Statistical Inference for Decentralized Federated Learning

Jia Gu

Center for Statistical Science, Peking University

July 5, 2023

Federated Learning

- **Federated Learning** (FL) was introduced in McMahan et al. (2017) : " … the learning task is solved by a loose federation of participating devices (clients) which are coordinated by a central server."
- **Heterogeneous distributed data across different clients** and **highly restrictive inter-block communication** are two defining characteristics and challenges in the FL (Li et al., 2020; Kairouz and McMahan, 2021).

- The canonical FL framework requires a central server for data aggregation: **heavy computation and communication burden**;
- Decentralized FL (DFL) paradigm is gaining popularity, where edge devices exchange their parameter estimates or gradient information **only with their neighboring devices** (Yuan et al., 2016; Lian et al., 2017; Sirb and Ye, 2018; Liu et al., 2022).

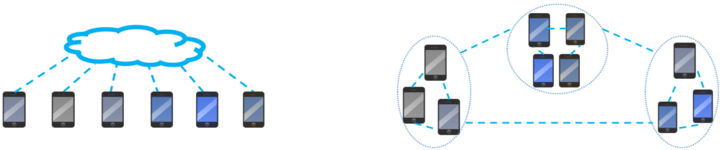


Figure 1: Star network (left) and decentralized network (right).

Problem setup (1)

- A typical FL setting consists of K clients;
- $\mathcal{D}_k = \{\boldsymbol{\xi}_i^k\}_{i=1}^{n_k}$ is the local data of client k, consists of IID observations from an unknown distribution $\mathcal{P}_k$.
- Let $f_k(\cdot;\boldsymbol{\xi}_k)$ and $F_k(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{P}_k} f_k(\boldsymbol{\theta};\boldsymbol{\xi}_k)$ be the k-th client specific loss function and risk function.
- One wants to minimize the **federated risk function**

$$F(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k F_k(\boldsymbol{\theta}) \tag{1}$$

  where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the interested parameter, w_k is the pre-specified weight.
- We allow different $\{\mathcal{P}_k\}_{k=1}^{K}$ to accommodate **heterogeneity** in FL.

Problem Setup (2)

- The FL is to estimate a true parameter

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

- If **full data communication** is available, one can minimize the empirical version of (2), and the corresponding full sample M-estimator of $\boldsymbol{\theta}_K^*$

$$\hat{\boldsymbol{\theta}}_K = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{argmin} \sum_{k=1}^{K} w_k \sum_{i=1}^{n_k} f_k(\boldsymbol{\theta}; \boldsymbol{\xi}_i^k). \tag{3}$$

- What if the full data communication is not available ?

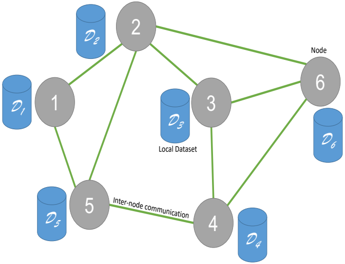
Local Connection Network

1. The decentralized FL has a **connection network** defined by an undirected graph (V, E) where $V = \{v_k\}_{k=1}^{K}$ represents the set of K clients and

$$E = \{(i, j) | \text{client } i \text{ and client } j \text{ are connected } \}$$

   specifies the edge set.

2. $\boldsymbol{C}(c_{ij}) \in \mathbb{R}^{K \times K}$ is a symmetric connection matrix defined on (V, E) where $c_{ij} > 0$ if and only if $(i, j) \in E$ and $\sum_{j=1}^{K} c_{ij} = 1$ for all i. (column-wise probability matrix)



$$C = \begin{pmatrix} 11/20 & 1/5 & 0 & 0 & 1/4 & 0 \\ 1/5 & 1/5 & 1/5 & 0 & 1/5 & 1/5 \\ 0 & 1/5 & 3/10 & 1/4 & 0 & 1/4 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/5 & 0 & 1/4 & 3/10 & 0 \\ 0 & 1/5 & 1/4 & 1/4 & 0 & 3/10 \end{pmatrix}$$

Figure 2: a connection network with 6 nodes (left) and the connection matrix $\boldsymbol{C}$ (right) where $c_{ij} = 1/\left(max\{d_i, d_j\} + 1\right)$ for $i \neq j$ and $c_{ii} = 1 - \sum_{j=1, j\neq i}^{K} c_{i,j}$, according to the Metropolis-Hastings rule (Boyd et al., 2006).

Local SGD: a communication-efficient algorithm for DFL

- $\hat{\boldsymbol{\theta}}_t^k$– the local estimate on the k-th data block at t-th step;
- Estimate matrix of all clients

$$\widehat{\boldsymbol{\Theta}}_t = \left(\hat{\boldsymbol{\theta}}_t^1, \quad \hat{\boldsymbol{\theta}}_t^2, \quad \cdots, \quad \hat{\boldsymbol{\theta}}_t^K\right) \in \mathbb{R}^{d \times K}$$

- η_t – the step size, the weighted SG matrix

$$\widehat{\boldsymbol{G}}_t = K\left(w_1 \nabla f_1(\hat{\boldsymbol{\theta}}_t^1; \boldsymbol{\xi}_t^1), w_2 \nabla f_2(\hat{\boldsymbol{\theta}}_t^2; \boldsymbol{\xi}_t^2), \cdots, w_K \nabla f_K(\hat{\boldsymbol{\theta}}_t^K; \boldsymbol{\xi}_t^K)\right) \tag{4}$$

- For each k, $\{\boldsymbol{\xi}_t^k\}_{t \geq 1}$ is chosen sequentially without replacement from the local dataset $\mathcal{D}_k$.

At $t = 0$, all local estimates are initialized as $\hat{\boldsymbol{\theta}}_0 \in \mathbb{R}^d$. For $t = 1, \cdots, T-1$ and some positive integer $\tau > 1$,

- if $t+1$ is divisible by τ, synchronize local estimates among neighbors according to $\boldsymbol{C}$

$$\widehat{\boldsymbol{\Theta}}_{t+1} = \left(\widehat{\boldsymbol{\Theta}}_t - \eta_t \widehat{\boldsymbol{G}}_t\right) \boldsymbol{C}$$

- otherwise update the estimates locally by $\widehat{\boldsymbol{\Theta}}_{t+1} = \widehat{\boldsymbol{\Theta}}_t - \eta_t \widehat{\boldsymbol{G}}_t$.

Local SGD reduces the communication cost by $(1 - 1/\tau) \times 100\%$ as compared with classical SGD ($\tau = 1$).

- Robbins and Monro (1951) (Ann. Math Stats) suggested to estimate θ^* by recursively updating (RM procedure)

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \eta_t \nabla f_1(\hat{\theta}_t; \boldsymbol{\xi}_t^1). \tag{5}$$

Lemma 1 (Chung (1954) (Ann Math Stats))

If $\eta_t = Dt^{-\alpha}$ for some $D > 0$ and $1/2 < \alpha \leq 1$, then

$$T^{\alpha/2}(\hat{\theta}_T - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\alpha, D)\right) \quad \text{as } T \to \infty,$$

where $\sigma_\theta^2 = \mathrm{Var}_{\mathcal{P}_1}\left(\nabla f_1(\theta; \boldsymbol{\xi}^1)\right)$ and

$$\sigma^2(\alpha, D) = \begin{cases} D\sigma_{\theta^*}^2 / (2\nabla^2 F_1(\theta^*)) & \text{if } 1/2 < \alpha < 1, \\ D^2\sigma_{\theta^*}^2 / (2\nabla^2 F_1(\theta^*) D - 1) & \text{if } \alpha = 1, \end{cases} \tag{6}$$

When $\alpha = 1$,

- $\hat{\theta}_T$ is $\sqrt{T}$-**consistent**, which is the same as the convergence rate of a regular full sample M-estimator;
- **Inefficient** unless $D = 1/\nabla^2 F_1(\theta^*)$;
- This requires extra information on the Hessian $\nabla^2 F_1(\theta^*)$.

When $\alpha < 1$,

- $\hat{\theta}_T$ converges at a slower rate of $T^{\alpha/2}$.
- Asymptotically, $\hat{\theta}_T - \theta^*$ is a weighted average of only the last $C(T) = O(T^\alpha log(T))$ gradient noises (Ruppert, 1988);
- This fact leads to less efficient estimation.

Averaged Stochastic Gradient

- The weak serial dependence when $\alpha < 1$ of $\{\hat{\theta}_t\}_{t \geq 0}$ motivates one to take an average of the SGD iterate to improve statistical efficiency:

$$\text{ASGD:} \;\; \hat{\hat{\theta}}_T = \frac{1}{T} \sum_{t=0}^{T-1} \hat{\theta}_t;$$

- Such an averaging procedure is referred to as the Polyak-Ruppert averaging (PR) (Polyak and Juditsky, 1992; Ruppert, 1988).

- It is proved that when $1/2 < \alpha < 1$,

$$\sqrt{T}(\hat{\hat{\theta}}_T - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2_{\theta^*}}{(\nabla^2 F_1(\theta^*))^2}\right). \tag{7}$$

The decentralized structure

Assumption 1 (Decentralized structure) *The K-dimensional connection matrix $\boldsymbol{C}$ satisfies $\boldsymbol{C}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{C}^T = \boldsymbol{C}$ whose largest eigenvalue is 1 and the other eigenvalues are strictly less than 1, namely $max\{|\lambda_k(\boldsymbol{C})||k = 2, 3, \cdots, K\} \leq \rho < \lambda_1(\boldsymbol{C}) = 1$ for some $0 < \rho < 1$, where $\lambda_k(\boldsymbol{C})$ denotes the k-th largest eigenvalue of $\boldsymbol{C}$.*

Remark

- This assumption made in Xiao and Boyd (2003) is a sufficient and necessary condition for $\underset{s\to\infty}{lim} \boldsymbol{C}^s = \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$, which implies that $\underset{k\to\infty}{lim} (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_K)\, \boldsymbol{C}^s = \bar{\mathbf{a}}_K\mathbf{1}_K^T$, where $\mathbf{a}_k \in \mathbb{R}^d$, $\bar{\mathbf{a}}_K = \frac{1}{K}\sum_{k=1}^K \mathbf{a}_k$.
- And ensures

$$\underset{s\to\infty}{lim} \widehat{\boldsymbol{G}}_t\boldsymbol{C}^s = \left(\sum_{k=1}^{K} w_k \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right) \mathbf{1}_K^T.$$

Assumption 2 *There exist positive constants* $b_1 < 1 < b_2$ *and* $b_4 > b_3$ *such that* $b_1 \leq \frac{n_{k_1}}{n_{k_2}} \leq b_2$ *for all* (k_1, k_2) *pairs satisfying* $k_1, k_2 \leq K$ *and* $\frac{b_3}{K} \leq w_k \leq \frac{b_4}{K}$ *for all* $1 \leq k \leq K$.

Assumption 3 *Objective function* $F_k(\cdot)$ *is differentiable,* L*-smooth and* μ*-strongly convex with positive constants* L *and* μ *such that for any* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,

$$\frac{L}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \geq F_k(\boldsymbol{\theta}_1) - F_k(\boldsymbol{\theta}_2) - \langle \nabla F_k(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \geq \frac{\mu}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2$$

Assumption 4 *The step size* η_t *is constant within each iteration:* $\eta_t = \tilde{\eta}_j$ *for* $(j-1)\tau \leq t \leq j\tau - 1$; *and* $\tilde{\eta}_j = D(j+\gamma)^{-\alpha}$ *for positive constants* D, γ *and* α.

Regularity conditions (2)

Assumption 5 (Gradient noise variance) *There exists non-negative constants* $L_{\boldsymbol{\xi}}$ *and* σ^2 *such that the gradient noise* $\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) = \nabla f_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) - \nabla F_k(\boldsymbol{\theta})$ *satisfies* $\mathbb{E}_{\mathcal{P}_k} \|\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)\|_2^2 \leq \sigma^2 + L_{\boldsymbol{\xi}} \|\nabla F_k(\boldsymbol{\theta})\|_2^2$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^d$.

Assumption 5 allows variance of the gradient noise $\boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$ to grow quadratically with the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_K^*$.

Assumption 6 (Bounded Heterogeneity) *There is a positive* κ $\sum_{k=1}^K w_k \|\nabla F(\boldsymbol{\theta}) - \nabla F_k(\boldsymbol{\theta})\|_2^2 \leq \kappa^2$ *for any* $\boldsymbol{\theta} \in \mathbb{R}^d$.

Key Quantities

Three types of averaging

- Across the clients at t: $\hat{\bar{\boldsymbol{\theta}}}_t = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_t^k$;
- For a client k over time: $\hat{\bar{\boldsymbol{\theta}}}_T^k = \frac{1}{T}\sum_{t=0}^{T-1}\hat{\boldsymbol{\theta}}_t^k$
- Spatial-temporal average (**The PR-estimator in DFL**): $\hat{\bar{\bar{\boldsymbol{\theta}}}}_T = \frac{1}{TK}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}_t^k$.

Two types of estimation errors

- Consensus error $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(\|\hat{\boldsymbol{\theta}}_t^k - \hat{\bar{\boldsymbol{\theta}}}_t\|_2^2\right)$.
- Mean square error (MSE) $\mathbb{E}\left(\|\hat{\bar{\boldsymbol{\theta}}}_t - \boldsymbol{\theta}_K^*\|_2^2\right)$

Outline of results

- We start by establishing the consensus error bound, which exists only in decentralized setting;
- We then generalizes the MSE bound of the non-distributed SGD algorithm to our setting: **An extra term on the bound can dominate when K is large;**
- We establishes the asymptotic normality of the PR-estimator in the decentralized FL: **Efficiency is attained at a cost.**

Consensus error bound

Proposition 1

Let $\mathcal{F}_{t-}^K = \sigma(\{\boldsymbol{\xi}_s^k | 0 \le s \le t-1, 1 \le k \le K\})$ be the σ-algebra generated by $\{\boldsymbol{\xi}_s^k\}_{s<t, 1\le k \le K}$ and $Q = \underset{t\ge 1, K \ge 1}{sup} \frac{1}{K} \mathbb{E} \| \mathbb{E}\left(\widehat{\boldsymbol{G}}_t | \mathcal{F}_{t-}^K\right) \|_F^2$ where $\widehat{\boldsymbol{G}}_t$ is defined in (4). Then, under Assumptions 1 - 6, Q is of order $\mathcal{O}(\kappa^2+1) < \infty$ and

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left(\|\hat{\boldsymbol{\theta}}_t^k - \bar{\hat{\boldsymbol{\theta}}}_t\|_2^2\right) \le c_0 \eta_t^2, \quad \text{where} \tag{8}$$

$$c_0 = 2a\tau Q\left(\left(L_{\boldsymbol{\xi}} + \tau\right) c(2\alpha, \rho^2) + \frac{\tau}{1-\rho} c(2\alpha, \rho)\right) + 2ab_4\tau\sigma^2 c(2\alpha, \rho^2),$$

$c(\alpha, \rho) = \sum_{s=0}^{\infty} \rho^s (1+s)^\alpha < \infty.$

MSE bound

Theorem 2

If $\alpha \leq 1$, $D > \frac{2}{\mu}$ *and* $\gamma > 0$ *such that* $\eta_1 \leq \frac{1}{\mu}$, *then under the conditions of Proposition 1*

$$\mathbb{E}\left(\|\hat{\bar{\boldsymbol{\theta}}}_T - \boldsymbol{\theta}^*_K\|_2^2\right) \leq v_1 \frac{\eta_T}{K} + v_2 \eta_T^2, \tag{9}$$

where

$$\begin{aligned} v_1 &= Db_4(\sigma^2 + 3L_{\boldsymbol{\xi}}\kappa^2)/(D\mu - 1) \ \textit{and} \\ v_2 &= \max\{4Db_4(L+\mu)c_0/(D\mu - 2), \frac{\gamma^2}{D^2}\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*_K\|_2^2\}. \end{aligned}$$

Moreover, for each K, $\hat{\bar{\boldsymbol{\theta}}}_T \xrightarrow{as} \boldsymbol{\theta}^*_K$ *as* $T \to \infty$.

Remark

- Compared with the result of the non-distribued SGD (Bottou et al., 2018), there is an extra $v_2\eta_T^2$ term in (9).
- The effect of the decentralized structure $\boldsymbol{C}$ on the above MSE bound is of the second-order and is asymptotically negligible since the ρ factor only appears in v_2 through c_0.
- The heterogeneity factor κ^2 enlarges the v_1 term only when the $L_{\boldsymbol{\xi}}$ factor appeared in Assumption 5 is positive, namely when the variance of the gradient noise is unbounded.

Assumption 7 (L-average smoothness) *For $k = 1, 2, \cdots, K$, the objective function $f_k(\cdot;\cdot)$ is L-average smooth with a positive constant L_a such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,*

$$\mathbb{E}_{\mathcal{P}_k}\|\nabla f_k(\boldsymbol{\theta}_1; \boldsymbol{\xi}^k) - \nabla f_k(\boldsymbol{\theta}_2; \boldsymbol{\xi}^k)\|_2^2 \leq L_a\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2. \tag{10}$$

Remark

This assumption is stronger than the smoothness condition in Assumption 3, and holds for objective functions such as those for the linear regression, ridge regression or logistic regression if $\boldsymbol{\xi}^k$ has certain bounded moments.

Asymptotic normality (AN): conditions 2

Assumption 8 (Regularity of gradient noise) *There exist positive constants ℓ_{cov} and δ such that for all $k = 1, 2, \cdots, K$,*

$$\boldsymbol{S}_k = \mathbb{E}_{\mathcal{P}_k} \boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k) \boldsymbol{\epsilon}_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)^T \ \text{satisfies}$$
$$\boldsymbol{S}_k \succeq \ell_{cov} \boldsymbol{I} \ \text{and} \ \sup_{K \geq 1} \mathbb{E}_{\mathcal{P}_k} \|\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\|_2^{2+\delta} < \infty,$$

where $\boldsymbol{\epsilon}(\boldsymbol{\theta}) = \sqrt{K} \sum_{k=1}^K w_k \boldsymbol{\epsilon}_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k)$.

Assumption 9a (Second-order smoothness) *$F(\boldsymbol{\theta}) = \sum_{k=1}^K w_k F_k(\boldsymbol{\theta})$ is second-order differentiable, and there exists $L_H > 0$ such that*

$$\|\nabla^2 F(\boldsymbol{\theta}) - \nabla^2 F(\boldsymbol{\theta}_K^*)\|_2 \leq L_H \|\boldsymbol{\theta} - \boldsymbol{\theta}_K^*\|_2$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $K \geq 1$.

AN of Spatial-Temporal (PR) Average

Theorem 3

Under Assumptions required in Theorem 2 and Assumptions 7, 8 and 9a, if $K = o(T^{2\alpha-1})$ *with* $T = \min_{1\le k\le K} n_k$, $\alpha \in (1/2, 1)$ *and* $\sup_{K\ge 1} \|\boldsymbol{\theta}_K^*\|_2 < \infty$, *we have*

$$\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}\left(\bar{\bar{\hat{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}_K^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \text{ as } T \to \infty, \tag{11}$$

where $\boldsymbol{H} = \nabla^2 F(\boldsymbol{\theta}_K^*)$ *is population Hessian and* $\boldsymbol{S} = \mathbb{E}\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*)\boldsymbol{\epsilon}(\boldsymbol{\theta}_K^*)^T$ *is the covariance of the aggregated gradient noise.*

- The statistical efficiency comes with stronger restriction on K.

$\boldsymbol{H} = \nabla^2 F(\boldsymbol{\theta}_K^*)$

We can directly estimate the local Hessian matrix $\nabla^2 F_k(\boldsymbol{\theta}_K^*)$ by using a small fraction of data stored in each data node k, say $\{\boldsymbol{\xi}_{T-s}^k\}_{s=0}^{a(T)-1}$, where $a(\cdot) : \mathbb{N}_+ \to \mathbb{N}_+$ is a non-decreasing function. The estimator is defined as

$$\hat{\boldsymbol{H}}_k = \frac{1}{a(T)} \sum_{s=0}^{a(T)-1} \nabla^2 f_m(\hat{\bar{\boldsymbol{\theta}}}_t^k; \boldsymbol{\xi}_{T-s}^k), \quad \hat{\boldsymbol{H}} = \sum_{k=1}^{K} p_k \hat{\boldsymbol{H}}_k. \tag{12}$$

FL: Large K helps

- We do not need to consistently estimate $\boldsymbol{H}_k$.
- Small $a(T)$ suffices.
- The law of large numbers takes effect as $K \to \infty$ and thus we can derive the consistency of $\sum_{k=1}^{K} p_k \hat{\boldsymbol{H}}_k$ as a whole.

$$\boldsymbol{S} = K\mathbb{E}\left(\sum_{k=1}^{K} w_k \nabla f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\right)\left(\sum_{k=1}^{K} w_k \nabla f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\right)^T$$

- The **infeasible centralized estimator** due to the cross terms:

$$\widetilde{\widehat{\boldsymbol{S}}} = \frac{K}{T}\sum_{t=0}^{T-1}\left(\sum_{k=1}^{K} w_k \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)\left(\sum_{k=1}^{K} w_k \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)$$

- Define

$$\begin{aligned}\widehat{\boldsymbol{S}}_k &= \frac{1}{T}\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)^T \\ &\quad -\frac{1}{T^2}\left(\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)\left(\sum_{t=0}^{T-1} \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\right)^T,\end{aligned}$$

  then an estimator of $\boldsymbol{S}$ can be defined as $\widehat{\boldsymbol{S}} = K\sum_{k=1}^{K} w_k^2 \widehat{\boldsymbol{S}}_k$.

- $\{\nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k)\}$ are readily available when running the DFL Algorithm and $\hat{\boldsymbol{S}}_k$ can be updated iteratively.

Validity of the confidence region (CR): condition

Assumption 9b (Second-order smoothness) *For all $k = 1, 2, \cdots, K$, we assume that the objective function $f_k(\boldsymbol{\theta}; \boldsymbol{\xi})$ is second-order differentiable with respect to $\boldsymbol{\theta} \in \mathbb{R}^d$, and there exists positive constants ℓ_H and H, such that*

$$\sqrt{\mathbb{E}_{\mathcal{P}_m} \|\nabla^2 f_k(\boldsymbol{\theta}; \boldsymbol{\xi}^k) - \nabla^2 f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k)\|_2^2} \leq \ell_H \|\boldsymbol{\theta} - \boldsymbol{\theta}_K^*\|_2$$

and $\mathbb{E}_{\mathcal{P}_m} \|\nabla^2 f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}^k) - \nabla^2 F_k(\boldsymbol{\theta}_K^*)\|_2 \leq H$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\theta}_K^*$ is the true value defined in (2).

Validity of the CR

Theorem 4

Under Assumptions required in Theorem 2 and Assumptions 7, 8 and 9b , if $K = o(T^{2\alpha-1})$ *and* $Ka(T) \to \infty$ *with* $T = \min_{1\le k\le K} n_k$*,* $\alpha \in (1/2, 1)$*,* $\sup_{K\ge 1} \|\boldsymbol{\theta}_K^*\|_2 < \infty$ *and* $\sup_{K\ge 1} \max_{1\le k\le K} \|\nabla F_k(\boldsymbol{\theta}_K^*)\|_2 < \infty$*, then we have that* $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}\|_2 = o_p(1)$ *and*

$$\mathbb{P}\left(TK\left(\hat{\hat{\bar{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}_K^*\right)^T \widehat{\boldsymbol{\Sigma}}^{-1} \left(\hat{\hat{\bar{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}_K^*\right) \le \chi^2_{d,\beta} \right) \to 1 - \beta. \tag{13}$$

for any $\beta \in (0, 1)$*, where* $\chi^2_{d,\beta}$ *is the upper* β *quantile of the* χ^2_d *distribution,* $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{H}}^{-1}$.

Now we are ready for the construction of the $1-\beta$ CR for $\boldsymbol{\theta}_K^*$:

$$\left\{\theta \middle| TK\left(\hat{\hat{\bar{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}\right)^T \widehat{\boldsymbol{\Sigma}}^{-1} \left(\hat{\hat{\bar{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}\right) \le \chi^2_{d,\beta} \right\}.$$

Drawback of the PR-procedure in decentralized FL

- If a large step size is chosen with $\alpha = 1/2 + \epsilon$ as suggested in Ruppert (1988), where ϵ is a small positive constant, then only $M = o(T^{2\epsilon})$ data nodes are allowed to participate in the decentralized FL.
- This is not satisfying when the network is large in modern applications.

How to achieve statistical efficiency when $\alpha = 1$?

- When $\alpha = 1$, although $\hat{\hat{\bar{\boldsymbol{\theta}}}}_T$ is not statistically efficient , it is $\sqrt{TK}-$consistent as long as $K = o(T)$.

- We can thus improve the $\hat{\hat{\bar{\boldsymbol{\theta}}}}_T$ estimator based on the idea of one-step estimator (Bickel, 1975). That is, given the preliminary $\hat{\hat{\bar{\boldsymbol{\theta}}}}_T$, we define the one-step estimator as

$$\hat{\hat{\bar{\boldsymbol{\theta}}}}^{os}_T = \hat{\hat{\bar{\boldsymbol{\theta}}}}_T - \left(\widehat{\boldsymbol{H}}\right)^{-1} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^{K} w_k \nabla f_k(\hat{\boldsymbol{\theta}}_t^k; \boldsymbol{\xi}_t^k). \tag{14}$$

- Each part of the RHS of (14) is “handy”.

Theorem 5

Under Assumptions required in Theorem 2 and Assumptions 7, 8 and 9b, if $\alpha = 1$ *and* $\sup_{K \geq 1} \|\boldsymbol{\theta}_K^*\|_2 < \infty$, *then the one-step estimator* $\hat{\bar{\bar{\boldsymbol{\theta}}}}_T^{os}$ *defined in (14) satisfies*

$$\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}\left(\hat{\bar{\bar{\boldsymbol{\theta}}}}_T^{os} - \boldsymbol{\theta}_K^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \text{ as } T \to \infty \text{ and } K \to \infty$$

as long as $K = o(T)$.

- This establishes the asymptotic normality of the one-step estimator with **a relaxed constraint** on the number K of data nodes.

Remark on the condition

- The condition $K \to \infty$ as $T \to \infty$ is necessary to ensure the validity of the following first-order expansion of the estimator $\hat{\bar{\bar{\boldsymbol{\theta}}}}_T$:

$$
\begin{aligned}
& \| \left(\hat{\bar{\bar{\boldsymbol{\theta}}}}_T - \boldsymbol{\theta}_K^* \right) - \boldsymbol{H}^{-1} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^{K} w_k \left(\nabla f_k(\hat{\bar{\boldsymbol{\theta}}}_t; \boldsymbol{\xi}_t^k) - \nabla f_k(\boldsymbol{\theta}_K^*; \boldsymbol{\xi}_t^k) \right) \|_2 \\
= \quad & o_p\left(\frac{1}{\sqrt{TK}}\right),
\end{aligned}
$$

- Natural since we are considering a large-scale decentralized FL problem where many clients conduct statistical inference collaboratively.

Sparse network structure

- In real-world appilcations, the network is often sparse and the spectral gap $1-\rho \to 0$ as $K \to \infty$.
- Assumption 1 no longer holds, introduing larger bias.

How does the network sparseness affect our previous result?

Assumption 10 (Sparse network) *The connection matrix $\boldsymbol{C}$ is a K-dimensional matrix satisfying $\boldsymbol{C}\mathbf{1} = \mathbf{1}$ and $\boldsymbol{C}^T = \boldsymbol{C}$, and the eigenvlaues of C satisfy*

$$\max\{|\lambda_k(\boldsymbol{C})||k = 2, 3, \cdots, K\} \leq 1 - \frac{\rho^{'}}{K^q} < \lambda_1(\boldsymbol{C}) = 1$$

for some $0 < \rho^{'} < 1$ and $q \geq 0$ as $K \to \infty$, where $\lambda_k(\boldsymbol{C})$ denotes the k-th largest eigenvalue of C.

Theorem 6

Under Assumption 2 -8, Assumptions 9b and 10, if $\tau = 1, \alpha = 1$ and the parameter space is a compact subset of $\mathbb{R}^d$, the one-step estimator $\hat{\bar{\bar{\boldsymbol{\theta}}}}{}_T^{os}$ defined in (14) satisfies

$$\sqrt{TK}\boldsymbol{S}^{-1/2}\boldsymbol{H}\left(\hat{\bar{\bar{\boldsymbol{\theta}}}}{}_T^{os} - \boldsymbol{\theta}_K^*\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \text{ as } T \to \infty \text{ and } K \to \infty,$$

as long as $K = o(T^{\frac{1}{6q+1}})$.

Remark

The constraint on K relative to T is much stricter on a sparse network.

Some typical network structures

(a) Cyclic: $q = 2$ (b) Grid: $q = 1$ (c) s-connected :$q = 0$

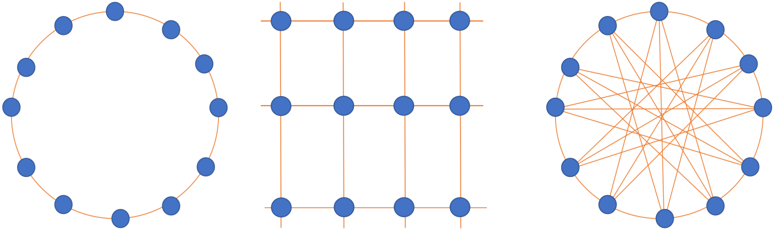


Figure 3: Three types of decentralized structure.

Table 1: Six typical types of sparse network and their corresponding maximal divergence rate of the network size K. Here the sparseness of the graph correspondes to the lazy Metropolis matrix $\tilde{\boldsymbol{C}} = \frac{1}{2}(\boldsymbol{C} + \boldsymbol{I})$, where $\boldsymbol{C}$ is constructed obeying the Metropolis-Hastings rule.

Graph Topology	Sparseness q	Network Size K
expander graph	0	$o(T)$
k-dimensional torus	$\frac{2}{k}$	$o(T^{\frac{k}{12+k}})$
2-D grid	1	$o(T^{1/7})$
star graph	2	$o(T^{1/13})$
cyclic graph	2	$o(T^{1/13})$
Erdős-Rényi random graph	0*	

* For the Erdős-Rényi random graph, the statement $q = 0$ holds with probability approaching 1 as $K \to \infty$.

Network setup

- Given a network size K, the nodes were denoted by their labels $1, 2, \cdots, K$, and a number $K_{neighbor}$ was used to denote the number of neighbors each node has, which controlled the connectivity of the network.
- Clients k and k' are connected if and only if $|k - k'| \leq \frac{K_{neigh}}{2}$ or $|k - k'| \geq K - \frac{K_{neigh}}{2}$.
- Thus, for a given K, a larger (smaller) $K_{neighbor}$ means a tightly (loosely) connected network.

Data Generating Process

- The local data sets were generated as follows: For each client k,

$$\boldsymbol{X}_{k,t} \overset{i.i.d}{\sim} \mathcal{N}\left(\mathbf{0}_{(d-1)\times 1}, \mathbf{I}_{(d-1)\times(d-1)}\right),\ \varepsilon_{k,t} \overset{i.i.d}{\sim} \Gamma(1,1) - 1 \ \text{ and}$$
$$Y_{k,t} = \begin{pmatrix}1, & \boldsymbol{X}_{k,t}^T\end{pmatrix} \boldsymbol{\phi}_k^* + \varepsilon_{k,t}.$$

- The dimension of the parameter $d = 6$.

- The true parameter $\boldsymbol{\theta}_K^*$ was $\boldsymbol{\theta}_K^* = \sum_{k=1}^{K} w_k \boldsymbol{\phi}_k^*$ where $w_k \equiv 1/K$.

- $\phi_{k,j}^* = \delta_{gap}\left((k-1) - (K-1)/2\right)$ for a $\delta_{gap} > 0$. This made the true parameter $\boldsymbol{\theta}_K^* = \mathbf{0}_d$.

- The parameter δ_{gap} quantifies the amount of heterogeneity across the data blocks.

Consensus error

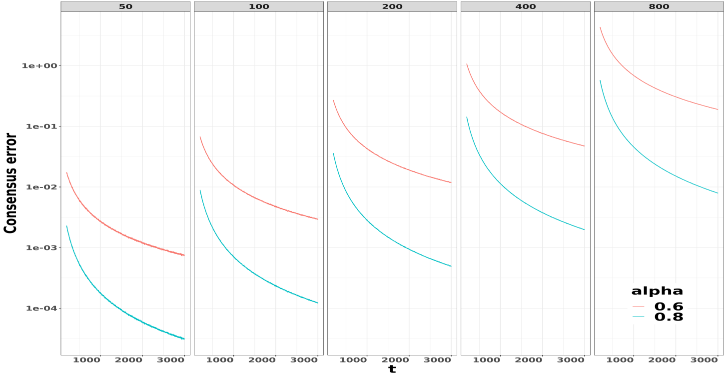


Figure 4: Average consensus error of the averaged estimator $\bar{\hat{\boldsymbol{\theta}}}_T$ under different numbers of block size K ($K = 50, 100, 200, 400$ and 800) with respect to the number of SGD steps t ($t \leq T$, where T was the local sample size) when the rate α of the step size was 0.6 and 0.8, respectively.

Mean square error

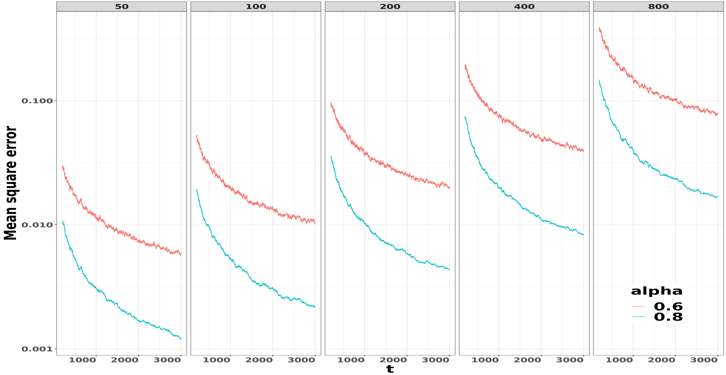


Figure 5: Average mean square error of the averaged estimator $\hat{\bar{\boldsymbol{\theta}}}_T$ under different numbers of block size K ($K = 50, 100, 200, 400$ and 800) with respect to the number of SGD steps t ($t \leq T$, where T was the local sample size) when the rate α of the step size was 0.6 and 0.8, respectively.

Absolute coverage error

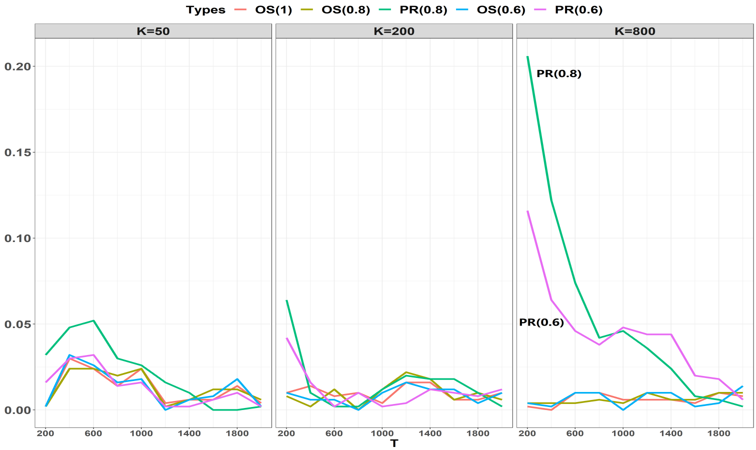


Figure 6: Absolute coverage errors of the 95% confidence regions based on the asymptotic normality of the one-step estimator (OS, $\alpha = 1, 0.8, 0.6$) and the Polyak-Ruppert averaged estimators (PR, $\alpha = 0.8, 0.6$). The gap parameter δ_{gap} was 0.2.

Summary

- the mean square error bound and the consensus error bound are established in decentralized FL
- the asymptotic normality of the PR-estimator in the decentralized FL setting is established, which attains the same efficiency as the full-sample estimator at the cost of heavier network size constraint;
- A one-step estimator is proposed to mitigate the problem;
- The confidence regions based on both the PR-averaged estimator and the proposed one-step estimator are constructed;
- The effect of the decentralized connection network's sparseness on the one-step estimator's statistical property is also derived.

Future works

Algorithm perspective:

- Acceleration: Qian (1999); Johnson and Zhang (2013);
- Bias correction technique: *Gradient Tracking* methods (Nedic et al., 2016) and *Exact Diffusion* (Yuan et al., 2020).

Application perspective:

- Non-response of the clients;
- Malicious clients;

Bickel, P. (1975). One-Step Huber Estimates in the Linear Model. *Journal of the American Statistical Association*, 70:428–434.

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

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530.

Chung, K. (1954). On a stochastic approximation method. *Annals of Mathematical Statistics*, 25:463–483.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems (NIPS)*, 1:315–323.

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

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

Lian, X., Zhang, C., Zhang, H., et al. (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.

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

McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of Machine Learning Research*, 54:1273–1282.

Nedic, A., Olshevsky, A., and Shi, W. (2016). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27.

Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks: the official journal of the International Neural Network Society*, 12:145–151.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Ruppert, D. (1988). Efficient Estimations from a Slowly Convergent Robbins-Monro Process. *Technical report, Cornell University Operations Research and Industrial Engineering*.

Sirb, B. and Ye, X. (2018). Decentralized Consensus Algorithm with Delayed and Stochastic Gradients. *SIAM Journal on Optimization*, 26:1835–1854.

Xiao, L. and Boyd, S. (2003). Fast Linear Iterations for Distributed Averaging. *Systems & Control Letters*, 53:65–78.

Yuan, K., Alghunaim, S. A., Ying, B., and Sayed, A. H. (2020). On the Influence of Bias-Correction on Distributed Stochastic Optimization. *IEEE Transactions on Signal Processing*, 68:4352–4367.

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

Thanks!