Sharp Gaussian Approximations for Decentralized Federated Learning

Soham Bonnerjee[1] Sayar Karmakar[2] Wei Biao Wu[1]

[1]University of Chicago; [2]University of Florida

Introduction: DFL & local SGD

- **Setting.** K clients to jointly solve $\theta^\star = \arg\min_\theta \sum_{k=1}^K w_k F_k(\theta)$. Can we apply SGD?
- ***Local*** **SGD**: K is large. Sharing of local gradients only happens periodically.
- ***Decentralized*** **Learning**: Clients may not share local gradients with everybody else. Instead sharing happens through connection matrix C.
- Let $\mathbf{\Theta}_t = (\theta_t^1, \ldots, \theta_t^K) \in \mathbb{R}^{d\times K}$ be client-wise iterates.

Two statistical targets:

- Inference for **PR-averaged** iterate $\bar{Y}_n := K^{-1}\sum_{k=1}^K n^{-1}\sum_{t=1}^n \theta_t^k$.
- Inference for **Entire trajectory**.

Why? and what's new?

- **Known**: Convergence rates, central limit theory.
- **Key practical goals:**
  1. Finite sample results + Multiplier Bootstrap-based inference without needing to estimate asymptotic covariance.
  2. Attack detection by establishing control over entire local SGD trajectory.

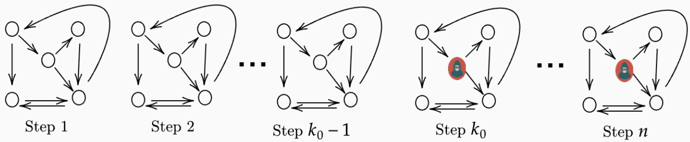

Figure 1: client(s) may turn malicious at some step.
Goal: identification of this step as well as the malicious client

Key Result 1: Berry Esseen

To enable Bootstrap, we require control over

$$d_{\mathcal{C}}\big(\sqrt{n}(\bar{Y}_n - \theta_K^\star), Z\big) := \sup_{A \text{ convex}} \Big|\mathbb{P}(\sqrt{n}(\bar{Y}_n - \theta_K^\star) \in A) - \mathbb{P}(Z \in A)\Big|.$$

We provide the first Berry–Esseen for local SGD. Step size $\eta_t = \eta t^{-\beta}$.

Berry–Esseen (PR-averaged). Under standard strong convexity/smoothness and graph assumptions,

$$d_{\mathcal{C}}\big(\sqrt{n}(\bar{Y}_n - \theta_K^\star), Z\big) \lesssim \frac{1}{\sqrt{n}K} + n^{\frac{1}{2}-\beta}\sqrt{K} + \frac{n^{-\frac{\beta}{2}}}{\sqrt{K}},$$

for a suitable Gaussian Z with covariance Σ_n (finite-sample scaling).

Key Result 1: Berry Esseen

Berry–Esseen (PR-averaged). Under standard strong convexity/smoothness and graph assumptions,

$$d_{\mathcal{C}}(\sqrt{n}(\bar{Y}_n - \theta_K^\star), Z) \;\lesssim\; \frac{1}{\sqrt{n}K} \;+\; n^{\frac{1}{2}-\beta}\sqrt{K} \;+\; \frac{n^{-\frac{\beta}{2}}}{\sqrt{K}},$$

for a suitable Gaussian Z with covariance Σ_n (finite-sample scaling).

- $d_{\mathcal{C}}$ goes to zero as long as $K = o(n^{2\beta-1})$; previously observed in Gu & Chen (2024), but not explicitly quantified.
- Replacing Σ_n by global limit Σ yields error

$$d_{\mathcal{C}}(\sqrt{n}(\bar{Y}_n - \theta_K^\star), N(0, K^{-1}\Sigma)) \;\lesssim\; \sqrt{K}(n^{\frac{1}{2}-\beta} + n^{\beta-1}).$$

- If $K = o(\sqrt{n})$, optimal β is $\beta^\star = \frac{3}{4}$.
- More details in **Section 2** of the camera-ready version.

Key Result 2: time-uniform approximations

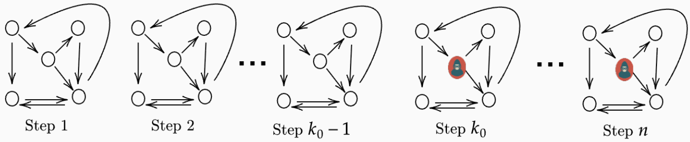

Figure 2: client(s) may turn malicious at some step.
Goal: identification of this step as well as the malicious client

- Any relevant attack detection mechanism will depend on distribution of the entire trajectory; see **Section 3.1** and **Algorithm 2** in the camera-ready version.
- Establish statistical control over trajectory in absence of attackers.

Key Result 2: time-uniform approximations

Establish statistical control over trajectory in absence of attackers.

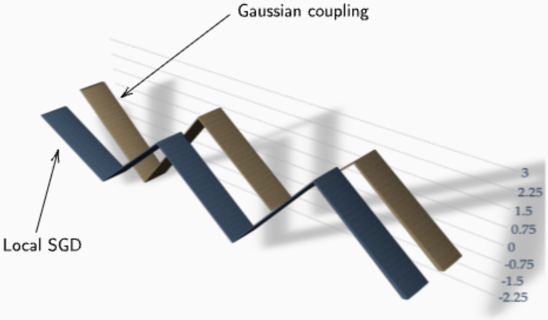

Figure 3: Establish valid Gaussian-process "twins" or couplings to the local SGD process.

Key Result 2: time-uniform approximations

Aggr-GA. Let $Y_t = K^{-1}\sum_{k=1}^{K}\theta_t^k$. Let A be the Hessian. There exists i.i.d. Gaussian variables Z_t such that

$$Y_{t,1}^G = (I - \eta_t A)Y_{t-1,1}^G + \eta_t Z_t K^{-1/2}, \quad Y_{0,1}^G = 0,$$

such that

$$\max_{1\le t\le n}\left|\sum_{s=1}^{t}(Y_s - \theta_K^\star - Y_{s,1}^G)\right| = O_{\mathbb{P}}\big(n^{1-\beta}\big) + o_{\mathbb{P}}\Big(n^{1/p}K^{-1/2}\log n\Big).$$

- We also discuss a client-level time-uniform approximation called **Client-GA**.

Use. Covariance-explicit construction $\Rightarrow$ Gaussian multiplier bootstrap for max/CUSUM-type statistics.

Simulations

An easy alternative: why not prove functional CLT and use the corresponding Gaussian coupling?

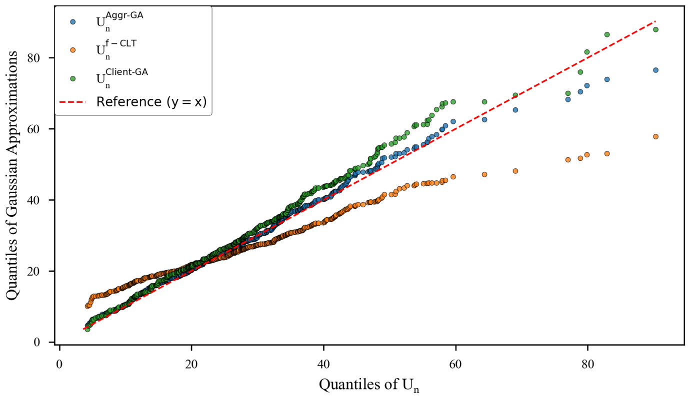

Figure 4: U_n denotes the test statistic for attack detection; x-axis plots the theoretical quantiles, y-axis plots the quantiles based on Gaussian coupling.

Thank You!

Contact `sohambonnerjee@uchicago.edu` for any questions.