Problem-Parameter-Free Federated Learning

Wenjing Yan
The Chinese University of Hong Kong

(Joint work with Kai Zhang, Xiaolu Wang, Xuanyu Cao)

ICLR 2025

Outline

1 Background and Motivation

2. Algorithm Design

3. Simulation Results

Outline

1 Background and Motivation

2. Algorithm Design

3. Simulation Results

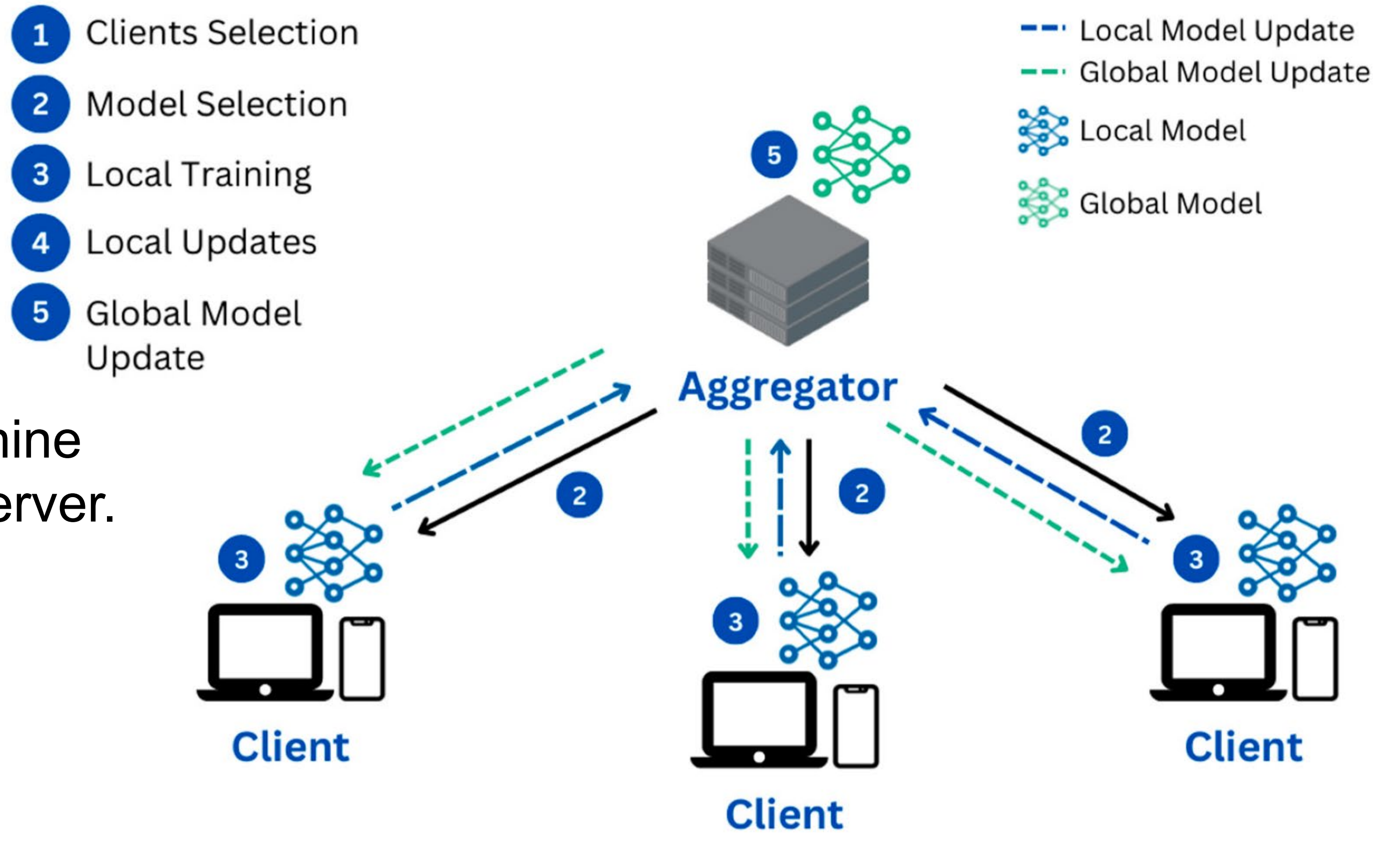
Problem Formulation of Federated Learning (FL)

- Federated Learning:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad f(\boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

$$\text{where} \quad f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i \sim \mathcal{D}_i} \left[F\left(\boldsymbol{\theta}; \boldsymbol{\xi}_i\right) \right]$$

- ➢ Multiple clients collaboratively train a machine learning model with the help of a central server.
- ➢ Each client performs multiple local update based on private data
- ➢ Server aggregates the global model



Advantages:

- **Ensures privacy by avoiding raw data sharing**
- **Offers scalability and communication efficiency**

- Figure from https://encyclopedia.pub/entry/48625

Problem Formulation of Federated Learning (FL)

- Federated Learning:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad f(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

$$\text{where} \quad f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}\left[F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)\right]$$

FedAvg Key Steps:

➢ K steps local update at client i:

$$\boldsymbol{g}_i^{t,k} = \nabla F\left(\boldsymbol{\theta}_i^{t,k};\boldsymbol{\xi}_i^{t,k}\right)$$

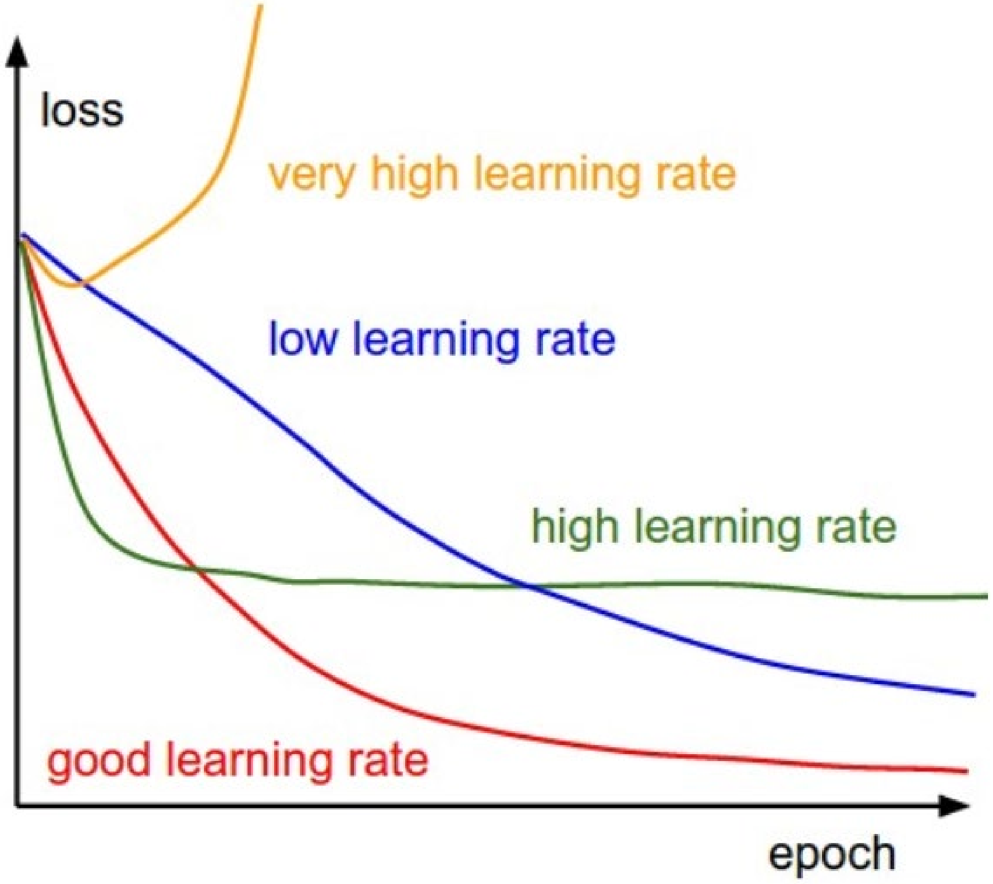
$$\boldsymbol{\theta}_i^{t,k+1} = \boldsymbol{\theta}_i^{t,k} - \eta_t \boldsymbol{g}_i^{t,k}$$

➢ Global model aggregation at server:

$$\boldsymbol{g}^t = \frac{1}{NK}\sum_{i=1}^{N}(\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,K})$$

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \gamma \boldsymbol{g}^t$$

Stepsize setting is crucial



1

Background and Motivation

Challenge 1: Problem-Specific Hyperparameter Tuning

- Federated Learning:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad f(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

$$\text{where} \quad f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}\left[F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)\right]$$

- Problem-Specific Constants:
  - L: Smoothness constant
  - σ^2: Stochastic gradient variance
  - L, σ_h^2: coefficients on gradient dissimilarity bound
  - $\Delta := f(\boldsymbol{\theta}_0) - f^*$: Initial suboptimality gap

Assumptions

***L*-Smoothness**

$$\left\|\nabla F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right) - \nabla F\left(\boldsymbol{\delta};\boldsymbol{\xi}_i\right)\right\| \le L\|\boldsymbol{\theta}-\boldsymbol{\delta}\|$$

Stochastic Gradient Variance

$$\mathbb{E}_{\boldsymbol{\xi}_i}\left\|\nabla F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right) - \nabla f_i(\boldsymbol{\theta})\right\|^2 \le \sigma^2$$

Bounded Gradient Dissimilarity

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\boldsymbol{\theta})\right\|^2 \le B\left\|\nabla f(\boldsymbol{\theta})\right\|^2 + \sigma_h^2$$

1 Background and Motivation

Challenge 1: Problem-Specific Hyperparameter Tuning

- Federated Learning:

$$\min_{\theta\in\mathbb{R}^d} f(\theta) := \frac{1}{N}\sum_{i=1}^{N} f_i(\theta)$$

where $f_i(\theta) := \mathbb{E}_{\xi_i\sim D_i}[F(\theta;\xi_i)]$

- Problem-Specific Constants:
  - ➢ L: Smoothness constant
  - ➢ σ^2: Stochastic gradient variance
  - ➢ L, σ_h^2: coefficients on gradient dissimilarity bound
  - ➢ $\Delta := f(\theta_0) - f^*$: Initial suboptimality gap

Assumptions

L-Smoothness

$$\|\nabla F(\theta;\xi_i) - \nabla F(\delta;\xi_i)\| \le L\|\theta-\delta\|$$

Stochastic Gradient Variance

$$\mathbb{E}_{\xi_i}\|\nabla F(\theta;\xi_i) - \nabla f_i(\theta)\|^2 \le \sigma^2$$

Bounded Gradient Dissimilarity

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\theta)\|^2 \le B\|\nabla f(\theta)\|^2 + \sigma_h^2$$

- **Algorithms require careful tuning of learning rates, momentum, and other coefficients**
- **Hyperparameter Tuning depends on problem-specific factors (e.g., smoothness constants L, gradient variances σ^2, etc.)**
- **Estimating those constants is difficult in FL due to privacy and data constraints.**
- **Problem-specific tuning limits applicability in dynamic environments (e.g., IoT, edge devices).**

1

Background and Motivation

Challenge 2: Requirement on Data Heterogeneity Bounds

- Federated Learning:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad f(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

$$\text{where} \quad f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}\left[F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)\right]$$

- Impact of Data Heterogeneity:
  - Clients often have non-IID
  - Leads to inconsistent local updates and "client drift"

Assumptions

L-Smoothness

$$\left\|\nabla F\left(\theta;\xi_i\right)-\nabla F\left(\delta;\xi_i\right)\right\|\leq L\|\theta-\delta\|$$

Stochastic Gradient Variance

$$\mathbb{E}_{\xi_i}\left\|\nabla F\left(\theta;\xi_i\right)-\nabla f_i(\theta)\right\|^2\leq\sigma^2$$

Bounded Gradient Dissimilarity

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\boldsymbol{\theta})\right\|^2\leq B\left\|\nabla f(\boldsymbol{\theta})\right\|^2+\sigma_h^2$$

Requirement on Data Heterogeneity Bounds

- Federated Learning:

$$\min_{\theta \in \mathbb{R}^d} \quad f(\theta) := \frac{1}{N} \sum_{i=1}^{N} f_i(\theta)$$

$$\text{where} \quad f_i(\theta) := \mathbb{E}_{\xi_i \sim D_i} [F(\theta; \xi_i)]$$

Assumptions

L-Smoothness

$$\|\nabla F(\theta;\xi_i) - \nabla F(\delta;\xi_i)\| \leq L\|\theta - \delta\|$$

- **Quantifying gradient dissimilarity bound is difficult in FL due to privacy and data constraints.**
- **Data Heterogeneity Bounds limit the applicability of FL dynamic environments with varying data distributions.**

- Impact of Data Heterogeneity:
  - Clients often have non-IID
  - Leads to inconsistent local updates and "client drift"

Bounded Gradient Dissimilarity

$$\frac{1}{N} \sum_{i=1}^{N} \|\nabla f_i(\theta)\|^2 \leq B \|\nabla f(\theta)\|^2 + \sigma_h^2$$

Our Method: Problem-Parameter-Free Federated Learning

- Federated Learning:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \quad f(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

$$\text{where} \quad f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i\sim\mathcal{D}_i}\left[F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)\right]$$

Independent of all problem-specific parameters, enabling **tuning-free**

Eliminating the requirement on data heterogeneity bounds

Assumptions

L-Smoothness

$$\left\|\nabla F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)-\nabla F\left(\boldsymbol{\delta};\boldsymbol{\xi}_i\right)\right\| \le L\|\boldsymbol{\theta}-\boldsymbol{\delta}\|$$

Stochastic Gradient Variance

$$\mathbb{E}_{\boldsymbol{\xi}_i}\left\|\nabla F\left(\boldsymbol{\theta};\boldsymbol{\xi}_i\right)-\nabla f_i(\boldsymbol{\theta})\right\|^2 \le \sigma^2$$

Bounded Gradient Dissimilarity

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\boldsymbol{\theta})\right\|^2 \le B\left\|\nabla f(\boldsymbol{\theta})\right\|^2+\sigma_h^2$$

Outline

1 Background and Motivation

2. Algorithm Design

3. Simulation Results

2

Problem-Parameter-Free Federated Learning

Key Techniques: Normalized Gradient Descent

- Traditional gradient descent:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)$$

- The stepsize constraint is:

$$e.g.\ \eta \leq 1/L$$

The step size η must be small enough to account for the gradient's magnitude, which is governed by L. Large gradients can lead to overshooting and instability.

- Normalized gradient descent:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|}$$

The gradient is normalized, so the step size η no longer depends on how large the gradient is (irrespective of L). Consistent step sizes allows NGD to navigate flat regions, steep regions, and saddle points more effectively.

Steep region → NGD scales down the update step → prevents overshooting

Flat region → NGD amplifies the update step → avoids stagnation

Key Techniques: Normalized SGD

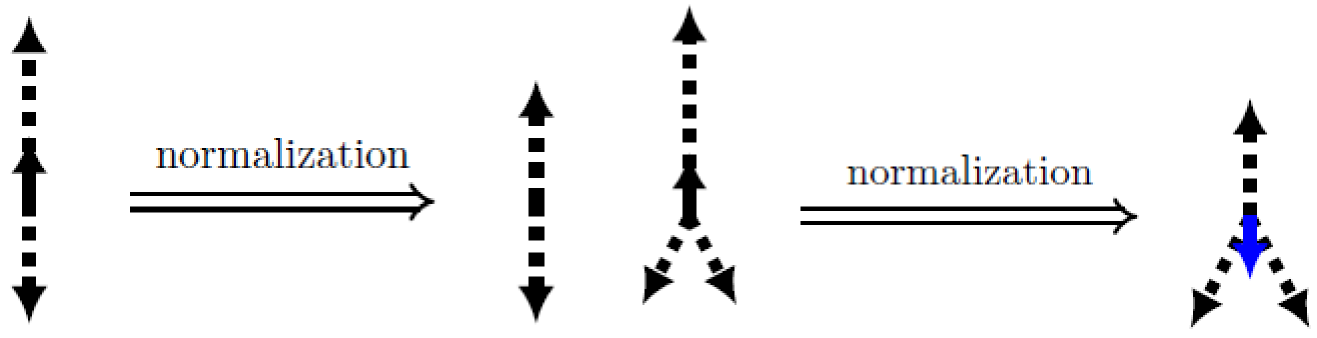
Normalized SGD: The stepsize is normalized by the current gradient.

- K steps local update at client i:

$$\boldsymbol{g}_i^{t,k} = \nabla F\left(\boldsymbol{\theta}_i^{t,k};\boldsymbol{\xi}_i^{t,k}\right)$$

$$\boldsymbol{\theta}_i^{t,k+1} = \boldsymbol{\theta}_i^{t,k} - \eta_t \frac{\boldsymbol{g}_i^{t,k}}{\|\boldsymbol{g}_i^{t,k}\|}$$

Problems for Stochastic Gradients *



(a) example 1 (b) example 2

Normalization losses magnitude information, inefficient for stochastic gradients

*Yang J, Li X, Fatkhullin I, et al. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods [J]. Advances in Neural Information Processing Systems, 2024, 36.

Key Techniques: Momentum

How to maintain the descent direction for normalized SGD: **Momentum**

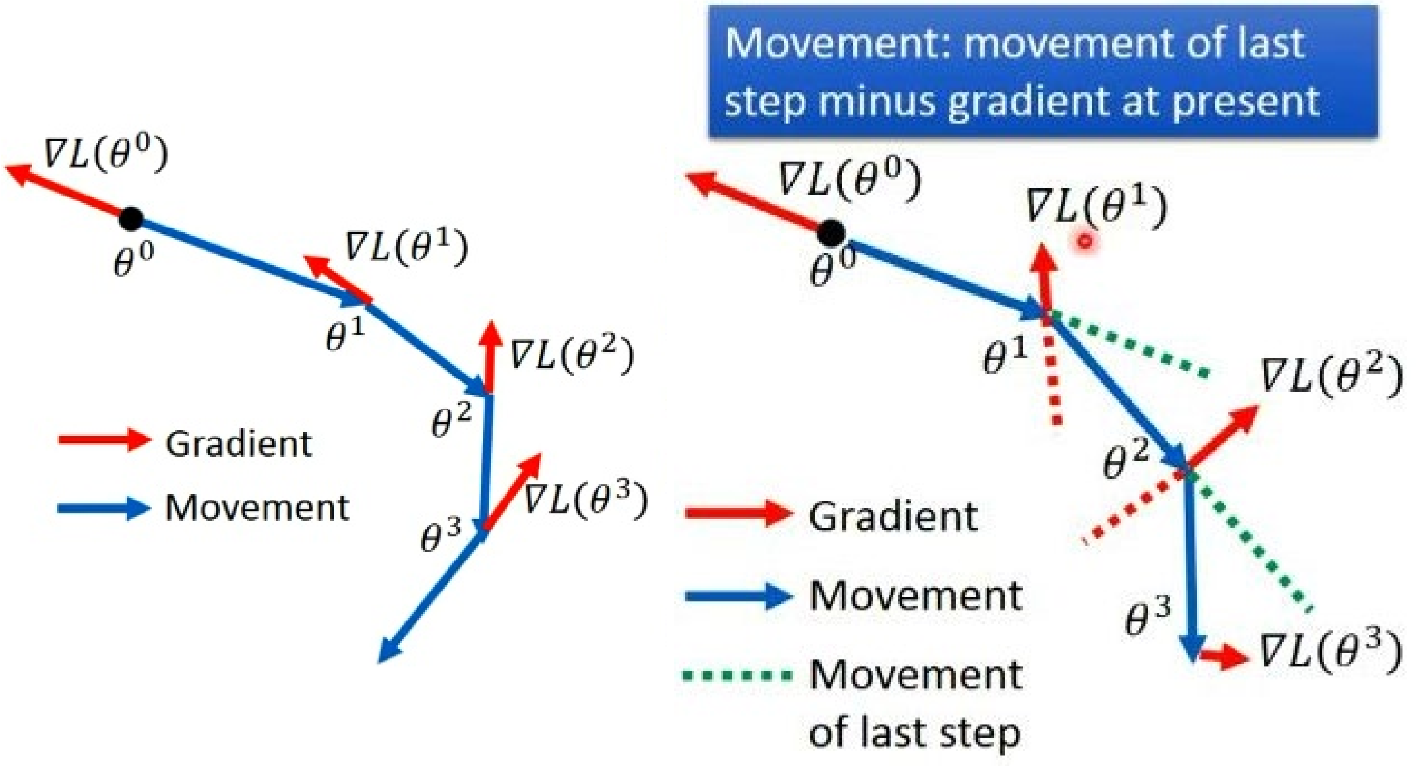
- Local update at client i:

$$\boldsymbol{g}_i^{t,k} = \beta \nabla F\left(\boldsymbol{\theta}_i^t; \boldsymbol{\xi}_i^t\right) + (1-\beta)\boldsymbol{g}^t$$

$$\boldsymbol{\theta}_i^{t,k+1} = \boldsymbol{\theta}_i^{t,k} - \eta_t \frac{\boldsymbol{g}_i^{t,k}}{\|\boldsymbol{g}_i^{t,k}\|}$$

- Global aggregation at server:

$$\boldsymbol{g}^t = \frac{1}{NK}\sum_{i,k} \boldsymbol{g}_{i,k}^t$$



Momentum: Accumulating past gradients across iterations and clients

Problem-Parameter-Free Federated Learning

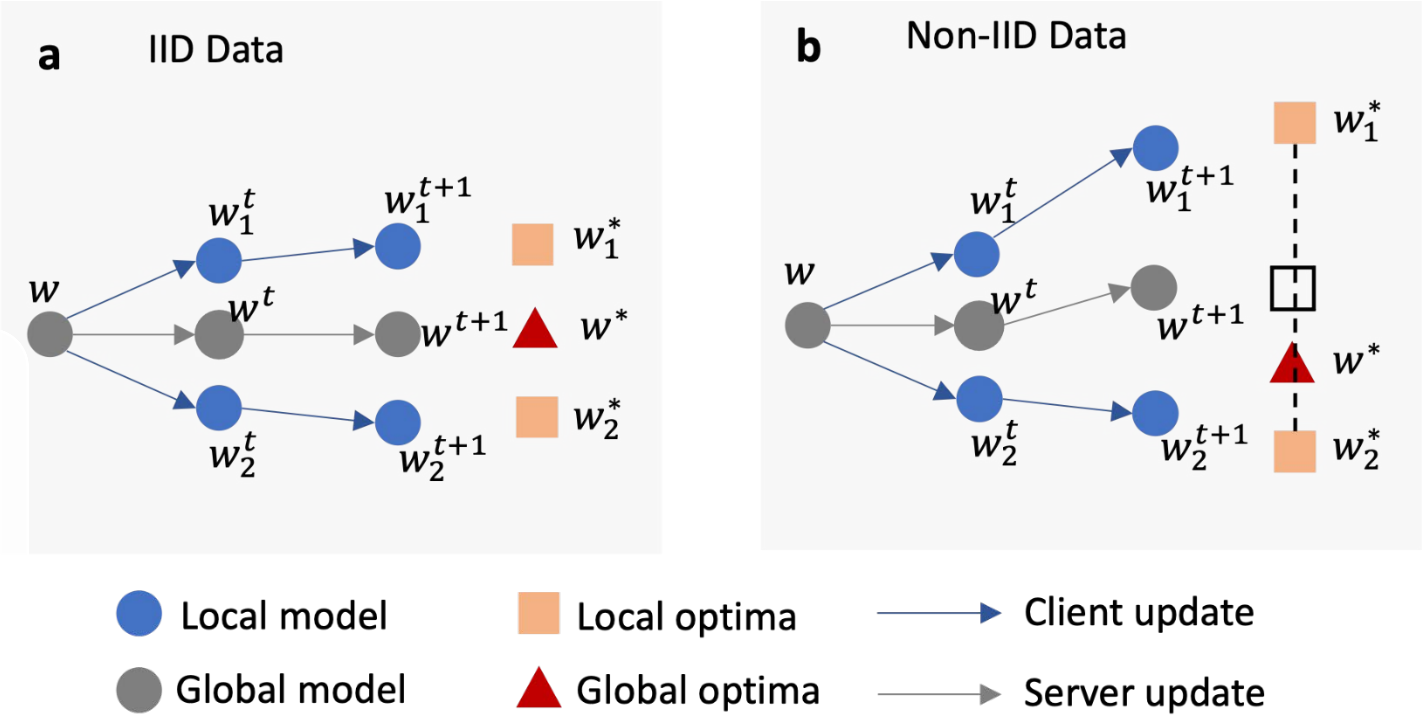
Key Techniques: Momentum

- Data Heterogeneity in FL:

$$f_i(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}_i \sim \mathcal{D}_i}\left[F\left(\boldsymbol{\theta}; \boldsymbol{\xi}_i\right)\right]$$

$$\mathcal{D}_i \neq \mathcal{D}_j \text{ for any } i \neq j$$

"Client drift": Local updates from individual clients diverge significantly from one another and from the global objective



- **Figure from https://ar5iv.labs.arxiv.org/html/2103.00710**

Key Techniques: Momentum

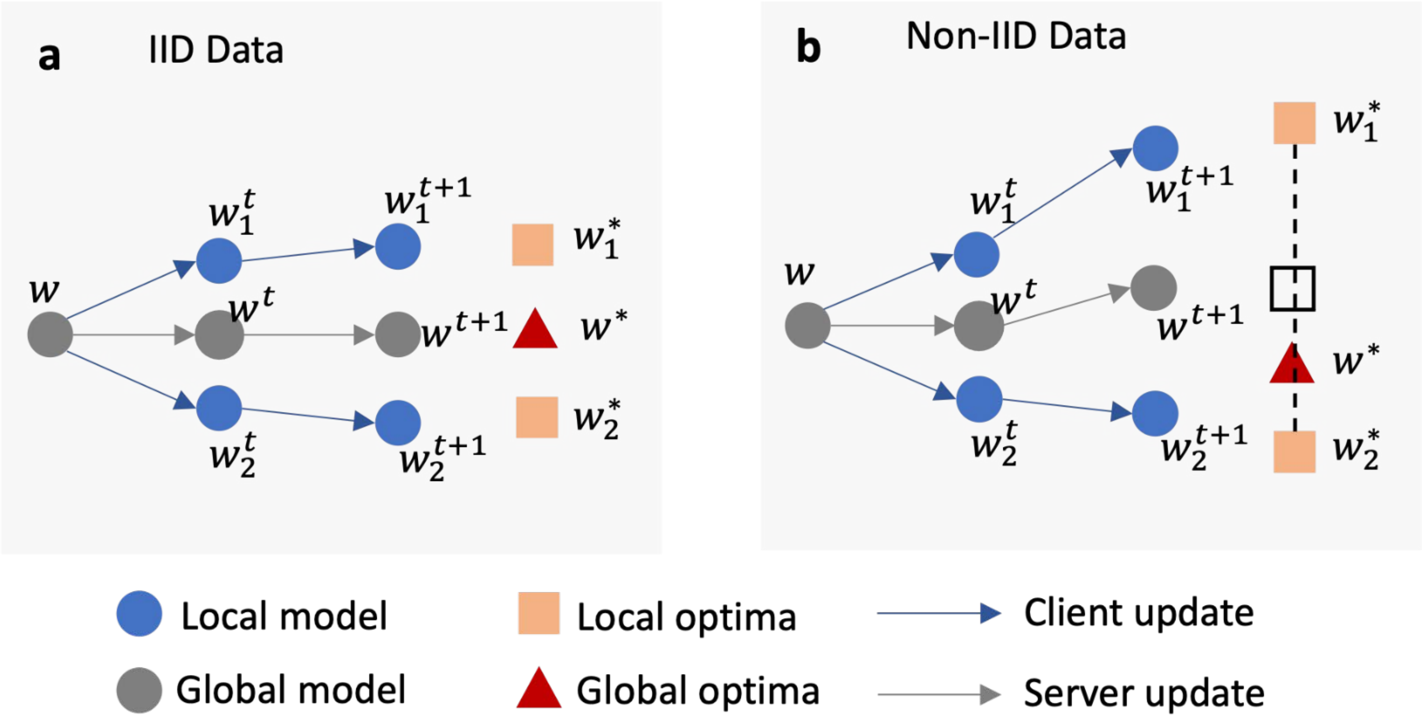
Momentum helps eliminating data heterogeneity bounds *

$$g_i^{t,k} = \beta \nabla F\left(\theta_i^t; \xi_i^t\right) + (1-\beta) g^t$$

$$\theta_i^{t,k+1} = \theta_i^{t,k} - \eta_t g_i^{t,k}$$

$$g^t = \frac{1}{NK} \sum_{i,k} g_{i,k}^t$$

"Client drift": Local updates from individual clients diverge significantly from one another and from the global objective



Momentum: Accumulating past gradients across iterations and clients

- * Cheng, Z., Huang, X., Wu, P., & Yuan, K. Momentum Benefits Non-iid Federated Learning Simply and Provably. In The Twelfth International Conference on Learning Representations (ICLR 2024).
- Figure from https://ar5iv.labs.arxiv.org/html/2103.00710

Key Techniques: Momentum

However…

$$\boldsymbol{g}_i^{t,k} = \beta \nabla F\left(\boldsymbol{\theta}_i^t; \boldsymbol{\xi}_i^t\right) + (1-\beta)\boldsymbol{g}^t$$

$$\boldsymbol{\theta}_i^{t,k+1} = \boldsymbol{\theta}_i^{t,k} - \eta_t \frac{\boldsymbol{g}_i^{t,k}}{\left\|\boldsymbol{g}_i^{t,k}\right\|}$$

$$\boldsymbol{g}^t = \frac{1}{NK}\sum_{i,k} \boldsymbol{g}_{i,k}^t$$

To achieve parameter-free, our derivative needs to handle the following unavoidable term:

$$\left\|\boldsymbol{g}_i^{t,k} - \boldsymbol{g}^t\right\|, \forall i, k$$

$$= \beta \mathbb{E}\left\|\nabla F\left(\boldsymbol{\theta}_i^{t,k}; \boldsymbol{\xi}_i^{t,k}\right) - \frac{1}{NK}\sum_{i,k} \nabla F\left(\boldsymbol{\theta}_i^{t,k}; \boldsymbol{\xi}_i^{t,k}\right)\right\|$$

Direct reflection of gradient dissimilarity

Momentum is insufficient to eliminate gradient dissimilarity bounds for normalized SGD case to achieve parameter-parameter-free

Key Techniques: Control Variates

Add control variates* to control data heterogeneity

- Local update at client i:

$$\boldsymbol{g}_i^{t,k} = \beta\left(\nabla F\left(\boldsymbol{\theta}_i^{t,k}; \boldsymbol{\xi}_i^{t,k}\right) - \boldsymbol{c}_i^{t-1} + \boldsymbol{c}^{t-1}\right) + (1-\beta)\boldsymbol{g}^{t-1}$$

$$\boldsymbol{c}_i^t = \frac{1}{K}\sum_{k=0}^{K-1} \nabla F\left(\boldsymbol{\theta}_i^{t,k}; \boldsymbol{\xi}_i^{t,k}\right)$$

- Global aggregation at server:
  - Full participation: $\boldsymbol{c}^t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}_i^t$
  - Partial participation: $\boldsymbol{c}^t = \boldsymbol{c}^{t-1} + \frac{1}{N}\sum_{i\in\mathcal{S}_t} \left(\boldsymbol{c}_i^t - \boldsymbol{c}_i^{t-1}\right)$

* Karimireddy, Sai Praneeth, et al. "Scaffold: Stochastic controlled averaging for federated learning." International conference on machine learning. PMLR, 2020.

2

Problem-Parameter-Free Federated Learning

Algorithm 1 PAdaMFed: A Problem-Parameter-Agnostic Algorithm for Nonconvex FL

Require: initial model $\boldsymbol{\theta}^0$, control variates $\boldsymbol{c}_i^{-1} = \frac{1}{K}\sum_{k=0}^{K-1} \nabla F\left(\theta^0; \xi_i^{-1,k}\right)$ for any i, $\boldsymbol{c}^{-1} = \frac{1}{N}\sum_i \boldsymbol{c}_i^{-1}$, momentum $\boldsymbol{g}^{-1} = \boldsymbol{c}^{-1}$, global learning rate γ, local learning rate η, and momentum parameter β

for $t = 0, \cdots, T-1$ **do**

  Central Server: Uniformly sample clients $\mathcal{S}_t \subseteq \{1, \cdots, N\}$ with $|\mathcal{S}_t| = S$

  for each client $i \in \mathcal{S}_t$ in parallel **do**

    Initialize local model $\boldsymbol{\theta}_i^{t,0} = \boldsymbol{\theta}^t$

    for $k = 0, \cdots, K-1$ **do**

      Compute $\boldsymbol{g}_i^{t,k} = \beta\left(\nabla F\left(\boldsymbol{\theta}_i^{t,k}; \xi_i^{t,k}\right) - \boldsymbol{c}_i^{t-1} + \boldsymbol{c}^{t-1}\right) + (1-\beta)\boldsymbol{g}^{t-1}$

      Update local model $\boldsymbol{\theta}_i^{t,k+1} = \boldsymbol{\theta}_i^{t,k} - \eta\frac{\boldsymbol{g}_i^{t,k}}{\|\boldsymbol{g}_i^{t,k}\|}$

    end for

    Update control variate $\boldsymbol{c}_i^t = \frac{1}{K}\sum_{k=0}^{K-1} \nabla F\left(\boldsymbol{\theta}_i^{t,k}; \xi_i^{t,k}\right)$ (set $\boldsymbol{c}_i^t = \boldsymbol{c}_i^{t-1}$ for $i \notin \mathcal{S}_t$)

    Upload $\boldsymbol{\theta}_i^{t,K}$ and $\boldsymbol{c}_i^t$ to central server

  end for

  Central server:

  Aggregate local updates $\overline{\boldsymbol{g}}^t = \frac{1}{\eta S K}\sum_{i\in\mathcal{S}_t}\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,K}\right)$

  Update global model $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \gamma\overline{\boldsymbol{g}}^t$

  Aggregate control variate $\boldsymbol{c}^t = \boldsymbol{c}^{t-1} + \frac{1}{N}\sum_{i\in\mathcal{S}_t}\left(\boldsymbol{c}_i^t - \boldsymbol{c}_i^{t-1}\right)$

  Aggregate momentum $\boldsymbol{g}^t = \beta\left(\frac{1}{S}\sum_{i\in\mathcal{S}_t}\left(\boldsymbol{c}_i^t - \boldsymbol{c}_i^{t-1}\right) + \boldsymbol{c}^{t-1}\right) + (1-\beta)\boldsymbol{g}^{t-1}$

  Download $\boldsymbol{\theta}^{t+1}$, $\beta\boldsymbol{c}^t + (1-\beta)\boldsymbol{g}^t$ to all clients

end for

3

Theoretical Results

Convergence of Algorithm 1

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Let the local and global learning rates of PAdaMFed be* $\eta = \frac{1}{K\sqrt{T}}$ *and* $\gamma = \frac{(SK)^{1/4}}{T^{3/4}}$*, respectively, the momentum parameter be* $\beta = \sqrt{\frac{SK}{T}}$*, and* $\{\theta^t\}_{t\geq 0}$ *be the iterates generated by Algorithm* $\boxed{1}$*. Then, it holds for all* $T \geq 1$ *that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f\left(\theta^t\right)\right\| \leq \mathcal{O}\left(\frac{\Delta + L + \sigma + \sqrt{L\sigma}}{(SKT)^{\frac{1}{4}}} + \frac{\sqrt{SK}\sigma + L}{\sqrt{T}}\right),$$

where $\Delta := f\left(\theta^0\right) - \min_{\theta} f(\theta)$.

Remark 1

All hyperparameters (η: local learning rate, γ: global learning rate, and β: momentum parameter) in Algorithm 1 are explicated determined by system-predefined constants: S (the number of participation clients), K (local update times), T (iteration times)

3 Theoretical Results

Enhanced convergence by STORM* variance reduction

$$\boldsymbol{g}_i^{t,k} = \nabla F\left(\boldsymbol{\theta}_i^{t,k}; \boldsymbol{\xi}_i^{t,k}\right) + \beta\left(\boldsymbol{c}^{t-1} - \boldsymbol{c}_i^{t-1}\right) + (1-\beta)\left(\boldsymbol{g}^{t-1} - \nabla F\left(\boldsymbol{\theta}^{t-1}; \boldsymbol{\xi}_i^{t,k}\right)\right)$$

Convergence of Algorithm 1 with variance reduction

Theorem 2. *Let the local and global learning rates of PAdaMFed-VR be* $\eta = \frac{1}{KT}$ *and* $\gamma = \frac{(SK)^{1/3}}{T^{2/3}}$, *respectively, the momentum parameter be* $\beta = \frac{(SK)^{1/3}}{T^{2/3}}$, *and* $\{\boldsymbol{\theta}^t\}_{t\geq 0}$ *be the iterates generated by Algorithm* $\boxed{2}$. *Then, it holds for all* $T \geq 1$ *that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f\left(\boldsymbol{\theta}^t\right)\right\| \leq \mathcal{O}\left(\frac{\Delta + L + \sigma}{(SKT)^{\frac{1}{3}}} + \frac{(L+\sigma)(SK)^{\frac{1}{3}}}{T^{\frac{2}{3}}}\right).$$

* Cutkosky, Ashok, and Francesco Orabona. "Momentum-based variance reduction in non-convex sgd." Advances in neural information processing systems 32 (2019).

3 Comparisons with Prior Work

Algorithms	Add. Assump.	Stepsize Restrictions	Stepsize-Related Problem-Parameters	Communication Complexity
SCAFFOLD (Karimireddy et al., 2020b)	–	$\gamma = \sqrt{S}, \eta \leq \frac{1}{24\gamma KL}\left(\frac{S}{N}\right)^{\frac{2}{3}}$	L	$\mathcal{O}\left(\left(\frac{N}{S}\right)^{\frac{1}{3}}\frac{L}{K\epsilon^4}\right)$
MIme (Karimireddy et al., 2020a)	BDH, BHD	$\eta = \sqrt{\frac{\Delta S}{L\tilde{G}TK^2}}, \tilde{G} = \sigma_h^2 + \frac{\sigma^2}{K}$	$L, \Delta, \sigma^2, \sigma_h^2$	$\mathcal{O}\left(\frac{1}{SK\epsilon^4}\right)$
FedSPS Sohom Mukherjee (2024)	BDH	$\eta_i^{t,k} = \min\left\{\frac{F\left(\theta_i^{t,k};\xi_i^{t,k}\right)-\ell_i^*}{c\left\Vert\nabla F\left(\theta_i^{t,k};\xi_i^{t,k}\right)\right\Vert^2}, \eta_b\right\}$ [1] $\eta_b \leq \min\left\{\frac{1}{2cL}, \frac{1}{25LK}\right\}$	$L, \ell_i^*, \forall i$	$\mathcal{O}\left(\frac{1}{NK\epsilon^4}\right)$
SCAFFOLD-M (Cheng et al., 2024)	–	$\beta = \min\left\{1, \frac{S}{N^{\frac{2}{3}}}, \sqrt{\frac{L\Delta SK}{\sigma^2 T}}, \sqrt{\frac{L\Delta S^2}{G_0 N}}\right\}$ [2] $\gamma = \frac{\beta}{L}, \eta KL \lesssim \min\left\{\frac{1}{\sqrt{S}}, \frac{1}{\beta K^{\frac{1}{4}}}, \frac{\sqrt{S}}{N}\right\}$	L, Δ, σ^2, G_0	$\mathcal{O}\left(\frac{1}{SK\epsilon^4}\right)$
PAdaMFed (This paper)	–	$\beta = \sqrt{\frac{SK}{T}}, \gamma = \frac{(SK)^{\frac{1}{4}}}{T^{\frac{3}{4}}}, \eta = \frac{1}{K\sqrt{T}}$	–	$\mathcal{O}\left(\frac{1}{SK\epsilon^4}\right)$
Variance Reduction				
FAFED (Wu et al., 2023)	BDH, BG	$\eta_t \propto \frac{N^{\frac{2}{3}}}{Lt^{\frac{1}{3}}}, \beta_t \propto \eta_t^2$	L	$\mathcal{O}\left(\frac{1}{SK\epsilon^3}\right)$
SCAFFOLD-M-VR (Cheng et al., 2024)	–	$\beta = \min\left\{\frac{S}{N}, \left(\frac{KL\Delta}{\sigma^2 T}\right)^{\frac{2}{3}}, S^{\frac{1}{3}}\right\}$ $\gamma L = \min\left\{1, \sqrt{\beta S}\right\}$ $\eta KL \lesssim \min\left\{\sqrt{\frac{\beta}{S}}, \left(\frac{\beta}{SK}\right)^{\frac{1}{4}}\right\}$	L, Δ, σ^2	$\mathcal{O}\left(\frac{1}{S\sqrt{K}\epsilon^4}\right)$
PAdaMFed-VR (This paper)	–	$\beta = \frac{(SK)^{\frac{1}{3}}}{T^{\frac{2}{3}}}, \gamma = \frac{(SK)^{\frac{1}{3}}}{T^{\frac{2}{3}}}, \eta = \frac{1}{KT}$	–	$\mathcal{O}\left(\frac{1}{SK\epsilon^3}\right)$

Shorthand notation:

BDH = Bounded data heterogeneity

BG = Bounded gradient that $\|\nabla f_i(\boldsymbol{\theta})\| \leq G, \ \forall i, \boldsymbol{\theta}$

BHD = Bounded hessian dissimilarity that $\left\|\nabla^2 f_i(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta})\right\|^2 \leq \delta, \ \forall i, \boldsymbol{\theta}$

[1] $\ell_i^* \leq \inf_{\boldsymbol{\xi}_i \in \mathcal{D}_i, \boldsymbol{\theta}} F(\boldsymbol{\theta}; \boldsymbol{\xi}_i)$ for any i, and c is a constant to balance adaptivity and accuracy.

[2] $G_0 := \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\boldsymbol{\theta}^0)\right\|^2$.

A point $\boldsymbol{\theta}$ is said to be ϵ-stationary if $\|\nabla f(\boldsymbol{\theta})\| \leq \epsilon$. For any ϵ-stationary point defined using $\|\nabla f(\boldsymbol{\theta})\|^2$, we have guarantee for $\|\nabla f(\boldsymbol{\theta})\|$ by:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}^t)\| \\
&= \frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\sqrt{\|\nabla f(\boldsymbol{\theta}^t)\|^2} \\
&\leq \frac{1}{T}\sum_{t=1}^{T-1}\sqrt{\mathbb{E}\|\nabla f(\boldsymbol{\theta}^t)\|^2} \\
&\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}^t)\|^2}
\end{aligned}
$$

Outline

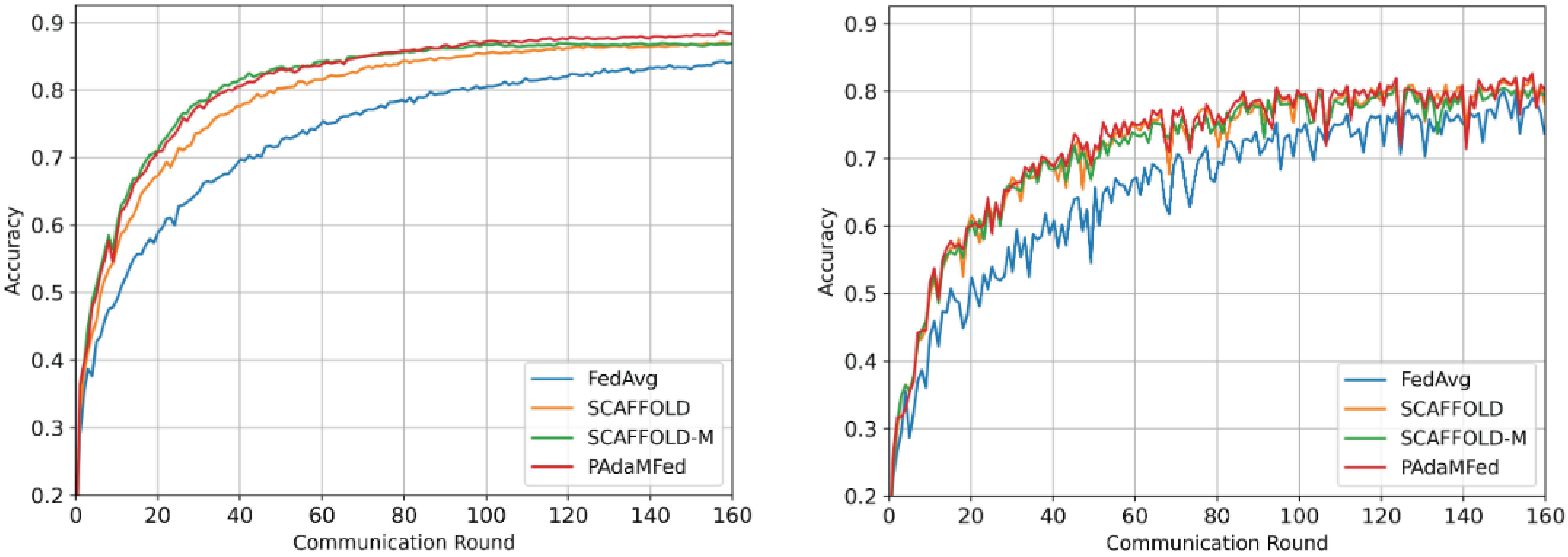
1 Background and Motivation

2. Algorithm Design

3. Simulation Results

3

Comparisons with Prior Work



(a) i.i.d

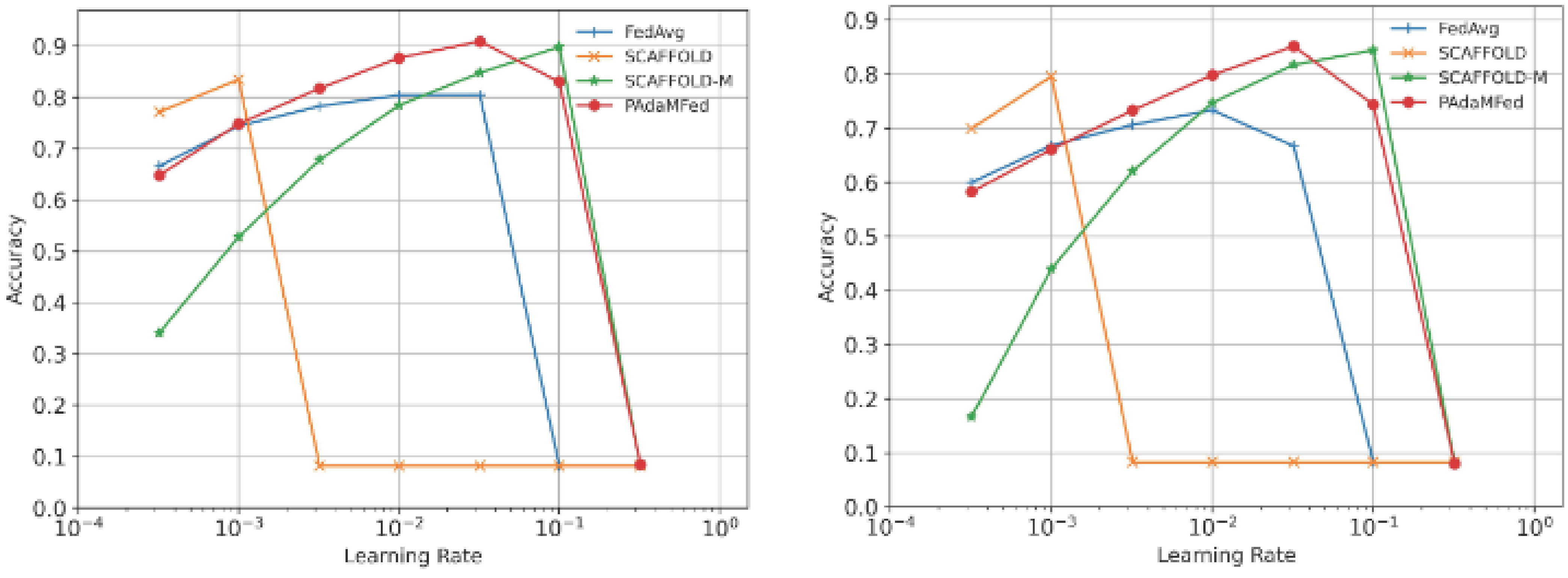
(b) non-i.i.d

Figure 1. Test accuracy versus the number of communication rounds on the CIFAR-10 dataset (ResNet 18, Dir(0.5)).

Our approach achieves start-of-the-art performance while eliminating the tedious stepsize tuning process

The stepsizes of our algorithm are set directly based on the guidance of Theorem 1. The stepsize of all baselines are perfectly tuned by grid search.

Comparisons with Prior Work



(a) i.i.d (b) non-i.i.d

Figure 1. Test accuracy versus learning rate on EMNIST dataset (RNN, 3 layer, Dir(1)).

Our algorithm demonstrates superior robustness to stepsize selection, maintaining stable performance across a significantly wider range of learning rates

THANKS

Thanks!