Fast Composite Optimization and Statistical Recovery in Federated Learning

Yajie Bao [1] Michael Crawshaw [2] Shan Luo [1] Mingrui Liu [2]

[1]School of Mathematical Sciences, Shanghai Jiao Tong University

[2]Department of Computer Science, George Mason University

June 26, 2022

SHANGHAI JIAO TONG UNIVERSITY

GEORGE MASON UNIVERSITY

Outline

Federated Learning Environment

Consider a federated learning problem with K clients,

- Local loss: $\mathcal{L}_k(\boldsymbol{w})$;
- Local weight: π_k;
- Non-smooth regularizer $h(\cdot)$.

We will consider two instances of the FL problem:

$$\arg\min_{\boldsymbol{w}\in\mathcal{W}} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w}) + h(\boldsymbol{w}) \right\}. \tag{1.1}$$

We denote the global loss as $\mathcal{L} = \sum_{k=1}^{K} \pi_k \mathcal{L}_k$ and the global composite objective as $\phi = \mathcal{L} + h$.

Examples: federated sparse linear regression, federated low-rank matrix estimation.

Federated Composite Optimization

For general **composite optimization** in FL,

- Local population distribution: $\mathcal{P}_k$;
- Local population loss: $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi \sim \mathcal{P}_k}[f(\boldsymbol{w}; \xi)]$;
- Non-smooth convex regularizer: $h(\cdot)$.

With corresponding objective

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{L}_k(\boldsymbol{w}) + h(\boldsymbol{w}) \right\}. \tag{1.2}$$

Matches previous analyses which ignore the statistical estimation problem.

Federated Statistical Recovery

For **statistical recovery** problem in FL,

- Local empirical distribution: $\mathcal{D}_k$;
- Local empirical loss: $\mathcal{L}_k(\boldsymbol{w}) = \mathbb{E}_{\xi\sim\mathcal{D}_k}[f(\boldsymbol{w};\xi)]$;
- Non-smooth norm regularizer: $\mathcal{R}(\cdot)$, which is decomposable [Negahban et al., 2012];
- The ground-truth parameter: $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \sum_{k=1}^K \pi_k \mathbb{E}_{\xi\sim\mathcal{P}_k}[f(\boldsymbol{w};\xi)]$.

We want to obtain the ground-truth parameter $\boldsymbol{w}^*$ through solving:

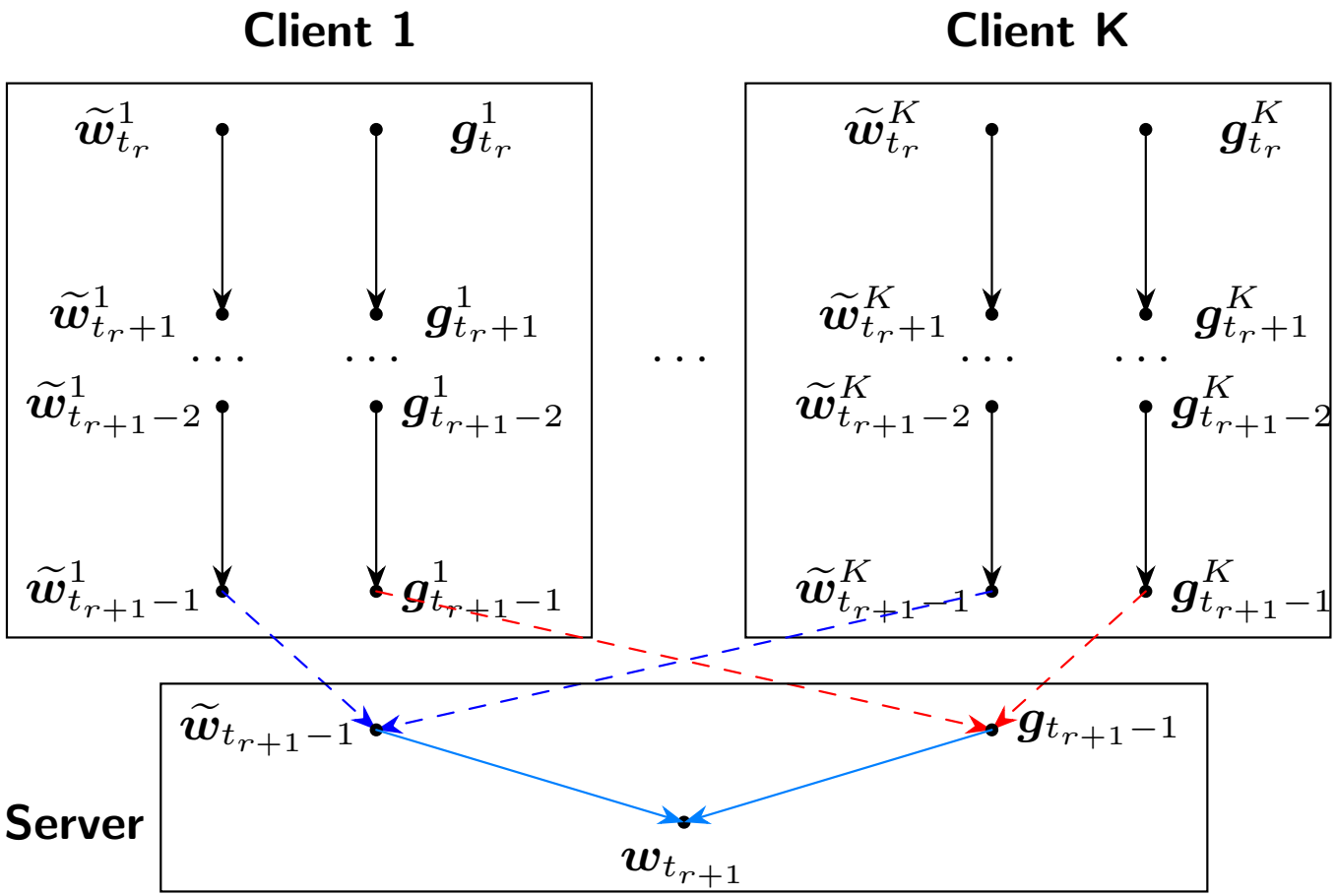
$$\widehat{\boldsymbol{w}}_{\text{opt}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \left\{ \sum_{k=1}^K \pi_k \mathcal{L}_k(\boldsymbol{w}) + \lambda_{\text{opt}} \mathcal{R}(\boldsymbol{w}) \right\}, \tag{1.3}$$

where λ_{opt} is the *optimal regularization parameter*.

Our goal is to design an algorithm to recover $\boldsymbol{w}^*$ with the optimal statistical precision $\|\widehat{\boldsymbol{w}}_{\text{opt}} - \boldsymbol{w}^*\| \lesssim \epsilon_{\text{stat}} := \Psi(\bar{\mathcal{M}})\lambda_{\text{opt}}/\mu$ [Negahban et al., 2012].

Outline

Fast-FedDA


Client 1
Client K
$\widetilde{\boldsymbol{w}}_{t_r}^1$
$\boldsymbol{g}_{t_r}^1$
$\widetilde{\boldsymbol{w}}_{t_r+1}^1$
$\boldsymbol{g}_{t_r+1}^1$
$\widetilde{\boldsymbol{w}}_{t_{r+1}-2}^1$
$\boldsymbol{g}_{t_{r+1}-2}^1$
$\widetilde{\boldsymbol{w}}_{t_{r+1}-1}^1$
$\boldsymbol{g}_{t_{r+1}-1}^1$
$\widetilde{\boldsymbol{w}}_{t_r}^K$
$\boldsymbol{g}_{t_r}^K$
$\widetilde{\boldsymbol{w}}_{t_r+1}^K$
$\boldsymbol{g}_{t_r+1}^K$
$\widetilde{\boldsymbol{w}}_{t_{r+1}-2}^K$
$\boldsymbol{g}_{t_{r+1}-2}^K$
$\widetilde{\boldsymbol{w}}_{t_{r+1}-1}^K$
$\boldsymbol{g}_{t_{r+1}-1}^K$
Server
$\widetilde{\boldsymbol{w}}_{t_{r+1}-1}$
$\boldsymbol{g}_{t_{r+1}-1}$
$\boldsymbol{w}_{t_{r+1}}$

Convergence Rate of Fast-FedDA

We impose the following assumptions:

1. $\mathcal{L}_k$ for $k \in [K]$ are μ-strongly convex and L-smooth.
2. $h : \mathcal{W} \to \mathbb{R}$ is a closed convex function.
3. Bounded heterogeneity: $\|\nabla \mathcal{L}_k(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{w})\| \leq H$ for any $\boldsymbol{w} \in \mathcal{W}$.
4. Bounded variance of stochastic gradient: $\mathbb{E}_{\xi \sim \mathcal{P}_k}[\|\nabla f(\boldsymbol{w}; \xi) - \nabla \mathcal{L}_k(\boldsymbol{w})\|^2] \leq \sigma^2$.

Theorem

Under Assumptions 1-4, with $\alpha_t = (t+a)^2$ *and* $\gamma = \mu a^3$ *with* $a \geq 4L/\mu$ *in Algorithm* `Fast-FedDA`,

$$\mathbb{E}_{\mathcal{P}}[\phi(\widehat{\boldsymbol{w}}_{\text{Fast-FedDA}}) - \phi(\widehat{\boldsymbol{w}})] \leq O\left(\frac{\sigma^2}{K\mu T}\right) \tag{2.1}$$

for equal-weighted case ($\pi_1 = \ldots = \pi_K = 1/K$).

Outline

RSC and RSM

Widely used in statistical recovery literature [Agarwal et al., 2012, Wang et al., 2014, Loh and Wainwright, 2015, Cai et al., 2020]. Denote

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') = \mathcal{L}_k(\boldsymbol{w}) - \mathcal{L}_k(\boldsymbol{w}') - \langle \nabla \mathcal{L}_k(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle.$$

Assumption (RSC)

The local loss functions $\mathcal{L}_k$ for $k = 1, ..., K$ are convex and there exist $\mu > 0$ and $\tau_k \geq 0$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$:

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2 - \tau_k \mathcal{R}^2(\boldsymbol{w} - \boldsymbol{w}').$$

.

Assumption (RSM)

For the local loss functions $\mathcal{L}_k$ for $k = 1, ..., K$, there exist $L > 0$ and $\nu_k \geq 0$ such that for any $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathcal{W}$:

$$\mathcal{T}_k(\boldsymbol{w}, \boldsymbol{w}') \leq \frac{L}{2} \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + \nu_k \mathcal{R}^2(\boldsymbol{w} - \boldsymbol{w}').$$

Multi-stage Constrained FedDA

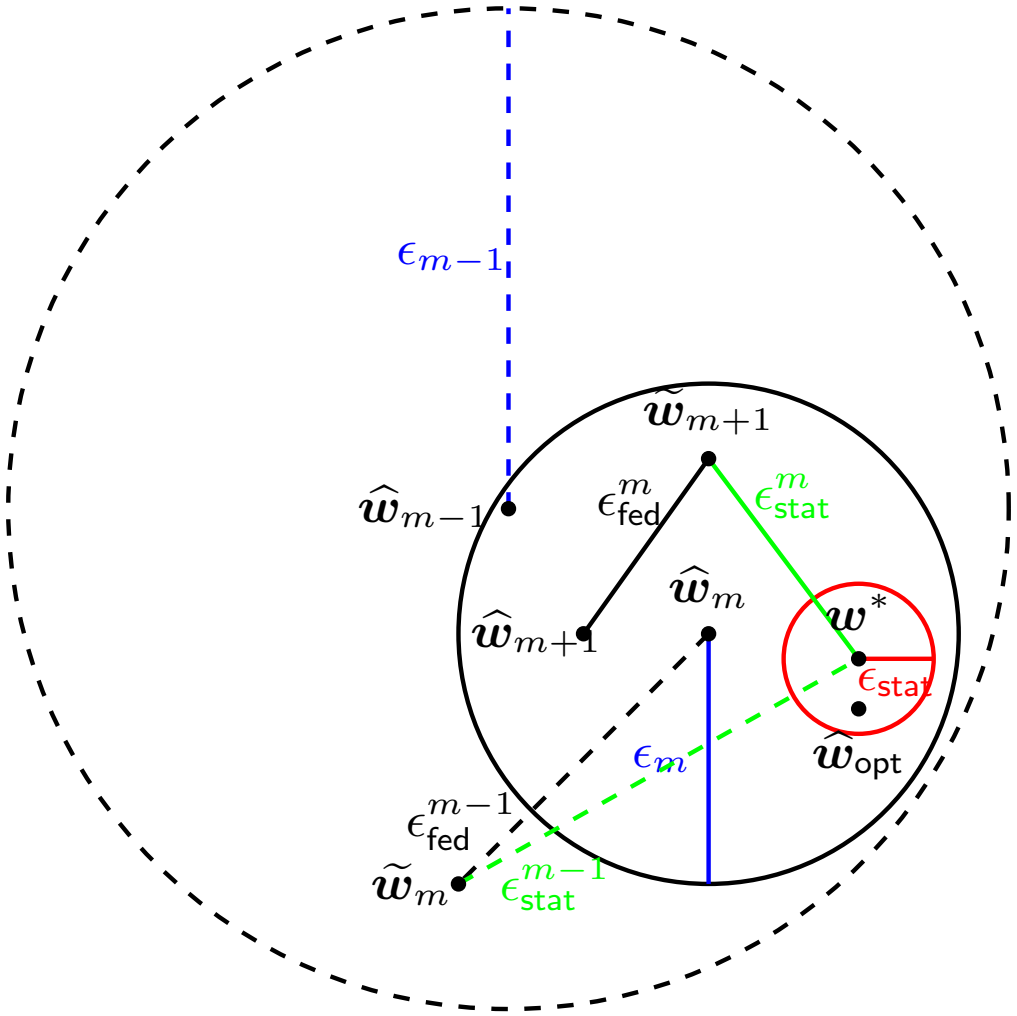


Figure: Optimization landscape in $\mathcal{R}$ norm at the m-th stage of `MC-FedDA`.

Complexity of MC-FedDA

With high probability, we can guarantee bounds on optimization error and estimation error:

- Optimization: $\phi(\widehat{\boldsymbol{w}}_{\text{MC-FedDA}}) - \phi(\widehat{\boldsymbol{w}}_{\text{opt}}) \leq \frac{\Psi^2(\bar{\mathcal{M}})\lambda_{\text{opt}}^2}{\mu}$.
- Estimation: $\|\widehat{\boldsymbol{w}}_{\text{MC-FedDA}} - \boldsymbol{w}^*\| \leq \frac{4\Psi(\bar{\mathcal{M}})\lambda_{\text{opt}}}{\mu} = 4\epsilon_{\text{stat}}$.

When $\pi_1 = \ldots = \pi_K = 1/K$:

- $\epsilon_{\text{stat}} = \Psi(\bar{\mathcal{M}})\lambda_{\text{opt}}/\mu$ converges to 0 as the total sample size $N \to \infty$.
- Let $\epsilon = \mu\epsilon_{\text{stat}}^2$, if the total number of iterations satisfies $T = \widetilde{\mathcal{O}}(\Psi(\bar{\mathcal{M}})^2\sigma^2/(K\mu\epsilon))$, then we are guaranteed that $\phi(\widehat{\boldsymbol{w}}_{\text{MC-FedDA}}) - \phi(\widehat{\boldsymbol{w}}_{\text{opt}}) \leq \epsilon$.
- The total communication complexity is bounded by $\tilde{\mathcal{O}}(T^{1/2}K^{1/2})$, matching the best known result of FedAvg for unconstrained problem [Woodworth et al., 2020, Stich, 2019].

Outline

Experiment Results

We compare our proposed algorithms with Federated Mirror Descent (`FedMiD`) and Federated Dual Averaging (`FedDA`) algorithms introduced in Yuan et al. [2021].

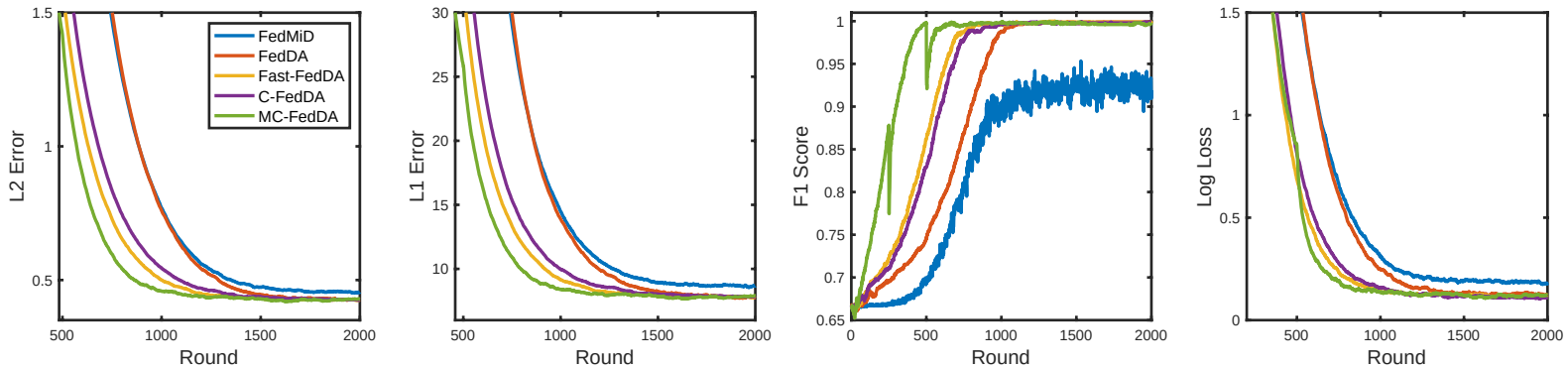


Figure: Recovery results for federated sparse linear regression problem.

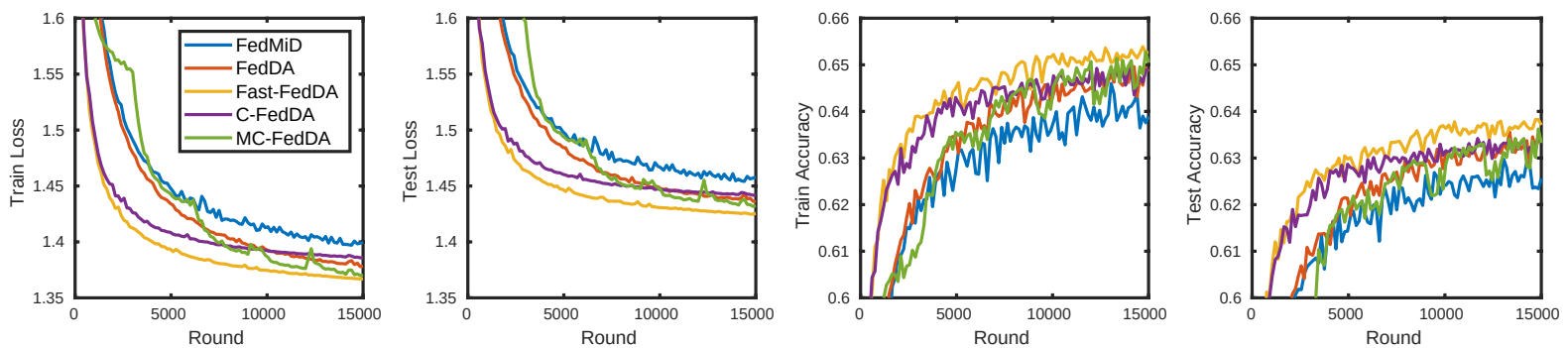


Figure: Results for federated sparse logistic regression on EMNIST-62 dataset.

Reference I

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, pages 2452–2482, 2012.

T. T. Cai, Y. Wang, and L. Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(1):559–616, 2015.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of statistics*, 42(6):2164, 2014.

B. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local SGD for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020.

H. Yuan, M. Zaheer, and S. Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.