# A Non-parametric View of FedAvg and FedProx: Beyond Stationary Points


Lili Su

Department of Electrical and Computer Engineering
Northeastern University

Joint work with
Jiaming Xu (Duke) and Pengkun Yang (Tsinghua)


PoDL,
Oct, 2023

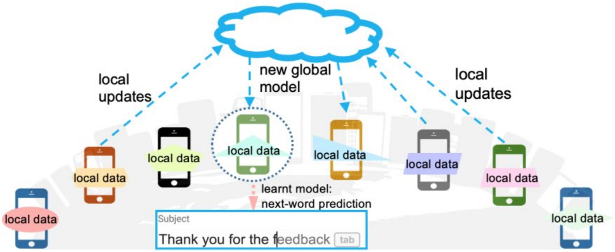
# Modern data generation and collection

- Machine learning at the edge (ML@Edge)
- Enabling technologies: edge devices + 5G

# Modern data generation and collection

- Machine learning at the edge (ML@Edge)
- Enabling technologies: edge devices + 5G
  - Privacy-preserving distributed machine learning

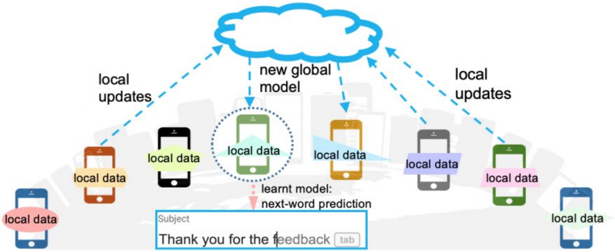
# Federated learning




Example: Gboard (Google keyboard) [HRM+18]

- Data privacy: training models without seeing your data

# Federated learning




Example: Gboard (Google keyboard) [HRM+18]

- Data privacy: training models without seeing your data
- Caveat: information leakage from model update!
  - This is unavoidable from an information theory prospective.

# “Perturbations” in federated learning

- Massive scale

# “Perturbations” in federated learning

- Massive scale
  $\rightarrow$ communication bottleneck

# “Perturbations” in federated learning

- Massive scale
  $\rightarrow$ communication bottleneck
  $\rightarrow$ communication-saving algorithms cause distortion

# “Perturbations” in federated learning

- Massive scale
  $\rightarrow$ communication bottleneck
  $\rightarrow$ communication-saving algorithms cause distortion
- Heterogeneity
  - Data distribution
  - Data volume
- Unreliable communication: connection and/or availability

# “Perturbations” in federated learning

- Massive scale
  $\rightarrow$ communication bottleneck
  $\rightarrow$ **communication-saving algorithms cause distortion**
- Heterogeneity
  - Data distribution
  - Data volume
- Unreliable communication: connection and/or availability

# Popular federated learning algorithms

For every communication round

- Parameter server (PS) broadcast latest model
- Clients update model based on local data
  - **FedAvg** [MMR+17]: run $s$ steps of local gradient descent
  - **FedProx** [LSZ+20]: solve a local program with a proximal term
- PS aggregates updated models from clients

# Popular federated learning algorithms

For every communication round

- Parameter server (PS) broadcast latest model
- Clients update model based on local data
  - **FedAvg** [MMR+17]: run $s$ steps of local gradient descent
  - **FedProx** [LSZ+20]: solve a local program with a proximal term
- PS aggregates updated models from clients

Many others FL algorithms: SCAFFOLD[KKM+20], FedNova[WLL+20], FedSplit[PW20], FedPD[ZHD+21] ...

# Failure of reaching stationary points

Linear regression: client $i$ holds local dataset $(X_i, y_i)$

$$X_i \in \mathbb{R}^{n_i \times d}, \quad y_i \in \mathbb{R}^{n_i}$$

- Objective function of ordinary least squares (OLS):

$$\min_{\theta} \; f(\theta) \triangleq \sum_{i=1}^{M} \|y_i - X_i\theta\|^2$$

# Failure of reaching stationary points

Linear regression: client $i$ holds local dataset $(X_i, y_i)$

$$X_i \in \mathbb{R}^{n_i \times d}, \quad y_i \in \mathbb{R}^{n_i}$$

- Objective function of ordinary least squares (OLS):

$$\min_{\theta} \ f(\theta) \triangleq \sum_{i=1}^{M} \|y_i - X_i \theta\|^2$$

- Closed form solution

$$\hat{\theta}_{\mathrm{OLS}} = \left( \frac{1}{M} \sum_{i=1}^{M} X_i^\top X_i \right)^{-1} \left( \frac{1}{M} \sum_{i=1}^{M} X_i^\top y_i \right)$$

# Failure of reaching stationary points

- For linear regression [Pathak-Wainwright'20]

$$\hat{\theta}_{\text{FedAvg}} = \left( \frac{1}{M} \sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell \right)^{-1} \left( \frac{1}{M} \sum_{i=1}^{M} \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell X_i^\top y_i \right)$$

$$\hat{\theta}_{\text{FedProx}} = \left( I - \frac{1}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} \right)^{-1} \left( \frac{\eta}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} X_i^\top y_i \right)$$

# Failure of reaching stationary points

- For linear regression [Pathak-Wainwright'20]

$$\hat{\theta}_{\mathrm{FedAvg}} = \left( \frac{1}{M} \sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell \right)^{-1} \left( \frac{1}{M} \sum_{i=1}^{M} \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell X_i^\top y_i \right)$$

$$\hat{\theta}_{\mathrm{FedProx}} = \left( I - \frac{1}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} \right)^{-1} \left( \frac{\eta}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} X_i^\top y_i \right)$$

- Failure of reaching stationary points: $\hat{\theta}_{\mathrm{Fed}} \neq \hat{\theta}_{\mathrm{OLS}}$

# Failure of reaching stationary points

- For linear regression [Pathak-Wainwright'20]

$$\hat{\theta}_{\mathrm{FedAvg}} = \left( \frac{1}{M} \sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell \right)^{-1} \left( \frac{1}{M} \sum_{i=1}^{M} \sum_{\ell=0}^{s-1} (I - \eta X_i^\top X_i)^\ell X_i^\top y_i \right)$$

$$\hat{\theta}_{\mathrm{FedProx}} = \left( I - \frac{1}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} \right)^{-1} \left( \frac{\eta}{M} \sum_{i=1}^{M} (I + \eta X_i^\top X_i)^{-1} X_i^\top y_i \right)$$

- Failure of reaching stationary points: $\hat{\theta}_{\mathrm{Fed}} \neq \hat{\theta}_{\mathrm{OLS}}$
- Many attempts to fix the optimization gap [KKM+20, PW20, GHR21,. . . ]

# Theory behind practice

**Question**

Do they really fail? In many applications, FedAvg and FedProx work quite well, despite the theoretical gap.

# Theory behind practice

**Question**

Do they really fail? In many applications, FedAvg and FedProx work quite well, despite the theoretical gap.

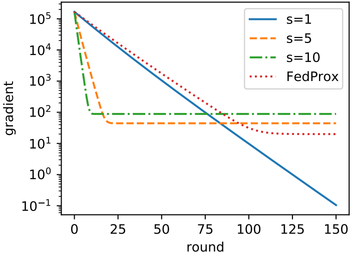
- Model: $y_i = X_i\theta^* + \xi_i$

# Theory behind practice

**Question**

Do they really fail? In many applications, FedAvg and FedProx work quite well, despite the theoretical gap.

- Model: $y_i = X_i\theta^* + \xi_i$

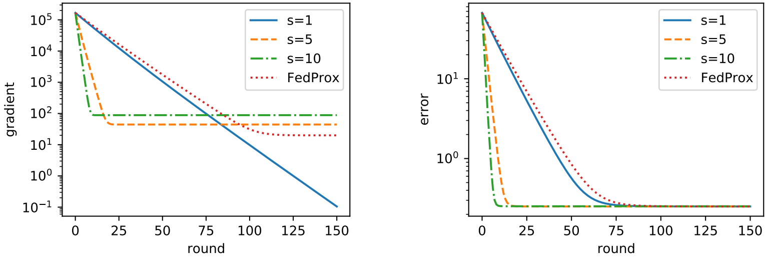


# Theory behind practice

## Question

Do they really fail? In many applications, FedAvg and FedProx work quite well, despite the theoretical gap.

- Model: $y_i = X_i\theta^* + \xi_i$

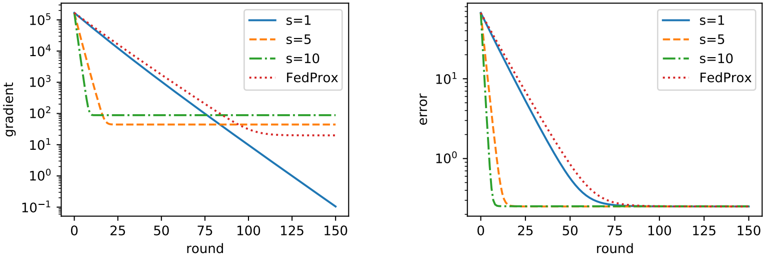


# Theory behind practice

**Question**

Do they really fail? In many applications, FedAvg and FedProx work quite well, despite the theoretical gap.

- Model: $y_i = X_i \theta^* + \xi_i$




- Why FedAvg and FedProx can achieve low estimation errors despite their failure of reaching stationary points?

# Statistical perspective: unbiasedness

$$\hat{\theta}_{\mathrm{OLS}} = \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top y_i\right)$$

Plugging the model $y_i = X_i\theta^* + \xi_i$

$$\Rightarrow \hat{\theta}_{\mathrm{OLS}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top \xi_i\right)$$

# Statistical perspective: unbiasedness

$$\hat{\theta}_{\mathrm{OLS}} = \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top y_i\right)$$

Plugging the model $y_i = X_i\theta^* + \xi_i$

$$\Rightarrow \hat{\theta}_{\mathrm{OLS}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \color{red}{\left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top \xi_i\right)}$$

Similarly, for $\hat{\theta}_{\mathrm{FedAvg}}$ and $\hat{\theta}_{\mathrm{FedProx}}$.

# Statistical perspective: unbiasedness

Plugging the model $y_i = X_i\theta^* + \xi_i$:

$$\hat{\theta}_{\mathrm{OLS}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top \xi_i\right)$$

$$\hat{\theta}_{\mathrm{FedAvg}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1}(I - \eta X_i^\top X_i)^\ell\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M}\sum_{\ell=0}^{s-1}(I - \eta X_i^\top X_i)^\ell X_i^\top \xi_i\right)$$

# Statistical perspective: unbiasedness

Plugging the model $y_i = X_i\theta^* + \xi_i$:

$$\hat{\theta}_{\text{OLS}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top \xi_i\right)$$

$$\hat{\theta}_{\text{FedAvg}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1}(I-\eta X_i^\top X_i)^\ell\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M}\sum_{\ell=0}^{s-1}(I-\eta X_i^\top X_i)^\ell X_i^\top \xi_i\right)$$

**Observation**

Both (and also FedProx) are unbiased estimator of $\theta^*$ with different variances.

# Statistical perspective: unbiasedness

Plugging the model $y_i = X_i\theta^* + \xi_i$:

$$\hat{\theta}_{\text{OLS}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i\right)^{-1} \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top \xi_i\right)$$

$$\hat{\theta}_{\text{FedAvg}} = \theta^* + \left(\frac{1}{M}\sum_{i=1}^{M} X_i^\top X_i \sum_{\ell=0}^{s-1}(I - \eta X_i^\top X_i)^\ell\right)^{-1}$$

$$\left(\frac{1}{M}\sum_{i=1}^{M}\sum_{\ell=0}^{s-1}(I - \eta X_i^\top X_i)^\ell X_i^\top \xi_i\right)$$

**Observation**

**All Roads Lead to Rome !!!**

# Understanding FedAvg and FedProx

**Model:** $f_i^* \in \mathcal{H}$ for some RKHS $\mathcal{H}$ on client $i \in [M]$,

$$y_{ij} = f_i^*(x_{ij}) + \xi_{ij} \quad j = 1, \ldots, n_i$$

Let $N = \sum_{i=1}^{M} n_i$ is the total number of data points

# Understanding FedAvg and FedProx

**Model:** $f_i^* \in \mathcal{H}$ for some RKHS $\mathcal{H}$ on client $i \in [M]$,

$$y_{ij} = f_i^*(x_{ij}) + \xi_{ij} \quad j = 1, \ldots, n_i$$

Let $N = \sum_{i=1}^{M} n_i$ is the total number of data points

**Algorithm:** at communication round $t$

- Parameter server (PS) broadcast $f_{t-1}$

# Understanding FedAvg and FedProx

**Model:** $f_i^* \in \mathcal{H}$ for some RKHS $\mathcal{H}$ on client $i \in [M]$,

$$y_{ij} = f_i^*(x_{ij}) + \xi_{ij} \quad j = 1, \ldots, n_i$$

Let $N = \sum_{i=1}^{M} n_i$ is the total number of data points

**Algorithm:** at communication round $t$

- Parameter server (PS) broadcast $f_{t-1}$
- Local update $f_{i,t}$ based on empirical risk function

$$\ell_i(f) = \frac{1}{2n_i} \sum_{j=1}^{n_i} (f(x_{ij}) - y_{ij})^2$$

# Understanding FedAvg and FedProx

**Model:** $f_i^* \in \mathcal{H}$ for some RKHS $\mathcal{H}$ on client $i \in [M]$,

$$y_{ij} = f_i^*(x_{ij}) + \xi_{ij} \quad j = 1, \ldots, n_i$$

Let $N = \sum_{i=1}^M n_i$ is the total number of data points

**Algorithm:** at communication round $t$

- Parameter server (PS) broadcast $f_{t-1}$
- Local update $f_{i,t}$ based on empirical risk function

$$\ell_i(f) = \frac{1}{2n_i} \sum_{j=1}^{n_i} (f(x_{ij}) - y_{ij})^2$$

- Global update by model averaging

$$f_t = \sum_{i=1}^M w_i f_{i,t}, \quad w_i = n_i/N$$

# Local updates of FedAvg and FedProx

**FedAvg:** one-step local gradient descent $\mathcal{G}_i(f) = f - \eta \nabla \ell_i(f)$

$$f_{i,t} = \mathcal{G}_i^s(f_{t-1}) \triangleq (\underbrace{\mathcal{G}_i \circ \cdots \circ \mathcal{G}_i}_{s \text{ times}})(f_{t-1})$$

**FedProx:**

$$f_{i,t} = \arg\min_{f \in \mathcal{H}} \ \ell_i(f) + \frac{1}{2\eta} \|f - f_{t-1}\|_{\mathcal{H}}^2$$

# Evolution of in-sample prediction

Representer in RKHS: $k_x = k(\cdot, x)$.
Let $(K_{\mathbf{x}})_{ij} = \frac{1}{N} k(x_i, x_j)$ be the normalized Gram matrix.

# Evolution of in-sample prediction

Representer in RKHS: $k_x = k(\cdot, x)$.
Let $(K_{\mathbf{x}})_{ij} = \frac{1}{N} k(x_i, x_j)$ be the normalized Gram matrix.

**Proposition (Su-Xu-Yang '23)**

$$f_t(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y,$$

*where $P \in \mathbb{R}^{N \times N}$ is a block-diagonal matrix whose $i$-th diagonal block of size $n_i \times n_i$ is*

$$P_{ii} = \begin{cases} \sum_{\tau=0}^{s-1} [I - \eta K_{\mathbf{x}_i}]^{\tau} & \text{for FedAvg,} \\ [I + \eta K_{\mathbf{x}_i}]^{-1} & \text{for FedProx.} \end{cases}$$

# Evolution of in-sample prediction

Representer in RKHS: $k_x = k(\cdot, x)$.
Let $(K_{\mathbf{x}})_{ij} = \frac{1}{N} k(x_i, x_j)$ be the normalized Gram matrix.

**Proposition (Su-Xu-Yang '23)**

$$f_t(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y,$$

*where $P \in \mathbb{R}^{N \times N}$ is a block-diagonal matrix whose $i$-th diagonal block of size $n_i \times n_i$ is*

$$P_{ii} = \begin{cases} \sum_{\tau=0}^{s-1} [I - \eta K_{\mathbf{x}_i}]^\tau & \text{for FedAvg,} \\ [I + \eta K_{\mathbf{x}_i}]^{-1} & \text{for FedProx.} \end{cases}$$

Key analysis challenge: eigenvalues of $I - \eta K_{\mathbf{x}} P$ (asymmetric)

# Proof idea of in-sample predictions

$$
\begin{aligned}
f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\
&= \left[I - \eta K_{\mathbf{x}} P\right] f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y
\end{aligned}
$$

# Proof idea of in-sample predictions

$$
\begin{aligned}
f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\
&= \left[I - \eta K_{\mathbf{x}} P\right] f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y
\end{aligned}
$$

For FedAvg:

- $\Psi = \frac{\eta}{N} \sum_{\tau=0}^{s-1} (\mathcal{L}_1^{\tau} k_{\mathbf{x}_1}, \ldots, \mathcal{L}_M^{\tau} k_{\mathbf{x}_M})$

$$
\mathcal{L}_i k_{\mathbf{x}_i} = (I - \eta K_{\mathbf{x}_i}) k_{\mathbf{x}_i} \quad \text{(kernel method)}
$$

# Proof idea of in-sample predictions

$$\begin{aligned} f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\ &= \left[I - \eta K_{\mathbf{x}} P\right] f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y \end{aligned}$$

For FedAvg:

- $\Psi = \frac{\eta}{N} \sum_{\tau=0}^{s-1} (\mathcal{L}_1^\tau k_{\mathbf{x}_1}, \ldots, \mathcal{L}_M^\tau k_{\mathbf{x}_M})$

$$\begin{aligned} \mathcal{L}_i k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i}) k_{\mathbf{x}_i} \quad \text{(kernel method)} \\ \Longrightarrow \mathcal{L}_i^\tau k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i})^\tau k_{\mathbf{x}_i} \end{aligned}$$

# Proof idea of in-sample predictions

$$
\begin{aligned}
f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\
&= \left[I - \eta K_{\mathbf{x}} P\right] f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y
\end{aligned}
$$

For FedAvg:

- $\Psi = \frac{\eta}{N} \sum_{\tau=0}^{s-1} (\mathcal{L}_1^\tau k_{\mathbf{x}_1}, \ldots, \mathcal{L}_M^\tau k_{\mathbf{x}_M})$

$$
\begin{aligned}
&\mathcal{L}_i k_{\mathbf{x}_i} = (I - \eta K_{\mathbf{x}_i}) k_{\mathbf{x}_i} \quad \text{(kernel method)} \\
\Longrightarrow\ &\mathcal{L}_i^\tau k_{\mathbf{x}_i} = (I - \eta K_{\mathbf{x}_i})^\tau k_{\mathbf{x}_i} \\
\Longrightarrow\ &\Psi(\mathbf{x}) = \eta K_{\mathbf{x}} P
\end{aligned}
$$

# Proof idea of in-sample predictions

$$
\begin{aligned}
f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\
&= \left[I - \eta K_{\mathbf{x}} P\right] f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y
\end{aligned}
$$

For FedAvg:

- $\Psi = \frac{\eta}{N} \sum_{\tau=0}^{s-1} (\mathcal{L}_1^\tau k_{\mathbf{x}_1}, \ldots, \mathcal{L}_M^\tau k_{\mathbf{x}_M})$

$$
\begin{aligned}
\mathcal{L}_i k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i}) k_{\mathbf{x}_i} \quad \text{(kernel method)} \\
\implies \mathcal{L}_i^\tau k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i})^\tau k_{\mathbf{x}_i} \\
\implies \Psi(\mathbf{x}) &= \eta K_{\mathbf{x}} P
\end{aligned}
$$

- Telescoping sum

$$
f - \mathcal{L}_i^s f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau f - \mathcal{L}_i^{\tau+1} f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau \left( \frac{\eta}{n_i} \sum_{j=1}^{n_i} f(x_{ij}) k_{x_{ij}} \right)
$$

# Proof idea of in-sample predictions

$$\begin{aligned} f_t(\mathbf{x}) &= \mathcal{L} f_{t-1}(\mathbf{x}) + \Psi(\mathbf{x}) y \\ &= [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y \end{aligned}$$

For FedAvg:

- $\Psi = \frac{\eta}{N} \sum_{\tau=0}^{s-1} (\mathcal{L}_1^\tau k_{\mathbf{x}_1}, \dots, \mathcal{L}_M^\tau k_{\mathbf{x}_M})$

$$\begin{aligned} \mathcal{L}_i k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i}) k_{\mathbf{x}_i} \quad \text{(kernel method)} \\ \Longrightarrow \mathcal{L}_i^\tau k_{\mathbf{x}_i} &= (I - \eta K_{\mathbf{x}_i})^\tau k_{\mathbf{x}_i} \\ \Longrightarrow \Psi(\mathbf{x}) &= \eta K_{\mathbf{x}} P \end{aligned}$$

- Telescoping sum

$$f - \mathcal{L}_i^s f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau f - \mathcal{L}_i^{\tau+1} f = \sum_{\tau=0}^{s-1} \mathcal{L}_i^\tau \left( \frac{\eta}{n_i} \sum_{j=1}^{n_i} f(x_{ij}) k_{x_{ij}} \right)$$

$$\Longrightarrow \; f(\mathbf{x}) - \mathcal{L} f(\mathbf{x}) = \Psi(\mathbf{x}) f(\mathbf{x}) = \eta K_{\mathbf{x}} P f(\mathbf{x})$$

# Convergence analysis

Key: eigenvalues of $I - \eta K_{\mathbf{x}} P$ (asymmetric)

# Convergence analysis

Key: eigenvalues of $I - \eta K_{\mathbf{x}} P$ (asymmetric)

- Analysis similar to graph Laplacians:

$$\text{eigenvalues of } K_{\mathbf{x}} P \Leftrightarrow \text{eigenvalues of } P^{1/2} K_{\mathbf{x}} P^{1/2}$$

# Convergence analysis

Key: eigenvalues of $I - \eta K_{\mathbf{x}} P$ (asymmetric)

- Analysis similar to graph Laplacians:

$$\text{eigenvalues of } K_{\mathbf{x}} P \Leftrightarrow \text{eigenvalues of } P^{1/2} K_{\mathbf{x}} P^{1/2}$$

- Stability:

$$\gamma \triangleq \eta \max_{i \in [M]} \|K_{\mathbf{x}_i}\| < 1 \implies \text{eigenvalues of } I - \eta K_{\mathbf{x}} P \in [0, 1]$$

# Convergence analysis

Key: eigenvalues of $I - \eta K_{\mathbf{x}} P$ (asymmetric)

- Analysis similar to graph Laplacians:

$$\text{eigenvalues of } K_{\mathbf{x}} P \Leftrightarrow \text{eigenvalues of } P^{1/2} K_{\mathbf{x}} P^{1/2}$$

- Stability:

$$\gamma \triangleq \eta \max_{i \in [M]} \|K_{\mathbf{x}_i}\| < 1 \implies \text{eigenvalues of } I - \eta K_{\mathbf{x}} P \in [0, 1]$$

- Condition number of $P$:

$$\|P\| \left\|P^{-1}\right\| \leq \kappa \triangleq \begin{cases} \frac{\gamma s}{1-(1-\gamma)^s} & \text{for FedAvg,} \\ 1+\gamma & \text{for FedProx.} \end{cases}$$

# Explicit convergence results

$$f_t(\mathbf{x}) = [I - \eta K_{\mathbf{x}} P]\, f_{t-1}(\mathbf{x}) + \eta K_{\mathbf{x}} P y,$$

- Convergence in either RKHS norm or the $L^2(\mathbb{P}_N)$ norm

$$\|f_t - f\|_N^2 \triangleq \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(f_t(x_{ij}) - f(x_{ij})\right)^2$$

- Explicit characterization of bias, variance, and heterogeneity
  - Covariate heterogeneity (a.k.a. covariate shift)
  - Response heterogeneity (a.k.a. concept shift)
  - Unbalanced data volume (a.k.a. quantity skew)

# Early stopping and optimal rates

**Theorem (Su-Xu-Yang '23 )**

*For any* $f \in \mathcal{H}$, $1 \leq t \leq T$,

$$\mathbb{E}_{\xi}[\|f_t - f\|_N^2] \leq \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\|\Delta_f\|^2,$$

*where* $T$ *is a carefully chosen stopping time, and*
$\Delta_f = (f_1^*(\mathbf{x}_1), f_2^*(\mathbf{x}_2), \cdots, f_M^*(\mathbf{x}_M)) - f(\mathbf{x})$.

# Early stopping and optimal rates

**Theorem (Su-Xu-Yang '23 )**

*For any* $f \in \mathcal{H}$, $1 \le t \le T$,

$$\mathbb{E}_\xi[\|f_t - f\|_N^2] \le \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\left\|\Delta_f\right\|^2,$$

*where* $T$ *is a carefully chosen stopping time, and* $\Delta_f = (f_1^*(\mathbf{x}_1), f_2^*(\mathbf{x}_2), \cdots, f_M^*(\mathbf{x}_M)) - f(\mathbf{x})$.

- Recover centralized rate (with $f_i^* = f^*$) [Raskutti-Wainwright-Yu'14]

# Early stopping and optimal rates

**Theorem (Su-Xu-Yang '23 )**

*For any $f \in \mathcal{H}$, $1 \leq t \leq T$,*

$$\mathbb{E}_{\xi}[\|f_t - f\|_N^2] \leq \frac{3\kappa}{2e\eta ts}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\|\Delta_f\|^2,$$

*where $T$ is a carefully chosen stopping time, and*
$\Delta_f = (f_1^*(\mathbf{x}_1), f_2^*(\mathbf{x}_2), \cdots, f_M^*(\mathbf{x}_M)) - f(\mathbf{x})$.

- Recover centralized rate (with $f_i^* = f^*$) [Raskutti-Wainwright-Yu'14]
- Example: polynomial decay $\lambda_i \lesssim i^{-2\beta}$ for $\beta > 1/2$

Error rate: $(\sigma^2/N)^{2\beta/(2\beta+1)}$

# Early stopping and optimal rates

**Theorem (Su-Xu-Yang '23 )**

*For any $f \in \mathcal{H}$, $1 \le t \le T$,*

$$\mathbb{E}_\xi[\|f_t - f\|_N^2] \le \frac{3\kappa}{2e\eta t s}\left(\|f_0 - f\|_{\mathcal{H}}^2 + 1\right) + \frac{3\kappa}{N}\left\|\Delta_f\right\|^2,$$

*where $T$ is a carefully chosen stopping time, and*
$\Delta_f = (f_1^*(\mathbf{x}_1), f_2^*(\mathbf{x}_2), \cdots, f_M^*(\mathbf{x}_M)) - f(\mathbf{x})$.

- Recover centralized rate (with $f_i^* = f^*$) [Raskutti-Wainwright-Yu'14]
- Example: polynomial decay $\lambda_i \lesssim i^{-2\beta}$ for $\beta > 1/2$

Error rate: $(\sigma^2/N)^{2\beta/(2\beta+1)}$

- Minimax $L^2(\mathbb{P})$ rate with iid data (empirical process theory)

# Convergence in RKHS norm for finite-rank kernels

**Theorem (Su-Xu-Yang '23)**

*Suppose kernel $k$ is of rank $d$. Then*

$$\mathbb{E}_\xi\left[\left\|f_t - \bar{f}\right\|_{\mathcal{H}}^2\right] \le \left(1 - \frac{s\eta\lambda_d}{\kappa}\right)^{2t} \left\|f_0 - \bar{f}\right\|_{\mathcal{H}}^2 + \sigma^2 \frac{\kappa d}{N\lambda_d},$$

*where* $\bar{f} = (\mathcal{I} - \mathcal{L})^{-1}\left(\left(f_1^*(\mathbf{x}_1), \ldots, f_M^*(\mathbf{x}_M)\right) \cdot \Psi\right).$

- $f_t$ converges exponentially to $\bar{f}$ that balances out heterogeneity

# Convergence in RKHS norm for finite-rank kernels

**Theorem (Su-Xu-Yang '23)**

*Suppose kernel $k$ is of rank $d$. Then*

$$\mathbb{E}_\xi\left[\left\|f_t-\bar{f}\right\|_{\mathcal{H}}^2\right] \leq\left(1-\frac{s \eta \lambda_d}{\kappa}\right)^{2 t}\left\|f_0-\bar{f}\right\|_{\mathcal{H}}^2+\sigma^2 \frac{\kappa d}{N \lambda_d},$$

*where* $\bar{f}=(\mathcal{I}-\mathcal{L})^{-1}\left(\left(f_1^*\left(\mathbf{x}_1\right), \ldots, f_M^*\left(\mathbf{x}_M\right)\right) \cdot \Psi\right)$.

- $f_t$ converges exponentially to $\bar{f}$ that balances out heterogeneity
- When $\lambda_d=\Omega(1)$, the estimation error is $O(d / N)$ and minimax-optimal

# Convergence in RKHS norm for finite-rank kernels

**Theorem (Su-Xu-Yang '23)**

*Suppose kernel $k$ is of rank $d$. Then*

$$\mathbb{E}_\xi\left[\left\|f_t - \bar{f}\right\|_{\mathcal{H}}^2\right] \le \left(1 - \frac{s\eta\lambda_d}{\kappa}\right)^{2t}\left\|f_0 - \bar{f}\right\|_{\mathcal{H}}^2 + \sigma^2\frac{\kappa d}{N\lambda_d},$$

*where* $\bar{f} = (\mathcal{I} - \mathcal{L})^{-1}\left(\left(f_1^*(\mathbf{x}_1), \ldots, f_M^*(\mathbf{x}_M)\right) \cdot \Psi\right)$.

- $f_t$ converges exponentially to $\bar{f}$ that balances out heterogeneity
- When $\lambda_d = \Omega(1)$, the estimation error is $O(d/N)$ and minimax-optimal
- We further show $\bar{f}$ stays within bounded distance to $f_j^*$.

# Federation gain

- $\hat{f}_j$ is an estimator based on the local data

$$R_j^{\mathsf{Loc}} = \inf_{\hat{f}_j} \sup_{f_j^*} \mathbb{E}_{\mathbf{x}_j, \xi_j} \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2$$

# Federation gain

- $\hat{f}_j$ is an estimator based on the local data

$$R_j^{\mathsf{Loc}} = \inf_{\hat{f}_j} \sup_{f_j^*} \mathbb{E}_{\mathbf{x}_j, \xi_j} \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2$$

- $f_t$ is the FL model after $t$ rounds

$$R_j^{\mathsf{Fed}} = \inf_{t \geq 0} \sup_{f_j^* \in \mathcal{H}_B} \mathbb{E}_{\mathbf{x}, \xi} \|f_t - f_j^*\|_{\mathcal{H}}^2,$$

# Federation gain

- $\hat{f}_j$ is an estimator based on the local data

$$R_j^{\mathsf{Loc}} = \inf_{\hat{f}_j} \sup_{f_j^*} \mathbb{E}_{\mathbf{x}_j, \xi_j} \|\hat{f}_j - f_j^*\|_{\mathcal{H}}^2$$

- $f_t$ is the FL model after $t$ rounds

$$R_j^{\mathsf{Fed}} = \inf_{t \geq 0} \sup_{f_j^* \in \mathcal{H}_B} \mathbb{E}_{\mathbf{x}, \xi} \|f_t - f_j^*\|_{\mathcal{H}}^2,$$

- **Federation gain** (quantify the benefit of joining FL)

$$\mathsf{FG}_j \triangleq \frac{R_j^{\mathsf{Loc}}}{R_j^{\mathsf{Fed}}}$$

# Federation gain versus model heterogeneity

- Linear regression $y_j = \mathbf{x}_j \theta_j^* + \xi_j$
- Diameter of model parameters $\Gamma = \max_{i,j \in [M]} \|\theta_i^* - \theta_j^*\|_2$

# Federation gain versus model heterogeneity

- Linear regression $y_j = \mathbf{x}_j \theta_j^* + \xi_j$
- Diameter of model parameters $\Gamma = \max_{i,j \in [M]} \|\theta_i^* - \theta_j^*\|_2$
- Theoretical lower bound

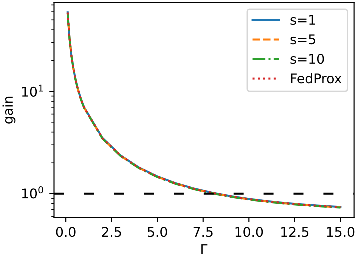
$$\mathsf{FG}_j \gtrsim \frac{\min\{\sigma^2 d/n_j, \|\theta_j^*\|^2\} + \max\{1 - n_j/d, 0\}\|\theta_j^*\|^2}{\sigma^2 d/N + \Gamma^2}$$

# Federation gain versus model heterogeneity

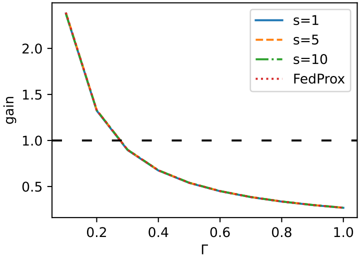
- Linear regression $y_j = \mathbf{x}_j \theta_j^* + \xi_j$
- Diameter of model parameters $\Gamma = \max_{i,j \in [M]} \|\theta_i^* - \theta_j^*\|_2$
- Theoretical lower bound

$$\mathsf{FG}_j \gtrsim \frac{\min\{\sigma^2 d/n_j, \|\theta_j^*\|^2\} + \max\{1 - n_j/d, 0\}\|\theta_j^*\|^2}{\sigma^2 d/N + \Gamma^2}$$

- $d = 100$, $n_i = 50$ (data scarce) or 500 (data rich)

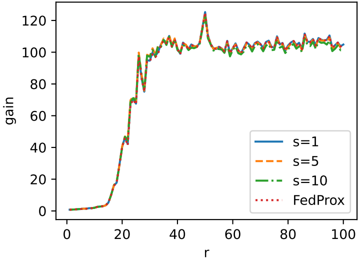



Data-scarce client
$\Gamma \approx \sqrt{1 - n_j/d}\|\theta_j^*\|$

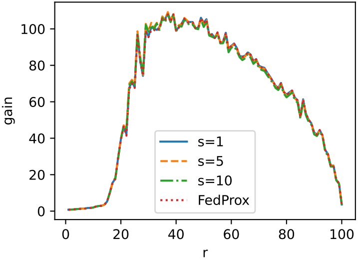



Data-rich client
$\Gamma \approx \sigma\sqrt{d/n_j}$

# Federation gain versus covariate heterogeneity




A data scarce client




A data rich client

# References


- Lili Su, Jiaming Xu, Y., *A Non-parametric View of FedAvg and FedProx: Beyond Stationary Points*, Journal of Machine Learning Research 2023.