Heterogeneity-aware Algorithms for Federated Optimization

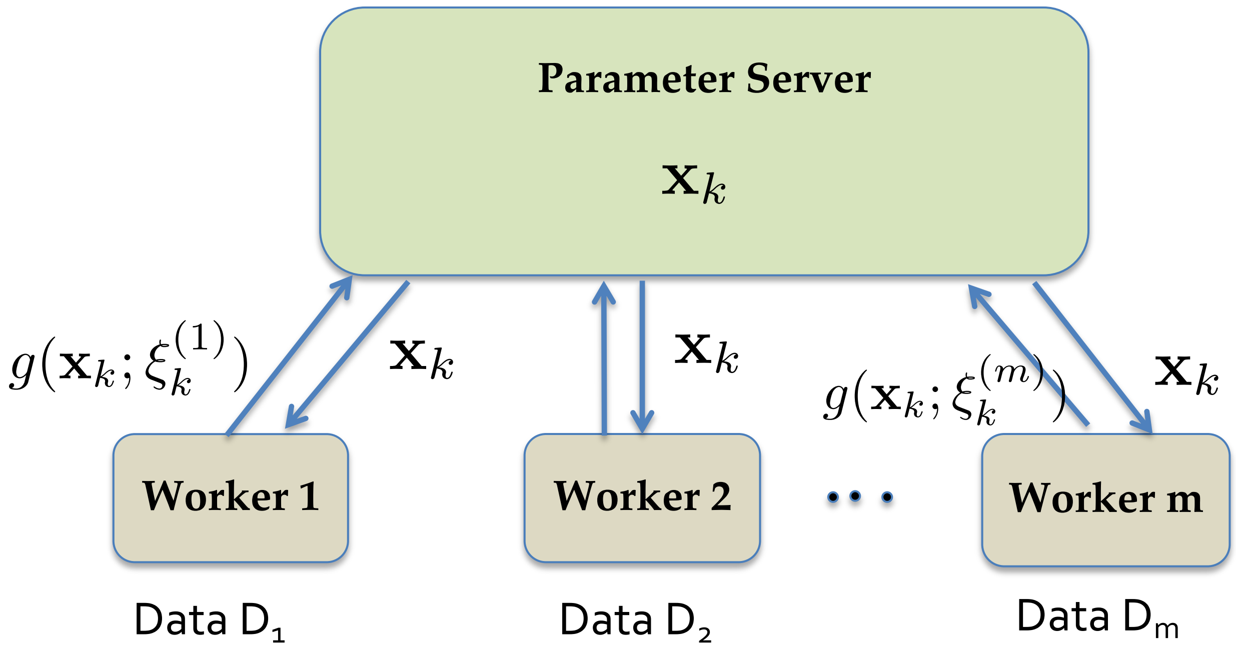
Gauri Joshi

Carnegie Mellon University

Joint work with

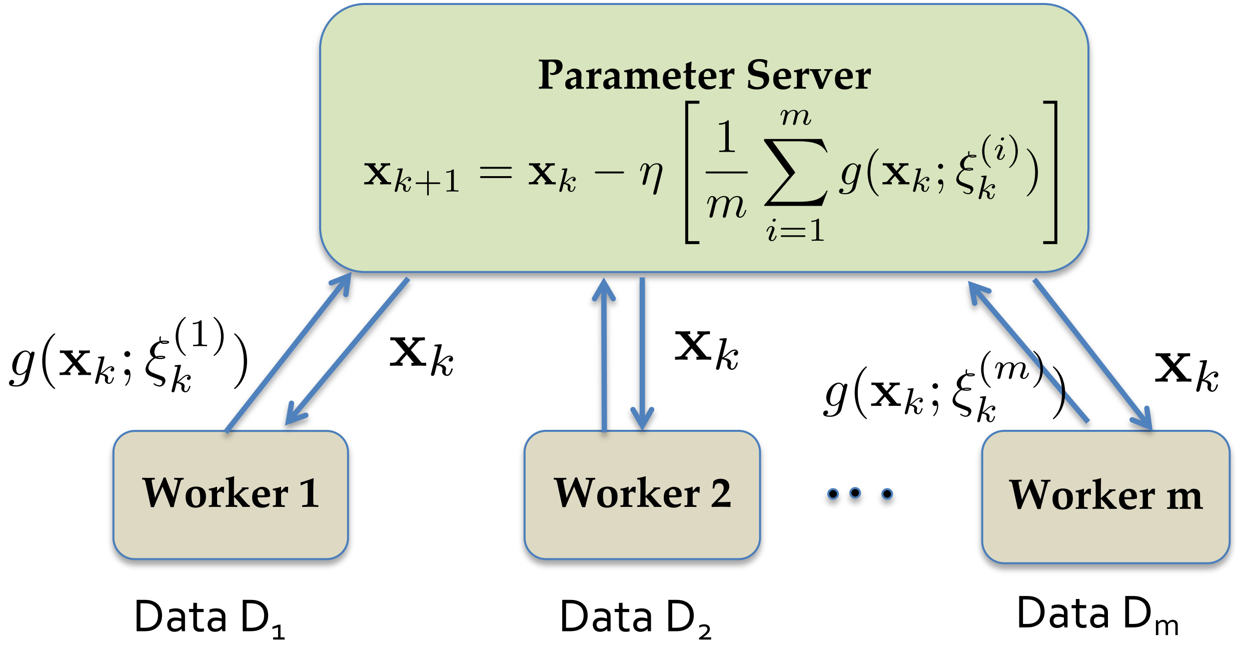
Yae Jee Cho, Divyansh Jhunjhunwala, Pranay Sharma,
Jianyu Wang, Aushim Nagarkatti, Shiqiang Wang (IBM),
Zheng Xu, Satyen Kale, Tong Zhang (Google) and others

Distributed SGD in the Data-Center Setting



- Dataset is shuffled and split equally across the worker nodes
- Parameter server waits to receive gradients from all nodes

Distributed SGD in the Data-Center Setting

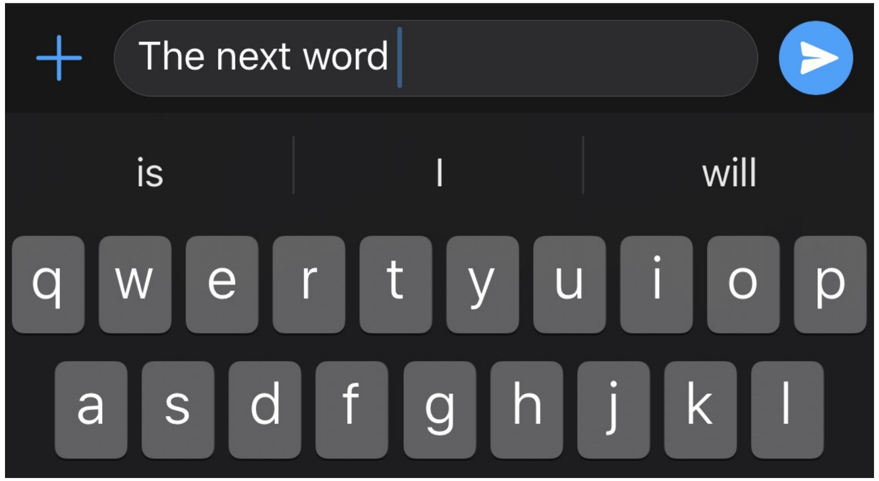


- Dataset is shuffled and split equally across the worker nodes
- Parameter server waits to receive gradients from all nodes
- Several works on improving the scalability of this framework

Async SGD, Local SGD, Overlap SGD etc.

Data Collection at the Edge

[McMahan et al 2017, Kairouz et al 2019]

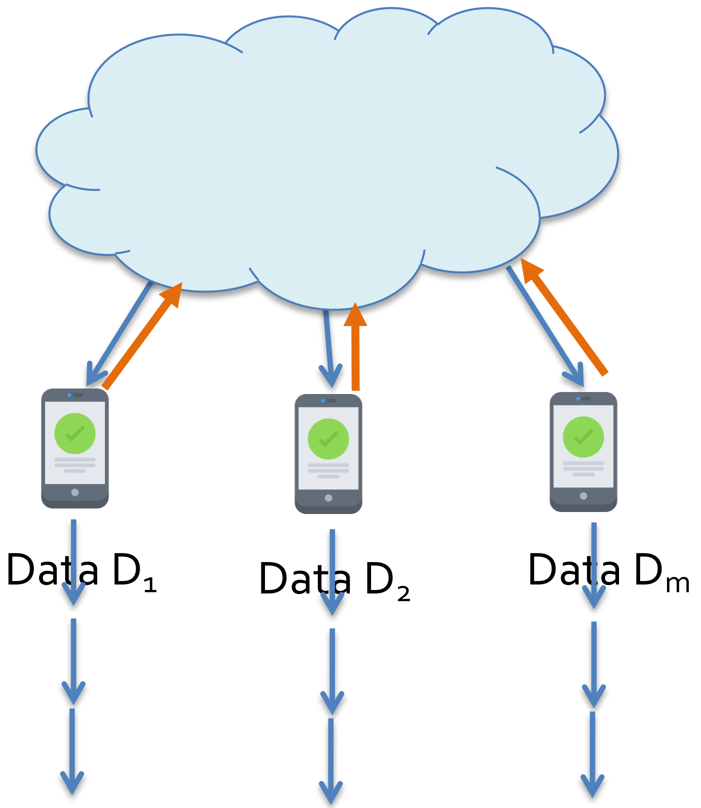


- Massive amounts of informative training data is being collected at edge devices such as cell phones, tablets, IoT sensors etc.
- Sending these data to the cloud can be too expensive and slow
- Privacy laws may also forbid data sharing with foreign cloud servers

Federated Learning: Bringing Training to the Edge

[McMahan et al 2017, Kairouz et al 2019]

- Data stays on the device, and model training is moved to the edge
- Each edge client performs a few local SGD updates, and the resulting models are aggregated by the central server
- Better communication-efficiency and privacy guarantees than sending all data to the cloud



Federated Optimization: Objectives and Notation

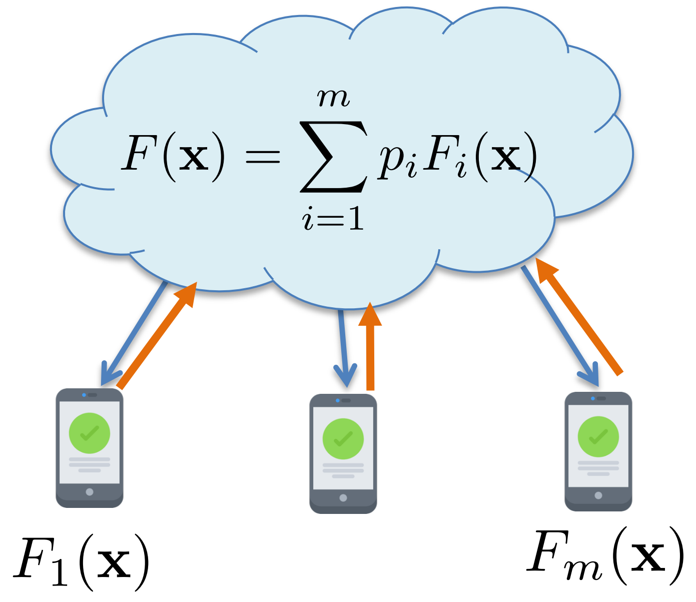
- Local Objective Function

$$F_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} f(\mathbf{x}; \xi)$$

- Global Objective is a weighted average of local objectives in proportional of data-sizes:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

where $p_i = \frac{|\mathcal{D}_i|}{\sum_{i=1}^{m} |\mathcal{D}_i|}$



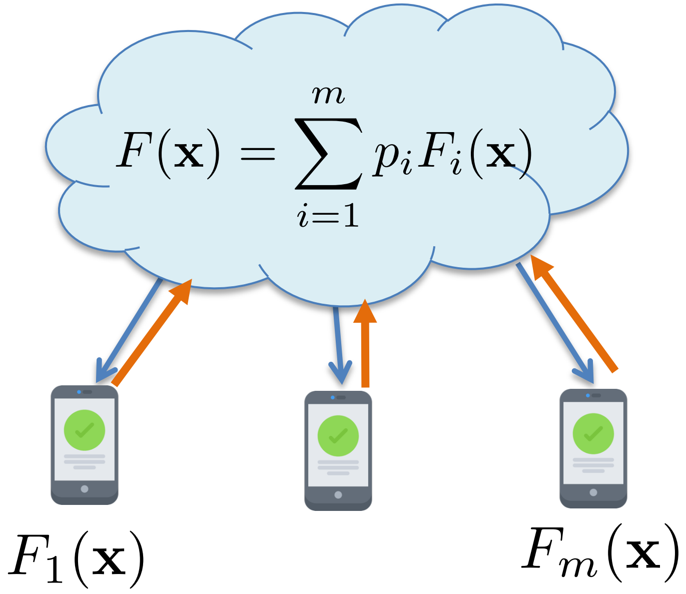
GOAL: Find ***x*** that minimizes the global objective

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

Federated Optimization: The FedAvg Algorithm

- Raw data at clients cannot be shared to the server due to privacy and communication constraints

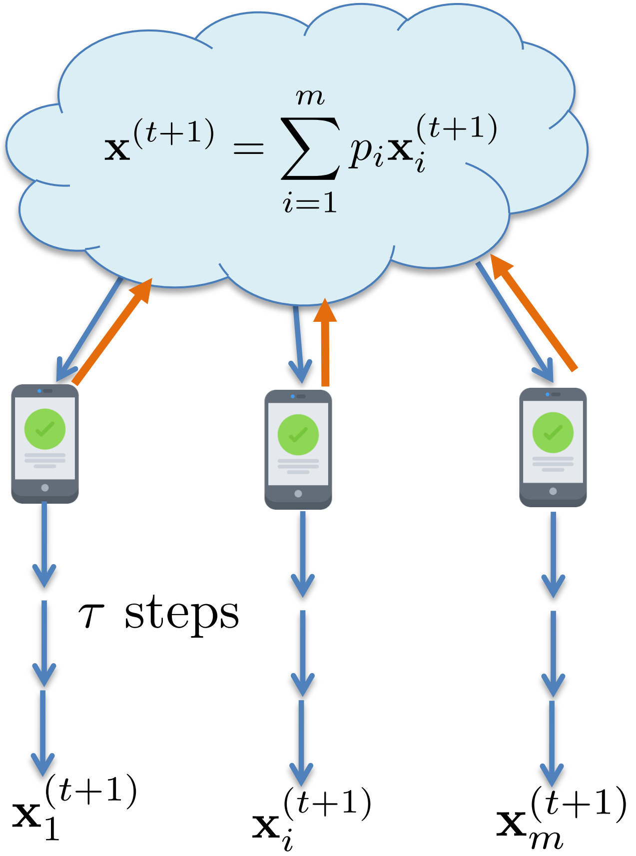
SOLUTION: Perform τ local updates at each client and only share the resulting model with the server



Federated Optimization: The FedAvg Algorithm

- Raw data at clients cannot be shared to the server due to privacy and communication constraints

SOLUTION: Perform τ local updates at each client and only share the resulting model with the server



FedAvg Algorithm

In each communication round:

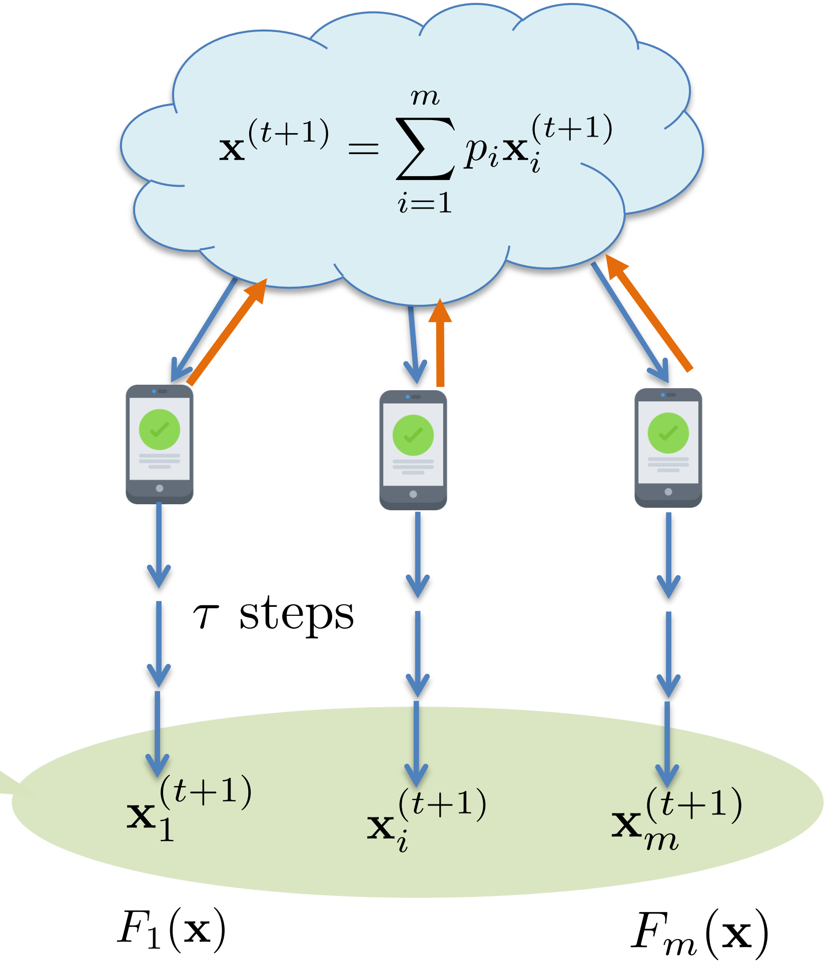
1. Send the current model to clients
2. Clients perform τ local updates using their data
3. Updated models are aggregated by the server

How is this algorithm affected by heterogeneity in the system?

Sources of Heterogeneity in Federated Learning

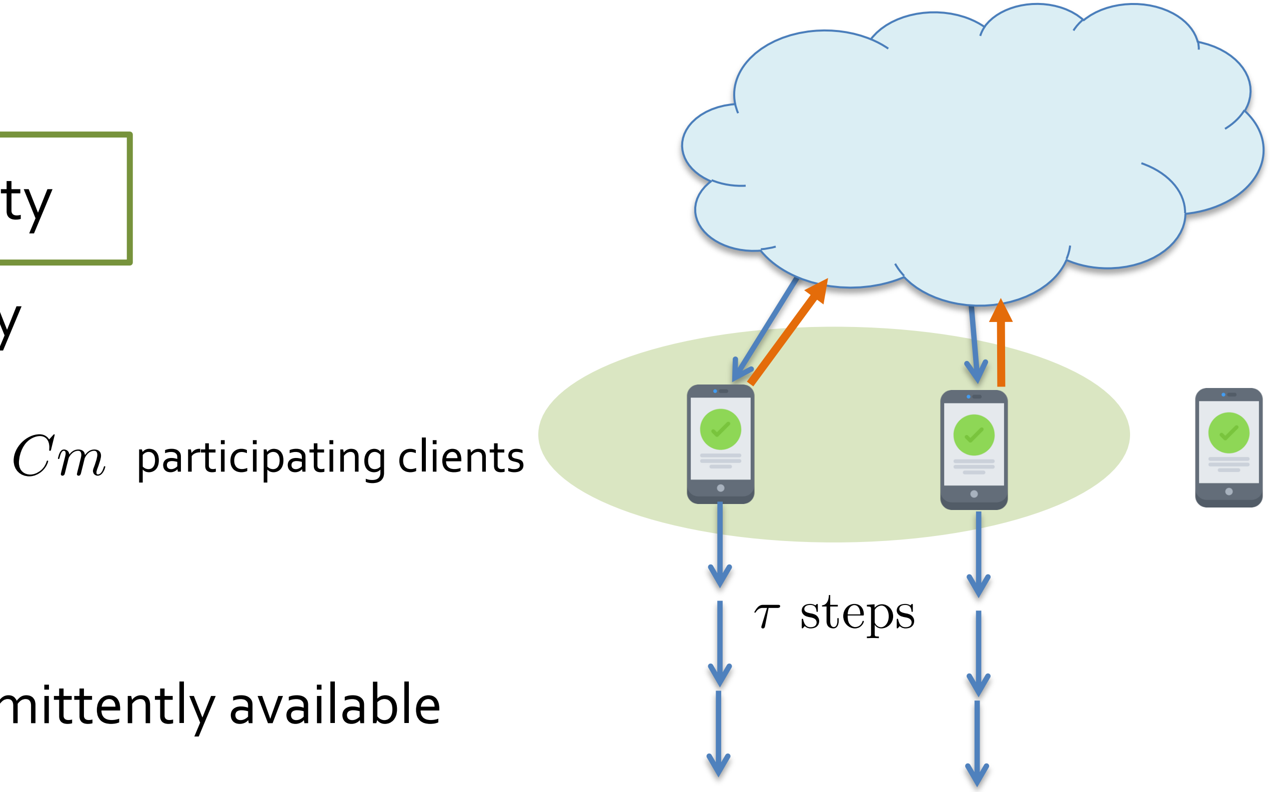
1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity

Due heterogeneous datasets and objective functions, local models drift apart as τ increases



Sources of Heterogeneity in Federated Learning

1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity

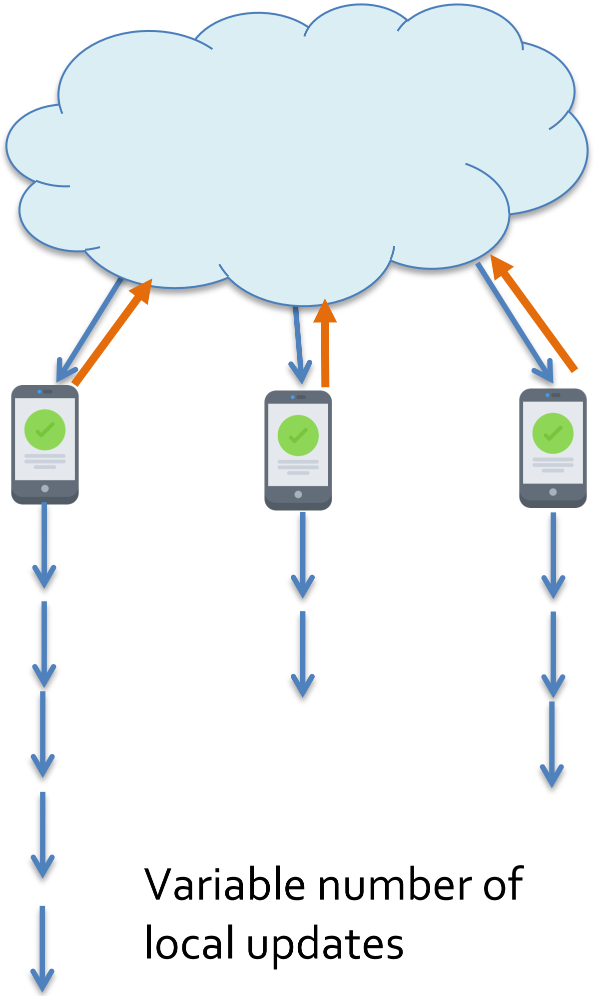


Thousands of clients that are intermittently available

SOLUTION: Partial participation of Cm clients, selected from among the available clients

Sources of Heterogeneity in Federated Learning

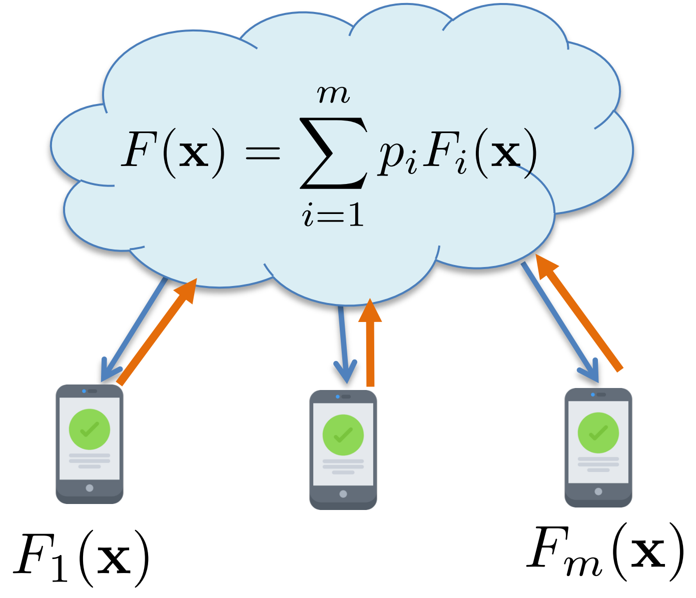
1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity



- Different computation speeds and memory
- Different learning rates or adaptive local optimizers

Sources of Heterogeneity in Federated Learning

1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity



THIS TALK

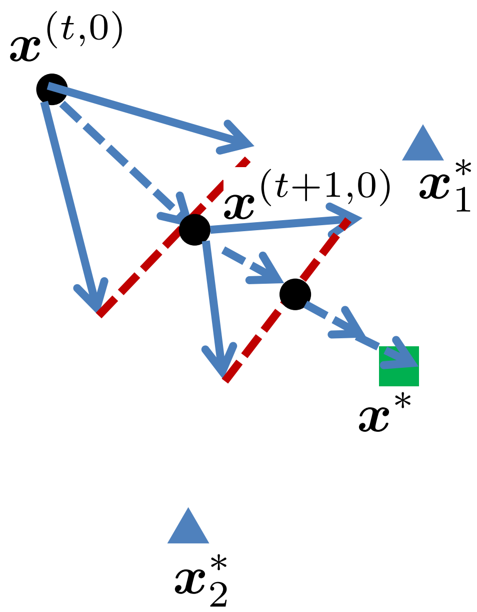
How do these sources of heterogeneity affect federated optimization algorithms and analyses?

GOAL: Find ***x*** that minimizes the global objective

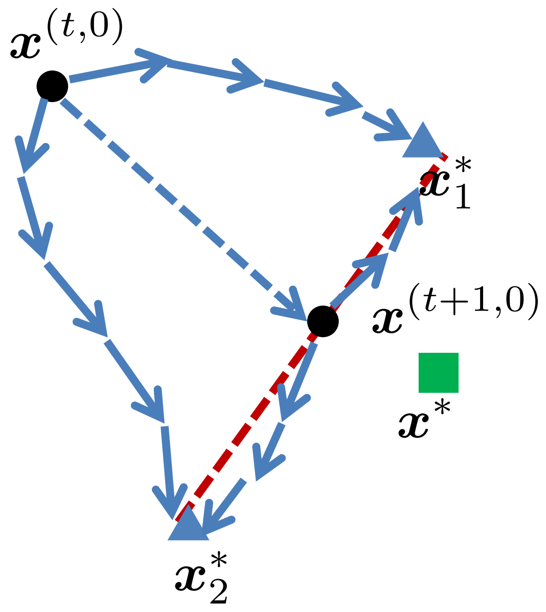
$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

1. Data Heterogeneity

τ = 1
case



Larger τ
case



Each client's gradient moves its model towards its minima, but the averaged gradient leads to the global optimum $\boldsymbol{x}^*$

The global model becomes the average of the local minima, which may differ from the true optimum $\boldsymbol{x}^*$

1. Data Heterogeneity

Client Drift Error and How to Mitigate it

For bounded gradient dissimilarity, that is, $\|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \sigma_g^2$ error for T communication rounds is:

$$\min_{t \in \{0,\ldots,T-1\}} \mathbb{E}[\|\nabla F(\mathbf{x}^{(t)})\|^2] \leq O\left(\frac{F(\mathbf{x}^{(0)}) - F^*}{\eta \tau T}\right) + O\left(\frac{\eta L \sigma^2}{m} + \eta^2 L^2 (\tau - 1)\sigma^2\right) + O\left(\eta^2 L^2 \tau(\tau - 1)\sigma_g^2\right)$$

Client Drift Error

Methods to Mitigate Client Drift Error

- Setting a small τ and/or small η
- Adapting τ over rounds [AdaComm 2019]
- Adding correction to pull back drifted models [FedProx,2020], [SCAFFOLD, 2021] etc

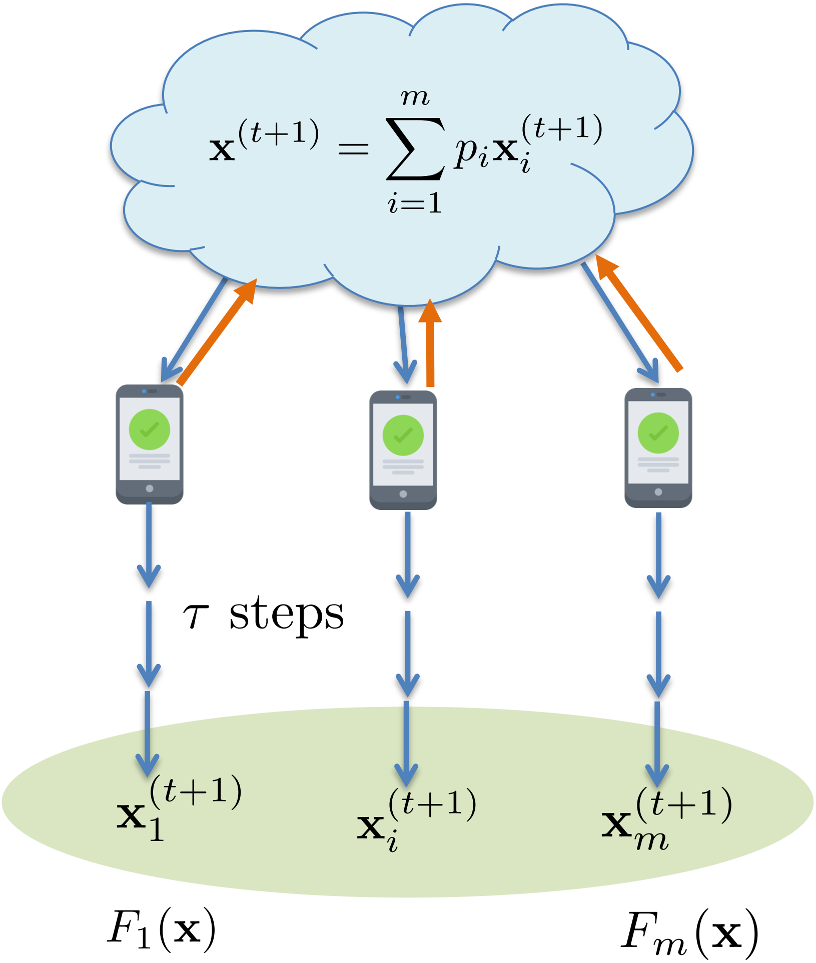
1. Data Heterogeneity

[Jhunjhunwala et al ICLR 2023, spotlight talk]

FedExP: Adapting Server Learning Rate

- Slowdown due to small client learning rate can be compensated by a larger server learning rate (default value 1 in FedAvg)
- We propose the following adaptive schedule based on the Extrapolated Parallel Projection method (EPPM) [Pierra 1984]

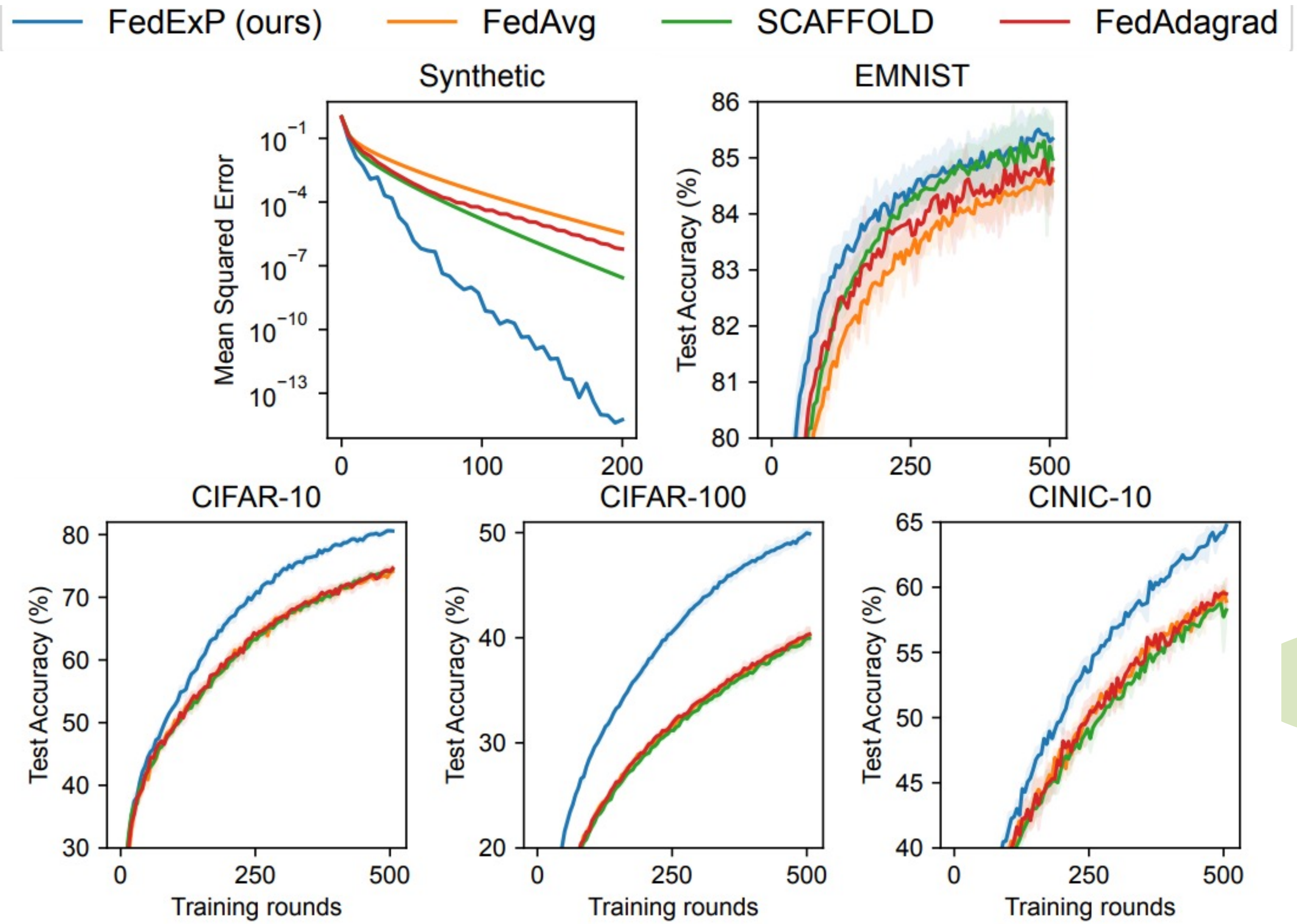
$$\eta_g = \max\left(1, \frac{\sum_{i=1}^{m} \|\mathbf{x} - \mathbf{x}_i^{(t)}\|^2}{2m(\|\mathbf{x} - \mathbf{x}_i^{(t)}\|^2 + \epsilon)}\right)$$

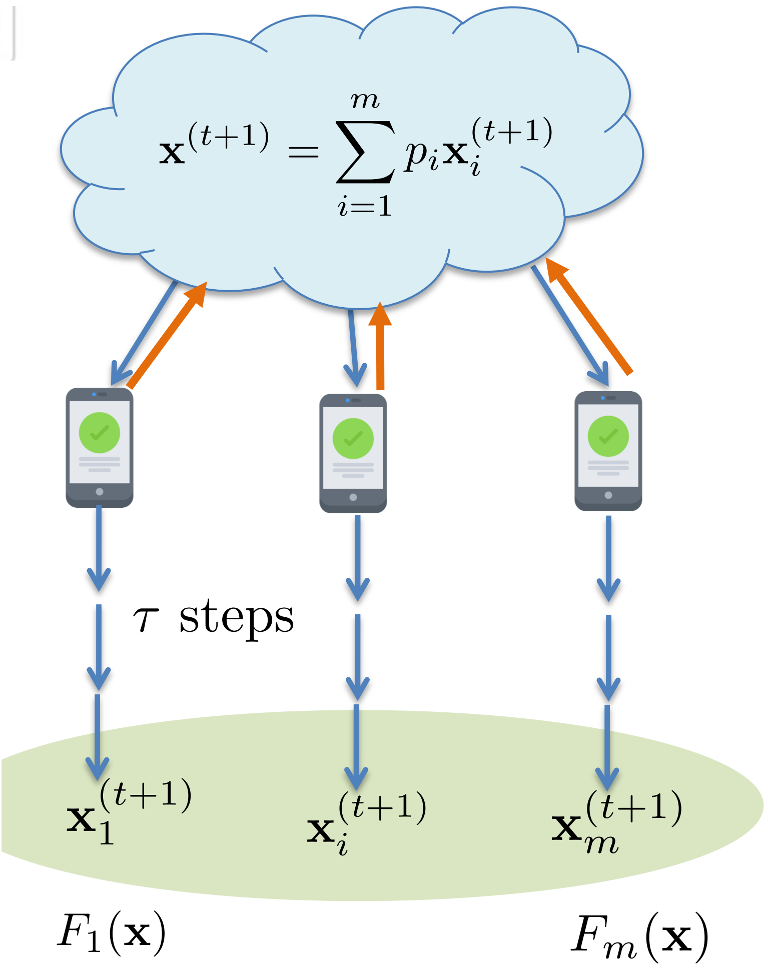


1. Data Heterogeneity

[Jhunjhunwala et al
ICLR 2023, spotlight talk]

FedExP: Adapting Server Learning Rate



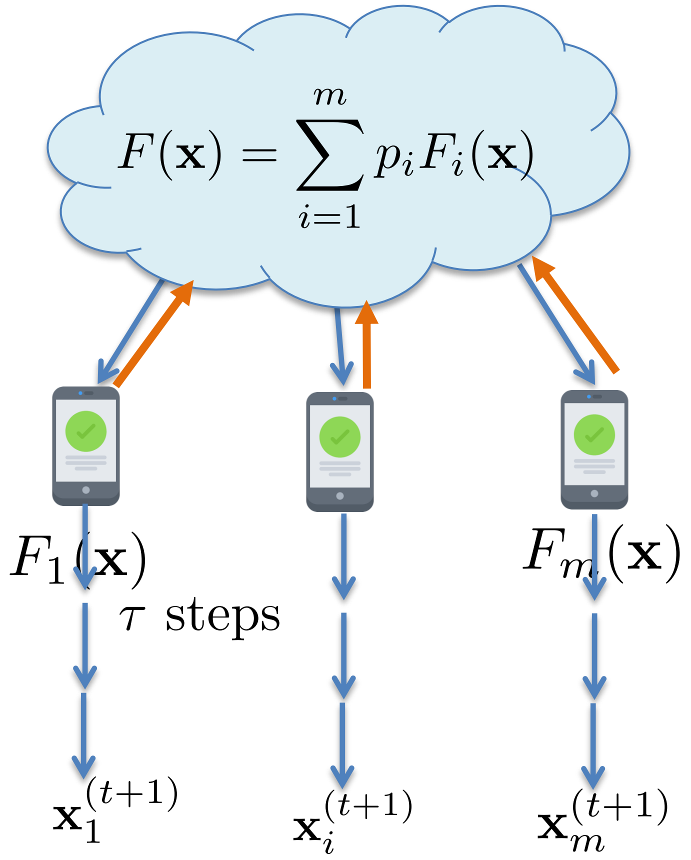


1. Data Heterogeneity

Open Questions and Ongoing Work

Q1: Is gradient dissimilarity assumption too pessimistic?

- In practice, FedAvg outperforms SGD, even though the client drift error increases with τ
- [Wang et al 2022] proposes a different data heterogeneity measure called average drift at optimum



1. Data Heterogeneity

Open Questions and Ongoing Work

For bounded gradient dissimilarity, that is, $\|\nabla F_i(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \sigma_g^2$ error for T communication rounds is:

$$\min_{t \in \{0,\ldots,T-1\}} \mathbb{E}[\|\nabla F(\mathbf{x}^{(t)})\|^2] \leq O\left(\frac{F(\mathbf{x}^{(0)}) - F^*}{\eta \tau T}\right) + O\left(\frac{\eta L \sigma^2}{m} + \eta^2 L^2 (\tau - 1)\sigma^2\right)$$
$$+ O\left(\eta^2 L^2 \tau(\tau - 1)\sigma_g^2\right)$$

Client Drift Error

Q2: Is client drift the dominant error term?

- With small learning rate, it decays faster than other terms
- Partial participation error is higher order (as we see next)

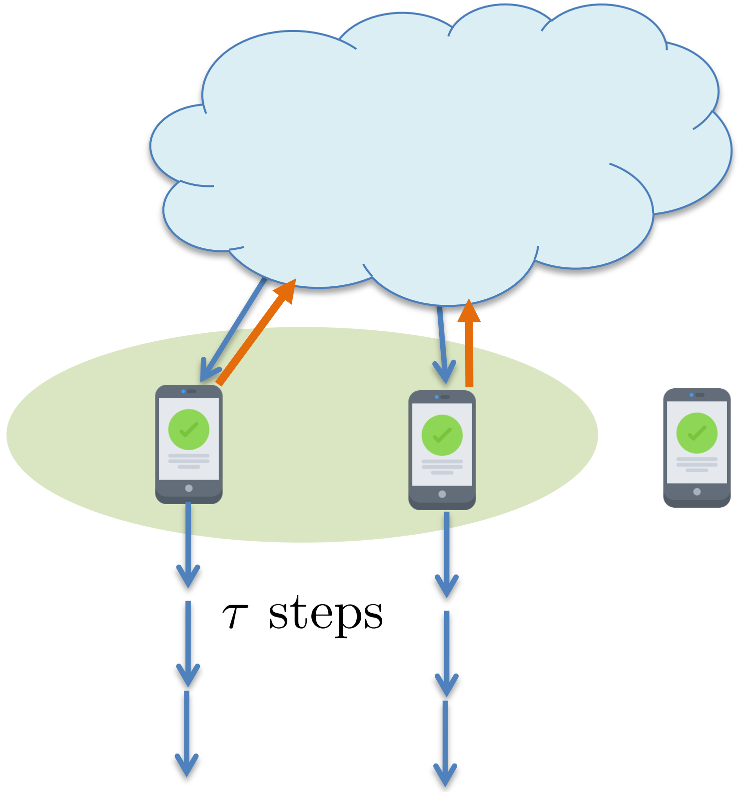
Sources of Heterogeneity in Federated Learning

1. Data Heterogeneity

2. Communication Heterogeneity

3. Computational Heterogeneity

Cm participating clients

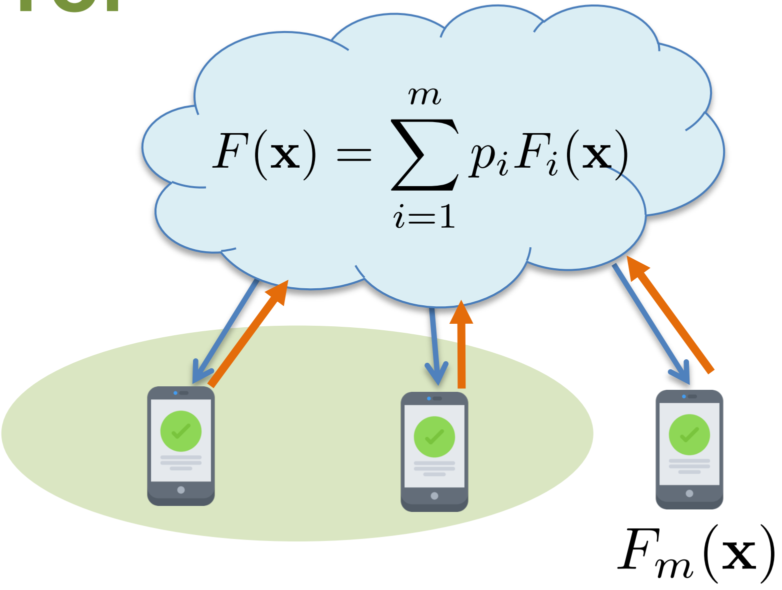


Thousands of clients that are intermittently available

SOLUTION: Partial participation of Cm clients, selected from among the available clients

2. Communication Heterogeneity

Partial Client Participation Error



- In FedAvg, fraction C of the clients chosen uniformly at random participate in each round
- The error after T communication rounds isas given by

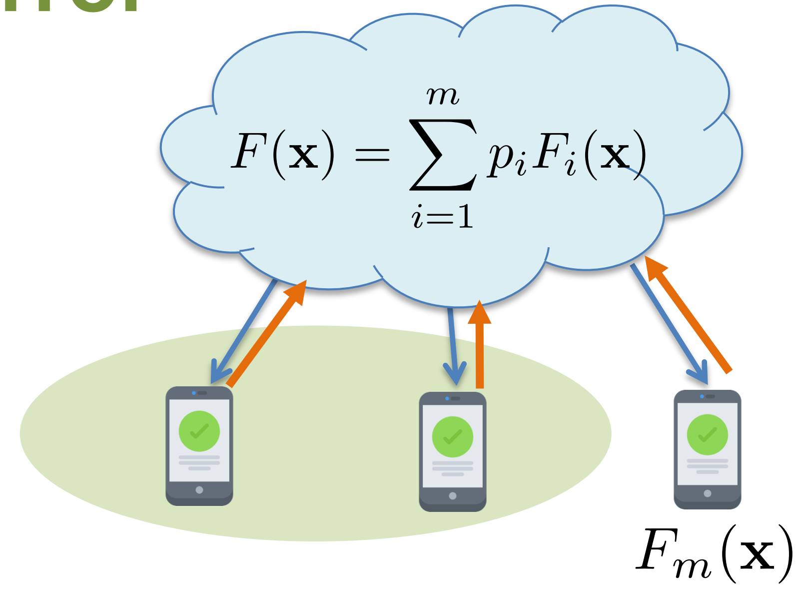
$$\min_{t\in\{0,\dots,T-1\}} \mathbb{E}\|\nabla F(\mathbf{x}^{(t)})\|^2 \leq O\left(\frac{F(\mathbf{x}^{(0)}) - F^*}{\eta\tau T}\right) + O\left(\frac{\eta L\sigma^2}{Cm} + \eta^2 L^2(\tau-1)\sigma^2\right) + O\left(\frac{\eta\tau L(1-C)\sigma_g^2}{C(m-1)}\right) + O\left(\eta^2 L^2 \tau(\tau-1)\sigma_g^2\right)$$

Partial Client Participation Error

Client Drift Error

2. Communication Heterogeneity

Partial Client Participation Error



- In FedAvg, fraction C of the clients chosen uniformly at random participate in each round
- The error after T communication rounds is
- After setting learning rate appropriately, we get

$$\min_{t \in [0,T]} \|\nabla f(\mathbf{w}^{(t)})\|^2 \leq \underbrace{\mathcal{O}\left(\frac{1}{\sqrt{M\tau T}}\right)}_{\text{stochastic noise}} + \underbrace{\mathcal{O}\left(\sqrt{\frac{\tau}{MT}}\right)}_{\text{partial participation}} + \underbrace{\mathcal{O}\left(\frac{1}{T}\right)}_{\text{client drift}}$$

Dominates

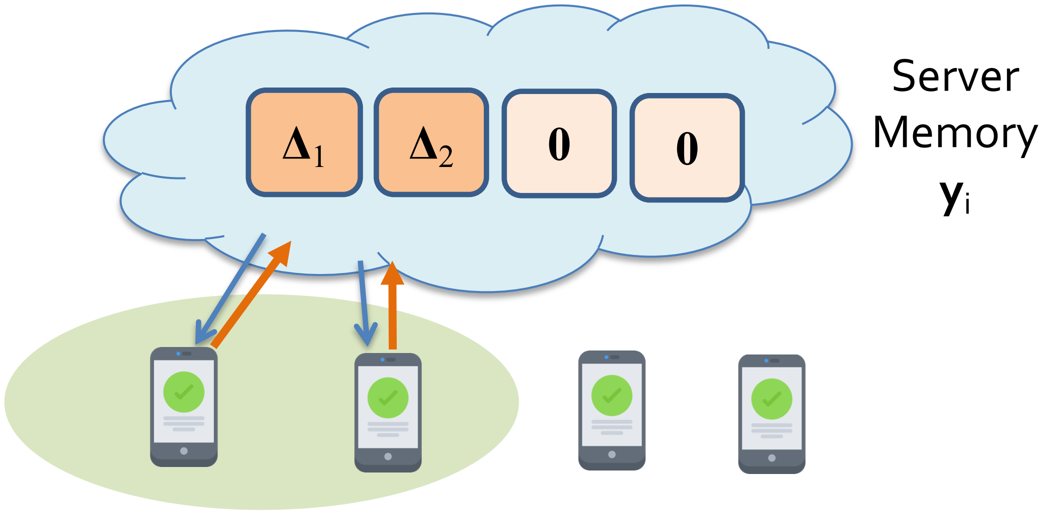
IDEA: Mitigate this term using correlation in client updates across rounds, like SAGA

2. Communication Heterogeneity

FedVARP: Reducing Participation Variance [FedVARP, UAI 2022]

1. Update node state in server memory

$$\mathbf{y}_j^{(t+1)} = \begin{cases} \Delta_i^{(t)} & \text{if } j \in \mathcal{S}^{(t)} \\ \mathbf{y}^{(t)} & \text{otherwise} \end{cases} \qquad \forall j \in [N]$$



$$\Delta_i^{(t)} = \mathbf{x}^{(t)} - \mathbf{x}_i^{(t,\tau)}$$

Key Idea: Use latest observed update $\{\mathbf{y}_j^{(t)}\}_{j=0}^{N}$ for each node as proxy for current update.

2. Communication Heterogeneity

FedVARP: Reducing Participation Variance [FedVARP, UAI 2022]

1. **Update node state in server memory**

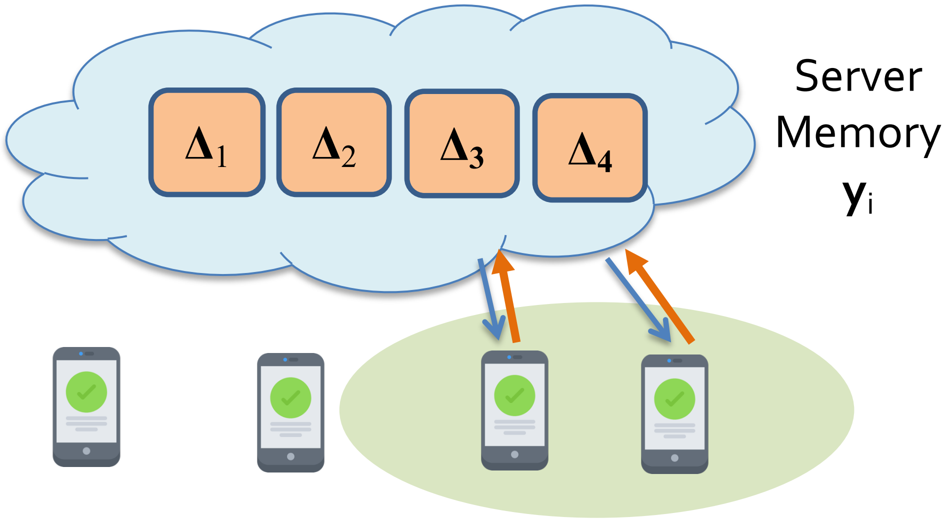
$$\mathbf{y}_j^{(t+1)} = \begin{cases} \Delta_i^{(t)} & \text{if } j \in \mathcal{S}^{(t)} \\ \mathbf{y}^{(t)} & \text{otherwise} \end{cases} \qquad \forall j \in [N]$$

2. **Compute variance-reduced update:**

$$\mathbf{v}^{(t)} = \frac{1}{|\mathcal{S}^{(t)}|} \sum_{i \in \mathcal{S}^{(t)}} (\Delta_i^{(t)} - \mathbf{y}_i^{(t)}) + \frac{1}{N} \sum_{j=1}^{N} \mathbf{y}_j^{(t)}$$

3. **Update global model:**

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_g \mathbf{v}^{(t)}$$



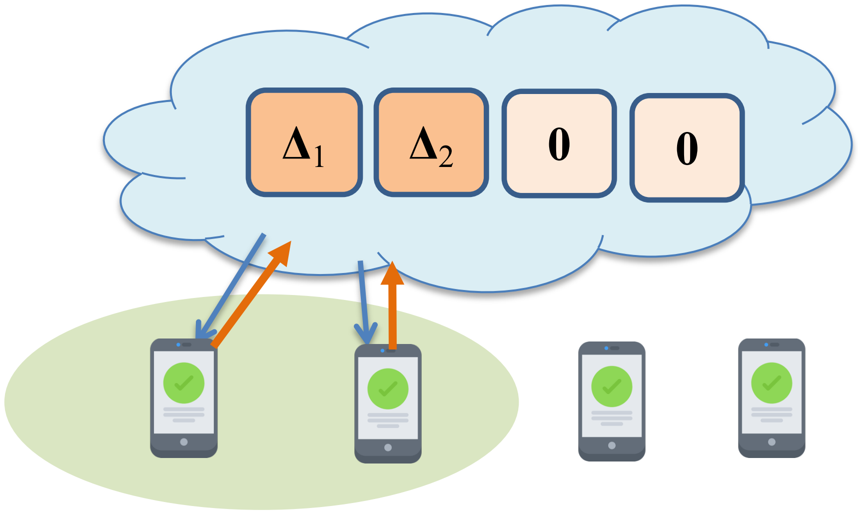
$$\Delta_i^{(t)} = \mathbf{x}^{(t)} - \mathbf{x}_i^{(t,\tau)}$$

Key Idea: Use latest observed update $\{\mathbf{y}_j^{(t)}\}_{j=0}^{N}$ for each node as proxy for current update.

2. Communication Heterogeneity

FedVARP: Convergence Analysis and Results

- By using the variance-reduced update that includes all clients' updates, FedVARP eliminates partial client participation error



$$\min_{t\in\{0,\dots,T-1\}} \mathbb{E}\|\nabla F(\mathbf{x}^{(t)})\|^2 \leq O\left(\frac{F(\mathbf{x}^{(0)}) - F^*}{\eta\tau T}\right) + O\left(\frac{\eta L \sigma^2}{Cm} + \eta^2 L^2 (\tau - 1)\sigma^2\right)$$

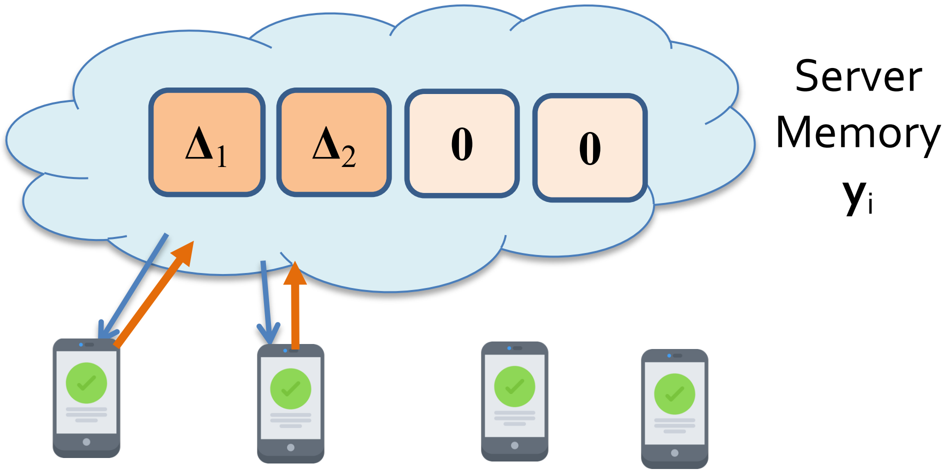
$$+ O\left(\frac{\eta\tau L(1-C)\sigma_g^2}{C(m-1)}\right) + O\left(\eta^2 L^2 \tau(\tau-1)\sigma_g^2\right)$$

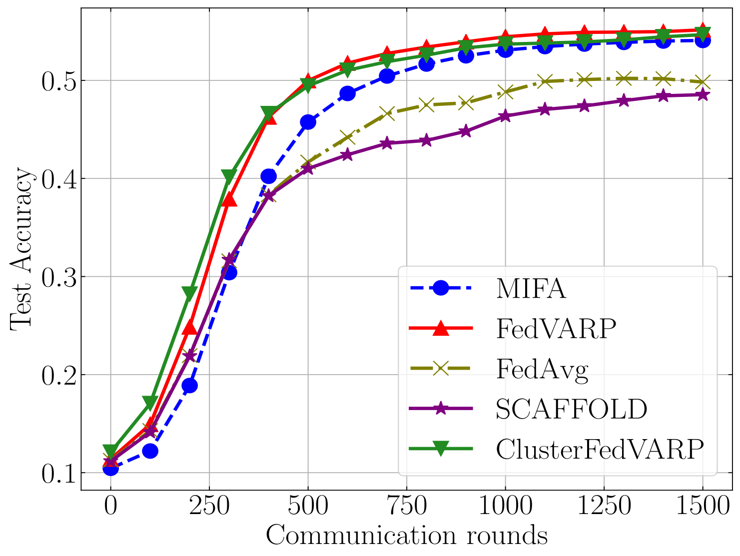
Partial Client Participation Error

2. Communication Heterogeneity

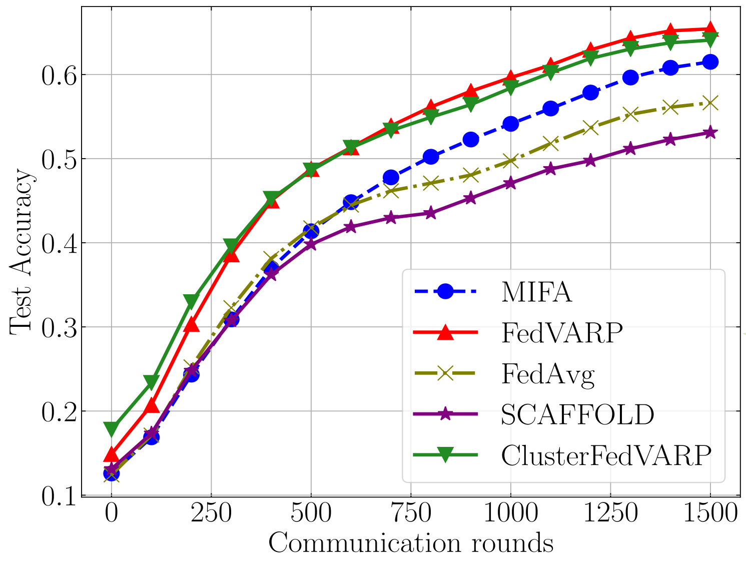
FedVARP: Convergence Analysis and Results

- Can we reduce server storage cost?
- Yes! **ClusterFedVARP** reduces storage by clustering clients and maintaining a single state per cluster.





i) Training LeNet-5 on CIFAR-10



ii) Training ResNet-18 on CIFAR-10

FedVARP outperforms FedAvg and client drift mitigation methods like SCAFFOLD

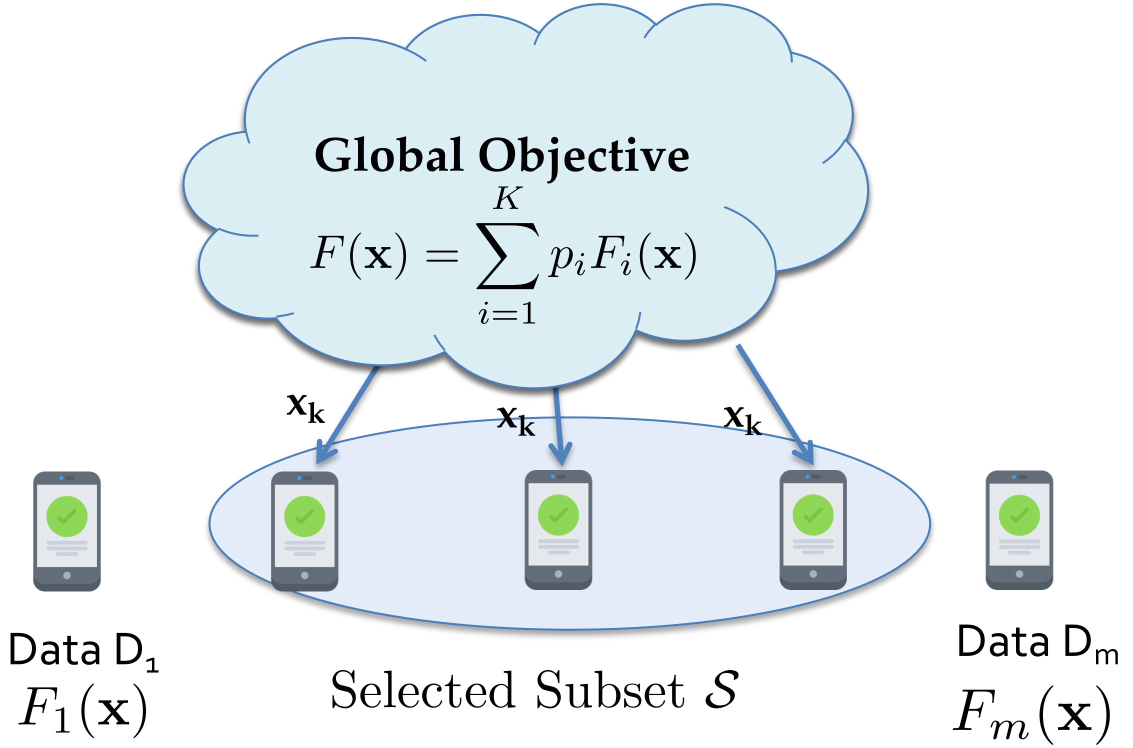
2. Communication Heterogeneity

Client Selection in Federated Learning

[Cho et al AISTATS 2022]

- **Unbiased Selection:** If we sample clients with probability p_k with replacement we have *unbiased* sampling, i.e. $\mathbb{E}[\tilde{F}(\boldsymbol{x})] = F(\boldsymbol{x})$

- Most works consider this scenario, and many prior results with full client participation can be extended to this setting

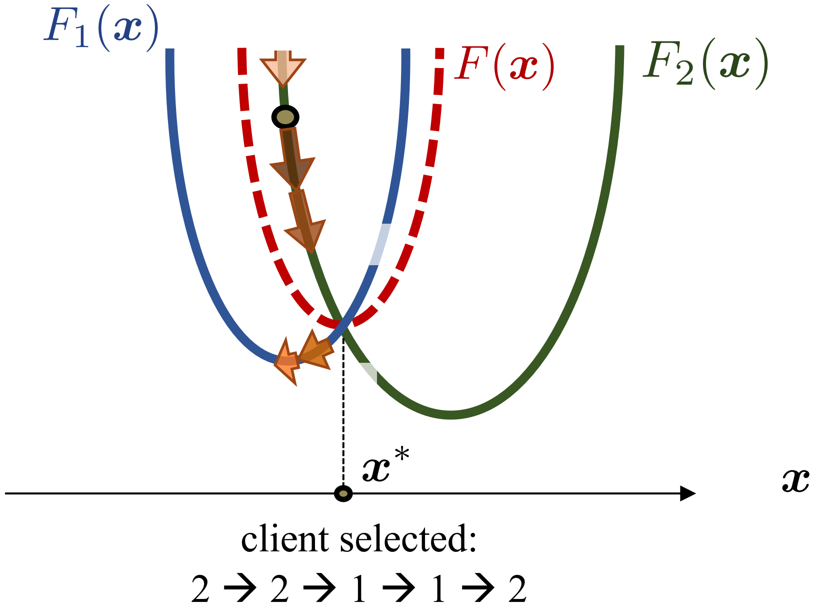


Can we improve convergence by biasing client selection towards higher loss clients??

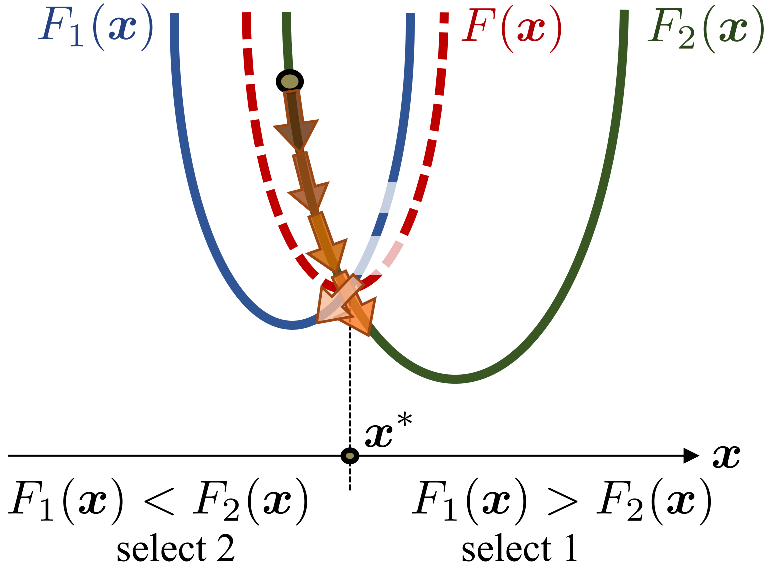
2. Communication Heterogeneity

Biased Client Selection can Speed-up Convergence

Unbiased Random Selection



Biasing Selection towards high loss clients



- Biasing towards higher loss clients gives faster convergence
- But will too much selection skew result in a higher solution bias?

2. Communication Heterogeneity

Measure Skew of a Client Selection Strategy

Selection strategy π maps the current global model to a selected set of clients $\mathcal{S}(\pi, \boldsymbol{x})$

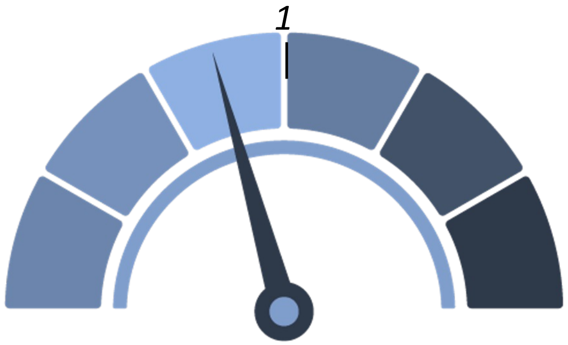
Selection Skew is measured in terms of the following quantities

$$\bar{\rho} = \min_{\boldsymbol{x}, \boldsymbol{x}'} \rho(\mathcal{S}(\pi, \boldsymbol{x}), \boldsymbol{x}') \qquad \tilde{\rho} = \max_{\boldsymbol{x}} \rho(\mathcal{S}(\pi, \boldsymbol{x}), \boldsymbol{x}^*)$$

where $\rho(\mathcal{S}(\pi, \boldsymbol{x}), \boldsymbol{x}') = \dfrac{\mathbb{E}_{\mathcal{S}}[\frac{1}{m}\sum_{k \in \mathcal{S}(\pi, \boldsymbol{x})}(F_k(\boldsymbol{x}') - F_k^*)]}{F(\boldsymbol{x}') - \sum_{k=1}^{K} p_k F_k^*} \geq 0$

unbiased

biased towards clients with lower loss



biased towards clients with higher loss

$\rho(\mathcal{S}(\pi, \boldsymbol{x}), \boldsymbol{x}')$

2. Communication Heterogeneity

Convergence with Biased Client Selection

Convergence guarantees for any client selection strategy for L-smooth and μ-strongly convex functions:

data heterogeneity

$$\text{Error after } T \text{ rounds} \leq \underbrace{O\left(\frac{1}{T\bar{\rho}}\right)}_{\textit{convergence rate}} + \underbrace{O\left(\Gamma\left(\frac{\tilde{\rho}}{\bar{\rho}} - 1\right)\right)}_{\textit{non-vanishing bias}}$$

Observations:

- More selection skew $\bar{\rho} > 0$ brings faster convergence
- But too much selection skew increases the non-vanishing bias term $\Gamma = 0$
- To get zero solution bias, we need $\widetilde{\rho} = \overline{\rho} = 1$ (homogeneous data) or an unbiased selection strategy

2. Communication Heterogeneity

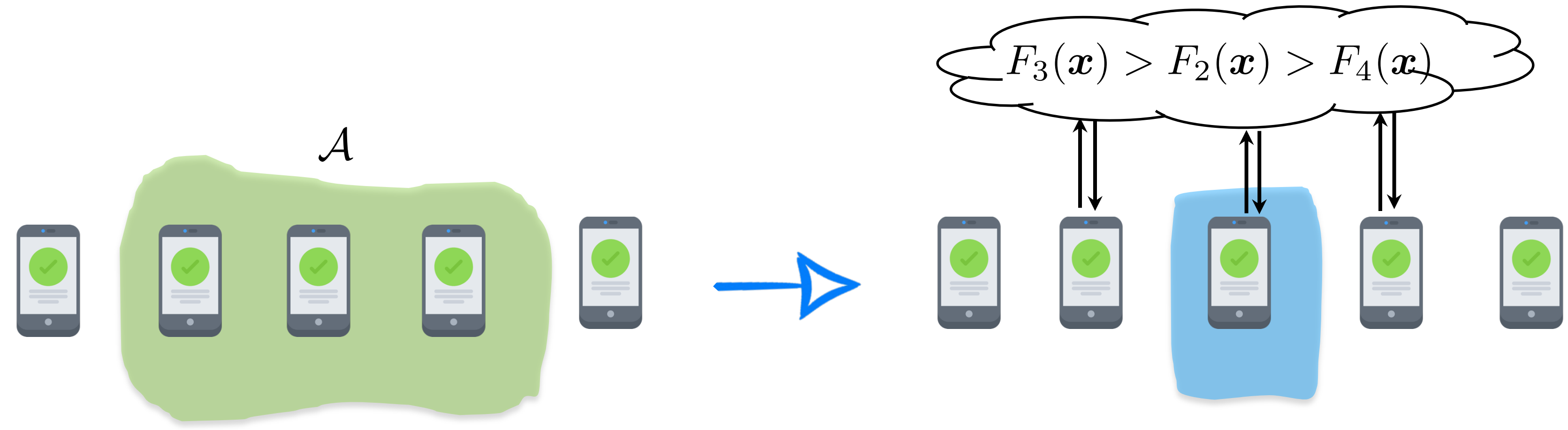
Power-of-d choices Client Selection

[Cho et al AISTATS 2022]

Step 1. Sample clients to candidate set $\mathcal{A}$ of size d with probability p_k

Step 2. Estimate Local Losses of clients in set $\mathcal{A}$ for current global model

Step 3. Select the Cm clients with the largest local losses



- Setting $d = Cm$ is equivalent to unbiased client selection
- Connected to mini-batch sampling techniques used in single-node SGD training

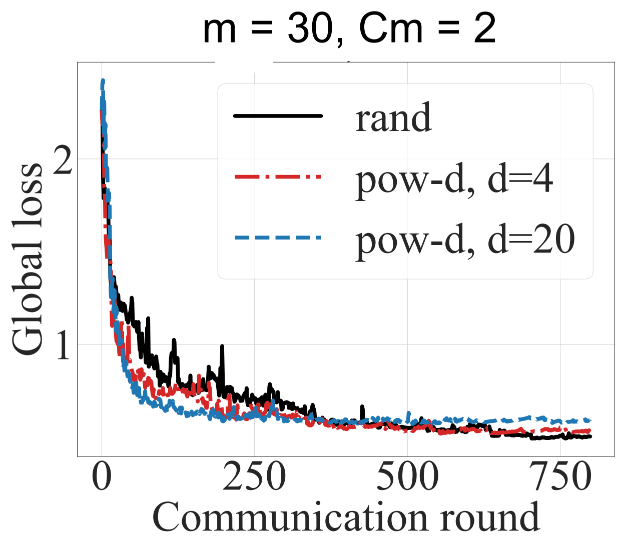
2. Communication Heterogeneity

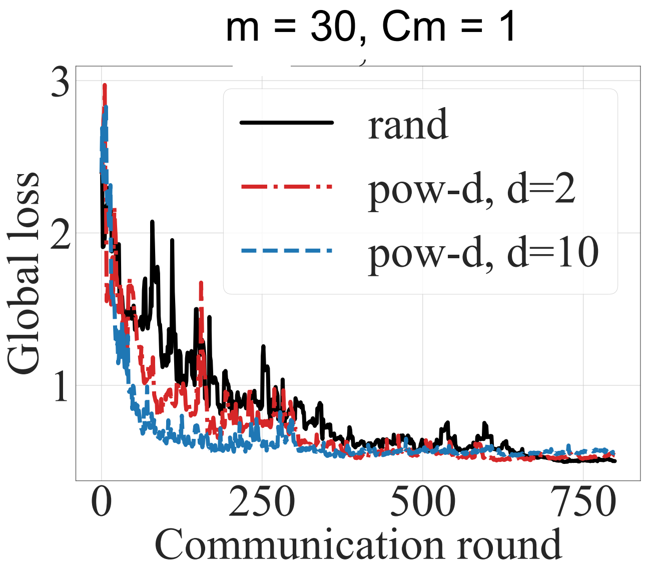
Power-of-d choices Client Selection

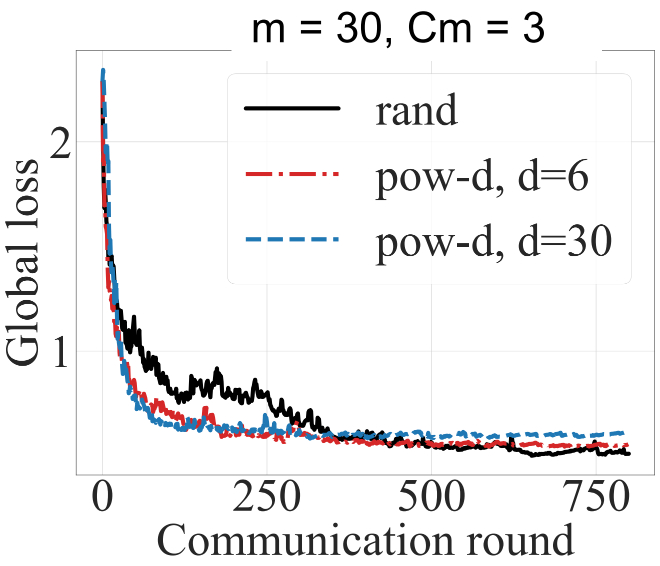
Step 1. Sample clients to candidate set $\mathcal{A}$ of size d with probability p_k

Step 2. Estimate Local Losses of clients in set $\mathcal{A}$ for current global model

Step 3. Select the Cm clients with the largest local losses







Larger d gives faster convergence, but slightly higher erorr floor

2. Communication Heterogeneity

Power-of-d choices Client Selection

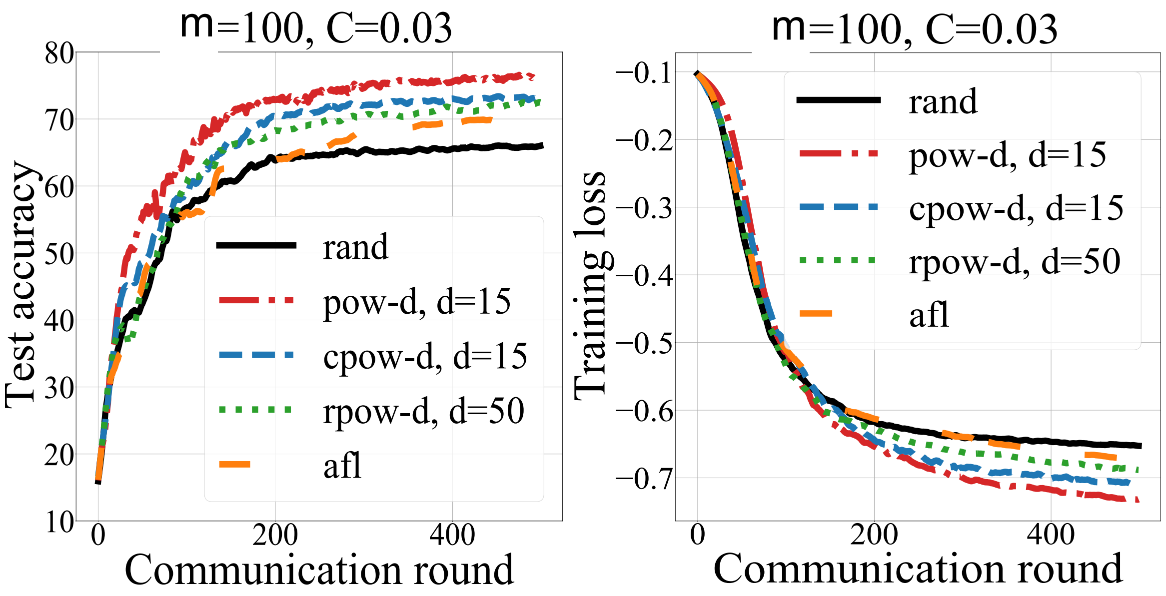
Step 1. Sample clients to candidate set $\mathcal{A}$ of size d with probability p_k

Step 2. Estimate Local Losses of clients in set $\mathcal{A}$ for current global model

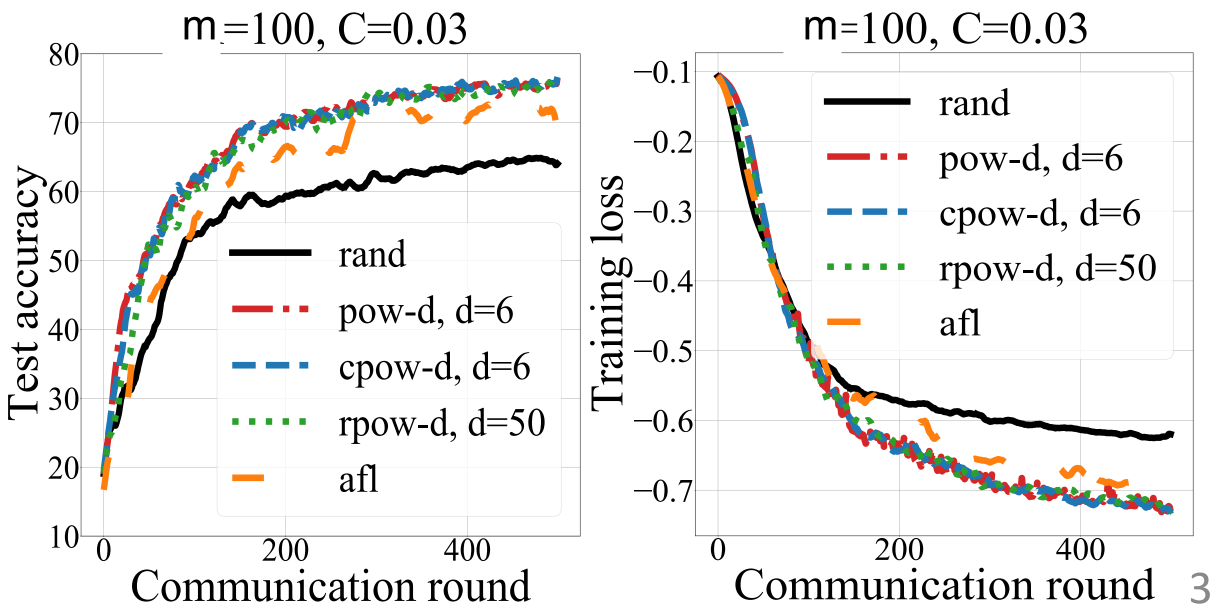
Step 3. Select the Cm clients with the largest local losses

We have comp. and comm. efficient variants of this step

Less Data Heterogeneity



More Data Heterogeneity



2. Communication Heterogeneity

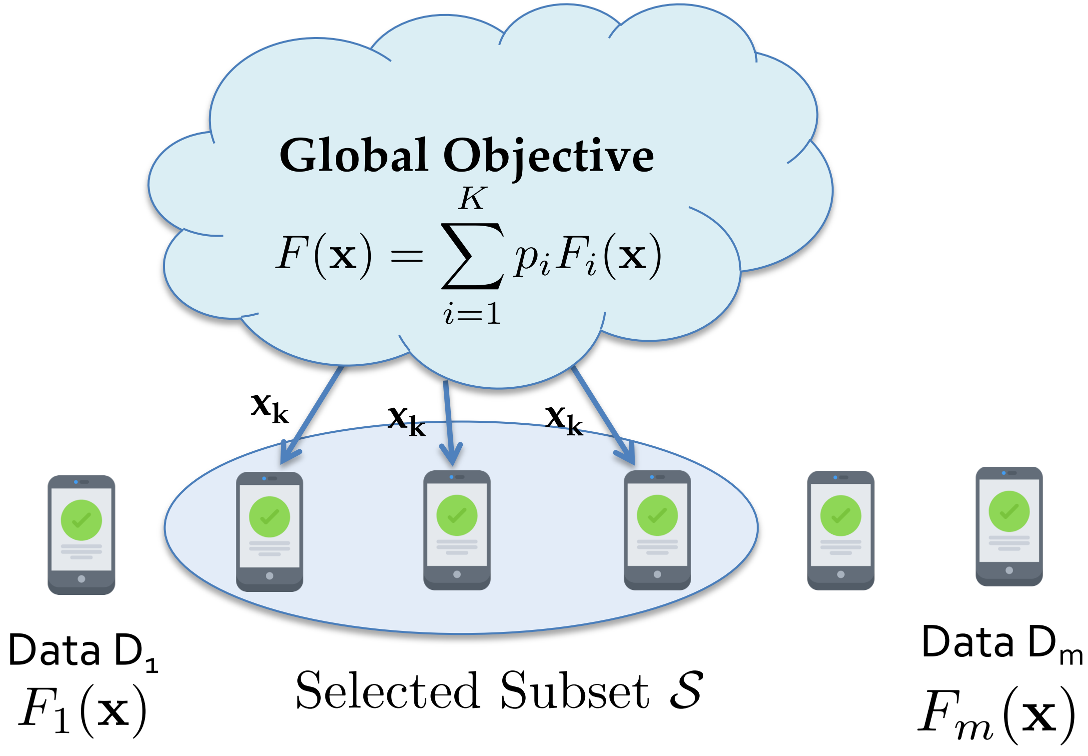
Open Directions in Client Selection

Q1: Can loss-aware and/or non-uniform client selection improve fairness?

- Yes, power-of-choice client selection improves fairness

Q2: Can loss-aware and/or non-uniform client selection improve robustness to adversarial clients?

- Yes, biasing towards lower loss clients can improve robustness

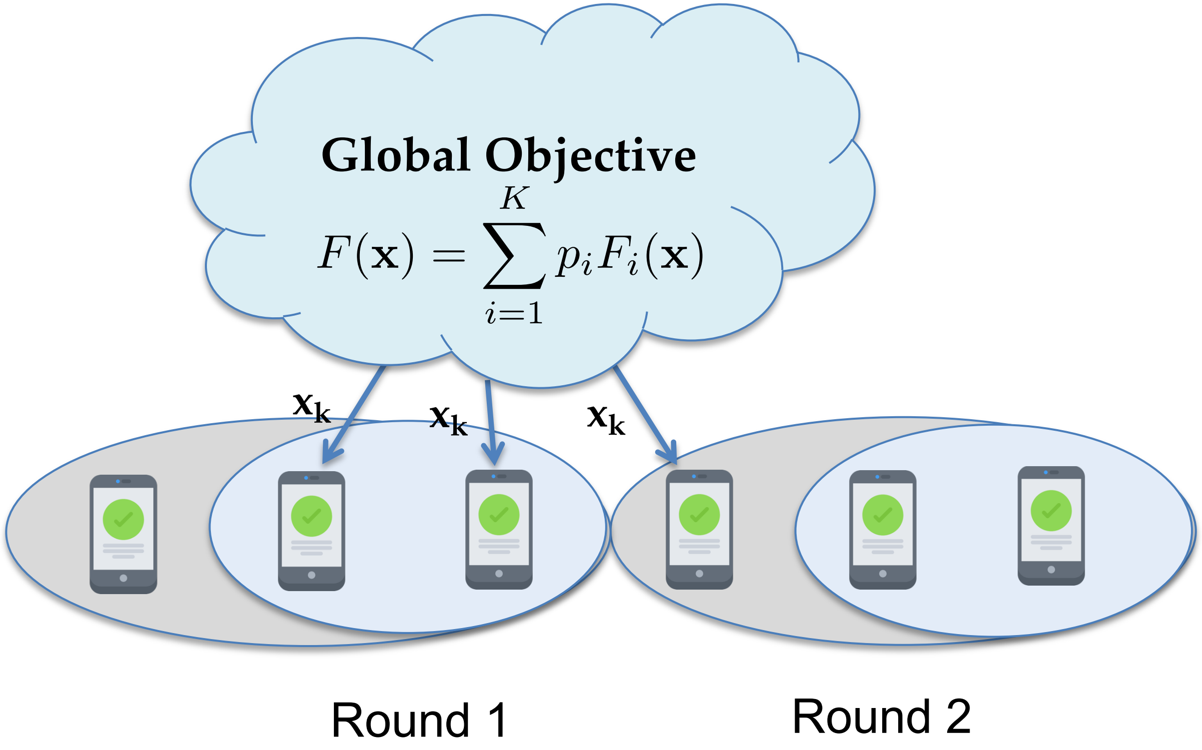


2. Communication Heterogeneity

Cyclic Client Participation

[Cho et al ICML 2023]

- Clients have a cyclic availability pattern based on location or timezone
- Defies the uniform sampling with replacement assumption made by most current FedAvg convergence analyses, which show an O(1/T) convergence with num. of comm. rounds T



Q: How does cyclic client participation affect FedAvg convergence?

2. Communication Heterogeneity

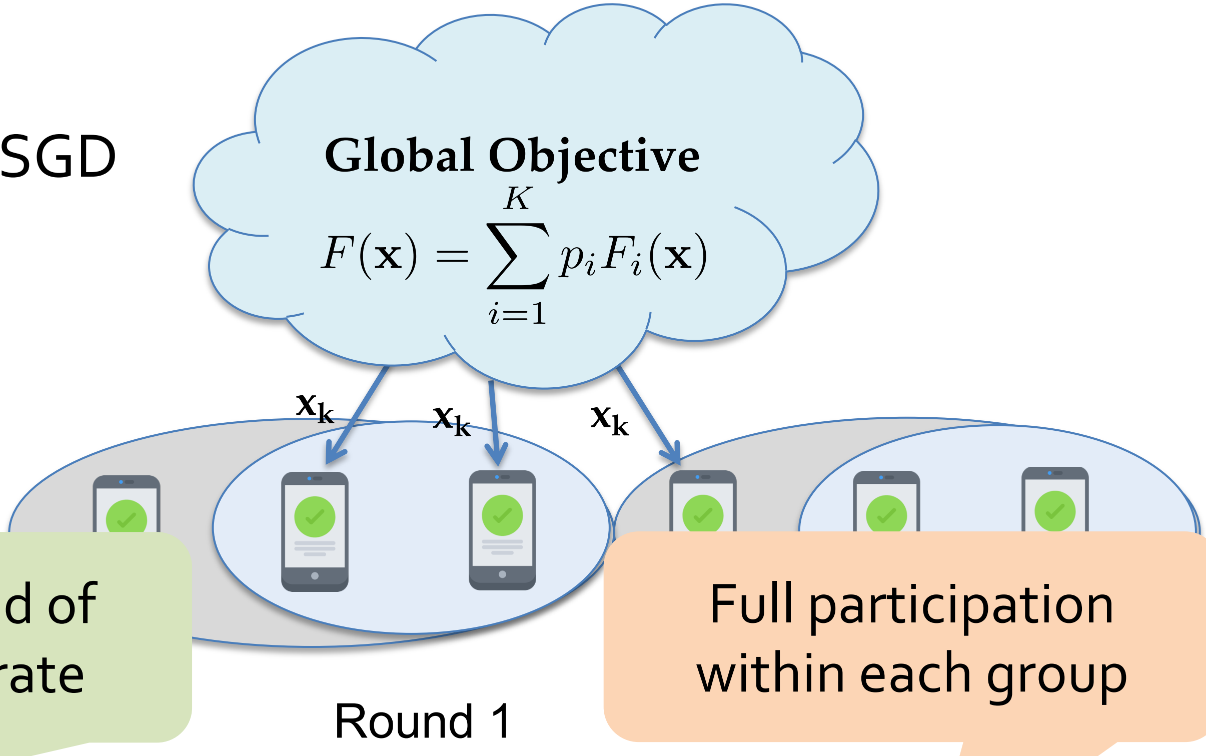
Cyclic Client Participation

[Cho et al ICML 2023]

First analysis with cyclic participation in FL

Technique based on [Yun 2022] on shuffled SGD

- Total number of clients: m
- Number of client groups: $\bar{K}$
- Clients selected per round: Cm



O(1/T²) instead of O(1/T) conv. rate

Full participation within each group

For local GD,

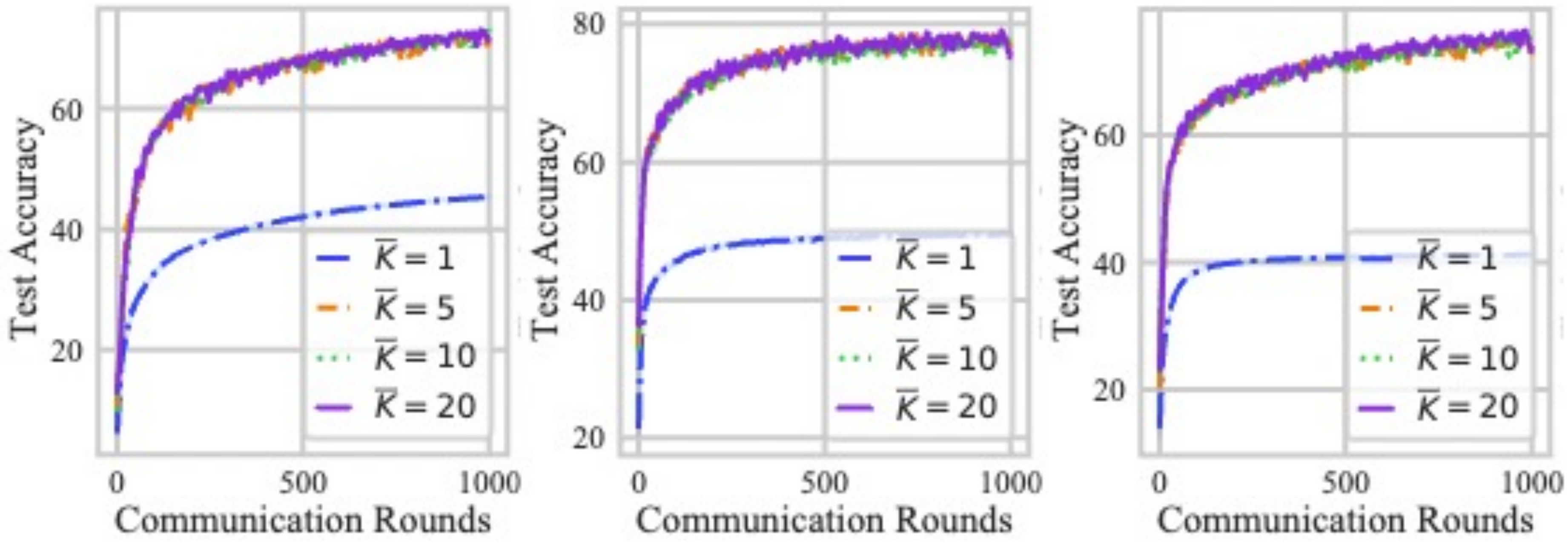
$$\mathbb{E}[F(\mathbf{x}^{(T)})] - F^* \leq \frac{\bar{K}^2(F(\mathbf{x}^{(0)}) - F^*)}{mT^2} + \tilde{O}\left(\frac{\kappa^2(\bar{K}-1)^2\alpha^2}{\mu T^2}\right) + \tilde{O}\left(\frac{\bar{K}\kappa\gamma^2}{\mu C T}\left(\frac{1 - \bar{K} - C}{m/\bar{K} - 1}\right)\right)$$

2. Communication Heterogeneity

Cyclic Client Participation

[Cho et al ICML 2023]

- EMNIST, Number of client groups: $\overline{K}$

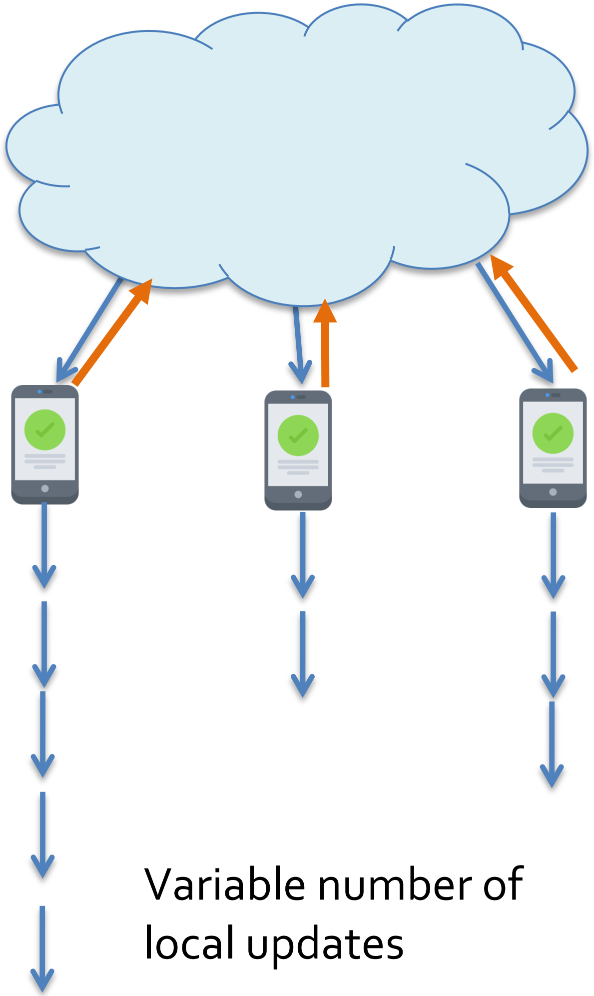


(a) GD (b) Local SGD (c) Shuffled SGD

Sources of Heterogeneity in Federated Learning

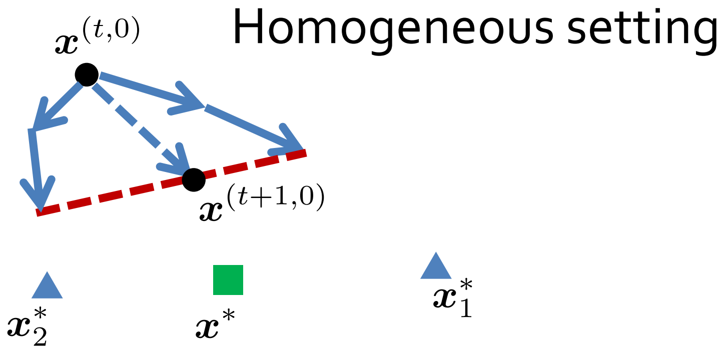
1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity

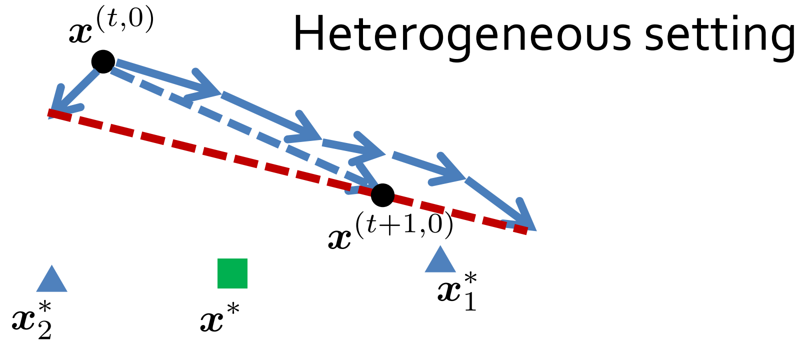
- Different computation speeds and memory
- Different learning rates or adaptive local optimizers



Variable number of local updates

3. Computational Heterogeneity





Cannot be fixed by the techniques used to tackle data heterogeneity

In [FedNova, NeurIPS 2020] we analyze a generalized FedAvg algorithm and show that

True Global Objective

$$F(\mathbf{x}) = \sum_{i=1}^{m} \frac{n_i}{n} F_i(\mathbf{x})$$
$$= \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

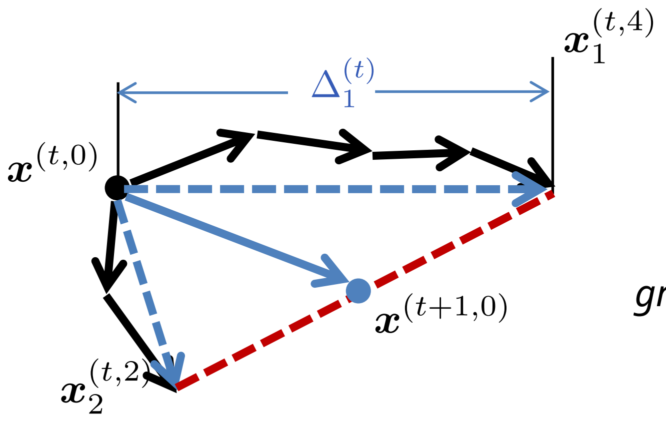
Mismatched Global Objective

$$\tilde{F}(\boldsymbol{x}) = \sum_{i=1}^{m} \frac{n_i \tau_i}{\sum_{i=1}^{m} n_i \tau_i} F_i(\boldsymbol{x})$$

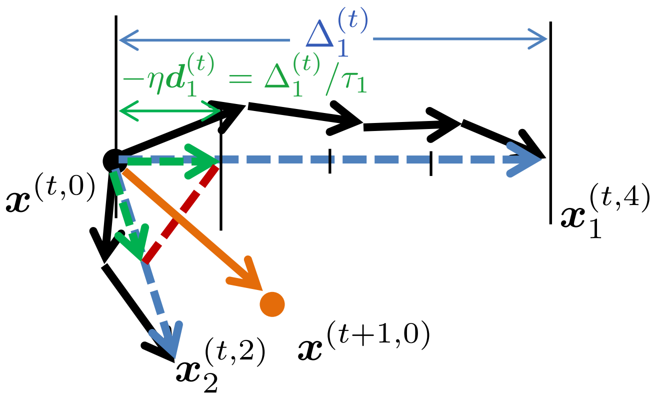
Need to fix the aggregation weights!

3. Computational Heterogeneity

A Generalized Version of the FedAvg algorithm



Average *normalized gradients* instead of local changes



FedAvg's Update Rule

$$\boldsymbol{x}^{(t+1,0)} = \boldsymbol{x}^{(t,0)} + \sum_{i=1}^{m} p_i \Delta_i^{(t)}$$

Optimizes $\widetilde{F}(\boldsymbol{x}) = \sum_{i=1}^{m} \frac{p_i \tau_i}{\sum_{i=1}^{m} p_i \tau_i} F_i(\boldsymbol{x})$

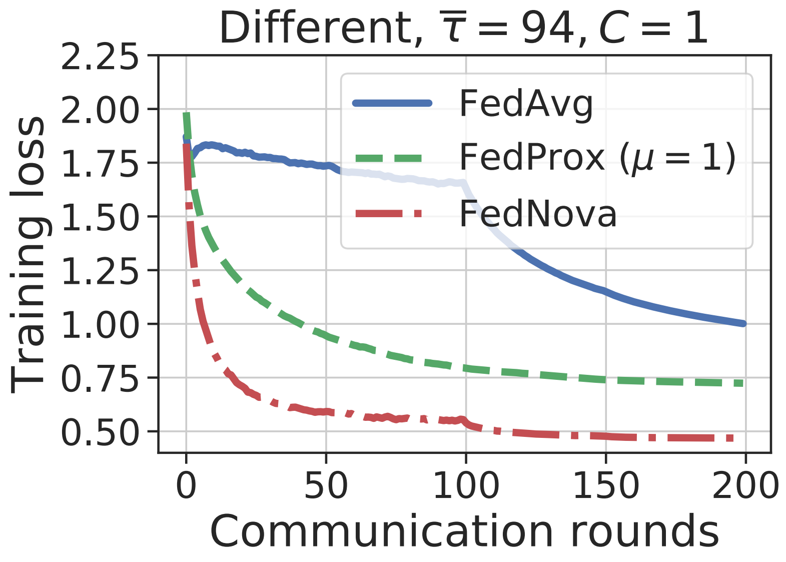
FedNova Update Rule

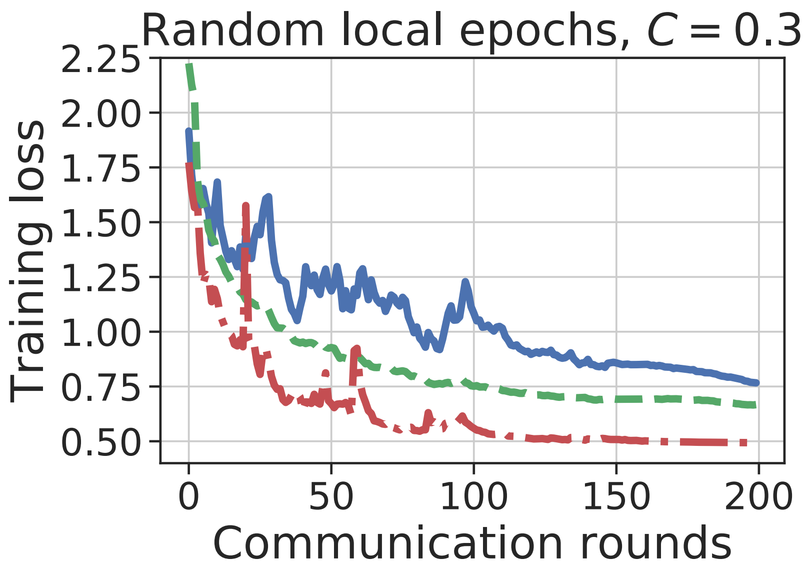
$$\boldsymbol{x}^{(t+1,0)} = \boldsymbol{x}^{(t,0)} + \tau_{\text{eff}} \sum_{i=1}^{m} p_i \frac{\Delta_i^{(t)}}{\tau_i}$$

$$\tau_{\text{eff}} = \bar{\tau} = \sum_{i=1}^{m} p_i \tau_i$$

Optimizes $\widetilde{F}(\boldsymbol{x}) = \sum_{i=1}^{m} p_i F_i(\boldsymbol{x})$

3. Computational Heterogeneity





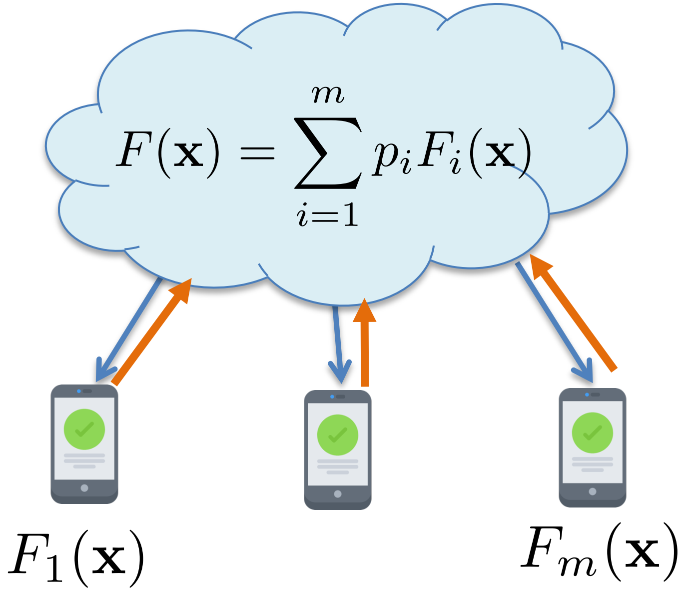
Local Epochs	Client Opt.	Test Accuracy % FedAvg	Test Accuracy % FedNova
$E_i = 2$ $(16 \leq \tau_i \leq 408)$	Vanilla	60.68±1.05	**66.31**±0.86
	Momentum	65.26±2.42	**73.32**±0.29
	Proximal [38]	60.44±1.21	**69.92**±0.34
$E_i^{(t)} \sim \mathcal{U}(2,5)$ $(16 \leq \tau_i^{(t)} \leq 1020)$	Vanilla	64.22±1.06	**73.22**±0.32
	Momentum	70.44±2.99	**77.07**±0.12
	Proximal [38]	63.74±1.44	**73.41**±0.45
	VR [20]	74.72±0.34	**74.72**±0.19
	Momen.+VR	Not Defined	**79.19**±0.17

FedNova works with various local solvers: vanilla SGD, proximal SGD, SCAFFOLD/VRL, Momentum etc.

We extend this to local adaptive optimizers in [Wang et al 2021]

Summary and Key Takeaways

1. Data Heterogeneity
2. Communication Heterogeneity
3. Computational Heterogeneity



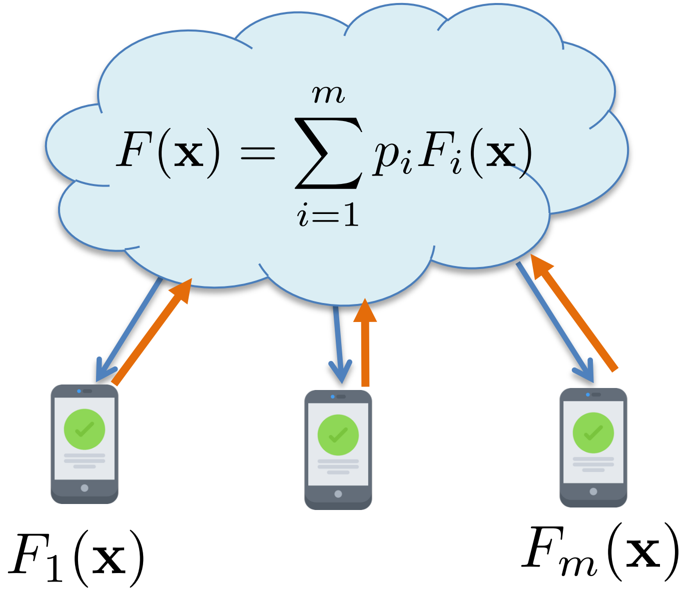
- Allowing heterogeneity makes the system more scalable and flexible
- Heterogeneity-aware algorithms can ensure fast convergence in the presence of heterogeneity

GOAL: Find ***x*** that minimizes the global objective

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

Some Ongoing and Future Directions

- Allowing Model Heterogeneity [Lin 2022, Cho 2022]
- Concept Drift at Clients [Jothimurugesan et al 2022]
- Personalized Federated Learning [Li 2021, Cho 2022]
- Incentivizing Clients to Participate [Cho et al 2022]



GOAL: Find ***x*** that minimizes the global objective

Modify this objective

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i F_i(\mathbf{x})$$

arXiv links to our papers

On the Convergence of Federated Learning with Cyclic Client Participation
https://arxiv.org/pdf/2302.03109, ICML 2023
Y-J Cho, P. Sharma, G. Joshi, Z. Xu, S. Kale, T. Zhang

FedExP: Speeding Up Federated Averaging via Extrapolation
https://arxiv.org/pdf/2301.09604, ICLR 2023
D. Jhunjhunwala, S. Wang, G. Joshi

FedVARP: Tackling the Variance Due to Partial Client Participation in Federated Learning
https://arxiv.org/abs/2010.01243, UAI 2022,
D. Jhunjhunwala, P. Sharma, A. Nagarkatti, G. Joshi

Client Selection in Federated Learning: Convergence Analysis and Adaptive Strategies
https://arxiv.org/abs/2010.01243, AISTATS 2022
Y. Cho, J. Wang, G. Joshi

arXiv links to our papers

Tackling the Obj. Inconsistency Problem in Heterogeneous Federated Optimization

https://arxiv.org/abs/2007.07481, NeurIPS 2020
J. Wang, Qinghua Liu, Hao Liang, G. Joshi, H. Vincent Poor

Local Adaptivity in Federated Learning: Convergence and Consistency

https://arxiv.org/abs/2106.02305, preprint
J. Wang, Z. Xu, Z. Garrett, Z. Charles, L. Liu, G. Joshi

To Federate or Not To Federate: Incentivizing Client Participation in Federated Learning

https://arxiv.org/abs/2205.14840, preprint
Y. Cho, D. Jhunjhunwala, T. Li, V. Smith , G. Joshi

Federated Learning under Distributed Concept Drift

https://arxiv.org/abs/2206.00799, AISTATS 2023
E. Jothimurugesan, K. Hsieh, J Wang, G. Joshi, P. Gibbons