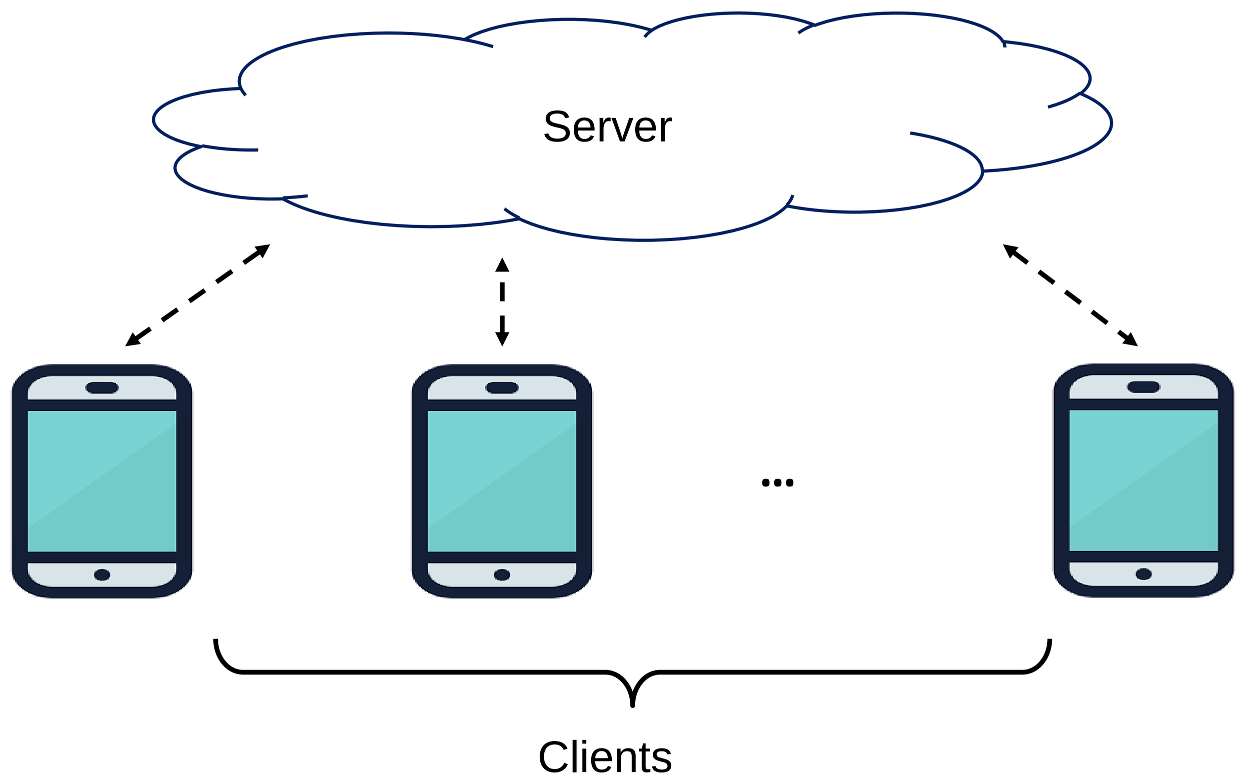
Linear Convergence in Federated Learning: Tackling Client Heterogeneity and Sparse Gradients

Aritra Mitra, Rayana Jaafar, George J. Pappas and Hamed Hassani
Department of Electrical and Systems Engineering, University of Pennsylvania
Email: {amitra20, rayanaj, pappasg, hassani}@seas.upenn.edu

NEURAL INFORMATION PROCESSING SYSTEMS

Federated Learning Setup



- Each client $i \in \mathcal{C}$ has a local loss function $f_i(x) = \mathbb{E}_{z\sim\mathcal{D}_i}[F_i(x;z)]$.
- **Goal:** Minimize

$$f(x) = \frac{1}{m}\sum_{i\in\mathcal{C}} f_i(x)$$

Q. How can we achieve linear convergence rates?

Q. How do such rates compare with a centralized baseline?

Challenges, Motivation, and Contributions

Challenges

- **Statistical Heterogeneity**: Clients' loss functions can have different minima.
- **Systems Heterogeneity:** Clients can have different operating speeds. Can lead to **stragglers** that slow down pace of computation.
- **Intermittent and Imprecise Communication:** Local steps & compression.

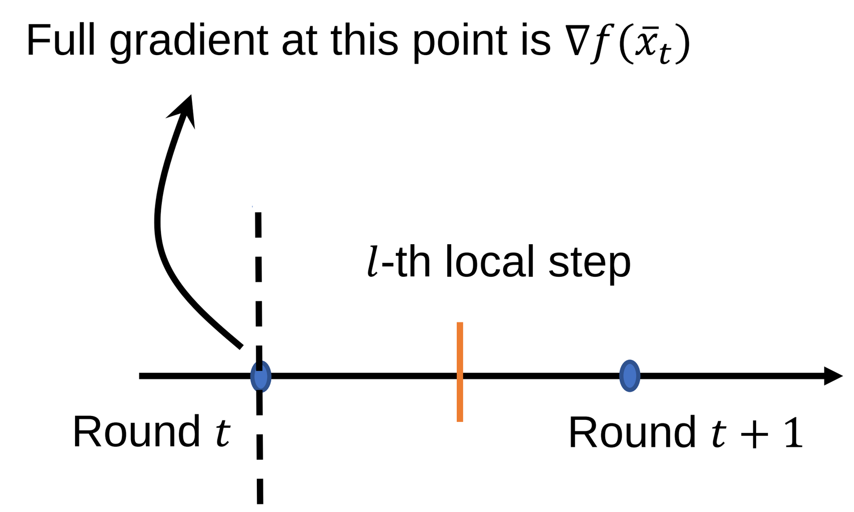
Existing Algorithms and Results

- **FedAvg**, **FedProx**, and **FedNova** all fail to guarantee linear convergence to the global minimum x^*. Suffer from **"speed-accuracy conflict"** .
- For **Scaffold** and **FedSplit**: No lower bounds; do not account for systems heterogeneity; no sparsification/compression.

Contributions

- Propose a new algorithmic framework called **FedLin** that tackles statistical+systems heterogeneity, and sparse gradients.
- Prove that **FedLin matches centralized rates** despite arbitrary heterogeneity.
- Provide the **first tight linear convergence rate analysis in FL**.
- Provide the **first analysis of biased gradient sparsification in FL**.
- **Key Takeaway:** Even mild statistical heterogeneity across clients' loss functions can hurt convergence rates.

Proposed Algorithm: FedLin



Idea: Use $\nabla f(\bar{x}_t)$ as guiding direction in round t

- FedLin **exploits memory** for objective heterogeneity, **client-specific learning rates** ($\eta_i \propto 1/\tau_i$) for systems heterogeneity, and **error-feedback** for gradient sparsification. The FedLin update rule is:

$$x^{(t)}_{i,\ell+1} \leftarrow x^{(t)}_{i,\ell} - \eta_i(\nabla f_i(x^{(t)}_{i,\ell}) - \nabla f_i(\bar{x}_t) + g_t)$$

- Here, g_t is an inexact version of $\nabla f(\bar{x}_t)$ due to sparsification.
- **Key Property: Global minimum is a fixed point of FedLin.**
- FedAvg, FedProx, FedNova, and Scaffold *do not* have this property.

Main Theoretical Results

- Suppose each $f_i(x)$ is L-smooth and μ- strongly convex.

Theorem I (Upper Bound for FedLin)

Suppose $\tau_i \geq 1,\ \eta_i = 1/(6L\tau_i)$. Then, after T comm. rounds, we have:

$$f(\bar{x}_{T+1}) - f(x^*) \leq \left(1 - \frac{1}{6\kappa}\right)^T (f(\bar{x}_1) - f(x^*))$$

Theorem II (Lower Bound for FedLin)

Suppose all clients perform H local steps. Given any $H \geq 2$, $\exists$ an initial condition $\bar{x}_1$, and an instance involving 2 clients, s.t. for FedLin,

$$f(\bar{x}_{T+1}) - f(x^*) \geq \exp(-4T)\,(f(\bar{x}_1) - f(x^*))$$

Main Takeaways:

- FedLin guarantees linear convergence to x^* despite **arbitrary** objective and systems heterogeneity.
- Convergence rate is **tight**. No benefits of performing multiple local steps. $\eta \propto 1/H$ is necessary.
- Lower bound holds even for **simple** instances: quadratic losses with **same minima!**

Additional Results and Simulations

Uplink/ Downlink	Error-Feedback	Convergence Rate ($\delta = d/k$ is the compression ratio)
Server	No	$f(\bar{x}_{T+1}) - f(x^*) \leq \left(1 - \frac{1}{2\delta_s(2+\sqrt{\delta_s})\kappa}\right)^T (f(\bar{x}_1) - f(x^*))$
Server	Yes	$f(\bar{x}_{T+1}) - f(x^*) \leq 2\kappa\left(1 - \frac{1}{96\delta_s\kappa}\right)^T (f(\bar{x}_1) - f(x^*))$
Clients	Yes	$f(\bar{x}_{T+1}) - f(x^*) \leq 2\kappa\left(1 - \frac{3}{4}\bar{\eta}\mu\right)^T (f(\bar{x}_1) - f(x^*)) + O(\bar{\eta}); \bar{\eta} \propto \frac{1}{\delta_c}$

Table: Summary of results for gradient sparsification

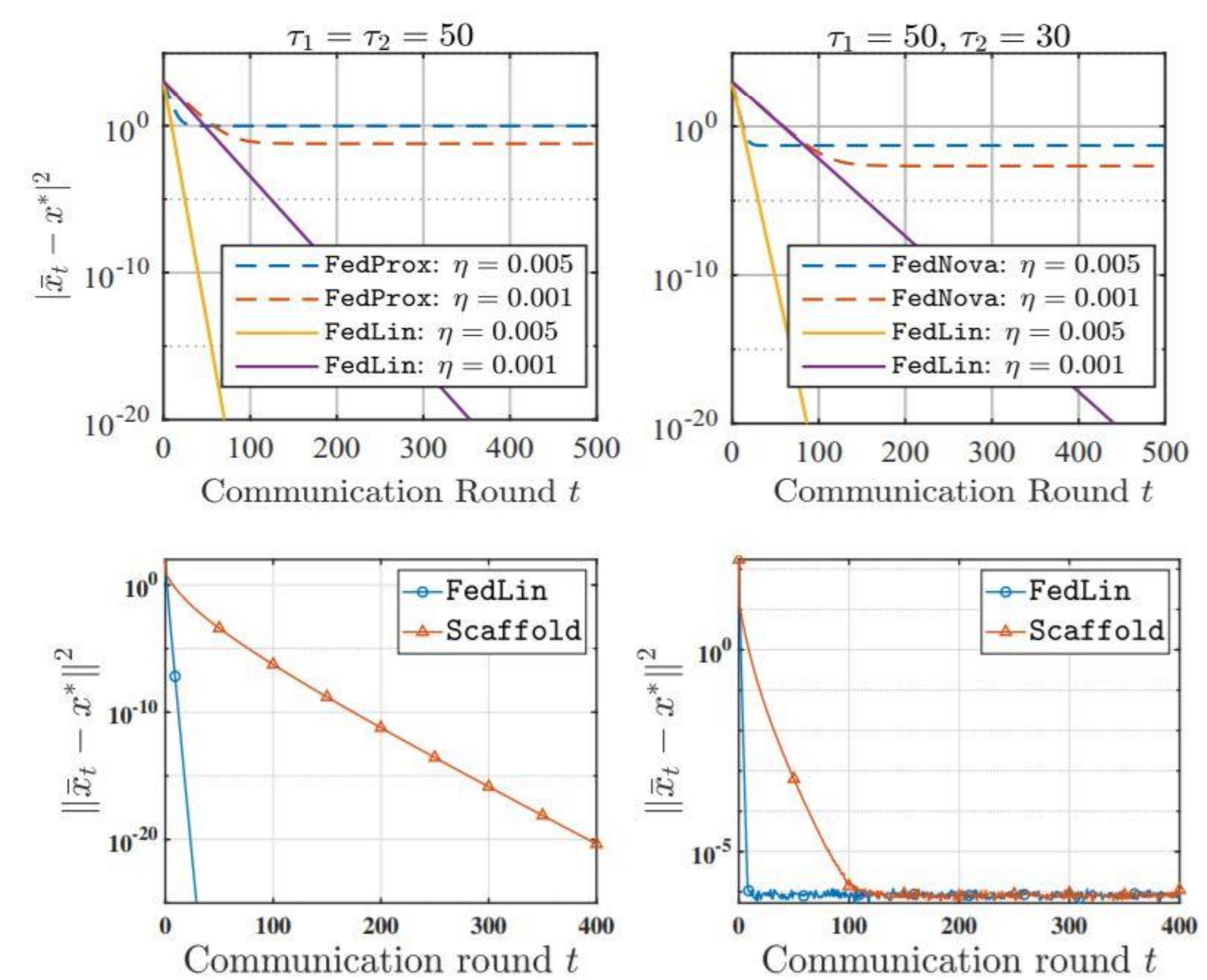


Figure: Simulations on a linear regression model

Algorithm	Linear Conv. to x^*	Lower Bounds	Systems Heterogeneity	Sparsification/ Compression
FedAvg	×	✓	×	×
FedProx	×	–	×	×
FedNova	×	–	✓	×
FedSplit	×	–	×	×
Scaffold	✓	–	×	×
FedLin	✓	✓	✓	✓

Table: Succinct comparison of FedLin with state-of-the-art FL algorithms

Penn
UNIVERSITY of PENNSYLVANIA

Funding: NSF Award 1837253, NSF CAREER award CIF 1943064, AFOSR-YIP award FA 9550-20-1-0111