Global Convergence of Federated Learning for Mixed Regression

Jiaming Xu

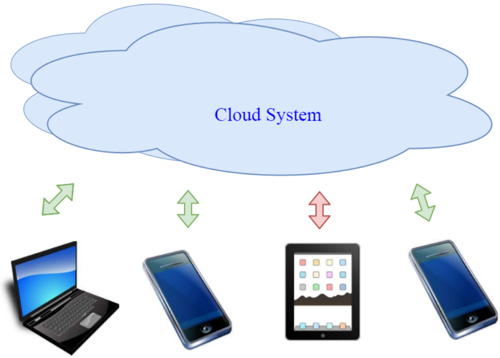
The Fuqua School of Business
Duke University

Joint work with
Lili Su (Northeastern) and Pengkun Yang (Tsinghua)

Allerton Conference, September 28, 2022

Data Heterogeneity in Federated Learning

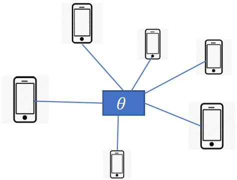
- Unbalanced data partition
- Non-identical data distribution



Leave training data on mobile devices

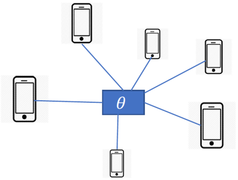
Existing Approaches for Data Heterogeneity

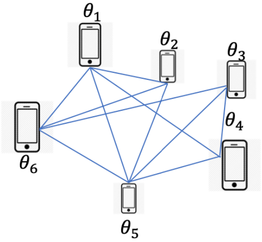
- Common model [Li-Sau-Zaheer-Sanjabi-Talwalkar-Smith '20, Su-X.-Yang '21,...]: only work with moderate heterogeneity



Existing Approaches for Data Heterogeneity

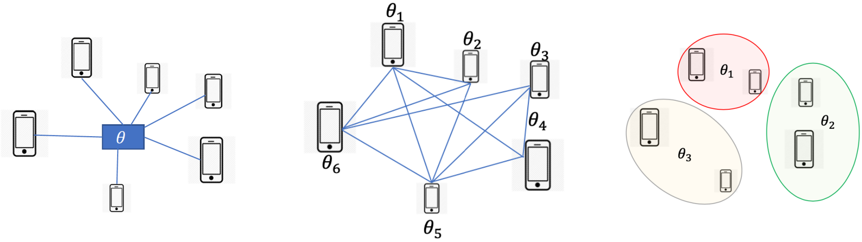
- Common model [Li-Sau-Zaheer-Sanjabi-Talwalkar-Smith '20, Su-X.-Yang '21,...]: only work with moderate heterogeneity
- Fully personalized model [Smith-Chiang-Sanjabi-Talwalkar '17, Marfoq-Neglia-Bellet-Kameni-Vidal '21, ...]: non-convex formulation, no convergence/generalization theory





Existing Approaches for Data Heterogeneity

- Common model [Li-Sau-Zaheer-Sanjabi-Talwalkar-Smith '20, Su-X.-Yang '21,...]: only work with moderate heterogeneity
- Fully personalized model [Smith-Chiang-Sanjabi-Talwalkar '17, Marfoq-Neglia-Bellet-Kameni-Vidal '21, ...]: non-convex formulation, no convergence/generalization theory
- Clustered models [Sattler-Müller-Samek '20, Ghosh-Hong-Yin-Ramchandran '19, Ghosh-Chung-Yin-Ramchandran '20,...]



Clustered Federated Learning

- Most previous works are heuristic and lack of convergence guarantees
- Limited theoretical study under stringent assumptions [Ghosh-Chung-Yin-Ramchandran '20]
  - Good initialization
  - Balanced and high-volume of local data
  - Sample splitting across iterations

Clustered Federated Learning

- Most previous works are heuristic and lack of convergence guarantees
- Limited theoretical study under stringent assumptions [Ghosh-Chung-Yin-Ramchandran '20]
  - Good initialization
  - Balanced and high-volume of local data
  - Sample splitting across iterations

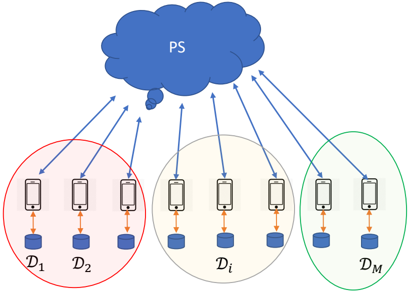
This talk

A new algorithm that achieves global convergence from any initialization despite of unbalanced cluster and data partitions

Outline of the Remainder

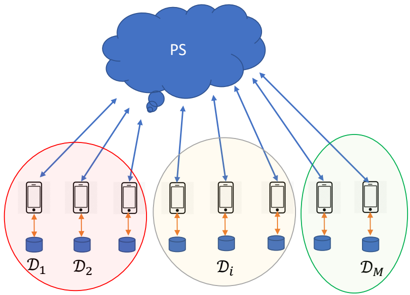
1. Model setup
2. Our two-phase algorithm
3. Theoretical guarantees
4. Summary and concluding remarks

Our Model: Mixed Regression



- One parameter server + M clients partitioned into k hidden clusters

Our Model: Mixed Regression

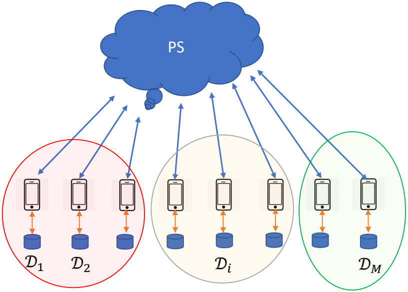


- One parameter server + M clients partitioned into k hidden clusters

- Each client i has n_i local data points $\mathcal{D}_i = \{x_{ij}, y_{ij}\}$:

$$y_{ij} = \langle x_{ij}, \theta^*_{z_i} \rangle + \zeta_{ij}, \ j \in [n_i]$$

Our Model: Mixed Regression



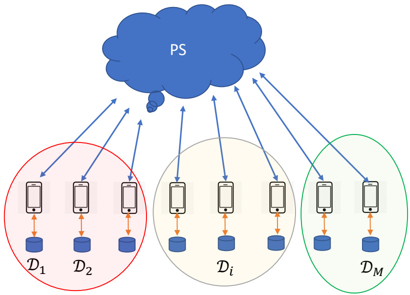
- One parameter server + M clients partitioned into k hidden clusters

- Each client i has n_i local data points $\mathcal{D}_i = \{x_{ij}, y_{ij}\}$:

$$y_{ij} = \langle x_{ij}, \theta^*_{z_i} \rangle + \zeta_{ij}, \; j \in [n_i]$$

- Model parameters $(\theta^*_1, \ldots, \theta^*_k)$
- Cluster label $z_i = \ell$ w.p. p_ℓ

Our Model: Mixed Regression



- One parameter server + M clients partitioned into k hidden clusters

- Each client i has n_i local data points $\mathcal{D}_i = \{x_{ij}, y_{ij}\}$:

$$y_{ij} = \langle x_{ij}, \theta^*_{z_i} \rangle + \zeta_{ij}, \; j \in [n_i]$$

- Model parameters $(\theta^*_1, \ldots, \theta^*_k)$
- Cluster label $z_i = \ell$ w.p. p_ℓ
- Feature vector $x_{ij} \in \mathbb{R}^d$: independent, sub-Gaussian,

$$\mathbb{E}\left[x_{ij} x_{ij}^\top\right] = \Sigma_\ell, \; \text{if } z_i = \ell$$

- noise ζ_{ij}: independent, sub-Gaussian

Our Two-phase FL algorithm

1. Obtain coarse estimates of model parameters $(\theta_1^*, \ldots, \theta_k^*)$ via Federated moment descent

2. Iteratively estimate cluster label and refine local model estimate via either FedAvg or FedProx

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]
- Goal: generate a sequence of estimators $\{\theta_{i,t}\}$ approaching $\theta^*_{z_i}$:

$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]
- Goal: generate a sequence of estimators $\{\theta_{i,t}\}$ approaching $\theta^*_{z_i}$:

$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

- Decrease

$$\mathrm{Var}\Bigg(\underbrace{\langle x_{ij}, \theta^*_{z_i} \rangle - \theta_{i,t}}_{\text{residual error}} \Bigg) = \|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|_2^2$$

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]
- Goal: generate a sequence of estimators $\{\theta_{i,t}\}$ approaching $\theta^*_{z_i}$:

$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

- Decrease

$$\mathrm{Var}\Bigg(\underbrace{\langle x_{ij}, \theta^*_{z_i} \rangle - \theta_{i,t}}_{\text{residual error}} \Bigg) = \|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|_2^2$$

- Choose $r_{i,t}$ positively correlated with $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]
- Goal: generate a sequence of estimators $\{\theta_{i,t}\}$ approaching $\theta^*_{z_i}$:

$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

- Decrease

$$\mathrm{Var}\Big(\underbrace{\langle x_{ij}, \theta^*_{z_i} \rangle - \theta_{i,t}}_{\text{residual error}} \Big) = \|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|_2^2$$

- Choose $r_{i,t}$ positively correlated with $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$
- $(y_{ij} - \langle x_{ij}, \theta_{i,t} \rangle)\, x_{ij}$ is an unbiased estimator of $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$

Moment Descent: General Idea

- Powerful idea for clustering under mixture model [Moitra-Valiant '10, Li-Liang '18]
- Goal: generate a sequence of estimators $\{\theta_{i,t}\}$ approaching $\theta^*_{z_i}$:

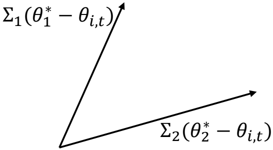
$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

- Decrease

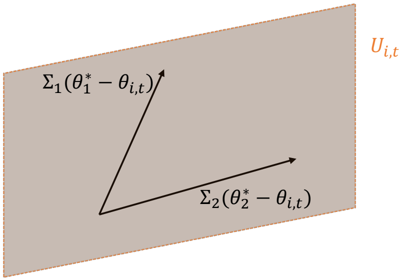
$$\mathrm{Var}\Big(\underbrace{\langle x_{ij}, \theta^*_{z_i}\rangle - \theta_{i,t}}_{\text{residual error}}\Big) = \|\Sigma_{z_i}^{1/2}(\theta^*_{z_i} - \theta_{i,t})\|_2^2$$

- Choose $r_{i,t}$ positively correlated with $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$
- $(y_{ij} - \langle x_{ij}, \theta_{i,t}\rangle)\, x_{ij}$ is an unbiased estimator of $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t})$
- However, need $\Omega(d)$ local data points at client i to well estimate $\Sigma_{z_i}(\theta^*_{z_i} - \theta_{i,t}) \Longrightarrow$ Unaffordable in FL with limited local data

Our idea: Federated Moment Descent



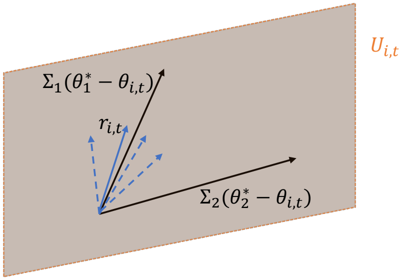
Our idea: Federated Moment Descent



1. Pool data from clients to estimate

$$U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t}) : \ell \in [k]\}$$

Our idea: Federated Moment Descent

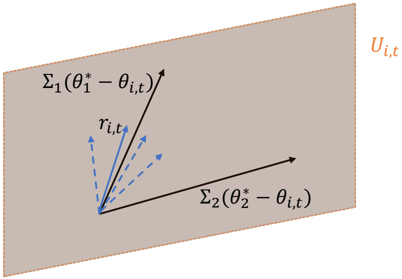


1. Pool data from clients to estimate

$$U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t}) : \ell \in [k]\}$$

2. Project local data onto $U_{i,t}$ to estimate $r_{i,t}$

Our idea: Federated Moment Descent



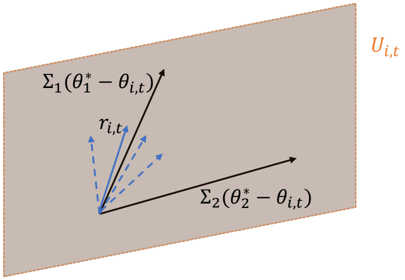
1. Pool data from clients to estimate

   $$U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t}) : \ell \in [k]\}$$

2. Project local data onto $U_{i,t}$ to estimate $r_{i,t}$

- Reduce estimation from d-dim to k-dim $\Rightarrow$ only need $\tilde{\Omega}(k)$ local data points (*anchor clients*)

Our idea: Federated Moment Descent



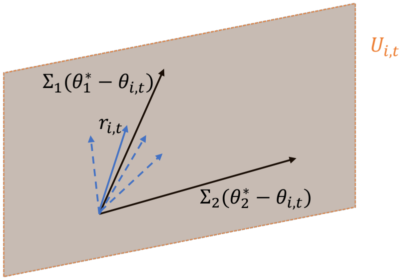
1. Pool data from clients to estimate

$$U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t}) : \ell \in [k]\}$$

2. Project local data onto $U_{i,t}$ to estimate $r_{i,t}$

- Reduce estimation from d-dim to k-dim $\Rightarrow$ only need $\tilde{\Omega}(k)$ local data points (*anchor clients*)
- To cover all k clusters, only need $\tilde{\Omega}(k)$ such anchor clients

Our idea: Federated Moment Descent



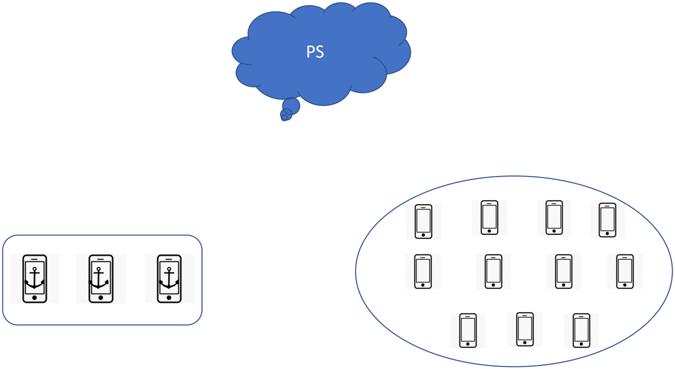
1. Pool data from clients to estimate

   $$U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta_\ell^* - \theta_{i,t}) : \ell \in [k]\}$$

2. Project local data onto $U_{i,t}$ to estimate $r_{i,t}$

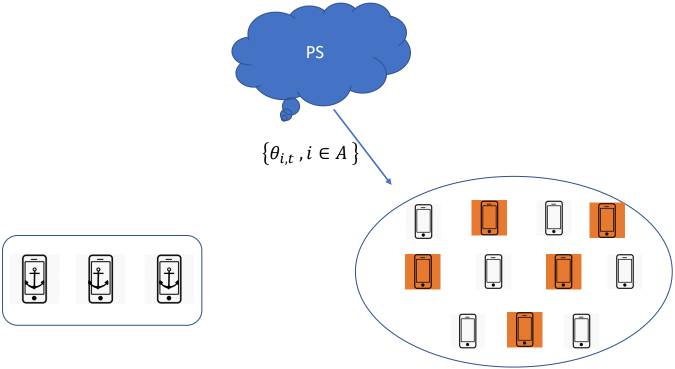
- Reduce estimation from d-dim to k-dim $\Rightarrow$ only need $\tilde{\Omega}(k)$ local data points (*anchor clients*)
- To cover all k clusters, only need $\tilde{\Omega}(k)$ such anchor clients
- Similar idea was used for meta-learning [Kong-Somani-Song-Kakade-Oh '20], but without using moment descent

Federated Moment Descent in Action



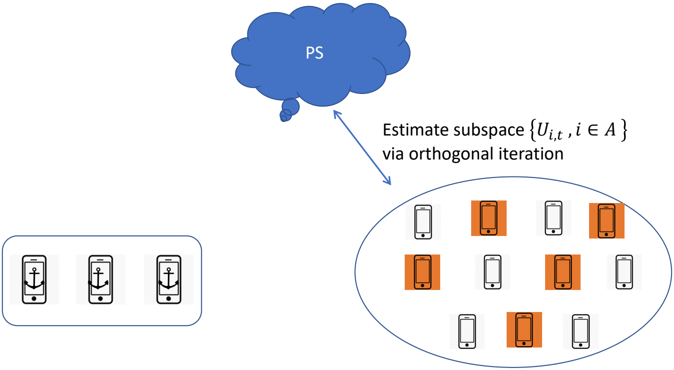
Step 1: Choose a set A of anchor clients at random

Federated Moment Descent in Action



Step 2: Broadcast $\{\theta_{i,t}, i \in A\}$ to a subset $\mathcal{S}_t$ of non-anchor clients

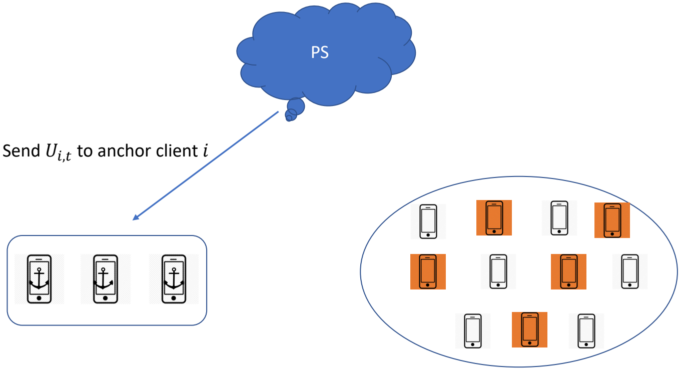
Federated Moment Descent in Action



Step 3: Estimate $U_{i,t} \approx \text{span}\{\Sigma_\ell(\theta^*_\ell - \theta_{i,t}) : \ell \in [k]\}$ based on the top-k singular vectors of

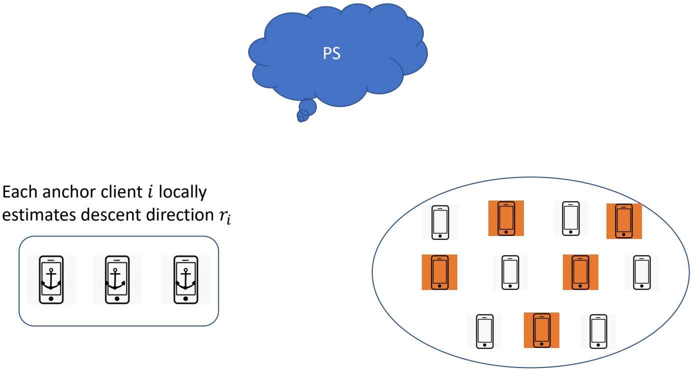
$$Y_{i,t} = \frac{1}{|\mathcal{S}_t|} \sum_{i' \in \mathcal{S}_t} \varepsilon(x_{i'1}, y_{i'1}, \theta_{i,t}) \varepsilon(x_{i'2}, y_{i'2}, \theta_{i,t})^\top, \ \varepsilon(x, y, \theta) \triangleq (y - \langle x, \theta \rangle)\, x$$

Federated Moment Descent in Action



Step 4: Send the estimated k-dim subspace $U_{i,t}$ to each anchor client i

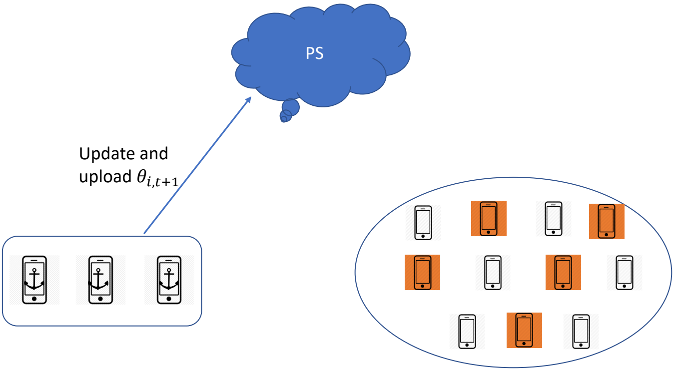
Federated Moment Descent in Action



Step 5: Let $r_{i,t} = U_{i,t}\beta_{i,t}$, where $\beta_{i,t}$ is the leading singular vector of

$$Z_{i,t} = \frac{1}{|\mathcal{D}_{i,t}|} \sum_{j \in \mathcal{D}_{i,t}} \left(U_{i,t}^\top \varepsilon(x_{ij}, y_{ij}, \theta_{i,t}) \right) \left(U_{i,t}^\top \varepsilon(x'_{ij}, y'_{ij}, \theta_{i,t}) \right)^\top$$

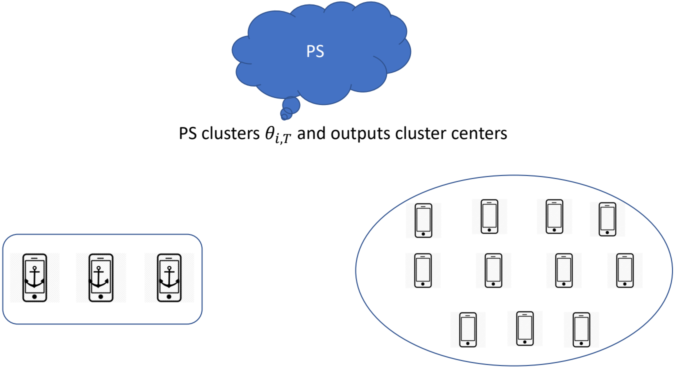
Federated Moment Descent in Action



Step 6: Update and upload

$$\theta_{i,t+1} = \theta_{i,t} + \eta_{i,t} r_{i,t}$$

Federated Moment Descent in Action



Step 7: Cluster $\{\theta_{i,T}, i \in A\}$ by thresholding on $\|\theta_{i,T} - \theta_{i',T}\|_2$ and output k cluster centers as $\hat{\theta}$

Theoretical Guarantee for Phase 1

$p_{\min} = \min_{\ell \in [k]} p_\ell, \quad M = \#$ of clients

$M_A = \#$ of anchor clients, $\quad n_A = \#$ of data points per anchor client

Theorem (Su-X.-Yang '22)

Let ϵ be a small but fixed constant. Suppose that

$$M \geq p_{\min}^{-2} \tilde{\Omega}(d), \quad M_A \geq \tilde{\Omega}(k), \quad n_A = \tilde{\Omega}(k).$$

With high probability, starting from any initialization θ_0, Phase 1 outputs $\hat{\theta}$ in $O(1)$ iterations:

$$d\left(\hat{\theta}, \theta^*\right) \triangleq \min_{\pi} \max_{\ell \in [k]} \|\hat{\theta}_{\pi(\ell)} - \theta_\ell^*\|_2 \leq \epsilon.$$

Phase 2: FedX + Clustering

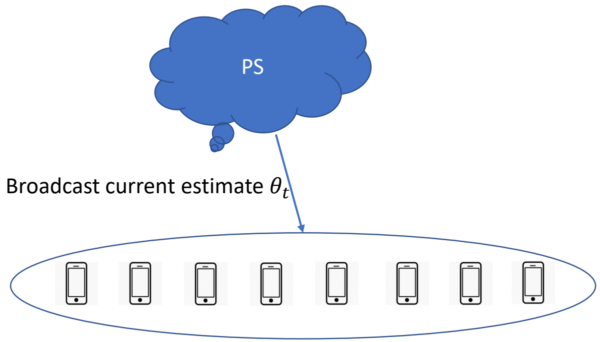
- All clients iteratively estimate their cluster label and refine their local model estimate via either FedAvg or FedProx

Phase 2: FedX + Clustering

- All clients iteratively estimate their cluster label and refine their local model estimate via either FedAvg or FedProx

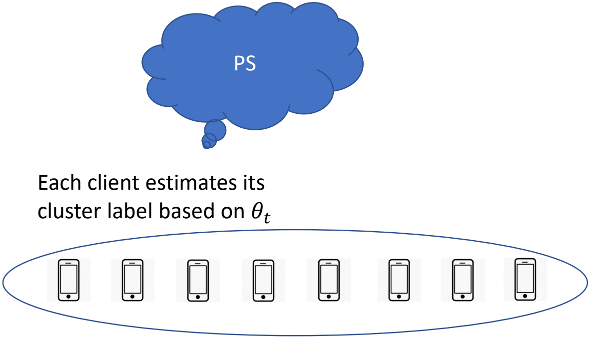
- At every iteration, clients reuse all local data, including those used in the first phase
  - Cruicial especially for data-scarce clients
  - Lead to sophisticated interdependency - significant analysis challenge

FedX + Clustering in Action



Step 1: Broadcast current estimate $\theta_t = (\theta_{1,t}, \ldots, \theta_{k,t})$ to all clients

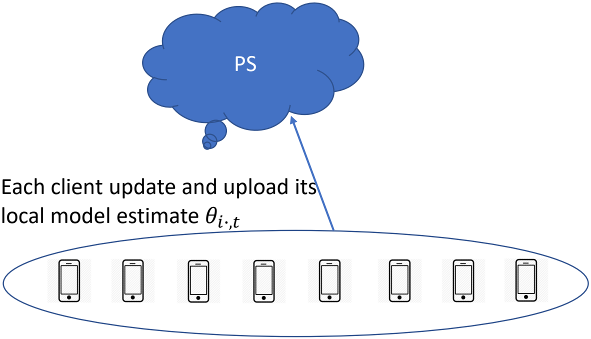
FedX + Clustering in Action



Step 2: Each client i estimates its cluster label via ML decoder:

$$z_{i,t} \in \arg\min_{\ell \in [k]} L_i(\theta_t, \ell) \triangleq \sum_{j=1}^{n_i} \left(y_{ij} - \langle x_{ij}, \theta_{\ell,t} \rangle\right)^2$$

FedX + Clustering in Action

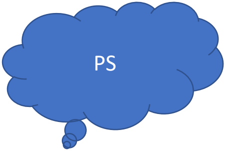


Step 3: Refine local model estimate $\theta_{i\cdot,t}$:

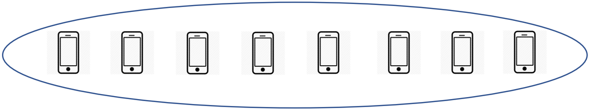
FedAvg(s): $\theta_{i\cdot,t} = \mathcal{G}_i^s(\theta_t)$, where $\mathcal{G}_i(\theta) = \theta - \eta \nabla L_i(\theta, z_{i,t})$

FedProx: $\theta_{i\cdot,t} \in \arg\min_{\theta} L_i(\theta, z_{i,t}) + \frac{1}{2\eta}\|\theta - \theta_t\|_2^2$

FedX + Clustering in Action



Update global model estimate θ_{t+1} based on $\{\theta_{i\cdot,t}\}$



Step 4: Update global model estimate

$$\theta_{t+1} = \sum_{i=1}^{M} w_i \theta_{i\cdot,t}, \quad \text{where } w_i = \frac{n_i}{\sum_i n_i}$$

Theoretical guarantee for Phase 2

Theorem (Su-X.-Yang '22)

Suppose $d(\theta_T, \theta^*) \le \epsilon$. *Then with high probability,*

$$d(\theta_{t+1}, \theta^*) \le (1 - C_1 p_{\min})\, d(\theta_t, \theta^*) + C_2 \nu \log \frac{1}{\nu}, \quad \forall t \ge T,$$

where

$$\nu \triangleq \underbrace{\sum_{i=1}^{M} k w_i e^{-C_3 n_i}}_{\text{avg clustering error}} + \underbrace{C_4 \sqrt{\frac{dk \log k}{M} (\chi^2(w) + 1)}}_{\text{uniform deviation}}$$

Theoretical guarantee for Phase 2

Theorem (Su-X.-Yang '22)

Suppose $d(\theta_T, \theta^*) \le \epsilon$. *Then with high probability,*

$$d(\theta_{t+1}, \theta^*) \le (1 - C_1 p_{\min})\, d(\theta_t, \theta^*) + C_2 \nu \log \frac{1}{\nu}, \quad \forall t \ge T,$$

where

$$\nu \triangleq \underbrace{\sum_{i=1}^{M} k w_i e^{-C_3 n_i}}_{\textit{avg clustering error}} + \underbrace{C_4 \sqrt{\frac{dk \log k}{M} (\chi^2(w) + 1)}}_{\textit{uniform deviation}}$$

- $\chi^2(w)$: chi-square divergence between w and uniform distribution
- If $\chi^2(w) = O(1)$, then the uniform deviation is $\tilde{O}(dk/M)$

Uniform Bounds on Clustering Errors

- Need to bound the total number of misclustered data points:

$$\sum_i n_i \mathbf{1}_{\{i \text{ is misclustered under } \theta_t\}} = \sum_i n_i \mathbf{1}_{\left\{f_{\theta_t}(x_i, y_i) \geq 0\right\}},$$

where $f_\theta(x, y)$ is quadratic in θ

Uniform Bounds on Clustering Errors

- Need to bound the total number of misclustered data points:

$$\sum_i n_i \mathbf{1}_{\{i \text{ is misclustered under } \theta_t\}} = \sum_i n_i \mathbf{1}_{\{f_{\theta_t}(x_i, y_i) \geq 0\}},$$

  where $f_\theta(x, y)$ is quadratic in θ

- Key challenge: θ_t and local data $\{x_i, y_i\}$ are heavily dependent!

Uniform Bounds on Clustering Errors

- Need to bound the total number of misclustered data points:

$$\sum_i n_i \mathbf{1}_{\{i \text{ is misclustered under } \theta_t\ \}} = \sum_i n_i \mathbf{1}_{\left\{f_{\theta_t}(x_i, y_i) \geq 0\right\}},$$

  where $f_\theta(x, y)$ is quadratic in θ

- Key challenge: θ_t and local data $\{x_i, y_i\}$ are heavily dependent!

- Establish uniform bound on

$$\sup_\theta \sum_i n_i \mathbf{1}_{\{f_\theta(x_i, y_i) \geq 0\}}$$

Uniform Bounds on Clustering Errors

- Need to bound the total number of misclustered data points:

$$\sum_i n_i \mathbf{1}_{\{i \text{ is misclustered under } \theta_t\}} = \sum_i n_i \mathbf{1}_{\left\{f_{\theta_t}(x_i, y_i) \geq 0\right\}},$$

  where $f_\theta(x, y)$ is quadratic in θ

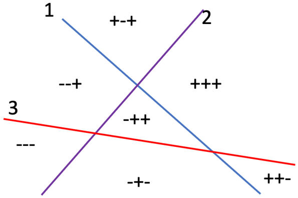
- Key challenge: θ_t and local data $\{x_i, y_i\}$ are heavily dependent!
- Establish uniform bound on

$$\sup_\theta \sum_i n_i \mathbf{1}_{\{f_\theta(x_i, y_i) \geq 0\}}$$

- Need to control the VC dimension of polynomial concept class

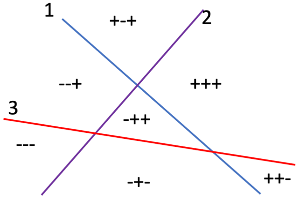
$$\left\{ \mathbf{1}_{\{f_\theta(x, y) \geq 0\}} : \theta \in \mathbb{R}^{dk} \right\}$$

Sign Patterns of Polynomials

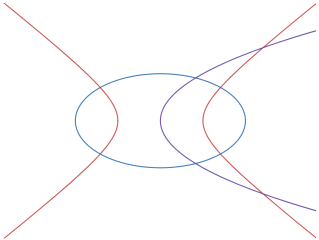


Linear: $\binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$

Sign Patterns of Polynomials

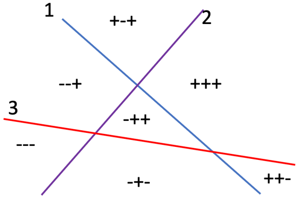


Linear: $\binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$

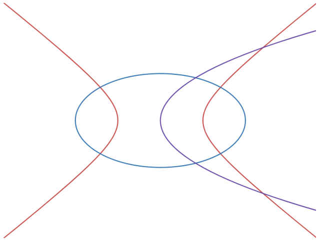


Quadratic: ?

Sign Patterns of Polynomials



Linear: $\binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$

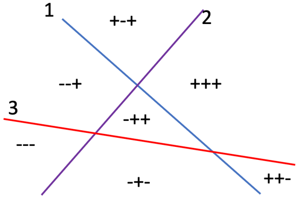


Quadratic: ?

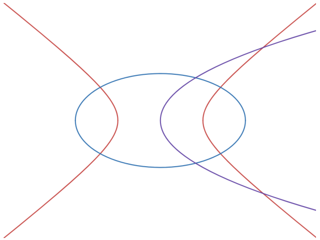
Theorem (Milnor-Thom theorem)

The number of sign patterns of m d-variate polynomials of degree D is at most $\left(\frac{50Dm}{d}\right)^d$.

Sign Patterns of Polynomials



Linear: $\binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$



Quadratic: ?

Theorem (Milnor-Thom theorem)

The number of sign patterns of m d-variate polynomials of degree D is at most $\left(\frac{50Dm}{d}\right)^d$.

Implication: Since $f_\theta(x, y)$ is quadratic in θ,

$$\mathrm{VC}\left\{\mathbf{1}_{\{f_\theta(x,y)\geq 0\}} : \theta \in \mathbb{R}^{dk}\right\} = O(dk)$$

Putting Two Phases Together: Global Convergence

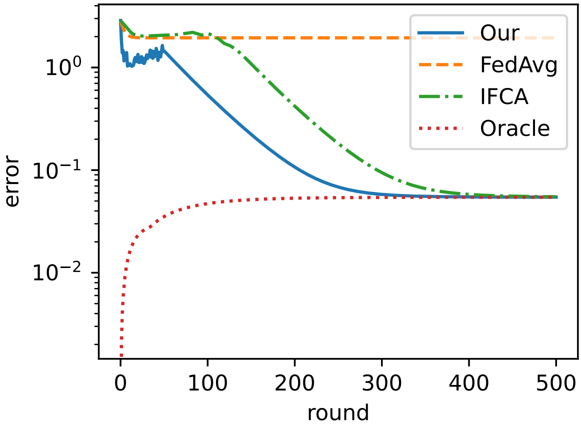
Theorem (Su-X.-Yang '22)

With high probability, starting from any initialization θ_0, running Phase 1 with $\Theta(1)$ iterations and followed by Phase 2 with $\Theta(\rho^{-1}\log(1/\nu))$ iterations outputs $\hat{\theta}$:

$$d(\hat{\theta}, \theta^*) \lesssim \frac{1}{p_{\min}} \nu \log \frac{1}{\nu}$$

- $p_{\min}$ captures the effect of unbalanced clusters
- ν captures the clustering error, which depends on the imbalance of data partition

Numerical Comparisons



- IFCA [Ghosh-Chung-Yin-Ramchandran '20]: stuck on error floor
- Oracle: IFCA initialized with true model parameters

Concluding Remarks

- Clustered federated learning under the mixed regression model
- Design a two-phase FL algorithm: Federated moment descent
- Prove the global convergence from any initialization even with unbalanced cluster and data partitions
- Uniform bound on clustering errors based on VC dimension of polynomial concept classes

Concluding Remarks

- Clustered federated learning under the mixed regression model
- Design a two-phase FL algorithm: Federated moment descent
- Prove the global convergence from any initialization even with unbalanced cluster and data partitions
- Uniform bound on clustering errors based on VC dimension of polynomial concept classes

Future work:

- General risk minimization setup beyond mixed regression
- System heterogeneity
- Security/privacy consideration

Concluding Remarks

- Clustered federated learning under the mixed regression model
- Design a two-phase FL algorithm: Federated moment descent
- Prove the global convergence from any initialization even with unbalanced cluster and data partitions
- Uniform bound on clustering errors based on VC dimension of polynomial concept classes

Future work:

- General risk minimization setup beyond mixed regression
- System heterogeneity
- Security/privacy consideration

Reference

- L. Su, J. Xu, & P. Yang, *Global Convergence of Federated Learning for Mixed Regression*. arXiv:2206.07279. To appear in *Proceedings of NeurIPS 2022*