\zeta^2}{M}$ in the convergence bound for partial participation, which denotes the error caused by partial participation. This term depends on L and not on D. This means that a small D cannot reduce the error caused by partial participation, which can be shown explicitly by Corollary 2.

3. Conclusion

In this paper, we bridge the gap between the pessimistic theoretical results and the good experimental performance for FL algorithms by introducing a new theoretical perspective of the data heterogeneity, which is shown by the proposed heterogeneity-driven Lipschitz constant D. Using the new assumption, we develop a novel convergence analysis for FedAvg and identify the regions where local updates can help to improve the convergence even when data are highly heterogeneous. Our convergence bounds for both full participation and partial participation are tighter compared to the state of the art in the literature. At the same time, the assumptions used in this paper are weaker. The proposed heterogeneity-driven Lipschitz condition can be applied to the non-convex analysis for FL algorithms (not limited to FedAvg) through a key step shared by the literature for non-convex analysis for FL algorithms (see details in related works). This key step shows the potential of extending the proposed analysis to other FL algorithms to reveal more insights. Future works include applying the proposed analysis in other federated algorithms and incorporating advanced sampling strategy for partial participation in our analysis.

References

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.

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.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10334–10343. PMLR, 13–18 Jul 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.

A. Related Works

There have been a considerable amount of works 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, $\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i)-\nabla f(\bar{\mathbf{x}})\right\|$, to the model divergence $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\bar{\mathbf{x}}\|$, 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 is often applied, which amplifies the effect of data heterogeneity. In this paper, the heterogeneity-driven Lipschitz constant D is applied in this step so that the convergence error is much smaller than that based on $\tilde{L}$, since it can be seen in Table 1 that D is often far smaller than $\tilde{L}$.

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. Therefore, the information about data heterogeneity contained in $\tilde{L}$ is not characterized in either work. The aim of (Wang et al., 2022) is 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)). The authors in (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 (Das et al., 2022), the authors introduce a parameter α in the process of relating the difference of gradients to the model divergence, which can be covered by D in this paper. But α cannot cover what D can show since they 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 $\alpha=0$.

B. Setup

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 their own local data distribution $\mathcal{D}_i$,

$$\mathbf{x}_i^{r,k+1}=\mathbf{x}_i^{r,k}-\gamma\mathbf{g}(\mathbf{x}_i^{r,k};\zeta_i), \tag{10}$$

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

$$\bar{\mathbf{x}}^{r+1}=\bar{\mathbf{x}}^r-\eta\cdot\frac{1}{N}\sum_{i=1}^{N}\Delta_i^r, \tag{11}$$

where η is the global learning rate. Let $\hat{\mathbf{x}}^{r,k}$ be the "virtual" averaged model during the local update phase and

$$\hat{\mathbf{x}}^{r,k+1}:=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{r,k+1}=\hat{\mathbf{x}}^{r,k}-\gamma\cdot\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^{r,k}), \tag{12}$$

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

$$\mathbf{x}_c^{r,k+1}:=\hat{\mathbf{x}}^{r,k}-\gamma\bar{\mathbf{g}}(\hat{\mathbf{x}}^{r,k}), \tag{13}$$

where $\mathbb{E}\left[\bar{\mathbf{g}}(\hat{\mathbf{x}}^{r,k})\right]=\nabla f(\hat{\mathbf{x}}^{r,k})$. The summary of FedAvg algorithm can be found in Algorithm 1. The following assumption is widely used in the literature.

[1]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 ease of analysis, and leave the consideration of the "actual" centralized model for future work.

Assumption 5 (Local Lipschitz Gradient)**.**

$$\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq \tilde{L} \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}, i. \tag{14}$$

There are also some works (Khaled et al., 2020) assuming that Lipschitz gradient condition holds for each data sample $\|\nabla \ell(\mathbf{x}; \xi) - \nabla \ell(\mathbf{y}; \xi)\| \leq L' \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}, \xi$, which is stronger and can imply local Lipschitz gradient condition.

C. Additional Results and Discussions

In this section, we reveal how D captures the information about the data heterogeneity from $\tilde{L}$ and why D can determine the error caused by local updates. Then we explicitly provide an example with $D = 0$. By analyzing this example, we identify a region where local SGD can outperform mini-batch SGD when ζ can be arbitrarily large.

Corollary 2 (Partial Participation with A Small D)**.** *Consider uniformly sampling M workers during each round in FedAvg. Given $c > 0$, when $D \leq \frac{c}{I}$, let $\gamma\eta = \frac{1}{c} \cdot \sqrt{\frac{M\mathcal{F}}{LIR(\sigma^2 + I\zeta^2)}}$ and $\gamma = \frac{1}{c} \cdot \frac{1}{\sqrt{RIN}}$, when R is sufficiently large, we have*

$$\min_{r \in [R]} \mathbb{E} \|\nabla f(\bar{\mathbf{x}}^r)\|^2 = \mathcal{O}\left(\sqrt{\frac{\mathcal{F}L\zeta^2}{RM}} + \sqrt{\frac{\mathcal{F}L\sigma^2}{RIM}} + \frac{\zeta^2 + \sigma^2/I}{RIN}\right). \tag{15}$$

It can be seen from Corollary 2 that when D is small, increasing I can still reduce the convergence error. However, the error caused by partial participation, which is represented by the first term in (15), cannot be reduced by increasing I. This is because that D characterizes the difference between the averaged model over all workers and the centralized model (we will formally explain this property in Section C). 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 in each round. The stochastic variance caused by the sampling strategy is not characterized by D. In addition, the dominant term in (15) becomes $\mathcal{O}\left(\sqrt{1/RM}\right)$. This means that given the sampling strategy, to achieve a small convergence error, performing a large number of aggregations is necessary. However, increasing I can still accelerate the convergence by reducing the other two terms.

Assumptions in this paper are weaker. In the following proposition, we show that Assumptions 3 and 4 are weaker than the commonly used Assumption 5 in the literature.

Proposition 1. *If Assumption 5 holds, then Assumption 3 holds by choosing $L = \tilde{L}$ and Assumption 4 holds by choosing $D = \tilde{L}$.*

Proposition 1 also shows how the information about the data heterogeneity contained in $\tilde{L}$ is captured. The information about the smoothness of the gradients remains in L, which does not change with the data heterogeneity, while D characterizes the effect of data heterogeneity. In addition, Proposition 1 implies that $L \leq \tilde{L}$ and $D \leq \tilde{L}$. However, as shown in Table 1, D can be much smaller than $\tilde{L}$. We examine the intricate relationship between D and $\tilde{L}$ through a deeper analysis of the quadratic[2] (potentially non-convex) objective function:

$$F_i(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{\mathrm{T}}\mathbf{x} + \mathbf{c}_i. \tag{16}$$

By (1), we directly obtain that the global objective function is given by $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + \mathbf{c}$, where $\mathbf{A} = \frac{1}{N}\sum_{i=1}^N \mathbf{A}_i$ and $\mathbf{b} = \frac{1}{N}\sum_{i=1}^N \mathbf{b}_i$. In this case, we can derive the explicit forms of D and $\tilde{L}$ as shown by the following proposition.

Proposition 2. *For quadratic objective functions defined in (16), Assumptions 5 and 4 hold with $\tilde{L} = \max_{i \in [N]} |\lambda(\mathbf{A}_i)|$, $D = 2 \cdot \max_{i \in [N]} |\lambda(\mathbf{A}_i - \mathbf{A})|$, respectively, where $|\lambda(\mathbf{A})|$ denotes the largest absolute value of the eigenvalues of $\mathbf{A}$.*

From Proposition 2, it can be seen that both D and $\tilde{L}$ capture the properties of Hessian matrices for quadratic objective functions. The heterogeneity-driven Lipschitz constant D characterizes the largest eigenvalue of the "deviation" of $\{\mathbf{A}_i\}$ from the global Hessian matrix $\mathbf{A}$, while $\tilde{L}$ characterizes the largest eigenvalue of $\{\mathbf{A}_i\}$ themselves. It can be observed that when $\mathbf{A}_i = \mathbf{A}, \forall i$, which means that the difference of local Hessian matrices is zero, Assumption 4 holds with $D = 0$. Note that, at the same time, we can pick an $\mathbf{A}_i$ such that $\tilde{L} = \max_{i \in [N]} |\lambda(\mathbf{A}_i)|$ is much larger than zero. This observation shows

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

that while the difference among Hessian matrices of local objective functions, shown by D, can be small, the eigenvalues of the individual Hessian matrix, shown by $\tilde{L}$ can still be very large.

Explanation of D. Assumption 4 captures the difference between the averaged model and centralized model, which can be seen from the following proposition. At the kth iteration of the rth round, we consider the virtual averaged model in (12) and the centralized model in (13).

Proposition 3. *Given the virtual averaged model at the rth round and kth iteration $\hat{\mathbf{x}}^{r,k}$, we have*

$$\left\|\mathbb{E}[\hat{\mathbf{x}}^{r,k+1}|\hat{\mathbf{x}}^{r,k}] - \mathbb{E}[\mathbf{x}_c^{r,k+1}|\hat{\mathbf{x}}^{r,k}]\right\|^2 \leq \gamma^2 \cdot \tfrac{D^2}{N} \textstyle\sum_{i=1}^N \left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2. \tag{17}$$

Proposition 3 shows that although the difference among local models, captured by $\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2$ (depends on both ζ and σ), can be large after multiple local iterations, the difference between the averaged model and centralized model can still be small if D is small. This means that while the variance among local models depends on ζ and σ, D determines how the averaged model is affected by this variance among local models, which is consistent with the theoretical results in Theorem 1. Now we show that when $D = 0$, ζ can be arbitrarily large.

Proposition 4. *For quadratic objective functions defined in (16), when $\zeta = 0$, Assumption 4 holds with $D = 0$, while when $D = 0$, ζ can be arbitrarily large.*

Proposition 4 shows that $D = 0$ 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 that as we have seen in Proposition 2, for quadratic objective functions, the key effect of heterogeneity on the local updates is shown on 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 $D = 0$ but $\zeta > 0$ is possible in practice.

Extended discussion about Local SGD v.s. Mini-batch SGD. In the following theorem, we consider the case of $D = 0$, by which we show that local SGD can outperform mini-batch SGD even when ζ is arbitrarily large. Instead of directly applying $D = 0$ to Theorem 1, we develop a new technique for Theorem 3. The difference on the techniques can be shown by the requirement on the learning rate, which no longer depends on I while in Theorem 1, it depends on I. In Theorem 1, it is shown that when $D = 0$, two-sided learning rates do not have advantage over a single learning rate for non-convex objective functions. Without loss of generality, we consider $\eta = 1$ in the following.

Theorem 3 (Special Case of $D = 0$)**.** *For quadratic objective functions defined in (16), with a common Hessian $\mathbf{A} = \mathbf{A}_i, \forall i$, if $\gamma \leq \frac{1}{|\lambda(\mathbf{A})|}$ and $\eta = 1$, for local SGD with I local iterations,*

$$\min_{r\in[R],k\in[I]} \mathbb{E}\left[\left\|\nabla f(\hat{\mathbf{x}}^{r,k})\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{F}}{\gamma RI} + \frac{\gamma L}{N}\sigma^2\right); \tag{18}$$

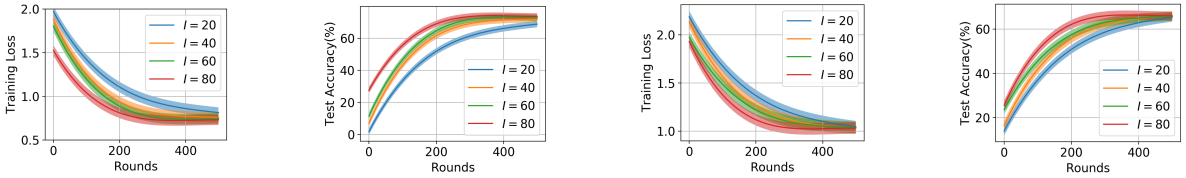
for mini-batch SGD with the batch size I,

$$\min_{r\in[R],k\in[I]} \mathbb{E}\left[\left\|\nabla f(\hat{\mathbf{x}}^{r,k})\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{F}}{\gamma R} + \frac{\gamma L}{NI}\sigma^2\right). \tag{19}$$

A fair comparison between local SGD and mini-batch SGD. In Theorem 3, the cost of communication and computation is the same for both local SGD and mini-batch SGD since the number of aggregations is R and the total number of gradients sampled is NRI for both algorithms. The restriction for the learning rate is also the same. Comparing (18) with (19), we see that the difference is on the place where I appears. For local SGD, I is in the first term of (18), which means that local SGD uses more computation to reduce the error caused by the initialization. For mini-batch SGD, I is in the second term of (19), which means that mini-batch SGD uses more computation to reduce the error caused by the stochastic variance. When the stochastic variance σ^2 is small (i.e., $\sigma^2 \to 0$), the first terms of (18) and (19) dominate, and it becomes beneficial to choose γ to be as large as possible, so we can choose $\gamma = \frac{1}{|\lambda(\mathbf{A})|}$ for both cases. Then, as $\sigma^2 \to 0$, the convergence rate of local SGD goes to $\mathcal{O}(\frac{1}{RI})$ while the convergence rate of mini-batch SGD goes to $\mathcal{O}(\frac{1}{R})$. This implies that when σ^2 is small, the speed of convergence for local SGD can be much faster than that for mini-batch SGD, which will also be validated in the experiments in the next section.

Table 2: Estimated D, $\tilde{L}$, L for a CNN model trained with the CIFAR-10 dataset.

Obj. Function	CNN			
Heterogeneity	25%	50%	75%	100%
$\tilde{L}$	447.59	898.49	1131.36	1662.24
D	0.96	1.21	1.63	2.15
L	323.35	323.35	323.35	323.35



(a) Training-50%. *(b)* Test-50%. *(c)* Training-75%. *(d)* Training-75%.

Figure 2: Results with CNN. The dataset is CIFAR-10. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.05$. Results for 50% of the heterogeneous data are shown in (a) and (b). Results for 75% of the heterogeneous data are shown in (c) and (d).

D. Experiments

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. Then we uniformly partition the remaining dataset into all workers and we say that the percentage of heterogeneous data on this worker is $X\%$.

Results with MNIST dataset. In Table 1, a two-layer neural network with cross-entropy loss and a linear regression model with mean squared error (MSE) is trained with the MNIST dataset (LeCun et al., 1998). The MNIST dataset is partitioned into 10 workers.

Results with CIFAR-10 dataset. A CNN model with cross-entropy loss is trained with the CIFAR-10 dataset (Krizhevsky & Hinton, 2009). The CIFAR-10 dataset is partitioned into 100 workers, and we randomly sample 10 workers in each round. The results for the general non-convex functions with partial participation are shown in Table 2 and Figure 2. In Table 2, it can be seen that D is far smaller than $\tilde{L}$. In Corollary 2, it is shown that when D is small, increasing I can reduce the convergence error. This is validated by experimental results in Figure 2. It can be observed that for both 50% and 75% of heterogeneous data, $I = 80$ is the best curve and increasing I can accelerate the convergence.

Results with synthetic data. For the special case of $D = 0$, we construct quadratic examples to validate the insights from Theorem 3. We construct the objective function as $F_i(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{U}\mathbf{x} - \mathbf{v}_i\right\|^2$, where $\mathbf{U} \in \mathbb{R}^{100\times 100}$, $\mathbf{v}_i \in \mathbb{R}^{100}$. Each column of $\mathbf{U}$ and $\mathbf{v}_i$ are sampled from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In this case, the gradient divergence is $\left\|\mathbf{U}(\mathbf{v}_i - \mathbf{v})\right\|^2 > 0$. Table 3 shows the results for quadratic objective functions. 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 1 shows that when $D = 0$, using two-sided learning rates does not have advantages over a single learning rate. This is validated by the experiments shown in Table 3, where there is no difference among the results with different learning rates when keeping the product of learning rates. Comparing results with $I = 1$, $I = 5$, and $I = 10$ with $s = 1$ in Table 3, it can be seen that more local updates can reduce the communication cost, which validates the results in Theorem 3. 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, which validates the discussion for Theorem 3.

Table 3: Special case of $D = 0$ with the quadratic objective functions. $I = 1$ is equivalent to mini-batch SGD. The number of rounds is the communication rounds needed for achieving a target function value of 0.8. For (η, γ), we fix $I = 10$ and for (I, s), we fix $\eta = 1$, $\gamma = 0.005$.

(η, γ)	$(1, 0.005)$	$(2, 0.0025)$	$(5, 0.001)$	$(10, 0.0005)$
Number of Rounds	86 ± 1	86 ± 1	86 ± 1	86 ± 1
(I, s)	$(1, 1)\&(1, 5)$	$(1, 10)$	$(5, 1)$	$(10, 1)$
Number of Rounds	927 ± 3	925 ± 1	187 ± 2	95 ± 2

E. Proofs

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

Algorithm 1: FedAvg with two-sided learning rates

Input: γ, η, $\bar{\mathbf{x}}^0$, 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
      $\mathbf{x}_i^{r,k+1} \leftarrow \mathbf{x}_i^{r,k} - \gamma \mathbf{g}_i(\mathbf{x}_i^{r,k})$;
      $k \leftarrow k + 1$;
    Send $\Delta_i^r \leftarrow \bar{\mathbf{x}}^r - \mathbf{x}_i^{r,I}$ to the server;
  `/* Global Update Phase */`
  Update the global model
  $\bar{\mathbf{x}}^{r+1} \leftarrow \bar{\mathbf{x}}^r - \eta \cdot \frac{1}{M} \sum_{i \in \mathcal{S}_r} \Delta_i^r$;

E.1. 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 1 (Local Gradient Deviation)**.** *With Assumption 2, 3 and 4, we have*

$$\frac{1}{N} \sum_{j=1}^N \left\| \frac{1}{N} \sum_{i=1}^N \nabla F_i(\mathbf{x}_i) - \nabla F_j(\mathbf{x}_j) \right\|^2 \leq 3(D^2 + L^2) \cdot \frac{1}{N} \sum_{j=1}^N \|\bar{\mathbf{x}} - \mathbf{x}_j\|^2 + 3\zeta^2. \tag{20}$$

Lemma 2 (Model Divergence)**.** *With* $\gamma \leq \frac{1}{30(D+L)I}$*, we have*

$$\sum_{k=0}^{I-1} \frac{D^2}{N} \sum_{i=1}^N \mathbb{E} \left\| \mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k} \right\|^2 \leq 3c(I-1)^3 \gamma^2 D^2 \zeta^2 + c(I-1)^2 \gamma^2 D^2 \sigma^2, \tag{21}$$

where $c = 3$ and $\hat{\mathbf{x}}^{r,k} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{r,k}$.

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

$$\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^{r}\right\|^2 \leq 5(I-1)\cdot\frac{\gamma^2\sigma^2}{N}+30I\gamma^2\sum_{k=0}^{I-1}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2+30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2. \tag{22}$$

E.2. Proof of Lemma 1

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

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i)-\nabla F_j\left(\mathbf{x}_j\right)\right\|^2\\
&=\frac{1}{N}\sum_{j=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i)-\nabla f(\bar{\mathbf{x}})+\nabla f(\bar{\mathbf{x}})-\nabla f(\mathbf{x}_j)+\nabla f(\mathbf{x}_j)-\nabla F_j\left(\mathbf{x}_j\right)\right\|^2\\
&\leq 3\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i)-\nabla f(\bar{\mathbf{x}})\right\|^2+3\cdot\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f(\bar{\mathbf{x}})-\nabla f(\mathbf{x}_j)\right\|^2+3\cdot\frac{1}{N}\sum_{j=1}^{N}\left\|\nabla f(\mathbf{x}_j)-\nabla F_j\left(\mathbf{x}_j\right)\right\|^2\\
&\overset{(a)}{\leq} 3\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i)-\nabla f(\bar{\mathbf{x}})\right\|^2+3L^2\cdot\frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}}-\mathbf{x}_j\right\|^2+3\zeta^2\\
&\overset{(b)}{\leq} 3\cdot\frac{D^2}{N}\sum_{i=1}^{N}\left\|\bar{\mathbf{x}}-\mathbf{x}_i\right\|^2+3L^2\cdot\frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}}-\mathbf{x}_j\right\|^2+3\zeta^2\\
&=3(D^2+L^2)\cdot\frac{1}{N}\sum_{j=1}^{N}\left\|\bar{\mathbf{x}}-\mathbf{x}_j\right\|^2+3\zeta^2,
\end{aligned}\tag{23}$$

where (a) is due to Assumption 2 and 3 and (b) is due to Assumption 4.

E.3. Proof of Lemma 2

At rth round, we have

$$\begin{aligned}
&\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\\
&=\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j(\mathbf{x}_j^{r,m})\right)\right\|^2\\
&=\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m})-\nabla F_i(\mathbf{x}_i^{r,m})+\nabla F_i(\mathbf{x}_i^{r,m})\right.\right.\\
&\quad\left.\left.-\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})+\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j(\mathbf{x}_j^{r,m})\right)\right\|^2\\
&\leq 2\cdot\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\nabla F_i(\mathbf{x}_i^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})\right)\right\|^2\\
&+2\cdot\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m})-\nabla F_i(\mathbf{x}_i^{r,m})+\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j(\mathbf{x}_j^{r,m})\right)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} 2\cdot\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\nabla F_i(\mathbf{x}_i^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})\right)\right\|^2\\
&\quad+2\cdot\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_i(\mathbf{x}_i^{r,m})-\nabla F_i(\mathbf{x}_i^{r,m})\right)\right\|^2\\
&\leq 2\cdot\frac{\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{m=0}^{k-1}\left(\nabla F_i(\mathbf{x}_i^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})\right)\right\|^2+2\gamma^2D^2k\sigma^2\\
&\leq 2k\cdot\frac{\gamma^2D^2}{N}\cdot\sum_{i=1}^{N}\sum_{m=0}^{k-1}\mathbb{E}\left\|\nabla F_i(\mathbf{x}_i^{r,m})-\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{r,m})\right\|^2+2\gamma^2D^2k\sigma^2\\
&\overset{(b)}{\leq} 2k\gamma^2D^2\sum_{m=0}^{k-1}\left(3(D^2+L^2)\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,m}-\mathbf{x}_k^{r,m}\right\|^2+3\zeta^2\right)+2\gamma^2D^2k\sigma^2\\
&=6k\gamma^2D^2(D^2+L^2)\sum_{m=0}^{k-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,m}-\mathbf{x}_i^{r,m}\right\|^2+6k^2\gamma^2D^2\zeta^2+2\gamma^2D^2k\sigma^2,
\end{aligned}\tag{24}
$$

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

$$
\begin{aligned}
&\sum_{k=1}^{I}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\\
&\leq\sum_{k=1}^{I}\left(6k\gamma^2D^2(D^2+L^2)\sum_{m=0}^{k-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,m}-\mathbf{x}_i^{r,m}\right\|^2+6k^2\gamma^2D^2\zeta^2+2\gamma^2D^2k\sigma^2\right)\\
&\leq 3\gamma^2D^2(D^2+L^2)I(I-1)\sum_{m=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,m}-\hat{\mathbf{x}}^{r,m}\right\|^2+6(I-1)^3\gamma^2D^2\zeta^2+2(I-1)^2\gamma^2D^2\sigma^2.
\end{aligned}\tag{25}
$$

Move the first term on RHS of (25) to LHS, we have

$$
\left(D^2-3\gamma^2D^2(D^2+L^2)I(I-1)\right)\sum_{k=0}^{I-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\leq 6(I-1)^3\gamma^2D^2\zeta^2+2(I-1)^2\gamma^2D^2\sigma^2.\tag{26}
$$

With $\gamma\leq\frac{1}{30(D+L)I}$, we have

$$
D^2-3\gamma^2D^2(D^2+L^2)I(I-1)>0.\tag{27}
$$

Since $\frac{2}{1-3\gamma^2(D^2+L^2)I(I-1)}<3$, we can choose $c=3$ such that

$$
\sum_{k=0}^{I-1}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\leq 3c(I-1)^3\gamma^2D^2\zeta^2+c(I-1)^2\gamma^2D^2\sigma^2.\tag{28}
$$

E.4. Proof of Lemma 3

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

$$
\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^{r}\right\|^2=0.\tag{29}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^{r}\right\|^{2} \\
&=\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k-1}-\frac{\gamma}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^{r,k-1})-\bar{\mathbf{x}}^{r}\right\|^{2} \\
&=\mathbb{E}\Bigg\|\hat{\mathbf{x}}^{r,k-1}-\bar{\mathbf{x}}^{r}-\gamma\Bigg(\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^{r,k-1})-\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k-1})+\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k-1}) \\
&\qquad -\nabla f(\hat{\mathbf{x}}^{r,k-1})+\nabla f(\hat{\mathbf{x}}^{r,k-1})-\nabla f(\bar{\mathbf{x}}^{r})+\nabla f(\bar{\mathbf{x}}^{r})\Bigg)\Bigg\|^{2} \\
&\overset{(a)}{\leq}\left(1+\frac{1}{2I-1}\right)\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k-1}-\bar{\mathbf{x}}^{r}\right\|^{2}+\frac{\gamma^2\sigma^2}{N}+6I\gamma^2\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k-1})-\nabla f(\hat{\mathbf{x}}^{r,k-1})\right\|^{2} \\
&\qquad +6I\gamma^2\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^{r,k-1})-\nabla f(\bar{\mathbf{x}}^{r})\right\|^{2}+6I\gamma^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{r})\|^{2} \qquad (30)\\
&\overset{(b)}{\leq}\left(1+\frac{1}{2I-1}+6I\gamma^2L^2\right)\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k-1}-\bar{\mathbf{x}}^{r}\right\|^{2}+\frac{\gamma^2\sigma^2}{N}+\frac{6I\gamma^2D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k-1}-\hat{\mathbf{x}}^{r,k-1}\right\|^{2} \\
&\qquad +6I\gamma^2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{r})\|^{2} \qquad (31)\\
&\overset{(c)}{\leq} 5(I-1)\cdot\frac{\gamma^2\sigma^2}{N}+30I\gamma^2\sum_{k=0}^{I-1}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^{2}+30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{r})\right\|^{2}, \qquad (32)
\end{aligned}
$$

where (a) is due to that $\|\mathbf{x}+\mathbf{y}\|^2 \leq (1+p)\|\mathbf{x}\|^2+(1+\frac{1}{p})\|\mathbf{y}\|^2, \forall p>0$, (b) is due to Assumption 3 and 4 and (c) is due to $(1+\frac{1}{q})^q < e, \forall q>0$, where e is the natural exponent.

E.5. Proof of Theorem 1

With Assumption 3, we have

$$
\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] &\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^{r}),\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\rangle+\frac{\gamma^2\eta^2L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^{2} \\
&=\mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^{r}),\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle+\frac{\gamma^2\eta^2L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^{2}. \qquad (33)
\end{aligned}
$$

The second term in the RHS of (33) can be computed as follows.

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

$$
\begin{aligned}
&\quad -I^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\Bigg\} \\
&\leq \frac{\gamma\eta}{2I}\Bigg\{2I\sum_{k=0}^{I-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}) - \nabla f(\hat{\mathbf{x}}^{r,k})\right\|^2 + 2I\sum_{k=0}^{I-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^{r,k}) - \nabla f(\bar{\mathbf{x}}^r)\right\|^2 \\
&\quad -I^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\Bigg\} \\
&\overset{(a)}{\leq} \frac{\gamma\eta}{2I}\Bigg\{\frac{2ID^2}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 + 2IL^2\sum_{k=0}^{I-1}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k} - \bar{\mathbf{x}}^r\right\|^2 \\
&\quad -I^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\Bigg\},
\end{aligned} \tag{34}
$$

where (a) is due to Assumption 3 and Assumption 4.

The third term in the RHS of (33) can be computed as follows.

$$
\begin{aligned}
&\frac{\gamma^2\eta^2 L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2 \\
&= \frac{\gamma^2\eta^2 L}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 + \frac{\gamma^2\eta^2 L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\left[\mathbf{g}_i(\mathbf{x}_i^{r,k}) - \nabla F_i(\mathbf{x}_i^{r,k})\right]\right\|^2 \\
&\leq \frac{\gamma^2\eta^2 L}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 + \frac{\gamma^2\eta^2 IL\sigma^2}{2N}.
\end{aligned} \tag{35}
$$

Substitute (34) and (35) to (33), we have

$$
\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] \\
&\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta D^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^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\eta}{2}\left(\frac{1}{I} - \gamma\eta L\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 IL\sigma^2}{2N} \\
&\overset{(a)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta D^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^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 IL\sigma^2}{2N} \\
&\overset{(b)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + \frac{\gamma\eta D^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 IL\sigma^2}{2N} \\
&+ \gamma\eta L^2 I\left(5(I-1)\frac{\gamma^2\sigma^2}{N} + 30I\gamma^2\sum_{k=0}^{I-1}\frac{D^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) \\
&\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + 2\gamma\eta\sum_{k=0}^{I-1}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k} - \hat{\mathbf{x}}^{r,k}\right\|^2 - \frac{\gamma\eta I}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 IL\sigma^2}{N} \\
&\overset{(c)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] + 2\gamma\eta\left[3c(I-1)^3\gamma^2 D^2\zeta^2 + c(I-1)^2\gamma^2 D^2\sigma^2\right] - \frac{\gamma\eta I}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 + \frac{\gamma^2\eta^2 IL\sigma^2}{N},
\end{aligned} \tag{36}
$$

where (a) is due to $\gamma\eta < \frac{1}{4IL}$, (b) is due to Lemma 3 and (c) is due to Lemma 2.

Rearrange the above inequality and average over r, we obtan

$$\min_{r\in[R]} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 \leq \frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 \leq \frac{4[f(\mathbf{x}^0)-f^*]}{\gamma\eta RI} + \frac{\gamma\eta L\sigma^2}{N} + 24c\gamma^2D^2(I-1)^2\zeta^2 + 8c\gamma^2D^2(I-1)\sigma^2. \tag{37}$$

Then we have

$$\min_{r\in[R]} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 = \mathcal{O}\left(\frac{f(\mathbf{x}^0)-f^*}{\gamma\eta RI} + \frac{\gamma\eta L\sigma^2}{N} + \gamma^2D^2(I-1)^2\zeta^2 + \gamma^2D^2(I-1)\sigma^2\right). \tag{38}$$

E.6. Proof of Theorem 2

Consider the partial participation shown in Algorithm 1. In each round, M workers are uniformly sampled with replacement. Then $\forall r, k$, we have

$$\mathbb{E}_{\mathcal{S}_r}\left[\frac{1}{M}\sum_{j\in\mathcal{S}_r}\nabla F_j(\mathbf{x}_j^{r,k})\right] = \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_j^{r,k}). \tag{39}$$

With Assumption 3, after one round of FedAvg, we have

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right] &\leq \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{M}\sum_{i\in\mathcal{S}_r}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\rangle + \frac{\gamma^2\eta^2L}{2}\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2 \\
&= \mathbb{E}\left[f(\bar{\mathbf{x}}^r)\right] - \gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle + \frac{\gamma^2\eta^2L}{2}\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2.
\end{aligned} \tag{40}$$

It can be seen that the inner-product term is the same as that in (34). So we have

$$\begin{aligned}
&-\gamma\eta\mathbb{E}\left\langle\nabla f(\bar{\mathbf{x}}^r), \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\rangle \\
&\leq \frac{\gamma\eta}{2I}\left\{\frac{2ID^2}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2 + 2IL^2\sum_{k=0}^{I-1}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^r\right\|^2\right. \\
&\left.-I^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\right\}.
\end{aligned} \tag{41}$$

In this case, $\mathbf{x}_i^{r,k}, i\notin\mathcal{S}_r$ is the virtual local model on worker i, which cannot be seen in the system. The virtual local model in mainly used for theoretical analysis. For the third term in the RHS of (40), we have

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

For simplicity, we use Q_i to denote the sum of expected gradients of worker i during rth round in the following. Then for the second term in the RHS of (42), we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2 &= \mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i - \frac{1}{N}\sum_{j=1}^{N}Q_j + \frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2 \\
&\overset{(a)}{=} \mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i - \frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2 + \mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2,
\end{aligned} \tag{43}$$

where (a) is due to $\mathbb{E}_{\mathcal{S}_r}\left[\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i\right] = \frac{1}{N}\sum_{j=1}^{N}Q_j$ by (39). Further we have

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i - \frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2 \\
=&\mathbb{E}\left[\frac{1}{M^2}\sum_{i\in\mathcal{S}_r}\left\|Q_i - \frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2 + \frac{1}{M^2}\sum_{i,j\in\mathcal{S}_r, i\neq j}\left\langle Q_i - \frac{1}{N}\sum_{m=1}^{N}Q_m, Q_j - \frac{1}{N}\sum_{m=1}^{N}Q_m\right\rangle\right] \\
\overset{(a)}{=}&\frac{1}{M^2}\sum_{i\in\mathcal{S}_r}\mathbb{E}\left\|Q_i - \frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2 \\
=&\frac{1}{M^2}\sum_{i\in\mathcal{S}_r}\left[\mathbb{E}\left\|Q_i\right\|^2 - \mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2\right] \\
=&\frac{1}{MN}\sum_{i=1}^{N}\mathbb{E}\left\|Q_i\right\|^2 - \frac{1}{M}\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2,
\end{aligned} \tag{44}$$

where (a) is due to that the sampling is with replacement so ith sampling and jth sampling are independent. Then we have

$$\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}Q_i\right\|^2 = \frac{1}{MN}\sum_{i=1}^{N}\mathbb{E}\left\|Q_i\right\|^2 + \frac{M-1}{M}\mathbb{E}\left\|\frac{1}{N}\sum_{j=1}^{N}Q_j\right\|^2. \tag{45}$$

Substituting above results back to (42), we obtain

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

For the second term of (46), we have

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

where (a) is due to Assumption 2 and Assumption 3. Substituting back and rearranging, we have

$$\begin{aligned}\frac{\gamma^2\eta^2 L}{2}\mathbb{E}\left\|\frac{1}{M}\sum_{i\in\mathcal{S}_r}\sum_{k=0}^{I-1}\mathbf{g}_i(\mathbf{x}_i^{r,k})\right\|^2 \leq{}& \frac{\gamma^2\eta^2 LI\sigma^2}{2M}+\frac{\gamma^2\eta^2 L(M-1)}{2M}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\\ &+\frac{2\gamma^2\eta^2 LI^2\zeta^2}{M}+\frac{2\gamma^2\eta^2 L^3 I}{MN}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\\ &+\frac{2\gamma^2\eta^2 L^3 I}{MN}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^{r}\right\|^2+\frac{2\gamma^2\eta^2 LI^2}{M}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2. \end{aligned}\tag{48}$$

Substituting all terms back to (40), we have

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right]\leq{}&\mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\left(\frac{\gamma\eta I}{2}-\frac{2\gamma^2\eta^2 LI^2}{M}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\\ &-\left(\frac{\gamma\eta}{2I}-\frac{\gamma^2\eta^2 L(M-1)}{2M}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{I-1}\nabla F_i(\mathbf{x}_i^{r,k})\right\|^2\\ &+\frac{\gamma^2\eta^2 LI\sigma^2}{2M}+\frac{2\gamma^2\eta^2 LI^2\zeta^2}{M}+\left(\gamma\eta D^2+\frac{2\gamma^2\eta^2 L^3 I}{M}\right)\cdot\frac{1}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\\ &+\left(\gamma\eta L^2+\frac{2\gamma^2\eta^2 L^3 I}{M}\right)\sum_{k=0}^{I-1}\mathbb{E}\left\|\hat{\mathbf{x}}^{r,k}-\bar{\mathbf{x}}^{r}\right\|^2. \end{aligned}\tag{49}$$

With $\gamma\eta\leq\frac{1}{4IL}$ and Lemma 3, we have

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right]\leq{}&\mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\left(\frac{\gamma\eta I}{2}-\frac{2\gamma^2\eta^2 LI^2}{M}\right)\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2+\frac{\gamma^2\eta^2 LI\sigma^2}{2M}+\frac{2\gamma^2\eta^2 LI^2\zeta^2}{M}\\ &+\left(\gamma\eta D^2+\frac{2\gamma^2\eta^2 L^3 I}{M}\right)\cdot\frac{1}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2\\ &+\left(\gamma\eta IL^2+\frac{2\gamma^2\eta^2 L^3 I^2}{M}\right)\\ &\cdot\left(5(I-1)\cdot\frac{\gamma^2\sigma^2}{N}+30I\gamma^2\sum_{k=0}^{I-1}\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2+30I(I-1)\gamma^2\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\right). \end{aligned}\tag{50}$$

With $\gamma\eta\leq\frac{1}{4IL}$ and $\gamma<\frac{1}{30(L+D)I}$, we have

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right]\leq{}&\mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\frac{\gamma\eta I}{8}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2+\frac{\gamma^2\eta^2 LI\sigma^2}{2M}+\frac{2\gamma^2\eta^2 LI^2\zeta^2}{M}+\frac{\gamma\eta\sigma^2}{N}\\ &+\left(2\gamma\eta D^2+\frac{2\gamma^2\eta^2 L^3 I}{M}\right)\cdot\frac{1}{N}\sum_{k=0}^{I-1}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2. \end{aligned}\tag{51}$$

With Lemma 2, we have

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}^{r+1})\right]\leq{}&\mathbb{E}\left[f(\bar{\mathbf{x}}^{r})\right]-\frac{\gamma\eta I}{8}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2+\frac{\gamma^2\eta^2 LI\sigma^2}{2M}+\frac{2\gamma^2\eta^2 LI^2\zeta^2}{M}+\frac{\gamma\eta\sigma^2}{N}\\ &+\left(2\gamma\eta D^2+\frac{2\gamma^2\eta^2 L^3 I}{M}\right)\cdot\left(3c(I-1)^3\gamma^2\zeta^2+c(I-1)^2\gamma^2\sigma^2\right). \end{aligned}\tag{52}$$

Then we obtain

$$\min_{r\in[R]}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\leq\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2\leq\frac{8(f^0-f^*)}{\gamma\eta IR}+\frac{4\gamma\eta L\sigma^2}{M}+\frac{16\gamma\eta LI\zeta^2}{M}$$

$$+\left(16D^2+\frac{16\gamma\eta L^3 I}{M}\right)\cdot\left(3c(I-1)^2\gamma^2\zeta^2+c(I-1)\gamma^2\sigma^2\right). \tag{53}$$

Rearrange,

$$\min_{r\in[R]}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^r)\right\|^2=\mathcal{O}\left(\frac{(f^0-f^*)}{\gamma\eta IR}+\frac{\gamma\eta L\sigma^2}{M}+\frac{\gamma\eta LI\zeta^2}{M}+\gamma^2D^2(I-1)\sigma^2+\gamma^2D^2(I-1)^2\zeta^2\right). \tag{54}$$

E.7. Proof of Proposition 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 5 implies Assumption 3 holds by choosing $L=\tilde{L}$.

Second, we can see that

$$\begin{aligned}\left\|\frac{1}{N}\sum_{i=1}^N\nabla F_i(\mathbf{x}_i)-\nabla f\left(\bar{\mathbf{x}}\right)\right\|^2&=\left\|\frac{1}{N}\sum_{i=1}^N[]\nabla F_i(\mathbf{x}_i)-\nabla F_i\left(\bar{\mathbf{x}}\right)]\right\|^2\\&\leq\frac{1}{N}\sum_{i=1}^N\left\|\nabla F_i(\mathbf{x}_i)-\nabla F_i\left(\bar{\mathbf{x}}\right)\right\|^2\\&\overset{(a)}{\leq}\frac{\tilde{L}^2}{N}\sum_{i=1}^N\left\|\mathbf{x}_i-\bar{\mathbf{x}}\right\|^2,\end{aligned} \tag{55}$$

where (a) is due to Assumption 5. By choosing $D=\tilde{L}$, Assumption 4 holds. □

E.8. Proof of Proposition 2

For quadratic functions, we have

$$\nabla F_i(\mathbf{x})=\mathbf{A}_i\mathbf{x}+\mathbf{b}_i,\mathbf{x}\in\mathbb{R}^d. \tag{56}$$

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

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^N\nabla F_i(\mathbf{x}_i)-\nabla f\left(\bar{\mathbf{x}}\right)\right\|^2\\&=\left\|\frac{1}{N}\sum_{i=1}^N\left(\mathbf{A}_i\mathbf{x}_i+\mathbf{b}_i\right)-\left(\bar{\mathbf{A}}\bar{\mathbf{x}}+\bar{\mathbf{b}}\right)\right\|^2\\&=\left\|\frac{1}{N}\sum_{i=1}^N\mathbf{A}_i\mathbf{x}_i-\bar{\mathbf{A}}\bar{\mathbf{x}}\right\|^2\\&=\left\|\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left(\mathbf{A}_i\mathbf{x}_i-\mathbf{A}_i\mathbf{x}_j\right)\right\|^2\\&=\left\|\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left[\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)-\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\right]\right\|^2\\&\leq2\left\|\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|^2+2\left\|\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\right\|^2\\&\leq\frac{2}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left\|\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|^2+\frac{2}{N^2}\sum_{i=1}^N\sum_{j=1}^N\left\|\left(\mathbf{A}_i-\bar{\mathbf{A}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\right\|^2\end{aligned}$$

$$\overset{(a)}{\leq} \frac{2|\lambda_{\text{diff}}|^2_{\max}}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^2+\frac{2|\lambda_{\text{diff}}|^2_{\max}}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\mathbf{x}_j-\bar{\mathbf{x}}\|^2$$
$$\leq \frac{4|\lambda_{\text{diff}}|^2_{\max}}{N^2}\sum_{i=1}^{N}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^2, \tag{57}$$

where (a) is due to Cauchy's inequality and $|\lambda_{\text{diff}}| := \max_{i\in[N]}|\lambda(\mathbf{A}_i-\mathbf{A})|$.

E.9. Proof of Proposition 3

Recall that $\hat{\mathbf{x}}^{r,k}$ is the virtual averaged model defined in (12). During one local iteration, we have

$$\mathbb{E}[\hat{\mathbf{x}}^{r,k+1}|\hat{\mathbf{x}}^{r,k}] = \hat{\mathbf{x}}^{r,k}-\gamma\cdot\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k}). \tag{58}$$

Using (13), if we use centralized update at this iteration, we have

$$\mathbb{E}[\mathbf{x}_c^{r,k+1}|\hat{\mathbf{x}}^{r,k}] = \hat{\mathbf{x}}^{r,k}-\gamma\nabla f(\hat{\mathbf{x}}^{r,k}). \tag{59}$$

Using Assumption 4, we obtain

$$\left\|\mathbb{E}[\hat{\mathbf{x}}^{r,k+1}|\hat{\mathbf{x}}^{r,k}]-\mathbb{E}[\mathbf{x}_c^{r,k+1}|\hat{\mathbf{x}}^{r,k}]\right\|^2 = \gamma^2\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^{r,k})-\nabla f\left(\hat{\mathbf{x}}^{r,k}\right)\right\|^2$$
$$\leq \gamma^2\cdot\frac{D^2}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i^{r,k}-\hat{\mathbf{x}}^{r,k}\right\|^2. \tag{60}$$

E.10. Proof of Theorem 3

It can be observed that for quadratic objective functions when $\mathbf{A}_i=\mathbf{A},\forall i$, we have $D=0$ and $L=|\lambda(\mathbf{A})|$. In this section, we use t to denote the index of the total number of iterations and $\hat{\mathbf{x}}^t$ is defined as

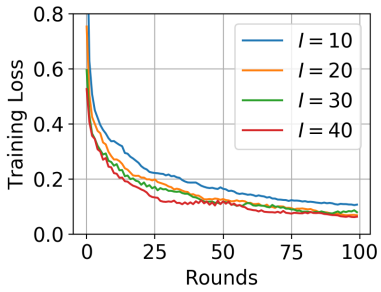
$$\hat{\mathbf{x}}^t = \begin{cases}\hat{\mathbf{x}}^{r,k}, t=rI+k, k\neq 0,\\ \bar{\mathbf{x}}^r, t=rI.\end{cases}$$

With Assumption 3, after one local iteration, we have

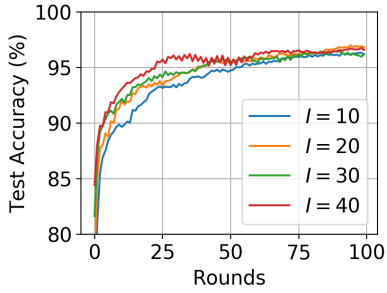
$$\mathbb{E}\left[f(\hat{\mathbf{x}}^{t+1})\right] \leq \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right]-\gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t),\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\rangle+\frac{\gamma^2L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2$$
$$= \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right]-\gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t),\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle+\frac{\gamma^2L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2. \tag{61}$$

For the second term in the RHS of (61), we have

$$-\gamma\mathbb{E}\left\langle\nabla f(\hat{\mathbf{x}}^t),\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\rangle$$
$$= \frac{\gamma}{2}\left(\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)-\nabla f(\hat{\mathbf{x}}^t)\right\|^2-\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2-\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2\right)$$
$$\leq \frac{\gamma}{2}\left(\frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t-\hat{\mathbf{x}}^t\right\|^2-\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2-\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2\right). \tag{62}$$



(a) Training Loss.



(b) Test Accuracy.

Figure 3: Results with MNIST dataset. The model is a two-layer neural network with the cross-entropy loss. The percentage of heterogeneous data is 50%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.1$.

For the third term of (61), we have

$$\frac{\gamma^2 L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\mathbf{x}_i^t)\right\|^2 \leq \frac{\gamma^2 L}{2}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L\sigma^2}{2N}. \tag{63}$$

Substitute (62) and (63) back to (61), we obtain

$$\begin{aligned}
&\mathbb{E}\left[f(\hat{\mathbf{x}}^{t+1})\right] \\
&\leq \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] + \frac{\gamma D^2}{2N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\gamma^2 L}{2}\right)\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{x}_i^t)\right\|^2 + \frac{\gamma^2 L\sigma^2}{2N} \\
&\overset{(a)}{\leq} \mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] + \frac{\gamma D^2}{2N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 - \frac{\gamma}{2}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 + \frac{\gamma^2 L\sigma^2}{2N},
\end{aligned} \tag{64}$$

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

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 &\leq \frac{2\mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - 2\mathbb{E}f(\hat{\mathbf{x}}^{t+1})}{\gamma} + \frac{D^2}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}_i^t - \hat{\mathbf{x}}^t\right\|^2 + \frac{\gamma L\sigma^2}{N} \\
&= \frac{2\mathbb{E}\left[f(\hat{\mathbf{x}}^t)\right] - 2\mathbb{E}f(\hat{\mathbf{x}}^{t+1})}{\gamma} + \frac{\gamma L\sigma^2}{N}.
\end{aligned} \tag{65}$$

Take the average over t on both sides, we obtain

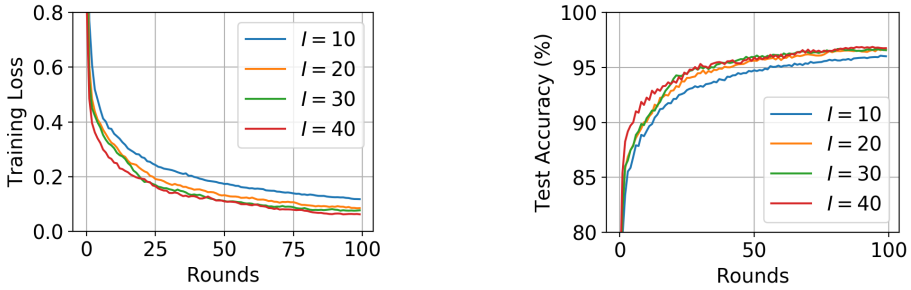
$$\min_{t\in[T]}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}}^t)\right\|^2 \leq \frac{2f(\hat{\mathbf{x}}^t) - 2f^*}{\gamma T} + \frac{\gamma L\sigma^2}{N}. \tag{66}$$

F. 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.

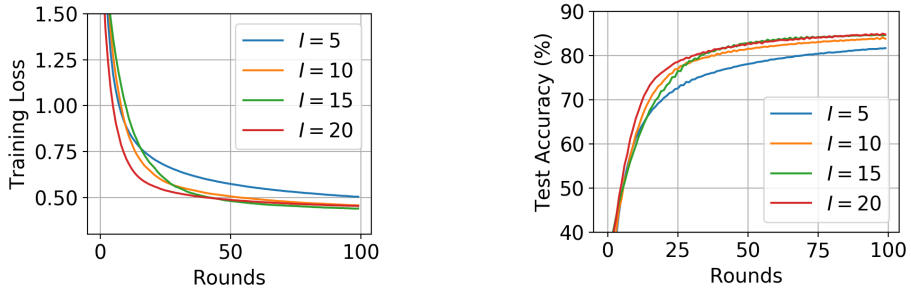
Environment. All our experiments are implemented in PyTorch and run on a server with four NVIDIA 2080Ti GPUs. The mini-batch size of SGD is 20. We run each experiment 5 times then plot their average.

Model. For experimental results with CIFAR-10 dataset in Section D, we use a CNN model. The structure of the CNN is $5 \times 5 \times 32$ Convolutional $\to 2 \times 2$ MaxPool $\to 5 \times 5 \times 32$ Convolutional $\to 2 \times 2$ MaxPool $\to 4096 \times 512$ Dense $\to 512 \times 128$ Dense $\to 128 \times 10$ Dense $\to$ 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.



(a) Training Loss. *(b)* Test Accuracy.

Figure 4: Results with MNIST dataset. The model is a two-layer neural network with the cross-entropy loss. The percentage of heterogeneous data is 75%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.1$.



(a) Training Loss. *(b)* Test Accuracy.

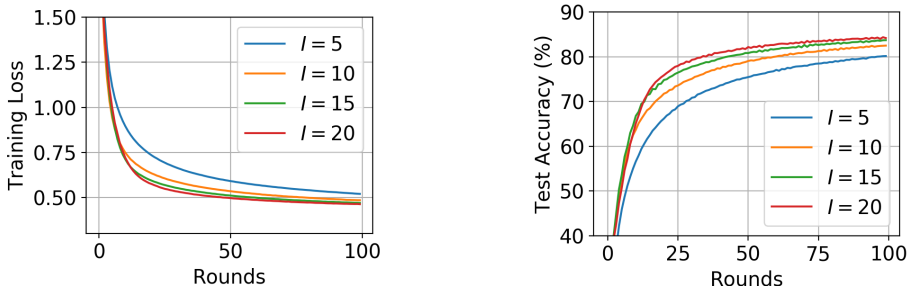
Figure 5: Results with MNIST dataset. The model is linear regression with the MSE loss. The percentage of heterogeneous data is 50%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.

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.

The estimate of D. 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 D using the following equations.

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

Starting from a global model that is close to convergence, we perform FedAvg for 10 rounds and estimate D^2 in each round. Then we use the averaged D^2 over 10 rounds as the estimate for D^2. The reason for starting from a global model that is



(a) Training Loss. *(b)* Test Accuracy.

Figure 6: Results with MNIST dataset. The model is linear regression with the MSE loss. The percentage of heterogeneous data is 75%. The learning rates are chosen as $\eta = 2$ and $\gamma = 0.01$.

close to convergence is that this can make the variance of the estimate smaller.

Additional Experimental Results. We partition the MNIST dataset into 10 workers. During each round, all workers will perform the local updates. Results with a two-layer neural network and the cross-entropy loss are shown in Figure 3 and 4. As shown in Table 1 of the main paper, D is very small in this case. In Corollary 1, with full participation, it is shown that when D is small, increasing I can improve the convergence even when data are highly heterogeneous. As shown in both Figure 3 and 4, the curve with the largest number of local iterations, $I = 40$, converges the fastest and achieves best accuracy, which validates Corollary 1. Results with linear regression and the MSE loss are shown in Figure 3 and 4. Since D and L are larger compared to that of the two-layer neural network, a smaller γ and smaller I's are chosen according to Corollary 1. It can be seen in both Figure 5 and 6, the curve with the largest number of local iterations, $I = 20$ converges the fastest and achieves the best accuracy.