Convergence Analysis of Sequential Federated Learning on Heterogeneous Data

Yipeng Li* and Xinchen Lyu

Beijing University of Posts and Telecommunications

liyipeng@bupt.edu.cn

November 2, 2023 → November 4, 2023

Brief introduction of Federated Learning

Federated Learning (FL) (McMahan et al., 2017) is a popular distributed machine learning paradigm for joint training across multiple clients.

The basic FL problem is to minimize a global objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ F(\mathbf{x}) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \left(F_m(\mathbf{x}) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}_m}[f_m(\mathbf{x};\xi)] \right) \right\},$$

where F_m, f_m and $\mathcal{D}_m$ denote the local objective function, the loss function and the local dataset of client m ($m \in [M]$), respectively. In particular, when $\mathcal{D}_m$ has finite data samples $\{\xi_m^i : i \in [|\mathcal{D}_m|]\}$, the local objective function can also be written as $F_m(\mathbf{x}) = \frac{1}{|\mathcal{D}_m|} \sum_{i=1}^{|\mathcal{D}_m|} f_m(\mathbf{x};\xi_m^i)$.

Motivation

There are two categories of methods in FL:

i) parallel FL (PFL), where models are trained in a parallel manner across clients with synchronization at intervals, e.g., Federated Averaging (`FedAvg`) (McMahan et al., 2017);

ii) sequential FL (SFL), where models are trained in a sequential manner across clients, e.g., Cyclic Weight Transfer (`CWT`) (Chang et al., 2018).

Convergence theory is critical for analyzing the learning performance of algorithms on heterogeneous data in FL. So far, there are numerous works to analyze the convergence of PFL (Khaled et al., 2020; Koloskova et al., 2020; Li et al., 2019) on heterogeneous data. However, the convergence theory of SFL on heterogeneous data has not been well investigated in the literature, with only limited preliminary empirical studies Gao et al. (2020, 2021).

Setup

Algorithm 1: Sequential FL

for *training round* $r = 0, 1, \ldots, R-1$ **do**
  Sample a permutation $\pi_1, \pi_2, \ldots, \pi_M$ of $\{1, 2, \ldots, M\}$
  for $m = 1, 2, \ldots, M$ **in sequence do**
    $\mathbf{x}_{m,0}^{(r)} = \begin{cases} \mathbf{x}^{(r)}, & m = 1 \\ \mathbf{x}_{m-1,K}^{(r)}, & m > 1 \end{cases}$
    for *local step* $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{\pi_m,k}^{(r)}$
  Global model: $\mathbf{x}^{(r+1)} = \mathbf{x}_{M,K}^{(r)}$

Algorithm 2: Parallel FL

for *training round* $r = 0, 1, \ldots, R-1$ **do**
  for $m = 1, 2, \ldots, M$ **in parallel do**
    $\mathbf{x}_{m,0}^{(r)} = \mathbf{x}^{(r)}$
    for *local step* $k = 0, \ldots, K-1$ **do**
      $\mathbf{x}_{m,k+1}^{(r)} = \mathbf{x}_{m,k}^{(r)} - \eta \mathbf{g}_{m,k}^{(r)}$
  Global model:
  $$\mathbf{x}^{(r+1)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_{m,K}^{(r)}$$

Contribution

The main contributions are: i) we establish the convergence guarantees of SFL and compare it against PFL.

- We derive convergence guarantees of SFL for strongly convex, general convex and non-convex objectives on heterogeneous data.
- We compare the convergence guarantees of PFL and SFL, and find a *counterintuitive* comparison result that the guarantee of SFL is better than that of PFL in terms of training rounds on heterogeneous data.

TOC

1. Introduction
2. Convergence analysis of SFL
3. PFL vs. SFL on heterogeneous data
4. Experiments
5. Conclusion

Assumptions

We consider three typical cases for convergence theory, i.e., the strongly convex case, the general convex case and the non-convex case.

[Assumption 1] Each local objective function F_m is L-smooth, $m \in \{1, 2, \ldots, M\}$, i.e., there exists a constant $L > 0$ such that $\|\nabla F_m(\mathbf{x}) - \nabla F_m(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

[Assumption 2: stochasticity] The variance of the stochastic gradient at each client is bounded:

$$\mathbb{E}_{\xi \sim \mathcal{D}_m} \left[\left\| \nabla f_m(\mathbf{x}; \xi) - \nabla F_m(\mathbf{x}) \right\|^2 \middle| \mathbf{x} \right] \leq \sigma^2, \quad \forall\, m \in \{1, 2, \ldots, M\} \tag{1}$$

[Assumption 3a: heterogeneity] There exist constants β^2 and ζ^2 such that

$$\frac{1}{M} \sum_{m=1}^{M} \left\| \nabla F_m(\mathbf{x}) - \nabla F(\mathbf{x}) \right\|^2 \leq \beta^2 \left\| \nabla F(\mathbf{x}) \right\|^2 + \zeta^2 \tag{2}$$

[Assumption 3b: heterogeneity] There exists one constant ζ_*^2 such that

$$\frac{1}{M} \sum_{m=1}^{M} \left\| \nabla F_m(\mathbf{x}^*) \right\|^2 = \zeta_*^2 \tag{3}$$

where $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x})$ is one global minimizer.

Theorem 1

For SFL (Algorithm 1), there exist a constant effective learning rate $\tilde{\eta} := MK\eta$ and weights w_r, such that $\bar{\mathbf{x}}^{(R)} := \frac{1}{W_R}\sum_{r=0}^{R} w_r \mathbf{x}^{(r)}$ ($W_R = \sum_{r=0}^{R} w_r$) satisfies the following upper bounds:

Strongly convex: *Under Assumptions 1, 2, 3b, there exist a constant effective learning rate $\frac{1}{\mu R} \leq \tilde{\eta} \leq \frac{1}{6L}$ and weights $w_r = (1 - \frac{\mu\tilde{\eta}}{2})^{-(r+1)}$, such that it holds that*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] \leq \frac{9}{2}\mu D^2 \exp\left(-\frac{\mu\tilde{\eta}R}{2}\right) + \frac{12\tilde{\eta}\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{M} \tag{4}$$

General convex: *Under Assumptions 1, 2, 3b, there exist a constant effective learning rate $\tilde{\eta} \leq \frac{1}{6L}$ and weights $w_r = 1$, such that it holds that*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] \leq \frac{3D^2}{\tilde{\eta}R} + \frac{12\tilde{\eta}\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\sigma^2}{MK} + \frac{18L\tilde{\eta}^2\zeta_*^2}{M} \tag{5}$$

Non-convex: *Under Assumptions 1, 2, 3a, there exist a constant effective learning rate $\tilde{\eta} \leq \frac{1}{6L(\beta+1)}$ and weights $w_r = 1$, such that it holds that*

$$\min_{0\leq r\leq R} \mathbb{E}\left[\|\nabla F(\mathbf{x}^{(r)})\|^2\right] \leq \frac{3A}{\tilde{\eta}R} + \frac{3L\tilde{\eta}\sigma^2}{MK} + \frac{27L^2\tilde{\eta}^2\sigma^2}{8MK} + \frac{27L^2\tilde{\eta}^2\zeta^2}{8M} \tag{6}$$

where $D := \left\|x^{(0)} - x^*\right\|$ for the convex cases and $A := F(\mathbf{x}^{(0)}) - F^*$ for the non-convex case.

Corollary 1

Applying the results of Theorem 1, we can obtain the convergence bounds for SFL as follows:

Strongly convex: *Under Assumptions 1, 2, 3b, there exist a constant effective learning rate* $\frac{1}{\mu R} \leq \tilde{\eta} \leq \frac{1}{6L}$ *and weights* $w_r = (1 - \frac{\mu\tilde{\eta}}{2})^{-(r+1)}$, *such that it holds that*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta_*^2}{\mu^2 MR^2} + \mu D^2 \exp\left(-\frac{\mu R}{12L}\right)\right) \tag{7}$$

General convex: *Under Assumptions 1, 2, 3b, there exist a constant effective learning rate* $\tilde{\eta} \leq \frac{1}{6L}$ *and weights* $w_r = 1$, *such that it holds that*

$$\mathbb{E}\left[F(\bar{\mathbf{x}}^{(R)}) - F(\mathbf{x}^*)\right] = \mathcal{O}\left(\frac{\sigma D}{\sqrt{MKR}} + \frac{\left(L\sigma^2 D^4\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L\zeta_*^2 D^4\right)^{1/3}}{M^{1/3}R^{2/3}} + \frac{LD^2}{R}\right) \tag{8}$$

Non-convex: *Under Assumptions 1, 2, 3a, there exist a constant effective learning rate* $\tilde{\eta} \leq \frac{1}{6L(\beta+1)}$ *and weights* $w_r = 1$, *such that it holds that*

$$\min_{0\leq r\leq R} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^{(r)})\right\|^2\right] = \mathcal{O}\left(\frac{\left(L\sigma^2 A\right)^{1/2}}{\sqrt{MKR}} + \frac{\left(L^2\sigma^2 A^2\right)^{1/3}}{(MK)^{1/3}R^{2/3}} + \frac{\left(L^2\zeta^2 A^2\right)^{1/3}}{M^{1/3}R^{2/3}} + \frac{L\beta A}{R}\right) \tag{9}$$

where $\mathcal{O}$ *omits absolute constants,* $\tilde{\mathcal{O}}$ *omits absolute constants and polylogarithmic factors,* $D := \left\|x^{(0)} - x^*\right\|$ *for the convex cases and* $A := F(\mathbf{x}^{(0)}) - F^*$ *for the non-convex case.*

TOC

1. Introduction
2. Convergence analysis of SFL
3. PFL vs. SFL on heterogeneous data
4. Experiments
5. Conclusion

PFL vs. SFL on heterogeneous data

Table 1: Upper bounds in the strongly convex case with absolute constants and polylogarithmic factors omitted. All results are for heterogeneous settings.

Method	Bound ($D = \|x^{(0)} - x^*\|$)
SGD (Stich, 2019)	$\frac{\sigma^2}{\mu MKR} + LD^2 \exp\left(-\frac{\mu R}{L}\right)$ (1)
PFL	
(Karimireddy et al., 2020)	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta^2}{\mu^2 R^2} + \mu D^2 \exp\left(-\frac{\mu R}{L}\right)$ (2)
(Koloskova et al., 2020)	$\frac{\sigma_*^2}{\mu MKR} + \frac{L\sigma_*^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + LKD^2 \exp\left(-\frac{\mu R}{L}\right)$ (3)
Theorem 2	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 KR^2} + \frac{L\zeta_*^2}{\mu^2 R^2} + \mu D^2 \exp\left(-\frac{\mu R}{L}\right)$
SFL	
Theorem 1	$\frac{\sigma^2}{\mu MKR} + \frac{L\sigma^2}{\mu^2 MKR^2} + \frac{L\zeta_*^2}{\mu^2 MR^2} + \mu D^2 \exp\left(-\frac{\mu R}{L}\right)$

TOC

1. Introduction
2. Convergence analysis of SFL
3. PFL vs. SFL on heterogeneous data
4. Experiments
5. Conclusion

Experiments on quadratic functions

To further catch the heterogeneity, in addition to Assumption 3b, we also use bounded Hessian heterogeneity in Karimireddy et al. (2020):

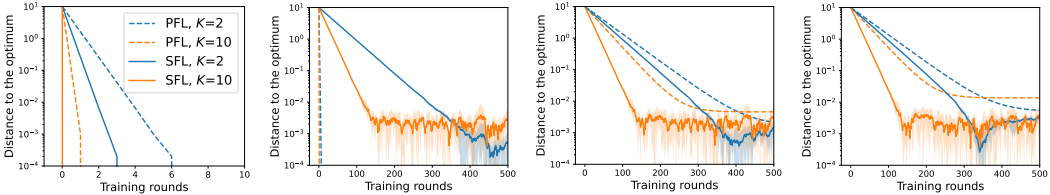
$$\max_m \left\|\nabla^2 F_m(\mathbf{x}) - \nabla^2 F(\mathbf{x})\right\| \leq \delta \ .$$

Choosing larger values of ζ_* and δ means higher heterogeneity.

Table 2: Settings of simulated experiments. Each group has two local objectives (i.e., $M = 2$) and shares the same global objective. The heterogeneity increases from Group 1 to Group 4.

	Group 1	Group 2	Group 3	Group 4
Settings	$\begin{cases} F_1(x) = \frac{1}{2}x^2 \\ F_2(x) = \frac{1}{2}x^2 \end{cases}$	$\begin{cases} F_1(x) = \frac{1}{2}x^2 + x \\ F_2(x) = \frac{1}{2}x^2 - x \end{cases}$	$\begin{cases} F_1(x) = \frac{2}{3}x^2 + x \\ F_2(x) = \frac{1}{3}x^2 - x \end{cases}$	$\begin{cases} F_1(x) = x^2 + x \\ F_2(x) = -x \end{cases}$
ζ_*, δ	$\zeta_* = 0, \delta = 0$	$\zeta_* = 1, \delta = 0$	$\zeta_* = 1, \delta = \frac{1}{3}$	$\zeta_* = 1, \delta = 1$

Experiments on quadratic functions



$\zeta_* = 0, \delta = 0$ $\zeta_* = 1, \delta = 0$ $\zeta_* = 1, \delta = \frac{1}{3}$ $\zeta_* = 1, \delta = 1$

Figure 1: Simulations on quadratic functions. It displays the experimental results from Group 1 to Group 4 in Table 2 from left to right. Shaded areas show the min-max values.

This in fact tells us that the comparison between PFL and SFL can be associated with data heterogeneity:

- When $\zeta_* = 0$ and $\delta = 0$, SFL outperforms PFL (Group 1).
- When $\zeta_* = 1$ and $\delta = 0$, the heterogeneity has no bad effect on the performance of PFL while hurts that of SFL significantly (Group 2).
- When the heterogeneity continues to increase to $\delta > 0$, SFL outperforms PFL with a faster rate and better result (Groups 3 and 4).

Experiments on quadratic functions

Intuitively, PFL updates the global model less frequently with more accurate gradients (with the global aggregation). In contrast, SFL updates the global model more frequently with less accurate gradients.

In homogeneous (gradients of both are accurate) and extremely heterogeneous settings (gradients of both are inaccurate), the benefits of frequent updates become dominant, and thus SFL outperforms PFL. In moderately heterogeneous settings, it's the opposite.

Experiments on real datasets

SFL outperforms PFL on extremely heterogeneous data.

Table 3: Test accuracy results in cross-device settings. We call $C = 1$ (where each client owns samples from one class) and $C = 2$ (where each client owns samples from two classes) as extremely heterogeneous data and moderately heterogeneous data, respectively.

Setup			$C = 1$			$C = 2$		
Dataset	Model	Method	$K = 5$	$K = 20$	$K = 50$	$K = 5$	$K = 20$	$K = 50$
CIFAR-10	VGG-9	PFL	67.61$\pm$4.02	62.00$\pm$4.90	45.77$\pm$5.91	78.42$\pm$1.47	78.88$\pm$1.35	78.01$\pm$1.50
		SFL	**78.43**$\pm$2.46	**72.61**$\pm$3.27	**68.86**$\pm$4.19	**82.56**$\pm$1.68	**82.18**$\pm$1.97	79.67$\pm$2.30
	ResNet-18	PFL	52.12$\pm$6.09	44.58$\pm$4.79	34.29$\pm$4.99	80.27$\pm$1.52	82.27$\pm$1.55	79.88$\pm$2.18
		SFL	**83.44**$\pm$1.83	**76.97**$\pm$4.82	**68.91**$\pm$4.29	**87.16**$\pm$1.34	**84.90**$\pm$3.53	79.38$\pm$4.49
CINIC-10	VGG-9	PFL	52.61$\pm$3.19	45.98$\pm$4.29	34.08$\pm$4.77	55.84$\pm$0.55	53.41$\pm$0.62	52.04$\pm$0.79
		SFL	**59.11**$\pm$0.74	**58.71**$\pm$0.98	**56.67**$\pm$1.18	**60.82**$\pm$0.61	**59.78**$\pm$0.79	**56.87**$\pm$1.42
	ResNet-18	PFL	41.12$\pm$4.28	33.19$\pm$4.73	24.71$\pm$4.89	57.70$\pm$1.04	55.59$\pm$1.32	46.99$\pm$1.73
		SFL	**60.36**$\pm$1.37	**51.84**$\pm$2.15	**44.95**$\pm$2.97	**64.17**$\pm$1.06	**58.05**$\pm$2.54	**56.28**$\pm$2.32

TOC

1. Introduction
2. Convergence analysis of SFL
3. PFL vs. SFL on heterogeneous data
4. Experiments
5. Conclusion

Future directions

In this paper, we have derived the convergence guarantees of SFL for strongly convex, general convex and non-convex objectives on heterogeneous data. Furthermore, we have compared SFL against PFL, showing that the guarantee of SFL is better than PFL on heterogeneous data. Experimental results validate that SFL outperforms PFL on extremely heterogeneous data in cross-device settings.

Future directions include:

i) lower bounds for SFL (this work focuses on the upper bounds of SFL),

ii) other potential factors that may affect the performance of PFL and SFL (this work focuses on data heterogeneity) and

iii) new algorithms to facilitate our findings (no new algorithm in this work).

Thanks for your attention!!!

References I

Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8):945–954, 2018.

Yansong Gao, Minki Kim, Sharif Abuadbba, Yeonjae Kim, Chandra Thapa, Kyuyeon Kim, Seyit A Camtep, Hyoungshick Kim, and Surya Nepal. End-to-end evaluation of federated learning and split learning for internet of things. In *2020 International Symposium on Reliable Distributed Systems (SRDS)*, pages 91–100. IEEE, 2020.

Yansong Gao, Minki Kim, Chandra Thapa, Sharif Abuadbba, Zhi Zhang, Seyit Camtepe, Hyoungshick Kim, and Surya Nepal. Evaluation and optimization of distributed machine learning techniques for internet of things. *IEEE Transactions on Computers*, 2021.

References II

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978. IEEE, 2022.

References III

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jDdzh5ul-d.