Decentralized Statistical Estimation and Inference in Federated Learning

January 29, 2022

Abstract

Keyword:

1 Introduction

Federated learning (FL), introduced in 2016 by McMahan et al. [9], is a machine learning setting where many clients (e.g. mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g. service provider), while keeping the training data decentralized [3]. In particular, heterogeneous or Non-IID distributed data across different data blocks and highly restrictive inter-block communication are two of the defining characteristics and challenges in the Federated Learning [3, 5].

A typical federated learning system considers a pool of M clients, in which the m-th client has a local dataset consisting of i.i.d. samples $\{\xi_{m,i}\}_{i=1}^{n_m}$ from some unknown distribution $\mathcal{P}_m$. Let $N = \sum_{m=1}^{M} n_m$ be the total number of samples. In particular, one wants to minimize the following objective function

$$F(w) = \sum_{m=1}^{M} p_m F_m(w) := \sum_{m=1}^{M} p_m \mathbb{E}_{\xi_m \sim \mathcal{P}_m} f_m(w; \xi_m), \tag{1}$$

where p_m is the weight of the m-th client with two natural settings being $p_m = 1/N$ or $p_m = n_m/N$, $f_m(\cdot; \xi_m)$ is the client-specific loss function and ξ_m is a random sample generated from unknown distribution $\mathcal{P}_m$. We do not assume $\{\mathcal{P}_m\}_{m=1}^M$ are identical. Indeed, it is natural to expect the existence of heterogeneity, especially for data stored in different locations or generated by different stochastic mechanism.

Many efficient algorithms have been proposed to cope with both statistical heterogeneity and expensive communication cost. Most of them are variants of the simple algorithm called *Local SGD* [11]. *Local SGD* runs stochastic gradient descent (SGD) independently in parallel on different clients and synchronizes the parameter estimates every K steps of SGD steps in each of the clients by taking averages in a central server . Let $\tilde{N}_R = KR$ be the number of samples the algorithm has accessed after R rounds of communication. When $\tilde{N}_R$ is fixed, choice of R controls the communication complexity. In particular, $R = 1$ is the most communication-efficient one [14] and is often referred as *One-shot averaging.* On the other extreme, $K = 1$ represents the situation when the algorithm alternates between one SGD step in parallel and one synchronization, which is statistically equivalent to single-machine SGD with mini-batches of size M and is often referred as *Mini-batch averaging*[13]. Li Xiang et al.[6] showed that the averaged iterates of *Local SGD* weakly converges to a rescaled Brownian motion and provided two iterative inference methods: One is based on a plug-in method which requires the estimation of hessian matrix [2]. The other is based on an online random scaling algorithm which only requires the SGD iterates path to facilitate an asymptotic pivotal statistic via random transformation for statistical inference [4], which is motivated by insights from time series regression in econometrics [1].

It is noted that the aforementioned centralized FL framework requires a central server for data aggregation, which can endure heavy communication burden when a large number of clients are deployed in the FL system. In some cases, there are no central servers at all. Also note that while SGD methods have been extended to the decentralized Federated Learning scenario [7, 8, 10, 12] with some convergence guarantee, few works have studied the limiting behaviour of the corresponding estimators and computationally efficient methods for statistical inference. In this work, we want to address the statistical estimation and inference aspect under the decentralized Federated Learning scenario.

2 Problem Formulation

In the following we will also refer to the clients as (edge) nodes to better present the decentralized scenario. We first illustrate the decentralized Federated Learning structure. Suppose the system is made of M edge nodes. The M nodes have the distributed datasets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_M$ with $\mathcal{D}_m = \{\xi_{m,i}\}_{i=1}^{n_m}$ owned by node m, where for each m, $\{\xi_{m,i}\}_{i=1}^{n_m}$ are IID data samples generated from unknown distribution $\mathcal{P}_m$. We use $\mathcal{D} = \cup_{m=1}^{M} \mathcal{D}_m$ to denote the global dataset of all nodes. We define $\mathbf{C} \in \mathbb{R}^{M \times M}$ as the confusion matrix which captures the network topology, and it is doubly stochastic, i.e. $\mathbf{C}\mathbf{1}_M = \mathbf{1}_M, \mathbf{C} = \mathbf{C}^T$. The element c_{ij} denotes the contribution of node j in model averaging at node i. The following Figure 1 gives an illustration of a decentralized Federated Learning system with 6 nodes. Note

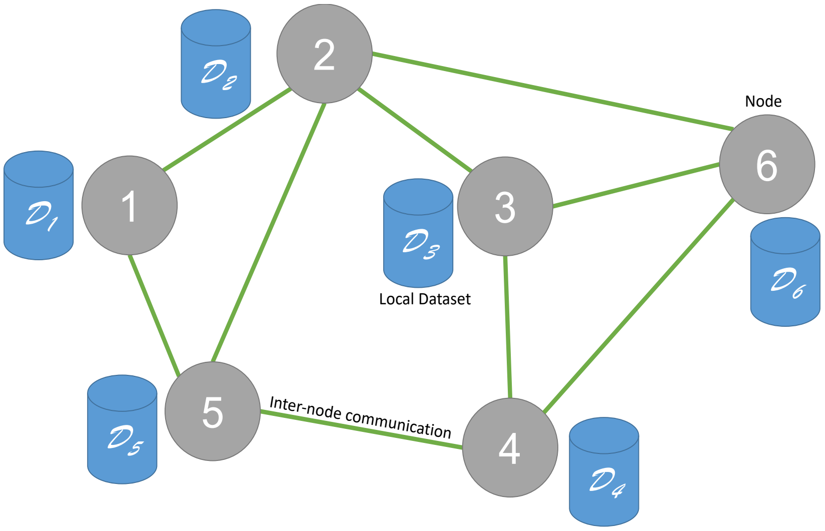

Figure 1: System of decentralized Federated Learning.

that when the graph is fully connected, i.e. $\mathbf{C} = \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^T$, the scenario is degenerated and is equivalent to a centralized Federated Learning scenario.

The purpose of decentralized Federated Learning is to obtain optimal model parameter w^* of all nodes without a central server, which is defined as follows:

$$w^* = \underset{w\in\mathbb{R}^d}{argmin}\ F(w), \tag{2}$$

where $F(w)$ is defined in Equation (1). In most cases, finding a closed-form solution of the optimization problem (2) is intractable, thus gradient-based method is generally used. In particular, we are going to analyze a *DFL* framework proposed by Liu et al.[8], which alternates between τ_1 independent steps of SGD in parallel and τ_2 steps of synchronization with neighboring nodes. When $\tau_2 = 1$, *DFL* is reduced to standard decentralized parallel SGD analyzed in [7]. In *DFL*, communication happens at $\mathcal{I} = \{t \in \mathbb{N}_+ | t \in [(k-1)\tau + \tau_1, k\tau), k \in \mathbb{N}_+\}$ if we denote $\tau = \tau_1 + \tau_2$. We summarize the procedure into the following recursive algorithm:

$$W_{t+1} = \begin{cases} W_t - \eta_t G_t & \text{if } t \notin \mathcal{I}, \\ W_t C & \text{if } t \in \mathcal{I}, \end{cases} \tag{3}$$

where C is confusion matrix, η_t is the gradient descent step size and we define the model parameter matrix $W_t \in \mathbb{R}^{d\times M}$ and the stochastic gradient matrix $G_t \in \mathbb{R}^{d\times M}$ as follows:

$$W_t = \begin{pmatrix} w_t^1 & w_t^2 & \cdots & w_t^M \end{pmatrix} \quad G_t = \left(\nabla f_1(w_t^1;\xi_t^1), \nabla f_2(w_t^2;\xi_t^2), \cdots, \nabla f_M(w_t^M;\xi_t^M)\right). \tag{4}$$

Here w_t^m represents the local parameter of node m at the t-th step and ξ_t^m is a sample uniformly sampled from dataset $\mathcal{D}_m$.

3 Target outline

1. Study the asymptotic behaviour of W_t or its transformation as $t \to \infty$ and how is its limiting distribution affected by those hyper-parameters described above and the

graph structure;

2. How to conduct valid statistical inference in a computationally-efficient and online fashion.

References

[1] Bunzel, H., Vogelsang, T., and Kiefer, N. (2000). Simple robust testing of regression hypotheses. *Econometrica*, 68:695–714.

[2] Chen, X., Lee, J., Tong, X., and Zhang, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48:251–273.

[3] Kairouz, P. and McMahan, H. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14:1–210.

[4] Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2021). Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling.

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

[6] Li, X., Liang, J., Chang, X., and Zhang, Z. (2021). Statistical Estimation and Inference via Local SGD in Federated Learning.

[7] Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[8] Liu, W., Chen, L., and Zhang, W. (2022). Decentralized Federated Learning: Balancing Communication and Computing Costs.

[9] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of Machine Learning Research*, 54:1273–1282.

[10] Sirb, B. and Ye, X. (2016). Decentralized Consensus Algorithm with Delayed and Stochastic Gradients. *SIAM Journal on Optimization*, 26:1835–1854.

[11] Stich, S. U. (2019). Local SGD converges fast and communicates little. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[12] Yuan, K., Ling, Q., and Yin, W. (2016). On the Convergence of Decentralized Gradient Descent. *SIAM Journal on Optimization*, 26:1835–1854.

[13] Zhang, J., Sa, C. D., Mitliagkas, I., and Ré, C. (2016). Parallel SGD: When does averaging help?

[14] Zinkevich, M., Weimer, M., Smola, A., and Li, L. (2010). Parallelized Stochastic Gradient Descent. *Advances in Neural Information Processing Systems*, 23:2595–2603.