Statistical Estimation and Online Inference via Local SGD

Xiang Li
Joint work with Jiadong Liang, Xiangyu Chang, and Zhihua Zhang

Peking University

2022/07/05

Federated Learning (FL)

- FL collaboratively trains a global model from data held by remote *clients* (e.g., mobile phones) [MMR+17].
- All local data are not allowed to be uploaded to the center and the central server has access only to intermediate quantities.
- Aim to protect sensitive information, such as personal identity information and state of health information, from unauthorized access of service providers.
- A typical application: Google Gboard word prediction.

Problem Formulation

- K clients with the k-th client has a local dataset consisting of i.i.d. samples from unknown distribution $\mathcal{D}_k$.
- The central server faces the following optimization problem:

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}) := \sum_{k=1}^{K} p_k \mathbb{E}_{\xi_k \sim \mathcal{D}_k} f_k(\boldsymbol{x}; \xi_k). \tag{1}$$

- Two consideration: (i) **Data heterogeneity**, i.e., different $\mathcal{D}_k$; (ii) **Communication efficiency**

Local SGD

- One of the simplest methods is *Local SGD* [Sti18].
- It runs SGD independently in parallel on different clients and averages the sequences only once in a while.
- Key idea: lower communication frequency to improve communication efficiency.

Local SGD

- $\boldsymbol{x}_t^k$ denotes the parameter held by the k-th client at iteration t.
- $\mathrm{g}_t^k = \nabla f_k(\boldsymbol{x}_t^k; \xi_t^k)$ is the unbiased stochastic gradient estimator of $\nabla f_k(\boldsymbol{x}_t^k)$ with $\xi_t^k \sim \mathcal{D}_k$.
- $\mathcal{I} = \{t_0, t_1, t_2, \cdots\}$ is the set of communication iterations with $E_m = t_{m+1} - t_m$ the m-th communication interval.
- Local SGD runs

$$\boldsymbol{x}_{t+1}^k = \begin{cases} \boldsymbol{x}_t^k - \eta_t \mathrm{g}_t^k & \text{if } t+1 \notin \mathcal{I}, \\ \sum_{k=1}^K p_k \left[\boldsymbol{x}_t^k - \eta_t \mathrm{g}_t^k\right] & \text{if } t+1 \in \mathcal{I}. \end{cases}$$

- When $t_m \leq t < t_{m+1}$, we abuse the notation and let $\eta_t = \eta_m$.

Our Target

Our goal is to

- obtain an efficient estimate of $\boldsymbol{x}^* = \mathrm{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ only through the SGD iterates $\{\boldsymbol{x}_{t_m}^k\}_{m\in[T],k\in[K]}$,
- provide asymptotic confidence intervals for further inference.

Three following questions:

1. how one constructs the estimator from Local SGD iterates;
2. how intermittent communication and non-iid data affect its asymptotic behavior;
3. how to quantify the variability and randomness of the estimator.

- It is known the averaged SGD estimator obtains the optimal asymptotic variance without any problem-dependent knowledge [PJ92].
- We are motivated to use the average of Local SGD iterates as the estimator,

$$\bar{\boldsymbol{y}}_T = \frac{1}{T}\sum_{m=1}^{T} \bar{\boldsymbol{x}}_{t_m} \quad \text{where} \quad \bar{\boldsymbol{x}}_{t_m} = \sum_{k=1}^{K} p_k \boldsymbol{x}_{t_m}^k.$$

- Two levels of average: (i) over devices k; (ii) over communication iterations t_m.

1 Introduction and Background

2 Statistical Estimation via Local SGD
- The Estimator
- Asymptotic Behavior

3 Statistical Inference via Local SGD

4 Conclusion

Asymptotic Normality

Let $E_m = t_{m+1} - t_m$ communication interval and assume

$$\frac{1}{T^2}(\sum_{m=1}^{T} E_m)(\sum_{m=1}^{T} E_m^{-1}) \to \nu(\nu \geq 1). \tag{2}$$

Theorem

If $\gamma_m = E_m \eta_m \propto m^{-\alpha}$ with $\alpha \in (0.5, 1)$, and E_m increases in m sufficiently slowly. Under some regularity conditions,

$$\sqrt{t_T}\,(\bar{\boldsymbol{y}}_T - \boldsymbol{x}^*) \xrightarrow{d} \mathcal{N}\left(0,\ \nu \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}\right),$$

where $\boldsymbol{G}$ is the Hessian matrix at $\boldsymbol{x}^*$, and $\boldsymbol{S}$ is the covariance matrix of aggregated gradient noise at $\boldsymbol{x}^*$.

Remark about Asymptotic Normality

$$\text{Variance} = \nu \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top} \text{ where } \frac{1}{T^2}(\sum_{m=1}^{T} E_m)(\sum_{m=1}^{T} E_m^{-1}) \to \nu(\nu \geq 1).$$

1. Optimal asymptotic variance when $\nu = 1$.
2. Data heterogeneity doesn't effect the variance, since Local SGD with small $\eta_m \approx$ parallel SGD with large $E_m\eta_m$.
3. Many diverging $\{E_m\}$ have $\nu = 1$ and vanishing asymptotic averaged communication frequency (ACF $= T/\sum_{m=0}^{T-1} E_m$).

$E_m(\geq 1)$	$\nu(\geq 1)$	ACF
E	1	E^{-1}
any $E_m \leq E$	1	$[E^{-1}, 1]$
$E \ln^{\beta} m$ $(\beta > 0)$	1	$E^{-1} \ln^{-\beta} T$
$E \ln^{\beta} \ln m$ $(\beta > 0)$	1	$E^{-1} \ln^{-\beta} \ln T$
Em^{β} $(\beta \in (0, 1))$	$(1-\beta^2)^{-1}$	$(1+\beta)E^{-1} T^{-\beta}$

The Plug-in Method

- The plug-in methodes timates $\boldsymbol{G}$ and $\boldsymbol{S}$ by its empirical version $\widehat{\boldsymbol{G}}_T$ and $\widehat{\boldsymbol{S}}_T$
- Both are in form of moving average and thus can be computed in an online manner [CLT+20].
- Under some regularity condition, $\widehat{\boldsymbol{G}}_T^{-1}\widehat{\boldsymbol{S}}_T\widehat{\boldsymbol{G}}_T^{-\top} \xrightarrow{p.} \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top}$.

Confidence Interval via Plug-in Method

- $(\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-\top})_{jj}$ can be estimated by $\widehat{\sigma}^2_{T,j} = (\widehat{\boldsymbol{G}}_T^{-1}\widehat{\boldsymbol{S}}_T\widehat{\boldsymbol{G}}_T^{-\top})_{jj}$
- Recall that $\bar{\boldsymbol{y}}_T = \frac{1}{T}\sum_{m=1}^{T}\bar{\boldsymbol{x}}_{t_m}$.
- To estimate j-th element $\boldsymbol{x}_j^*$ of $\boldsymbol{x}^*$, we can use

$$\mathbb{P}\left(\bar{\boldsymbol{y}}_{T,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\nu}_T}{t_T}}\widehat{\sigma}_{T,j} \le \boldsymbol{x}_j^* \le \bar{\boldsymbol{y}}_{T,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{\nu}_T}{t_T}}\widehat{\sigma}_{T,j}\right) \to 1-\alpha,$$

where $\widehat{\nu}_T \to \nu$ and $z_{\frac{\alpha}{2}}$ is $(1-\alpha/2)$-quantile of the standard normal distribution.

Drawbacks of The Plug-in Method

- Accessible Hessian information
- Formulation and sharing of each $\nabla^2 f_k(\bar{\boldsymbol{x}}_{t_m}; \xi_{t_m}^k)$ requires at least $O(d^2)$ memory and communication cost.

Random Scaling: Functional CLT

Random scaling aims to construct an asymptotically pivotal statistic using all information along the whole trajectory $\{\bar{\boldsymbol{x}}_{t_m}\}_{1\le m\le T}$ [LLSS21].

Theorem (Functional CLT)

Under the same conditions of previous theorem, as $T \to \infty$, the random function $\phi_T(\cdot)$ weakly converges to a scaled Brownian motion, i.e.,

$$\phi_T(r) := \frac{\sqrt{t_T}}{T} \sum_{m=1}^{h(r,T)} (\bar{\boldsymbol{x}}_{t_m} - \boldsymbol{x}^*) \Rightarrow \sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathrm{B}_d(r)$$

where $\mathrm{B}_d(\cdot)$ is the d-dimensional standard Brownian motion and $h(\cdot, T) : [0,1] \to [T]$ is the time scale function.

Random Scaling Estimator

- For any continuous functional f, $f(\phi_T(\cdot))$ will also weakly converge to $f(\sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}\mathrm{B}_d(\cdot))$.
- Key idea: set f be a self-standardization function to cancel out the scale $\sqrt{\nu}\boldsymbol{G}^{-1}\boldsymbol{S}^{1/2}$.
- Find some points $\{r_m\}$ and studentize $\phi_T(1)$ via $\Pi_T :=$

$$\sum_{m=1}^{T}\left(\phi_T(r_m)-\frac{m}{T}\phi_T(1)\right)\left(\phi_T(r_m)-\frac{m}{T}\phi_T(1)\right)^\top(r_m-r_{m-1}). \tag{3}$$

- $\phi_T(1)^\top \Pi_T^{-1}\phi_T(1)$ is asymptotically pivotal (not normal) since it weakly converges to

$$\mathrm{B}_d(1)^\top\left[\int_0^1(\mathrm{B}_d(r)-g(r)\mathrm{B}_d(1))(\mathrm{B}_d(r)-g(r)\mathrm{B}_d(1))^\top dr\right]^{-1}\mathrm{B}_d(1) \tag{4}$$

with $g:[0,1]\to[0,1]$ determined by $\{E_m\}$.

Random Scaling Estimator

Compared to previous work [LLSS21],

- The first to extend it to FL
- Weaker moments assumption on noises
- Better analysis on uniformly bounding decomposed errors.

Confidence Interval via Random Scaling Estimator

- Let $\widehat{\boldsymbol{V}}_T$ be the empirical estimate of Π_T
- $\widehat{\boldsymbol{V}}_T$ can be updated in an online manner.
- To estimate j-th element $\boldsymbol{x}_j^*$ of $\boldsymbol{x}^*$, we can use

$$\mathbb{P}\left(\left[\bar{\boldsymbol{y}}_{T,j} - q_{\frac{\alpha}{2},g}\sqrt{\widehat{\boldsymbol{V}}_{T,jj}} \le \boldsymbol{x}_j^* \le \bar{\boldsymbol{y}}_{T,j} + q_{\frac{\alpha}{2},g}\sqrt{\widehat{\boldsymbol{V}}_{T,jj}}\right]\right) \to 1-\alpha,$$

where $q_{\frac{\alpha}{2},g}$ is $(1-\alpha/2)$-quantile of the following random variable

$$B_1(1) \Big/ \left(\int_0^1 (B_1(r) - g(r)B_1(1))^2 dr\right)^{1/2} \tag{5}$$

with $B_1(\cdot)$ a one-dimensional standard Brownian motion.

- Only $O(d)$ computation and communication cost per round.

Conclusion

- We have established a (functional) central limit theorem for the averaged iterates of Local SGD.
- We present two fully online inference methods.
- Local SGD simultaneously achieves both statistical efficiency (i.e., optimal asymptotic variance) and communication efficiency (i.e., vanishing ACF).

Other directions:

- Non-smooth and non-strongly-convex counterparts.
- FCLT for proximal or accelerated methods.
- Weak assumptions on noises.

Thanks for listening!

References

[CLT+20] Xi Chen, Jason D Lee, Xin T Tong, Yichen Zhang, et al. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

[LLSS21] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. *arXiv preprint arXiv:2106.03156*, 2021.

[MMR+17] 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 (AISTATS)*, 2017.

[PJ92] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[Sti18] Sebastian U Stich. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.