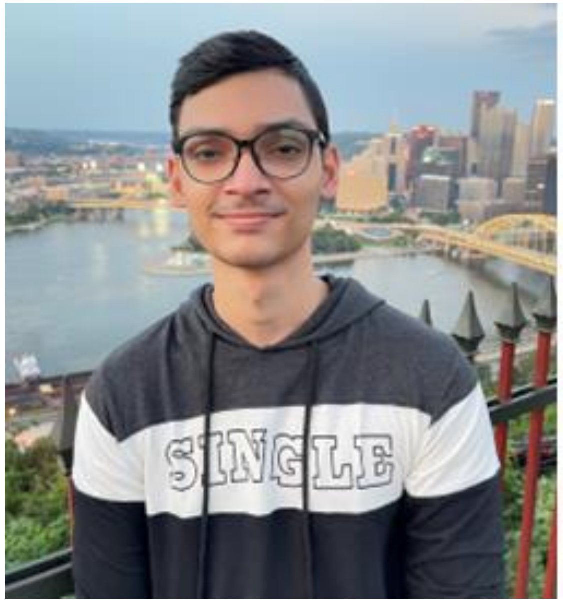
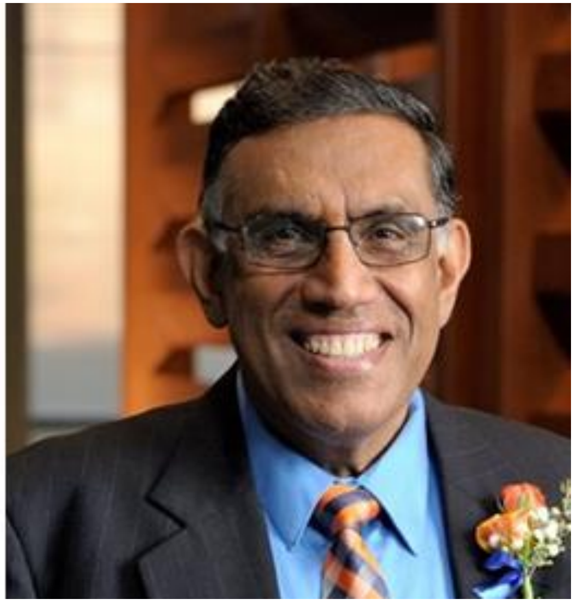
Federated Minimax Optimization

Improved Convergence Analyses and Algorithms

Pranay Sharma
Postdoc, ECE, Carnegie Mellon University





Rohan Panda[1] Gauri Joshi[1] Pramod K. Varshney[2]

[1]ECE, Carnegie Mellon University
[2]EECS, Syracuse University, NY

NSF grants CCF-2045694 and CNS-2112471.

Federated Minimax: Background

Federated Learning

Federated Minimax: Background

Federated Learning

Distributed learning with

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server
- Heterogeneous data across clients

Federated Minimax: Background

Federated Learning

Distributed learning with

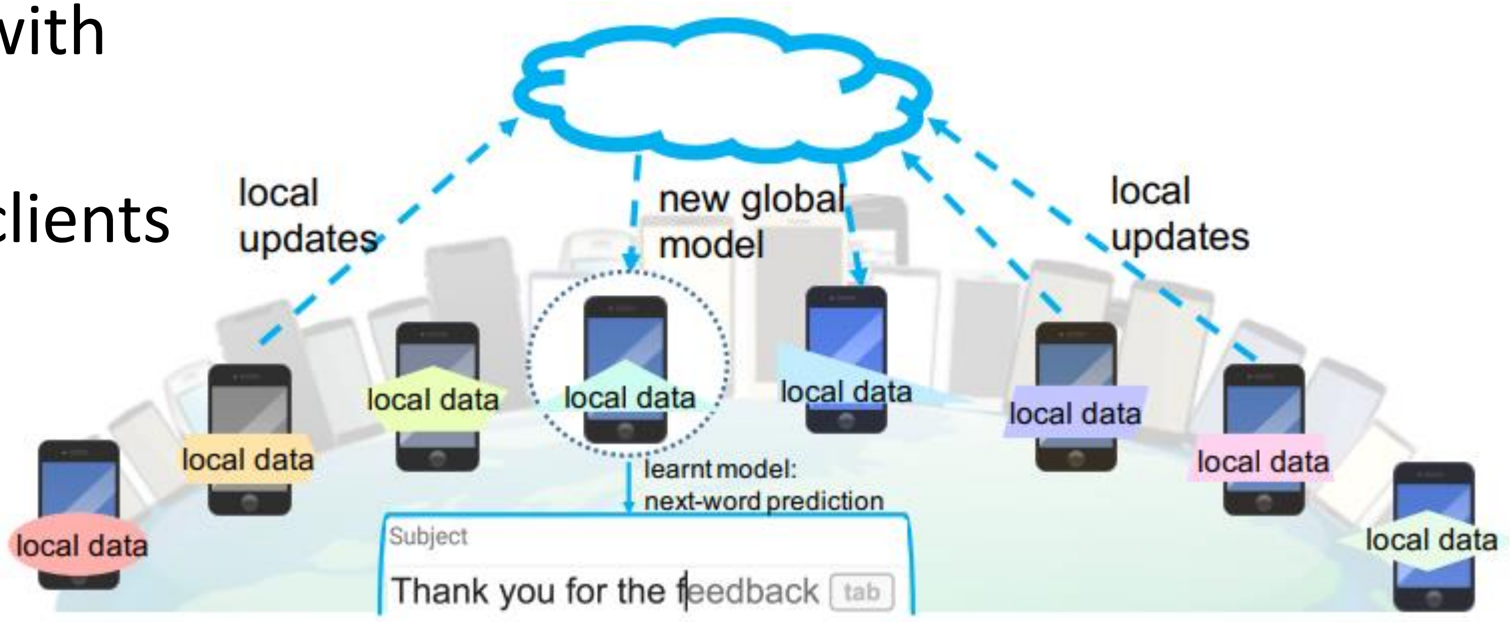
- **Infrequent** communication with the server
- Heterogeneous data across clients
- Client data privacy

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server
- Heterogeneous data across clients
- Client data privacy



Li et al., IEEE SPMag 2021.

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server
- Heterogeneous data across clients
- Ensures client data privacy

Minimax Optimization

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server
- Heterogeneous data across clients
- Ensures client data privacy

Minimax Optimization

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

- GANs, adversarial training of neural networks, reinforcement learning

Federated Minimax: Background

Federated Learning

Distributed learning with

- **Infrequent** communication with the server
- Heterogeneous data across clients
- Ensures client data privacy

Minimax Optimization

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

- GANs, adversarial training of neural networks, reinforcement learning
- f is often nonconvex in $\mathbf{x}$, nonconcave in $\mathbf{y}$

Problem and Algorithm

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

Problem and Algorithm

n - No. of clients

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

Problem and Algorithm

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

n - No. of clients

$f_i(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[L(\mathbf{x}, \mathbf{y}; \xi_i)]$

- $\mathcal{D}_i$ - Local data distribution

Problem and Algorithm

n - No. of clients

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f_i(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [L(\mathbf{x}, \mathbf{y}; \xi_i)]$

- $\mathcal{D}_i$ - Local data distribution

Local SGDA: at client i

Problem and Algorithm

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \right\}$$

n - No. of clients

$f_i(\mathbf{x},\mathbf{y}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[L(\mathbf{x},\mathbf{y};\xi_i)]$

- $\mathcal{D}_i$ - Local data distribution

Local SGDA: at client i

- $\mathbf{x}^i \leftarrow \mathbf{x}^i - \eta_x \widetilde{\nabla}_{\mathbf{x}} f_i(\mathbf{x}^i, \mathbf{y}^i)$
- $\mathbf{y}^i \leftarrow \mathbf{y}^i + \eta_y \widetilde{\nabla}_{\mathbf{y}} f_i(\mathbf{x}^i, \mathbf{y}^i)$

Problem and Algorithm

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \right\}$$

n - No. of clients

$f_i(\mathbf{x},\mathbf{y}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[L(\mathbf{x},\mathbf{y};\xi_i)]$

- $\mathcal{D}_i$ - Local data distribution

Local SGDA: at client i

- $\mathbf{x}^i \leftarrow \mathbf{x}^i - \eta_x \widetilde{\nabla}_{\mathbf{x}} f_i(\mathbf{x}^i, \mathbf{y}^i)$
- $\mathbf{y}^i \leftarrow \mathbf{y}^i + \eta_y \widetilde{\nabla}_{\mathbf{y}} f_i(\mathbf{x}^i, \mathbf{y}^i)$

Infrequent averaging

Problem and Algorithm

n - No. of clients

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \right\}$$

$f_i(\mathbf{x},\mathbf{y}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[L(\mathbf{x},\mathbf{y};\xi_i)]$

- $\mathcal{D}_i$ - Local data distribution

Local SGDA: at client i

- $\mathbf{x}^i \leftarrow \mathbf{x}^i - \eta_x \widetilde{\nabla}_{\mathbf{x}} f_i(\mathbf{x}^i, \mathbf{y}^i)$
- $\mathbf{y}^i \leftarrow \mathbf{y}^i + \eta_y \widetilde{\nabla}_{\mathbf{y}} f_i(\mathbf{x}^i, \mathbf{y}^i)$

Infrequent averaging

- Every τ iterations
- Restart with the average

Theoretical Results I

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

Theoretical Results I

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

- ϵ-approximate stationary point

Theoretical Results I

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \right\}$$

$f(\mathbf{x},\mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

- ϵ-approximate stationary point
- Sample Complexity:

$$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right) \text{ per node}$$

Theoretical Results I

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

- ϵ-approximate stationary point
- Sample Complexity:

$$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right) \text{ per node}$$

- Linear Speedup in n
- Optimal in ϵ

Theoretical Results I

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

- ϵ-approximate stationary point
- Sample Complexity:

$$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right) \text{ per node}$$

- Linear Speedup in n
- Optimal in ϵ

- Communication rounds: $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$

Comparison with Existing Work

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

	Sample Complexity	Communication Rounds
Lin et al., 2020 ($n = 1$)	$\mathcal{O}(\epsilon^{-4})$	-

- Needs $\mathcal{O}(\epsilon^{-2})$ batch-size

Comparison with Existing Work

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

	Sample Complexity	**Communication Rounds**
Lin et al., 2020 ($n = 1$)	$\mathcal{O}(\epsilon^{-4})$	-
Deng et al., 2021 ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^6}\right)$	$\mathcal{O}\left(\frac{1}{n^{1/4}\epsilon^4}\right)$

- Needs $\mathcal{O}(\epsilon^{-2})$ batch-size
- Suboptimal complexity in ϵ

Comparison with Existing Work

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

	Sample Complexity	**Communication Rounds**
Lin et al., 2020 ($n = 1$)	$\mathcal{O}(\epsilon^{-4})$	-
Deng et al., 2021 ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^6}\right)$	$\mathcal{O}\left(\frac{1}{n^{1/4}\epsilon^4}\right)$
Our Work ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$

- Needs $\mathcal{O}(\epsilon^{-2})$ batch-size
- Suboptimal complexity in ϵ
- Optimal in ϵ

Comparison with Existing Work

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

	Sample Complexity	**Communication Rounds**
Lin et al., 2020 ($n = 1$)	$\mathcal{O}(\epsilon^{-4})$	-
Deng et al., 2021 ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^6}\right)$	$\mathcal{O}\left(\frac{1}{n^{1/4}\epsilon^4}\right)$
Our Work ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$

- Needs $\mathcal{O}(\epsilon^{-2})$ batch-size

- Suboptimal complexity in ϵ

- Optimal in ϵ
- $\mathcal{O}(1)$ batch-size

Comparison with Existing Work

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$, **strongly concave** in $\mathbf{y}$

	Sample Complexity	**Communication Rounds**
Lin et al., 2020 ($n = 1$)	$\mathcal{O}(\epsilon^{-4})$	-
Deng et al., 2021 ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^6}\right)$	$\mathcal{O}\left(\frac{1}{n^{1/4}\epsilon^4}\right)$
Our Work ($n \geq 1$)	$\mathcal{O}\left(\frac{1}{n\epsilon^4}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$

- Needs $\mathcal{O}(\epsilon^{-2})$ batch-size

- Suboptimal complexity in ϵ

- Optimal in ϵ
- $\mathcal{O}(1)$ batch-size
- Linear speedup in n

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$

1. strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$

1. strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

- Local update algorithms

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is nonconvex in $\mathbf{x}$

1. strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

- Local update algorithms
- $\mathcal{O}(1)$ batch-size

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is smooth and

1. Strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

- Local update algorithms
- $\mathcal{O}(1)$ batch-size
- Optimal/SOTA complexity

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \right\}$$

$f(\mathbf{x}, \mathbf{y})$ is smooth and

1. Strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

- Local update algorithms
- $\mathcal{O}(1)$ batch-size
- Optimal/SOTA complexity
- Linear speedup in n

Summary

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}) \right\}$$

$f(\mathbf{x},\mathbf{y})$ is smooth and

1. Strongly concave in $\mathbf{y}$
2. PL in $\mathbf{y}$
3. Concave in $\mathbf{y}$
4. 1-Point-Concave in $\mathbf{y}$

- Local update algorithms
- $\mathcal{O}(1)$ batch-size
- Optimal/SOTA complexity
- Linear speedup in n

Poster - Hall E #605