Proof of Convergence of Grouped Federated Learning

Jiyao Liu
CIS Department (of Temple University)
Philadelphia, USA
jiyao.liu@temple.edu

***Abstract*—This is the final report for the course artificial intelligence. In this project, we investigated the grouped federated learning and prove its convergence. We mainly include how we reach the theoretical result. In the meantime, we also show the experimental data. The code can be found by the link in appendix. A how-to-run tutorial is included in the comments the of main.py file. Please be advised that some materials of this report is from my current research project, which is also the source of some materials submitted to some other courses. Those materials are modified from my own work.**

I. Team Members

1. Jiyao Liu

II. Introduction

Federated Learning [1] is proposed to preserve privacy and reduce communication cost when the data contributors do not want to share their data to the server, where the training process happens. The overall training is done in a iterative manner, each contains three steps: at the beginning of each iteration, server sends the model to all clients; then the clients train the model on their devices; finally, all clients upload the resulting models and server aggregates the models. This repeats until some preset criteria are met. Figure 1 [2] shows the typical workflow of a classic federated learning system.

Fig. 1. Federated Learning

In the edge computing scenario, many constraints may impose negative impact to federated learning. For a typical edge computing scenario, edge devices may suffer from low bandwidth, limited connectivity, weak computation ability, power limitation, etc. When the data contributors are edge devices, we must design new FL algorithms to avoid the negative effects. The aforementioned limitations in the edge computing scenario may pose obstacles in all the three steps of each training round. In the first step, model distribution, server may not select all clients to join this training round because of the limited connectivity: some devices may be unavailable at this time. Then, the bandwidth limitation requires edge devices communicate with server as few as possible, which means they need to do more local training before upload the model. This may backfire due because more local training means larger and distortion deviation from the global loss function, thus lower accuracy of the final model we get. When uploading the local models, edge devices may not be able to, and very likely cannot upload their models at the same time: they may own different number of training data, and have different computation ability, thus various time needed for each round. We can see in Figure 2 [3], a lot of time is wasted due to the asynchronous character of edge devices.

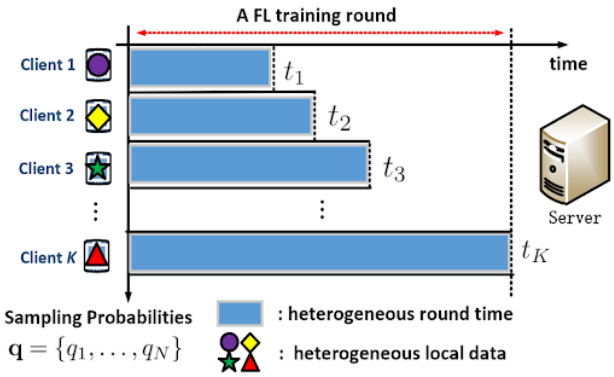

Fig. 2. Speed Discrepancy among Clients

Due to such systematical constraints, Hierarchical Federated Learning (HFL) is proposed to adjust FL to the edge environments. Figure 3 [4] shows the HFL structure. In HFL, local aggregators can be placed to the edge servers, and clients connected to the same edge server can be assigned to the same group. This simple method may solve many problems. for example, by allowing more local communication within an edge group, the global communication needed can be reduced.

Then, the statistical imbalance, i.e., the non-identical independent distribution (non-IID) data on clients. This is not a problem in classic machine learning process as all data are collected to the training machine. In federated learning, especially when clients are edge devices, one client may only possess partial categories of all data, which means that its local

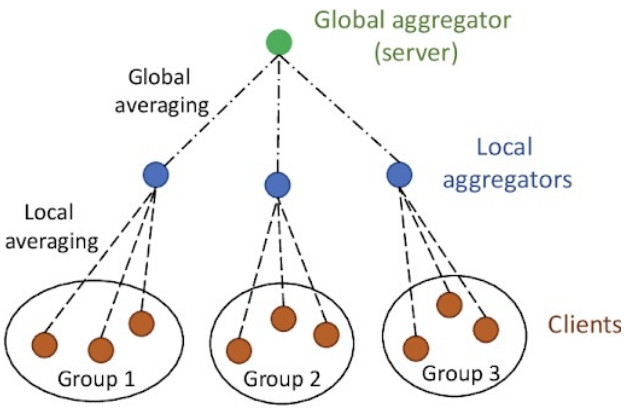

Fig. 3. Hierarchical Federated Learning

loss function is not identical to the global loss function. This makes the local optimizers have different optimize objectives, leading to worse result: typically the convergence speed is lower, sometimes the final testing accuracy also decreases, and in the worst case, the training may diverge. As we can see in the Figure 4, compared with the IID situation, when data distribution is non-IID, the convergence speed is severely impacted and the final testing accuracy decreases significantly.

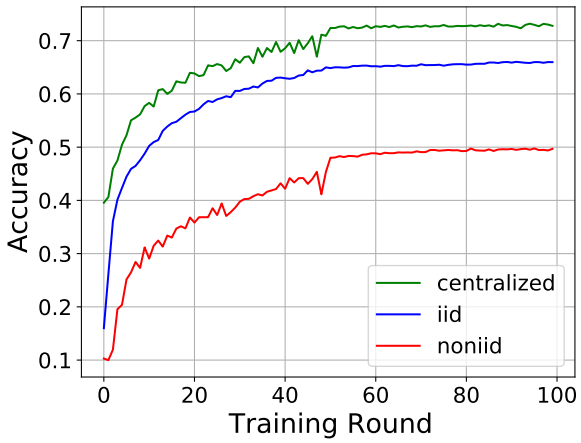

Fig. 4. IID v.s. Non-IID

Through initial experimental data, we find that grouping devices can reduce impact of the non-IID issue, as well as some systematic problems if properly designed. Further theoretical analysis is needed to make this project a comprehensive study, which is the key point of this report. We follow the main idea of the convergence proof for classic federated learning in [5], but turn it into a grouped version.

The remainder of this report is organized as the followings. Firstly, we introduce the common assumptions used in proof of convergence of federated learning. Then, we show the two lemmas and how we reach the final result using the two lemmas. After that, we also show how to prove the two lemmas. Finally, we give some experimental results to show that grouped federated learning do out perform the classic federated learning algorithm. Conclusion is also provided in the end of this paper.

III. Assumptions

For any type of federated learning, either those on IID data or non-IID data, we always assume all the local loss functions and the global loss function are smooth. By L-smoothness, we have

$$\mathbb{E}[f(\bar{\mathbf{x}}^t)] \leq \mathbb{E}[f(\bar{\mathbf{x}}^{t-1})] + \mathbb{E}[\langle \nabla f(\bar{\mathbf{x}}^{t-1}), \bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1} \rangle] + \frac{L}{2}\mathbb{E}[\|\bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\|] \tag{1}$$

For deep neural networks, they are usually not convex. For some other machine learning algorithms, like SVM, their loss function is convex. That means, for the models that have non-convex loss function, we usually cannot use the convexity assumption. Unfortunately, the most popular model in federated learning is exactly deep neural networks, which are not convex. However, because having convexity significantly reduce the complexity of proof of convergence, so many times we can still make this assumption if the proof is too hard. After we proof the convergence under the convexity assumption, we can next use experimental data to explain that the algorithm also works well for non-convex models. By μ-convex,

$$\mathbb{E}[f(\bar{\mathbf{x}}^t)] \geq \mathbb{E}[f(\bar{\mathbf{x}}^{t-1})] + \mathbb{E}[\langle \nabla f(\bar{\mathbf{x}}^{t-1}), \bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1} \rangle] \frac{\mu}{2}\mathbb{E}[\|\bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\|] \tag{2}$$

Because we do not use the full gradient method today, and use the mini-batch [6] instead, it introduces variance to the gradient. Fortunately, the variance is not too large and we can assume it is always bounded by a constant

$$\mathbb{E}_{\xi \sim \mathcal{D}_i}[\|\nabla F_i(\mathbf{x}; \xi) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2 \tag{3}$$

where ξ is the training data sampled by mini-batch and $\mathcal{D}_i$ is the data distribution on client i. Finally, the assumption about the non-IID data. If the data on clients are IID, then the local loss function is identical to the global loss function, and the gradient norm is bounded by a constant

$$\|\nabla f_i(\mathbf{x})\|^2 \leq G^2$$

however, when the data are non-IID, the gradient can be very large, and the above assumption does not apply any longer. In this case, we use another assumption

$$\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \zeta^2 \tag{4}$$

That means, although the gradient is not bounded, the deviation of local gradient is still bounded. This is what we adopt in our proof, instead of the previous one. The main reason is because that we want to demonstrate that the grouping method relieve the non-IID issue, so obviously we cannot use the assumption for the IID situation.

Finally, for grouped clients, within a group, we assume that their data are IID combined together. Then, by this grouping strategy,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_g(\mathbf{x})\|^2 \leq \kappa^2, \forall \mathbf{x}, \forall g \in \mathbb{G} \tag{5}$$

$$\|\nabla f_g(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \gamma^2 \tag{6}$$

where g is any group, $\mathbb{G}$ is the collection of all groups, f_g is the loss function within this group, and f is the global loss function. Under perfect grouping, $\kappa = \zeta$ and $\gamma = 0$.

IV. Main Result

Recall that we need to prove something like

$$\frac{1}{T}\sum_{t=1}^{T}(f^t(\bar{\mathbf{x}}) - f(\mathbf{x}^*)) \leq \frac{C_1}{T} + C_2$$

This means the loss of our trained model is tending to the optimal loss during the training. The final result we get is

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{TIE}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{e=1}^{E}F(\bar{\mathbf{x}}^{t,i,e}) - F(\mathbf{x}^*)\right]\\ &\leq \frac{\|\bar{\mathbf{x}}^{0,0,0} - \mathbf{x}^*\|^2}{2\eta TIE} + \frac{\eta\sigma^2}{N} + 25I^2E^2\eta^2L(\kappa+\gamma)^2\\ &\quad + 4IE\eta^2L\sigma^2\end{aligned} \tag{7}$$

where T, I, E are the total global, group, and local iteration numbers separately. N is the number of total clients. All other letters can be constants related to the training task. Then, we can see this result is consent to our expectation. It is a sketch here and latter we will see how we get it and what it means.

A. Lemmas

Lemma 1. Given assumptions 1-4,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{TIE}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{e=1}^{E}F(\bar{\mathbf{x}}^{t,i,e}) - F(\bar{\mathbf{x}}^*)\right]\\ &\leq \frac{1}{2\eta TIE}\|\bar{\mathbf{x}}^{(0,0,0)} - \mathbf{x}^*\|^2 + \frac{\eta\sigma^2}{N} +\\ &\frac{L}{NTIE}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{e=1}^{E}\mathbb{E}\left[\|\mathbf{x}_j^{(t,i,e)} - \bar{\mathbf{x}}^{(t,i,e)}\|^2\right]\end{aligned} \tag{8}$$

Lemma 2. Given assumptions 1-4,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}_j^{(t,i,e)} - \bar{\mathbf{x}}^{(t,i,e)}\|^2\right]\\ &\leq 25\eta^2I^2E^2(\kappa+\gamma)^2 + 4IE\eta^2\sigma^2\end{aligned} \tag{9}$$

B. Final Result

Theorem 1. Replacing lemma 2 back into lemma 1, we get theorem 1.

V. Proof of Lemmas

A. Proof of Lemma 1

According to the update scheme,

$$\bar{\mathbf{x}}^{(t,i,e+1)} = \bar{\mathbf{x}}^{t,i,e} - \eta\frac{1}{N}\sum_{j=1}^{N}\nabla F_j(\mathbf{x}_j^{t,i,e})$$

then,

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^{N}\left\langle\nabla F_j(\mathbf{x}_j^{t,i,e}), \bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}^*\right\rangle\\ &= \frac{1}{2\eta}(\|\bar{\mathbf{x}}^{t,i,e} - \mathbf{x}^*\|^2 - \|\bar{\mathbf{x}}^{(t,i,e+1)} - \bar{\mathbf{x}}^{(t,i,e)}\|^2\\ &\quad - \|\bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}^*\|^2)\end{aligned} \tag{10}$$

Here we simply expand the inner production.

$$\begin{aligned}&F_j(\bar{\mathbf{x}}^{(t,i,e+1)})\\ &\leq_a F_j(\mathbf{x}_j^{(t,i,e)}) + \left\langle\nabla F_i(\mathbf{x}_j^{(t,i,e)}), \bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\right\rangle\\ &\quad + \frac{L}{2}\|\bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\|^2\\ &\leq_b F_j(\mathbf{x}^*) + \left\langle\nabla F_j(\mathbf{x}_i^{(t,i,e)}), \bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\right\rangle\\ &\quad + \frac{L}{2}\|\bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\|^2\\ &\leq F_j(\mathbf{x}^*) + \left\langle\nabla F_j(\mathbf{x}_j^{(t,i,e)}), \bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\right\rangle\\ &\quad + L\|\bar{\mathbf{x}}^{(t,i,e+1)} - \mathbf{x}_j^{(t,i,e)}\|^2 + L\|\mathbf{x}_j^{(t,i,e)} - \bar{\mathbf{x}}^{(t,i,e)}\|^2\end{aligned} \tag{11}$$

Step a holds because of the L-smoothness assumption 1; step b is true simply because the loss at the optimal point is always smaller; step c keeps because we simply added a non-negative term.

Combine (10) and (11)

$$\begin{aligned}&f(\bar{\mathbf{x}}^{(t,i,e+1)}) - f(\bar{\mathbf{x}}^*)\\ &=_a \frac{1}{N}\sum_{j=1}^{N}(f_j(\bar{\mathbf{x}}^{(t,i,e+1)}) - f(\bar{\mathbf{x}}^*))\\ &\leq_b \frac{1}{N}\sum_{j=1}^{N}\left\langle\nabla f_i(\mathbf{x}_i^{(t,i,e)}) - \nabla F_i(\mathbf{x}_i^{(t,i,e)}), \bar{\mathbf{x}}^{(t,i,e+1)} - \bar{\mathbf{x}}^*)\right\rangle\end{aligned} \tag{12}$$

Step a follows by the definition of global loss function; step b keeps because of the L-smoothness assumption (1);

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\langle F_j(\mathbf{x}_j^{(t,i,e)})-\nabla F_j(\mathbf{x}_j^{(t,i,e)}),\bar{\mathbf{x}}^{(t,i,e+1)}-\mathbf{x}^*\right\rangle\right]\\
=_a\ &\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\langle F_j(\mathbf{x}_j^{(t,i,e)})-\nabla F_j(\mathbf{x}_j^{(t,i,e)}),\bar{\mathbf{x}}^{(t,i,e+1)}-\bar{\mathbf{x}}^{(t,i,e)}\right\rangle\right]\\
\leq_b\ &\eta\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\|F_j(\mathbf{x}_j^{(t,i,e)})-\nabla F_j(\mathbf{x}_j^{(t,i,e)})\|^2\right]\\
&+\frac{1}{4\eta}\mathbb{E}\left[\|\bar{\mathbf{x}}^{(t,i,e+1)}-\bar{\mathbf{x}}^{(t,i,e)}\|^2\right]\\
\leq\ &\frac{\eta\sigma^2}{N}+\frac{1}{4\eta}\mathbb{E}\left[\|\bar{\mathbf{x}}^{(t,i,e+1)}-\bar{\mathbf{x}}^{(t,i,e)}\|^2\right] \qquad (13)
\end{aligned}
$$

Step a follows because we know the expectation of the left hand of the inner production is 0. That means, we can replace the right hand by anything and the equation always stands; step b follows by the $Young$'s inequality; step c is true by introducing the bounded variance assumption (3). More details about the $Young$'s inequality are available in the appendix.

Plug the last inequality back to 12, we have

$$
\begin{aligned}
&f(\bar{\mathbf{x}}^{(t,i,e+1)})-f(\bar{\mathbf{x}}^*)\\
\leq\ &\frac{\eta\sigma^2}{N}-\frac{L}{N}\sum_{i=1}^{N}\|\mathbf{x}_i^{(t,i,e)}-\bar{x}^{(t,i,e)}\|^2 \qquad (14)
\end{aligned}
$$

Telescoping across all training rounds t, i, e,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{TIE}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{e=1}^{E}F(\bar{\mathbf{x}}^{t,i,e})-F(\mathbf{x}^*)\right]\\
&\leq\frac{1}{2\eta TIE}\left(\|\bar{\mathbf{x}}^{t,0,0}-\mathbf{x}^*\|^2-\mathbb{E}\left[\|\bar{\mathbf{x}}^{(t,i,e)}-\mathbf{x}^*\|^2\right]\right)\\
&+\frac{\eta\sigma^2}{N}+\frac{L}{NTIE}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{e=1}^{E}\mathbb{E}\left[\|\mathbf{x}_j^{(t,i,e)}-\bar{\mathbf{x}}^{(t,i,e)}\|^2\right] \qquad (15)
\end{aligned}
$$

B. Proof of Lemma 2

Suppose p, q are arbitrary two clients,

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_p^{(t,i,e+1)}-\mathbf{x}_q^{(t,i,e+1)}\|^2\right]\\
=_a\ &\mathbb{E}\left[\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\right.\\
&\left.-\eta\left(\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla F_q(\mathbf{x}_q^{(t,i,e)})\right)\|^2\right]\\
\leq_b\ &\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|^2\\
&-2\eta\langle\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla F_q(\mathbf{x}_q^{(t,i,e)}),\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\rangle\\
&+\eta^2\|\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla F_q(\mathbf{x}_q^{(t,i,e)})\|^2+2\eta^2\sigma^2 \qquad (16)
\end{aligned}
$$

Step a follows by the definition of local updates; step b expands the norm and introduce the bounded variance assumption (3).

$$
\begin{aligned}
&-\langle\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla F_q(\mathbf{x}_q^{(t,i,e)}),\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\rangle\\
\leq_a\ &-\langle\nabla f(\mathbf{x}_p^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)}),\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\rangle\\
&+2(\kappa+\gamma)\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|\\
\leq_b\ &-\frac{1}{L}\|\nabla f(\mathbf{x}_p^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)})\|^2\\
&+2(\kappa+\gamma)\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|\\
\leq_c\ &-\frac{1}{L}\|\nabla f(\mathbf{x}_p^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)})\|^2\\
&+\frac{1}{2\eta IE}\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|^2+2\eta IE(\kappa+\gamma)^2 \qquad (17)
\end{aligned}
$$

Step a follows by extracting the variance out; b follows by the smoothness assumption (1); step c follows by the AM-GM inequality, about which more details can be found in the appendix.

$$
\begin{aligned}
&\|\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla F_q(\mathbf{x}_q^{(t,i,e)})\|^2\\
=_a\ &\|\nabla F_p(\mathbf{x}_p^{(t,i,e)})-\nabla f_{gp}(\mathbf{x}_p^{(t,i,e)})+\nabla f_{gp}(\mathbf{x}_p^{(t,i,e)})\\
&-\nabla f(\mathbf{x}_p^{(t,i,e)})+\nabla f(\mathbf{x}_p^{(t,i,e)})\\
&-\nabla F_q(\mathbf{x}_q^{(t,i,e)})+\nabla f_{gq}(\mathbf{x}_q^{(t,i,e)})-\nabla f_{gq}(\mathbf{x}_q^{(t,i,e)})\\
&+\nabla f(\mathbf{x}_q^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)})\|^2\\
\leq_b\ &5(\|\nabla f(\mathbf{x}_p^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)})\|^2+2(\kappa^2+\gamma^2)) \qquad (18)
\end{aligned}
$$

where f_{gp} and f_{gq} mean the group loss function of the group where the clients p and q are in. Step a is true because we simply add and minus some same terms; step b follows by the "Sum in Norm Expansion" inequality, which can be found in the appendix.

Take 17 and 18 back to 16,

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_p^{(t,i,e+1)}-\mathbf{x}_q^{(t,i,e+1)}\|^2\right]\\
\leq_a\ &(1+\frac{1}{IE})\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|^2\\
&+(-\frac{2\eta}{L}+5\eta^2)\|\nabla f(\mathbf{x}_p^{(t,i,e)})-\nabla f(\mathbf{x}_q^{(t,i,e)})\|^2\\
&+4\eta^2IE(\kappa+\gamma)^2+10\eta^2(\kappa^2+\gamma^2)+2\eta^2\sigma^2\\
\leq_b\ &(1+\frac{1}{IE})\|\mathbf{x}_p^{(t,i,e)}-\mathbf{x}_q^{(t,i,e)}\|^2+14\eta^2IE(\kappa+\gamma)^2+2\eta^2\sigma^2
\end{aligned}
$$

Step a follows by using the same techniques when proving 18; for step b, we can assume that $(-\frac{2\eta}{L}+5\eta^2)$ is negative by adjusting the learning rate η to a small value.

Then, for any device j,

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_j^{(t,i,e)}-\bar{\mathbf{x}}^{(t,i,e)}\|^2\right]\\
\leq\ &(e-1)IE(14\eta^2IE(\kappa+\gamma)^2+2\eta^2\gamma^2)\\
\leq\ &25I^2E^2\eta^2L(\kappa+\gamma)^2+4IE\eta^2L\sigma^2 \qquad (19)
\end{aligned}
$$

Firstly, it is obvious that the difference between any two devices is the same as the difference between any device and the global average. Then, by telescoping across the whole training process, t, i, e, we can get the first inequality.

VI. Experiments

From the following picture, we can see that when in-group communication is allowed (Grouped in figure 5), the training process converges faster and the final result (testing loss) is as good as that of the classic federated learning algorithm (FedAvg in figure 5).

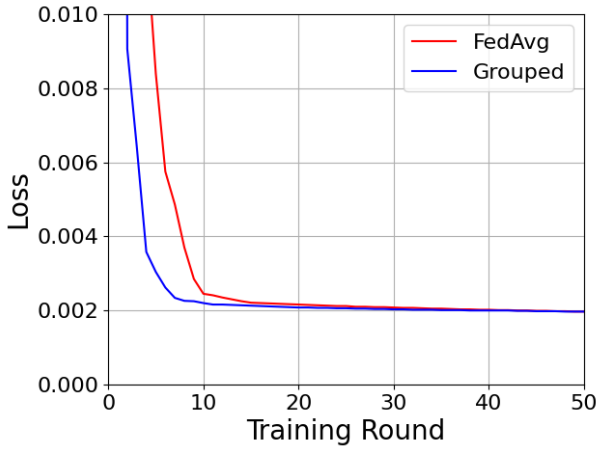

Fig. 5. Comparison of Classic and grouped federated learning

VII. Appendix

A. Equations and Inequalities

All equations and inequalities mentioned in this report can be found in another file, optimization.pdf.

B. Code

Code can be found in the zip file, as well as here. Please be adviced that the repository may be updated, but you can always find the version for this report by the commit tag "AI Report".

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] T. Li. (2019) Federated learning: Challenges, methods, and future directions. [Online]. Available: https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/

[3] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," *arXiv preprint arXiv:2112.11256*, 2021.

[4] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Local averaging helps: Hierarchical federated learning and convergence analysis," *arXiv preprint arXiv:2010.12998*, 2020.

[5] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data *et al.*, "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[6] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, no. 8, p. 2, 2012.