Heterogeneous Federated Learning on a Graph

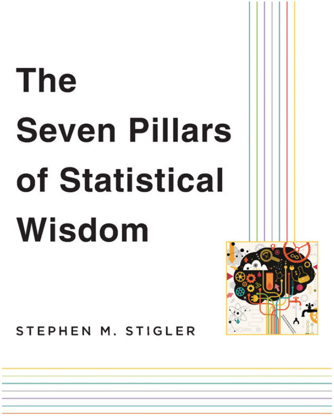
王惠远

第十二届全国概率统计会议

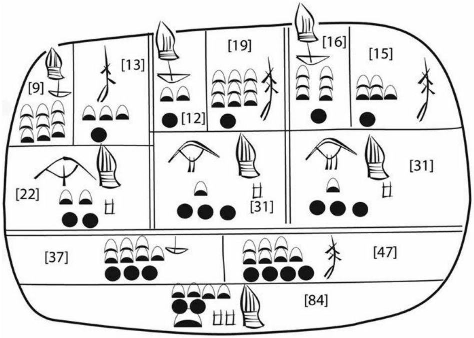
April 20, 2023

Aggregation

- *Aggregation*, or *combination of observations*, is not only the oldest but also the most radical pillar of statistical wisdom
- Gain information beyond individual data values
- Statistical summary is often sufficient



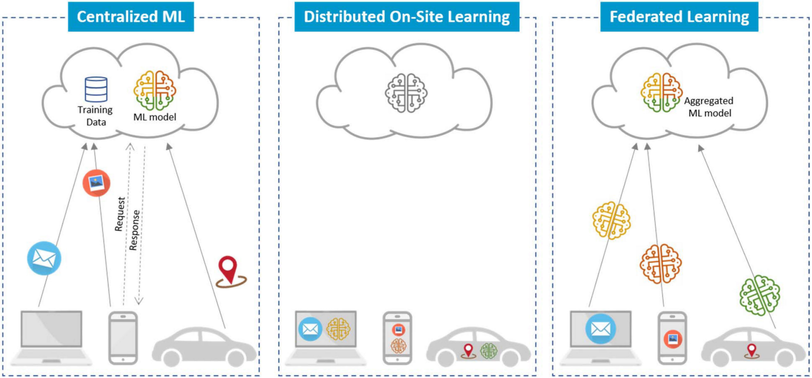
Early History of Aggregation



	Year 1	Year 2	Year 3	Total
Crop A	9	12	16	37
Crop B	13	19	15	47
Total	22	31	31	84

Sumerian tablet (ca. 3000 BCE) and modern contingency table

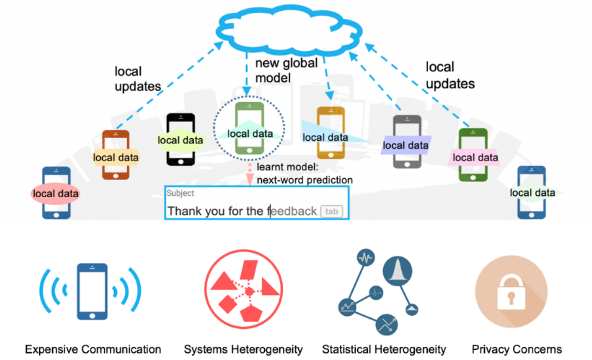
Modern Aggregation



Centralized, distributed, and federated learning

Heterogeneous Federated Learning

- Data heterogeneity; *personalized models* are desired
- Decentralized computation
- Communication heterogeneity



Challenges to Statistics: Heterogeneity of Individuals

- Why aggregation works: borrow strength from *similar* individuals
- Uniqueness of “me” renders $n = 0$: no genuine guinea pig for me (Li and Meng, 2021)
- Challenge to aggregation: individuals are intrinsically *heterogeneous*

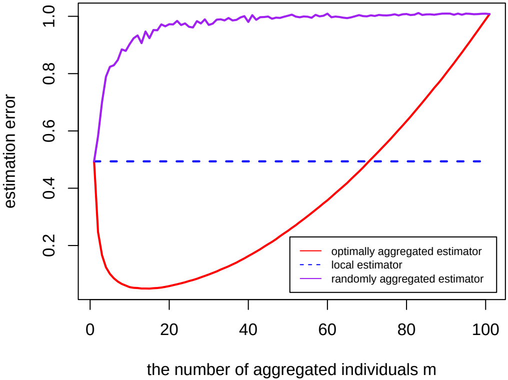
Heterogeneity of Individuals

- *IID* data: the more data aggregated, the more information gained
- *Non-IID* data: heterogeneity may counteract the sample size increase
- An illustrative example

$$y_i^{(k)} = \mu_k + \varepsilon_i^{(k)}, \qquad i = 1, \ldots, n_k,$$

where $\mu_k = 0.02k$, $k = 0, \ldots, 100$, $\varepsilon_i^{(k)} \sim N(0, 1)$, and $n_k = 50$

Trade-off Between Aggregation and Heterogeneity



Performance of *global* and *local* estimators
vs. the *optimally* aggregated estimator for $k = 0$

Problem Setup

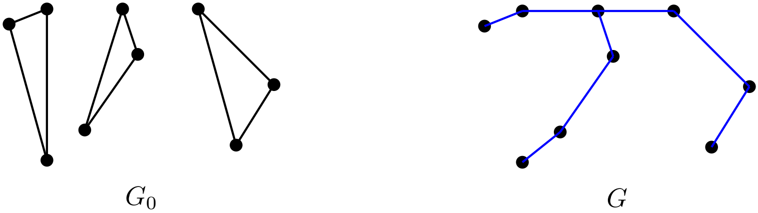
- Consider the general M-estimation problem

$$\boldsymbol{\theta}_u^* = \operatorname*{argmin}_{\boldsymbol{\theta}} E\ell_u(\mathbf{z};\boldsymbol{\theta}), \qquad u \in V$$

- This includes
  - ◇ Mean estimation: $\mathbf{z}_i^{(u)} = \boldsymbol{\theta}_u^* + \boldsymbol{\varepsilon}_i^{(u)}$
  - ◇ Linear regression: $y_i^{(u)} = (\boldsymbol{\theta}_u^*)^T \mathbf{x}_i^{(u)} + \varepsilon_i^{(k)}$
  - ◇ Logistic regression: $P(Y_i^{(u)} = 1 \mid \mathbf{x}_i^{(u)}) = 1/\{1 + \exp(-(\boldsymbol{\theta}_u^*)^T \mathbf{x}_i^{(u)})\}$
- *Characteristic graph* $G_0 = (V, E_0)$: $(u, v) \in E_0$ iff $\boldsymbol{\theta}_u^* = \boldsymbol{\theta}_v^*$

Problem Setup

- Characteristic graph explains heterogeneity, but generally *unknown*
- *Communication graph* $G = (V, E)$ given a priori as a surrogate for G_0
  - ◇ If G_0 is *completely* unknown, Zhao et al. (2023) proved the *minimax* estimation error scales with the same order as the local estimator (e.g. $O(n^{-1})$)



Methodology

- Consider the *network fusion* penalized M-estimator

$$\widehat{\mathbf{\Theta}} = \underset{\boldsymbol{\theta}_u}{\text{argmin}} \underbrace{\frac{1}{|V|} \sum_{u \in V} \frac{1}{n_u} \sum_{i=1}^{n_u} \ell_u(\mathbf{z}_i^{(u)}; \boldsymbol{\theta}_u)}_{\textit{Empirical risk}} + \underbrace{\lambda \sum_{(u,v) \in E} \phi(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v)}_{\textit{Regularization}},$$

where $\phi(\cdot)$ is a norm-based penalty on $\mathbb{R}^p$ such as the *group Lasso* $\phi(\cdot) = \|\cdot\|_1$

- Want to exploit prior information of G about G_0

Assumptions

- *Identifiability.* $\ell(\cdot)$ is convex and twice differentiable, the Hessian matrix $\mathbf{H}_u(\cdot)$ is Lipschitz continuous at $\boldsymbol{\theta}_u^*$
- *Sub-Gaussianity.* The score function $\boldsymbol{\psi}_u(\mathbf{z}_i^{(u)}; \boldsymbol{\theta}_u^*)$ is sub-Gaussian with parameter σ^2
- *Bounded conditional number.* The conditional number of $\widehat{\mathbf{H}}_u(\boldsymbol{\theta})$ is bounded by κ, or its population counterpart
- *Compatibility factor.* For $S = E \setminus E_0 \neq \emptyset$,

$$\kappa_S(\mathbf{D}) \equiv \inf_{\boldsymbol{\Theta}} \frac{\sqrt{|S|}\|\boldsymbol{\Theta}\|_F}{R\{(\mathbf{D}\boldsymbol{\Theta})_S\}} \geq \kappa_0 > 0,$$

where $R\{(\mathbf{D}\boldsymbol{\Theta})_S\} = \sum_{(u,v)\in S} \phi(\boldsymbol{\theta}_u - \boldsymbol{\theta}_v)$

Statistical Guarantees

- Deterministic result. Under appropriate conditions, the penalized M-estimator $\widehat{\boldsymbol{\Theta}}$ satisfies

$$\frac{1}{|V|}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 \leq 2\kappa^2\left(\rho^2 + \frac{4|S|}{\kappa_0}\lambda^2\right),$$

where

$$\rho = \frac{1}{\sqrt{|V|}}\|\Pi_{\mathsf{Ker}(\mathbf{D})}\widehat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^*)\|_F,$$

$$\lambda = \frac{1}{\sqrt{|V|}}R^*\{(\mathbf{D}^+)^T\widehat{\boldsymbol{\Psi}}(\boldsymbol{\Theta}^*)\}$$

  - $S = E \setminus E_0$ measures the bias introduced by aggregating 'wrong' devices
  - $\widehat{\boldsymbol{\Psi}}$: gradient of the empirical risk function
  - $R^*(\cdot)$: Fréchet dual of $R(\cdot)$

Implications

- Assuming sub-Gaussian noises, our rate:

$$\frac{1}{|V|}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 = O_p\left\{\frac{\sigma^2}{\kappa_0}\left(\frac{pK(G)}{n|V|} + \frac{p|E \setminus E_0|}{n|V|}\right)\right\}$$

  where $K(G)$ is the number of connected components of G

- The oracle rate:

$$\frac{1}{|V|}\|\widehat{\boldsymbol{\Theta}}^{\text{oracle}} - \boldsymbol{\Theta}^*\|_F^2 = O_p\left\{\sigma^2 \frac{pK(G_0)}{n|V|}\right\}$$

- Impact of G depends on the *graph fidelity*

$$\mathsf{GF}_{G_0}(G) \equiv \frac{K(G_0)}{K(G) + |E \setminus E_0|} \leq 1$$

  - ◇ $\mathsf{GF}_{G_0}(G) \not\to 0$, in which case $\widehat{\boldsymbol{\Theta}}$ achieves the *oracle* rate
  - ◇ *Aggregation–Heterogeneity trade-off*: $K(G)$ and $|E \setminus E_0|$ cannot be simultaneously small

Edge Selection

- To adapt to the *unknown* structure of G_0, we propose to test

$$H_{0e}: \boldsymbol{\theta}_u^* = \boldsymbol{\theta}_v^* \quad \text{vs.} \quad H_{1e}: \boldsymbol{\theta}_u^* \neq \boldsymbol{\theta}_v^*, \quad e = (u, v) \in E_0$$

- Construct the *Wald test* statistic

$$\widehat{W} = (\widehat{\boldsymbol{\theta}}_u^{\mathsf{loc}} - \widehat{\boldsymbol{\theta}}_v^{\mathsf{loc}})^T (\widehat{\boldsymbol{\Sigma}}_u + \widehat{\boldsymbol{\Sigma}}_v)^{-1} (\widehat{\boldsymbol{\theta}}_u^{\mathsf{loc}} - \widehat{\boldsymbol{\theta}}_v^{\mathsf{loc}})$$

  and select the edge set

$$\widehat{E} = \{e \in E_0 : \widehat{W} \leq \chi_p^2(\alpha/|E_0|)\}$$

- Theorem. Under appropriate conditions,

$$\liminf_{n\to\infty} P(\widehat{E} = E \cap E_0) \geq 1 - \alpha$$

FedADMM

- The augmented Lagrangian

$$L(\boldsymbol{\Theta}, \mathbf{B}, \mathbf{A}) = \frac{1}{|V|} \sum_{u \in V} \widehat{M}_u(\boldsymbol{\theta}_u) + \lambda \sum_{(u,v) \in E} \phi(\boldsymbol{\beta}_{uv} - \boldsymbol{\beta}_{vu})$$
$$- \sum_{(u,v) \in E} \{\boldsymbol{\alpha}_{uv}^T (\boldsymbol{\theta}_u - \boldsymbol{\beta}_{uv}) + \boldsymbol{\alpha}_{vu}^T (\boldsymbol{\theta}_v - \boldsymbol{\beta}_{vu})\}$$
$$+ \frac{\rho}{2} \sum_{(u,v) \in E} (\|\boldsymbol{\theta}_u - \boldsymbol{\beta}_{uv}\|_2^2 + \|\boldsymbol{\theta}_v - \boldsymbol{\beta}_{vu}\|_2^2)$$

FedADMM

1. Sample minibatches $B_u(t)$ on device u
2. *Node optimization step.* Update $\boldsymbol{\theta}_u$ on device u in the form of SGD
3. Broadcast $\boldsymbol{\theta}_u$ to neighboring devices
4. *Edge communication step.* On either device u or v such that $(u, v) \in E$,
   - Update $\boldsymbol{\beta}_{uv}$ and $\boldsymbol{\beta}_{vu}$
   - Update $\boldsymbol{\alpha}_{uv}$ and $\boldsymbol{\alpha}_{vu}$
5. Broadcast $(\boldsymbol{\beta}_{uv}, \boldsymbol{\beta}_{vu})$ and $(\boldsymbol{\alpha}_{uv}, \boldsymbol{\alpha}_{vu})$ to neighboring devices

FedADMM

- Convergence. Under appropriate conditions,

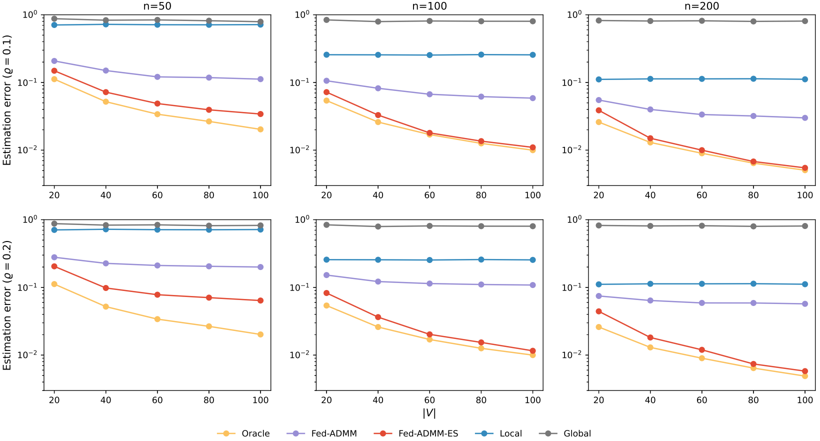
$$\frac{1}{|V|}E\|\boldsymbol{\Theta}_T - \widehat{\boldsymbol{\Theta}}\|_F^2 = O(T^{-1}\log T)$$

- Extension to *communication heterogeneity* by inverse probability weighting

$$\widehat{\mathbf{g}}_u = \frac{1}{|B_i(t)|}\sum_{i\in B_u(t)}\frac{R_u(t)}{\pi_u}\boldsymbol{\psi}_u(\mathbf{z}_i^{(u)};\boldsymbol{\theta}_u)$$

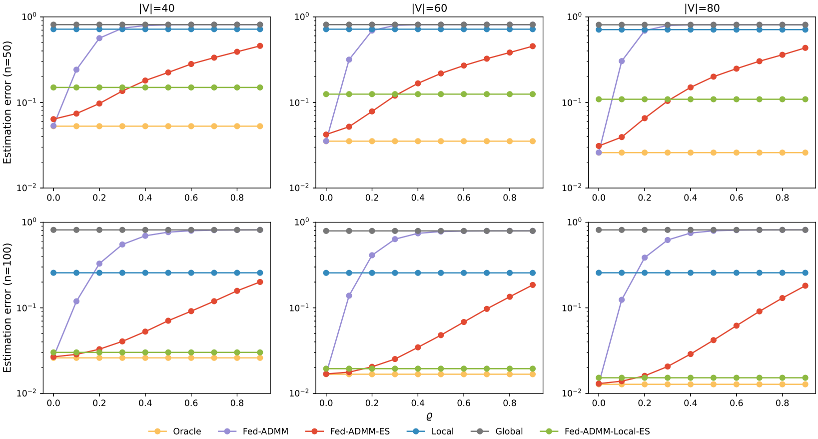
- Convergence rate $O((\pi_{\min}T)^{-1}\log T)$

Simulation Studies



Performance as the network grows

Simulation Studies



Sensitivity to graph corruption

Real Data Example

- 2020 U.S. presidential election data: 29 states with > 50 counties
- Prediction by *logistic regression* with 52 county-level predictors
- *Two thirds* of the counties for training and the rest for testing

Method	Local	Global	FedADMM-ES	FedADMM-Hist
Accuracy	0.741 (0.034)	0.752 (0.012)	0.793 (0.019)	0.742 (0.011)

Discussion

- Take-home message
  - Aggregation–heterogeneity trade-off is fundamental to federated learning
  - Simply pooling all data may not be optimal, and selective aggregation can be effective
  - Network topology plays a key role
- Future work
  - Aggregation–heterogeneity trade-off in multi-central distributed learning
  - Edge selection with error control
  - High-dimensional M-estimation
  - Beyond M-estimation, e.g., deep learning

- Wang, H., Zhao, X., and Lin, W. (2022). *Heterogeneous federated learning on a graph*. arXiv:2209.08737

Welcome discussion!