Federated Optimization in Heterogeneous Networks

Tian Li (CMU), Anit Kumar Sahu (BCAI), Manzil Zaheer (Google Research), Maziar Sanjabi (Facebook AI), Ameet Talwalkar (CMU & Determined AI), Virginia Smith (CMU)

tianli@cmu.edu

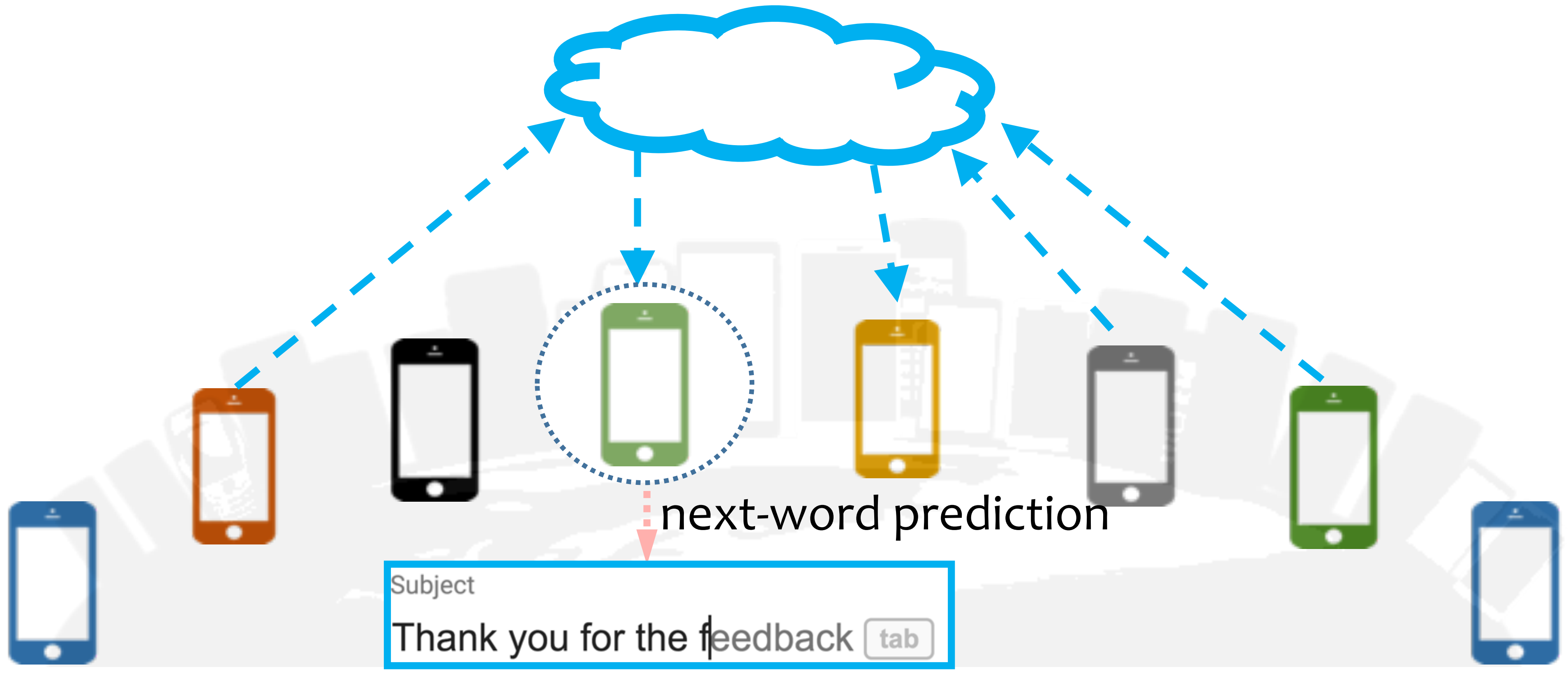
MLSys 2020

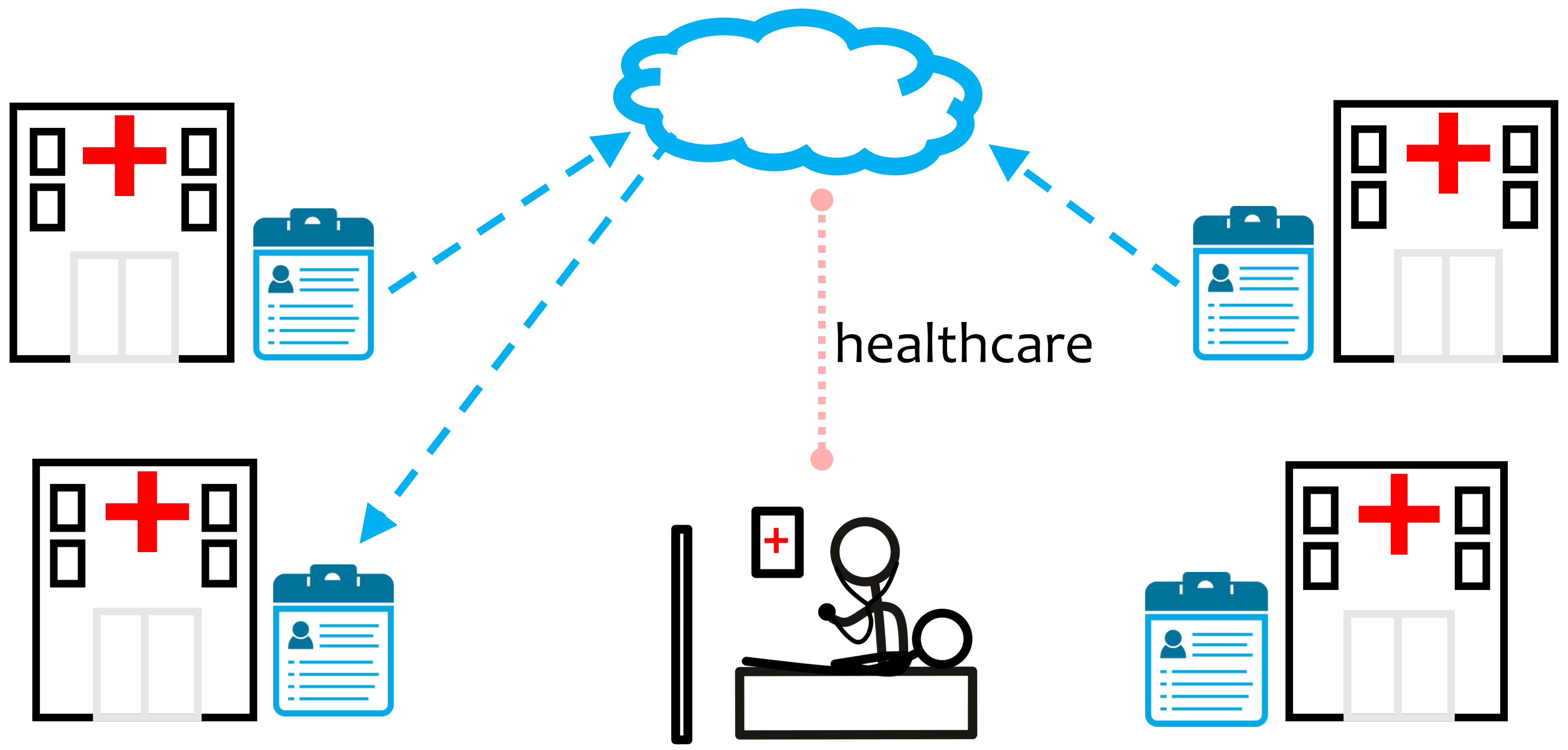
Federated Learning

Privacy-preserving training in heterogeneous, (potentially) massive networks

Networks of remote devices
e.g., cell phones

Networks of isolated organizations
e.g., hospitals





Example Applications

Voice recognition on mobile phones

Adapting to pedestrian behavior on autonomous vehicles

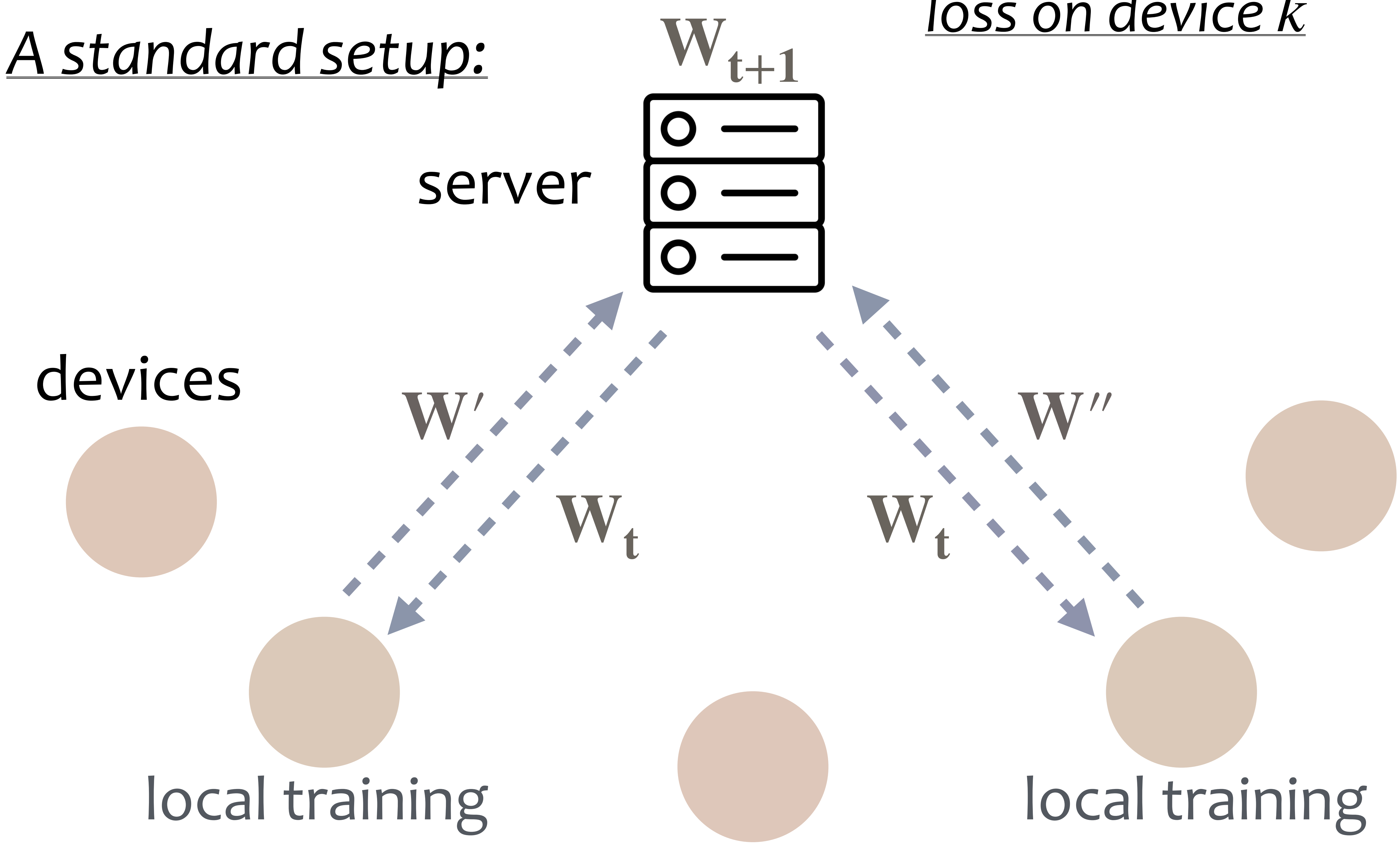
Personalized healthcare on wearable devices

Predictive maintenance for industrial machines

Workflow & Challenges

Objective: $$\min_{w} f(w) = \sum_{k=1}^{N} p_k \underbrace{F_k(w)}_{\text{loss on device } k}$$

A standard setup:



Systems heterogeneity
variable hardware, network connectivity, power, etc

Statistical heterogeneity
highly non-identically distributed data

Expensive communication
potentially massive network; wireless communication

Privacy concerns
privacy leakage through parameters

A Popular Method: Federated Averaging (FedAvg) [1]

At each communication roun

- Server randomly sele sends the current globa
- Each selected device k updat of SGD to optimize F_k & ends the new local model back
- Server aggregates local models to form a new global model w^{t+1}

What can go wrong?

vell in many settings ! (especially non-convex)

[1] McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." AISTATS, 2017.

What are the issues?

systems heterogeneity

stragglers

simply drop slo

statistical heterogeneity

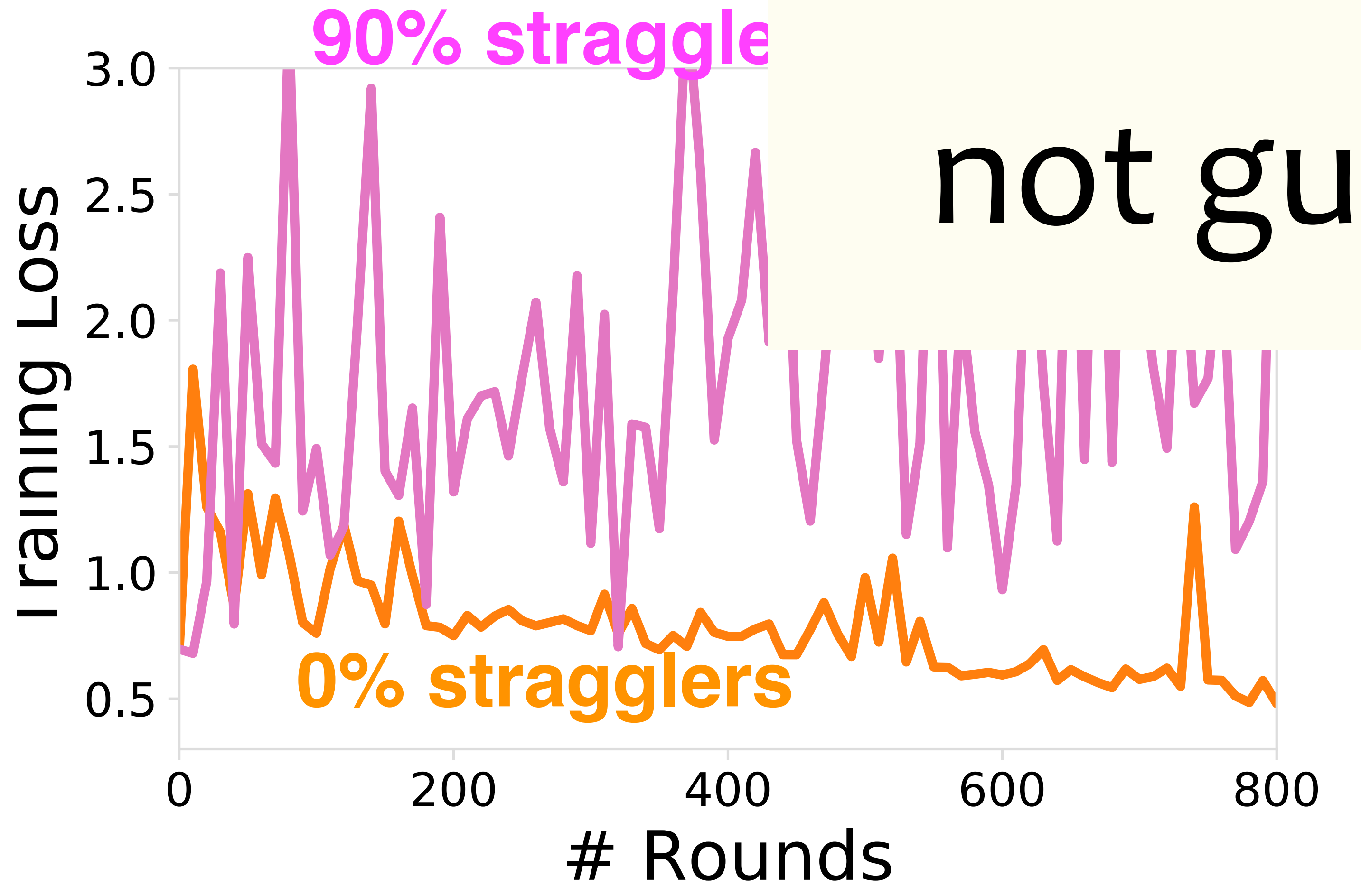
highly non-identically distributed data

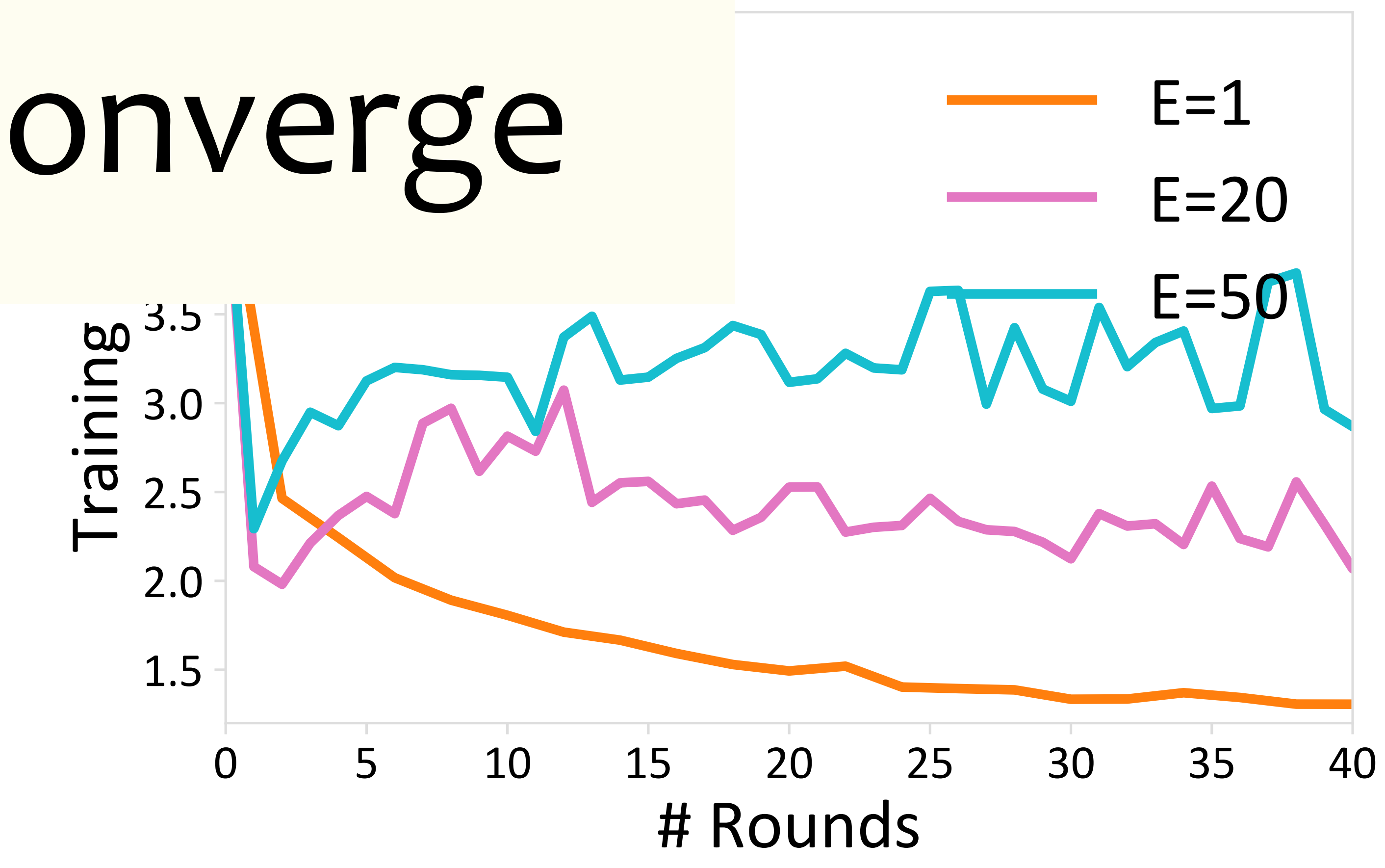
rage updates

FedAvg

heuristic method

not guaranteed to converge



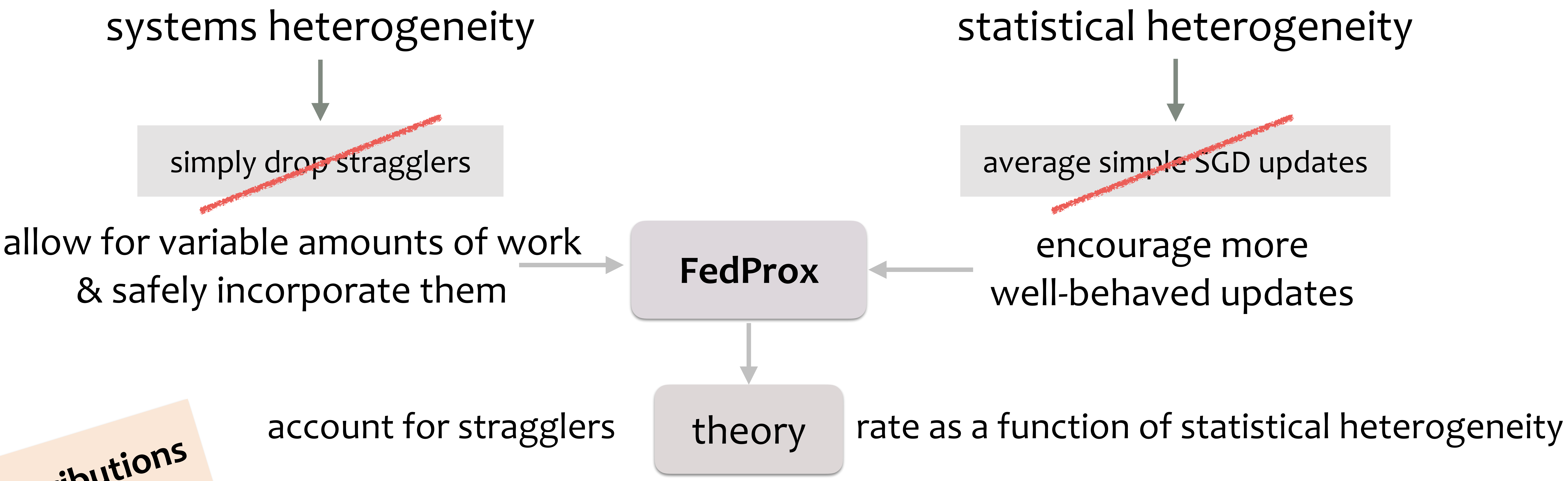


[2] Bonawitz, Keith, et al. "Towards Federated Learning at Scale: System Design." MLSys, 2019.

Outline

- Motivation
- **FedProx Method**
- Theoretical Analysis
- Experiments
- Future Work

FedProx — High Level



systems heterogeneity

~~simply drop stragglers~~

allow for variable amounts of work
& safely incorporate them

statistical heterogeneity

~~average simple SGD updates~~

encourage more
well-behaved updates

FedProx

account for stragglers

theory

rate as a function of statistical heterogeneity

Contributions

1. **convergence guarantees**
2. **more robust empirical performance**

for federated learning in heterogeneous networks

FedProx: A Framework For Federated Optimization

Objective:

$$\min_{w} f(w) = \sum_{k=1}^{N} p_k F_k(w)$$

At each communication round, local objective:

$$\min_{w_k} F_k(w_k)$$

Idea 1: Allow for variable amounts of work to be performed on local devices to handle stragglers

Idea 2: *Modified* Local Subproblem:

$$\min_{w_k} F_k(w_k) + \boxed{\frac{\mu}{2} \| w_k - w^t \|^2}$$

a proximal term

FedProx: A Framework For Federated Optimization

***Modified* Local Subproblem:** $\min_{w_k} F_k(w_k) + \frac{\mu}{2} \| w_k - w^t \|^2$

- The proximal term (1) safely incorporate noisy updates; (2) explicitly limits the impact of local updates
- Generalization of FedAvg
- Can use any local solver
- More robust and stable empirical performance
- Strong theoretical guarantees (with some assumptions)

Outline

- Motivation
- FedProx Method
- **Theoretical Analysis**
- Experiments
- Future Work

Convergence Analysis

Challenges: device subsampling, non-iid data, local updates

- High-level: **converges** despite these challenges
- Introduces notion of **B-dissimilarity** in to characterize statistical heterogeneity:

$$\mathbb{E}\left[\|\nabla F_k(w)\|^2\right] \leq \|\nabla f(w)\|^2 B^2$$

IID data: $B = 1$

non-IID data: $B > 1$

** used in other contexts, e.g., gradient diversity [3] to quantify the benefits of scaling distributed SGD*

[3] Yin, Dong, et al. "Gradient Diversity: a Key Ingredient for Scalable Distributed Learning." AISTATS, 2018.

Convergence Analysis

- Assumption 1: Dissimilarity is bounded
- Assumption 2: Modified local subproblem is convex & smooth
  - **Proximal term makes the method more amenable to theoretical analysis!**
- Assumption 3: Each local subproblem is solved to some accuracy
  - **Flexible communication/computation tradeoff**
  - **Account for partial work in the rates**

Convergence Analysis

[Theorem] Obtain suboptimality ε, after **T** rounds, with:

$$T = O\left(\frac{f(w^0) - f^*}{\rho\varepsilon}\right)$$

some constant, a function of $(B, \mu, \ldots)$

- **Rate is general:**
  - Covers both convex, and non-convex loss functions
  - Independent of the local solver; agnostic of the sampling method
- **The same asymptotic convergence guarantee as SGD**
  - Can converge much faster than distributed SGD in practice

Outline

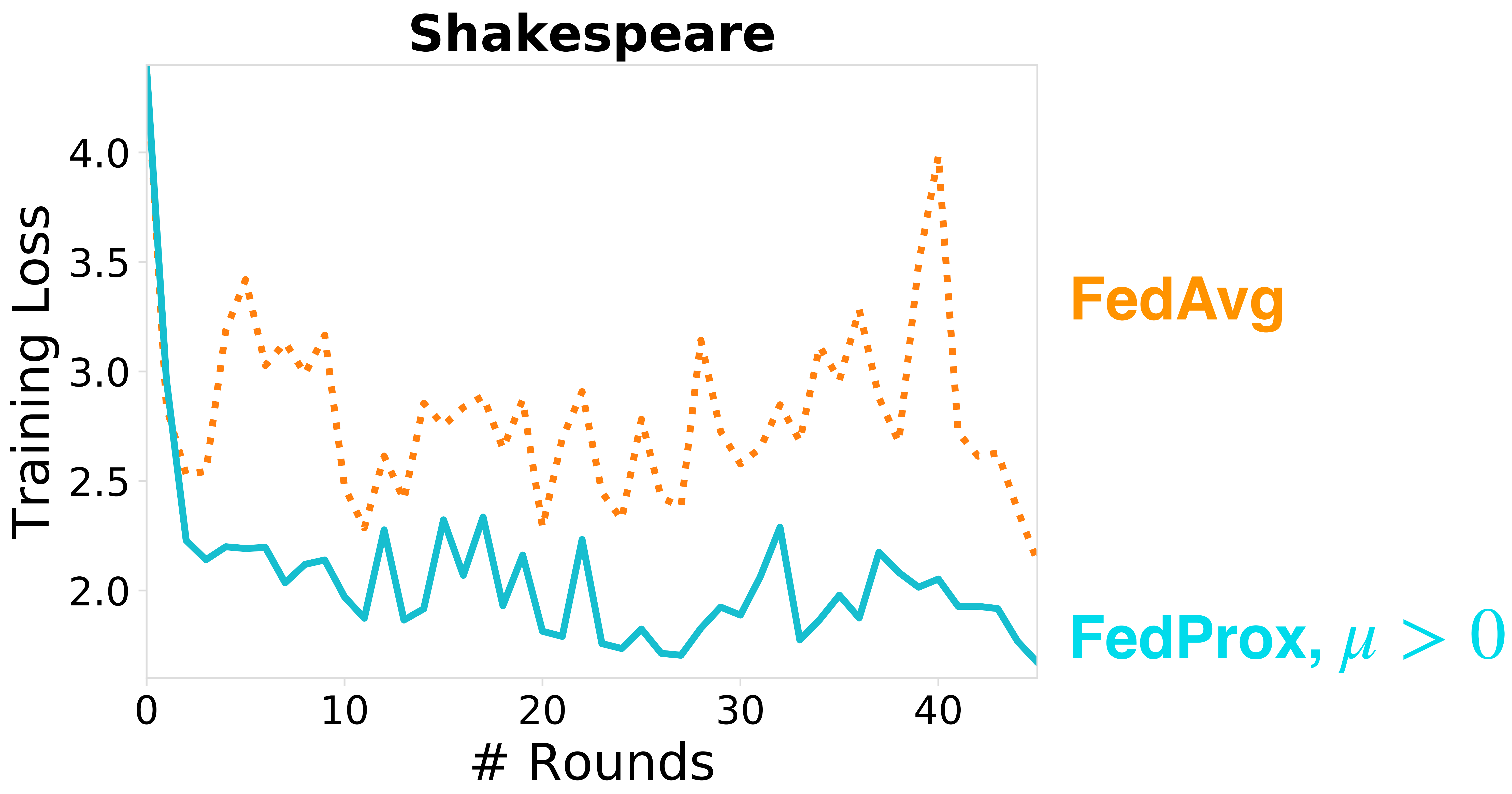
- Motivation
- FedProx Method
- Theoretical Analysis
- **Experiments**
- Future Work

Experiments

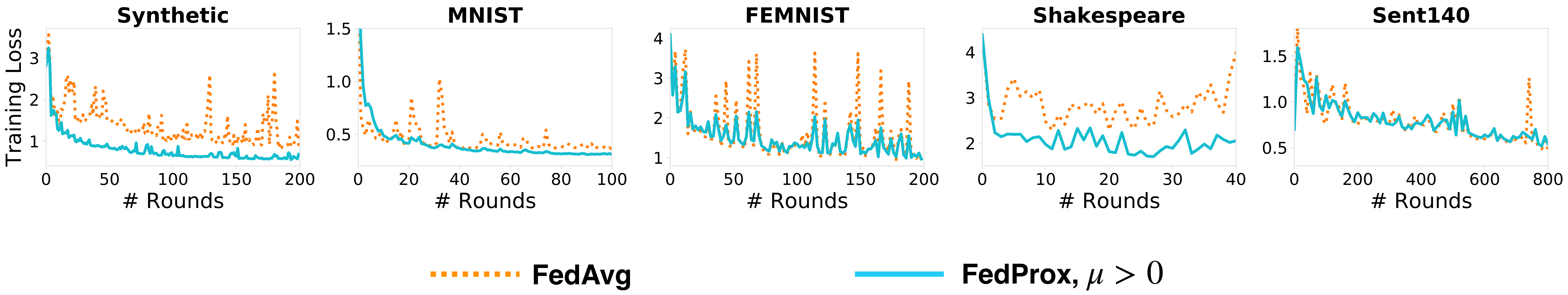
Zero Systems heterogeneity + Fixed Statistical heterogeneity

Benchmark:

LEAF *(leaf.cmu.edu)*



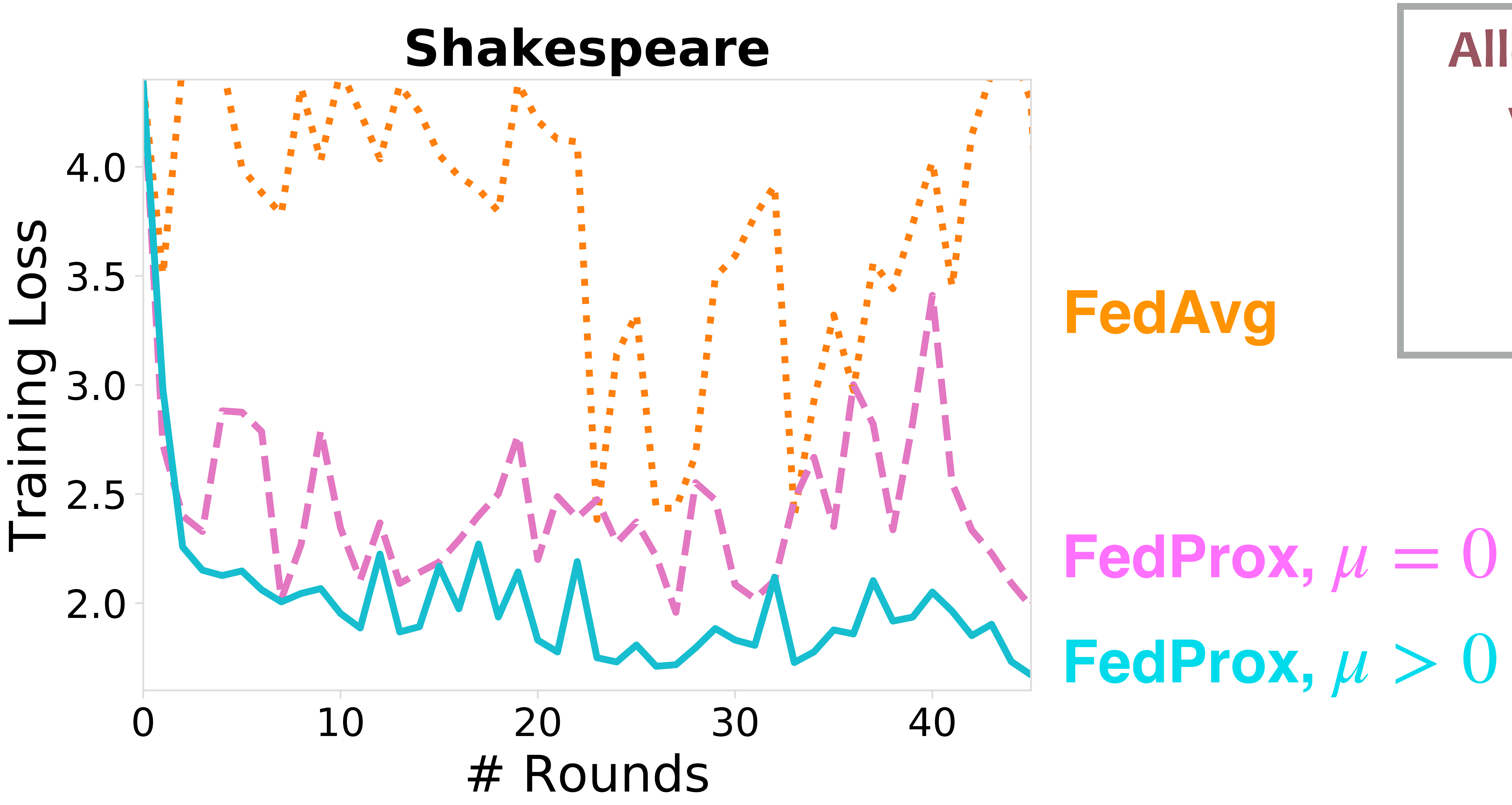
FedProx with $\mu > 0$ leads to more stable convergence under statistical heterogeneity



Similar benefits for all datasets

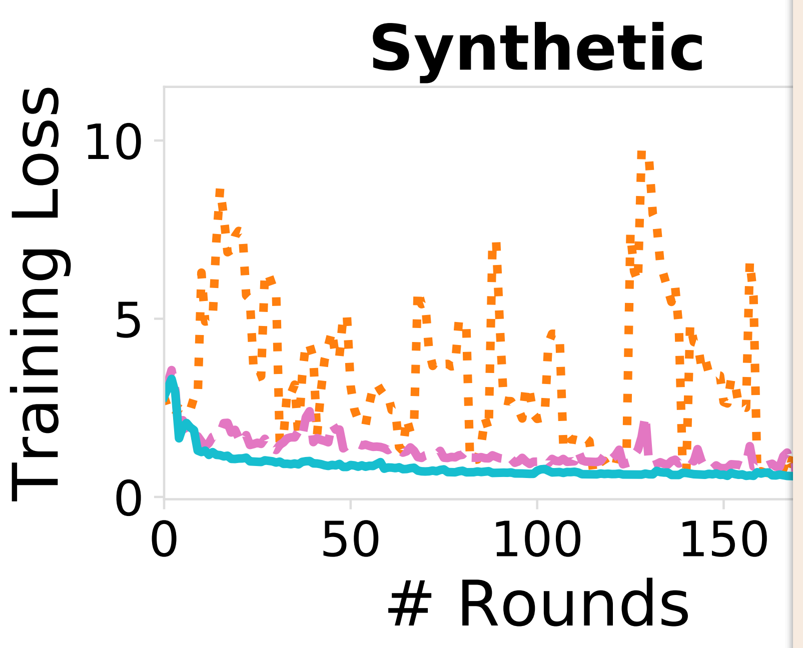
Experiments

High Systems heterogeneity + Fixed Statistical heterogeneity



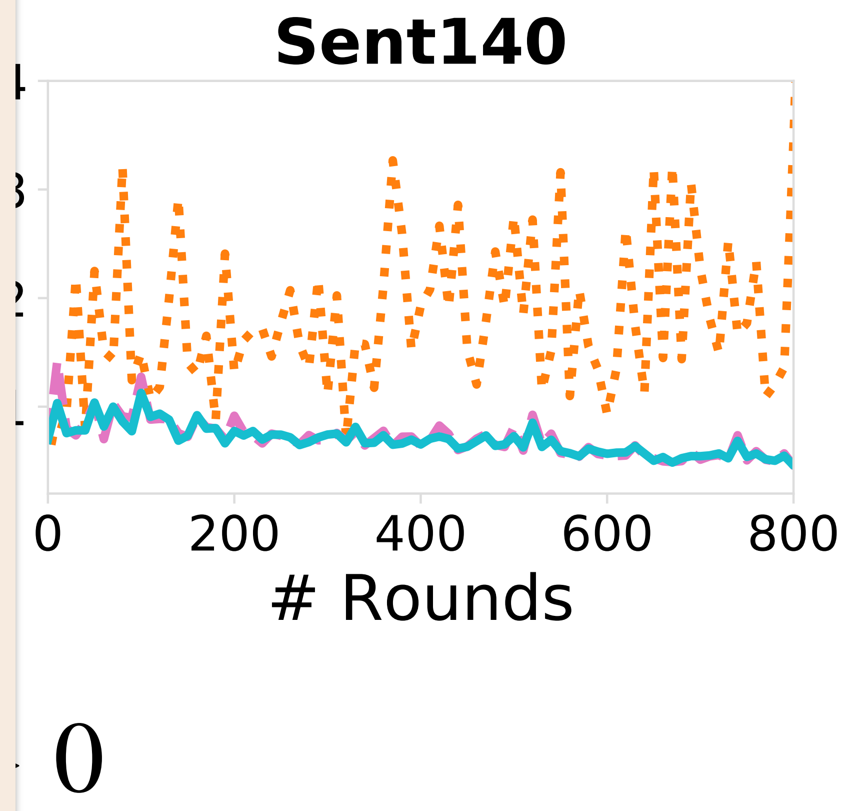
Allowing for variable amounts of work to be performed helps convergence in the presence of systems heterogeneity

FedProx with $\mu > 0$ leads to more stable convergence under statistical & systems heterogeneity



In terms of test accuracy:

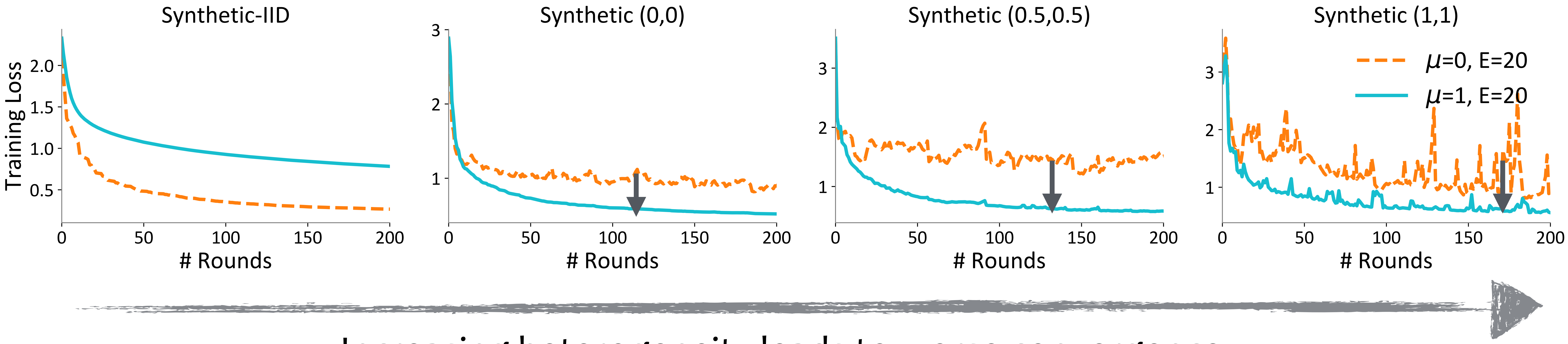
on average, 22% absolute accuracy improvement compared with FedAvg in highly heterogeneous settings



Similar benefits for all datasets

Experiments

Impact of Statistical Heterogeneity



Increasing heterogeneity leads to worse convergence

Setting μ > 0 can help to combat this

In addition, B-dissimilarity captures statistical heterogeneity (see paper)

Outline

- Motivation
- FedProx Method
- Theoretical Analysis
- Experiments
- **Future Work**

Future Work

Hyper-parameter tuning

- Set μ automatically

Diagnostics

- Determining heterogeneity a priori
- Leveraging the heterogeneity for improved performance

Privacy & security

- Better privacy metrics & mechanisms

Personalization

- Automatic fine-tuning

Productionizing

- Cold start problems

White paper: *Federated Learning: Challenges, Methods, and Future Directions*, *IEEE Signal Processing Magazine*, 2020. *(also on ArXiv)*

Thanks!

Poster: # 3, this room

On-device Intelligence Workshop, Wednesday, this room

Benchmark: leaf.cmu.edu

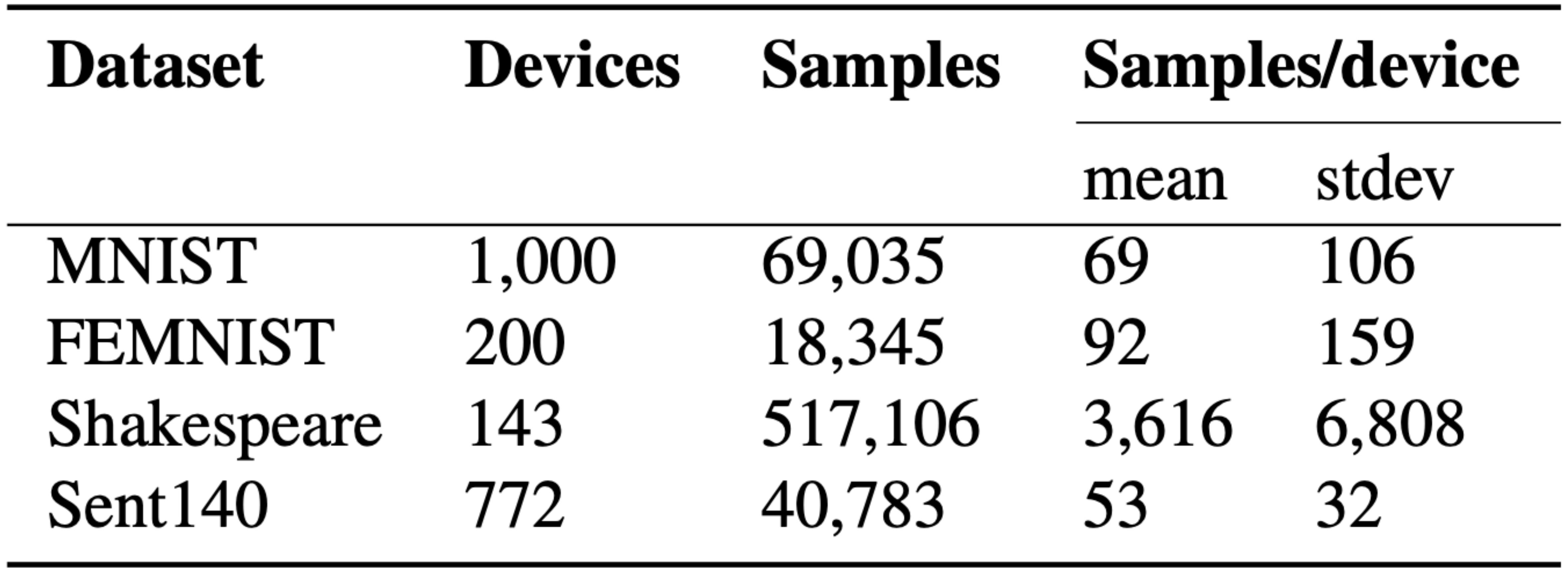
Paper & code: cs.cmu.edu/~litian/

Backup 1

- **Relations with previous works**
  - **proximal term**
    - Elastic SGD: employs a more complex moving average to update parameters; limited to SGD as a local solver; only been analyzed for quadratic problems
    - DANE and inexact DANE: adds an additional gradient correction term, assume full device participation (unrealistic); discouraging empirical performance
      - *FedDANE: A Federated Newton-Type Method, Arxiv.*
    - Other works: different purposes such as speeding up SGD on a single machine; different analysis assumptions (IID, solving subproblems exactly)
  - **B-dissimilarity term**
    - For other purposes, such as quantifying the benefit in scaling SGD for IID data

Backup 2

- **Data statistics**

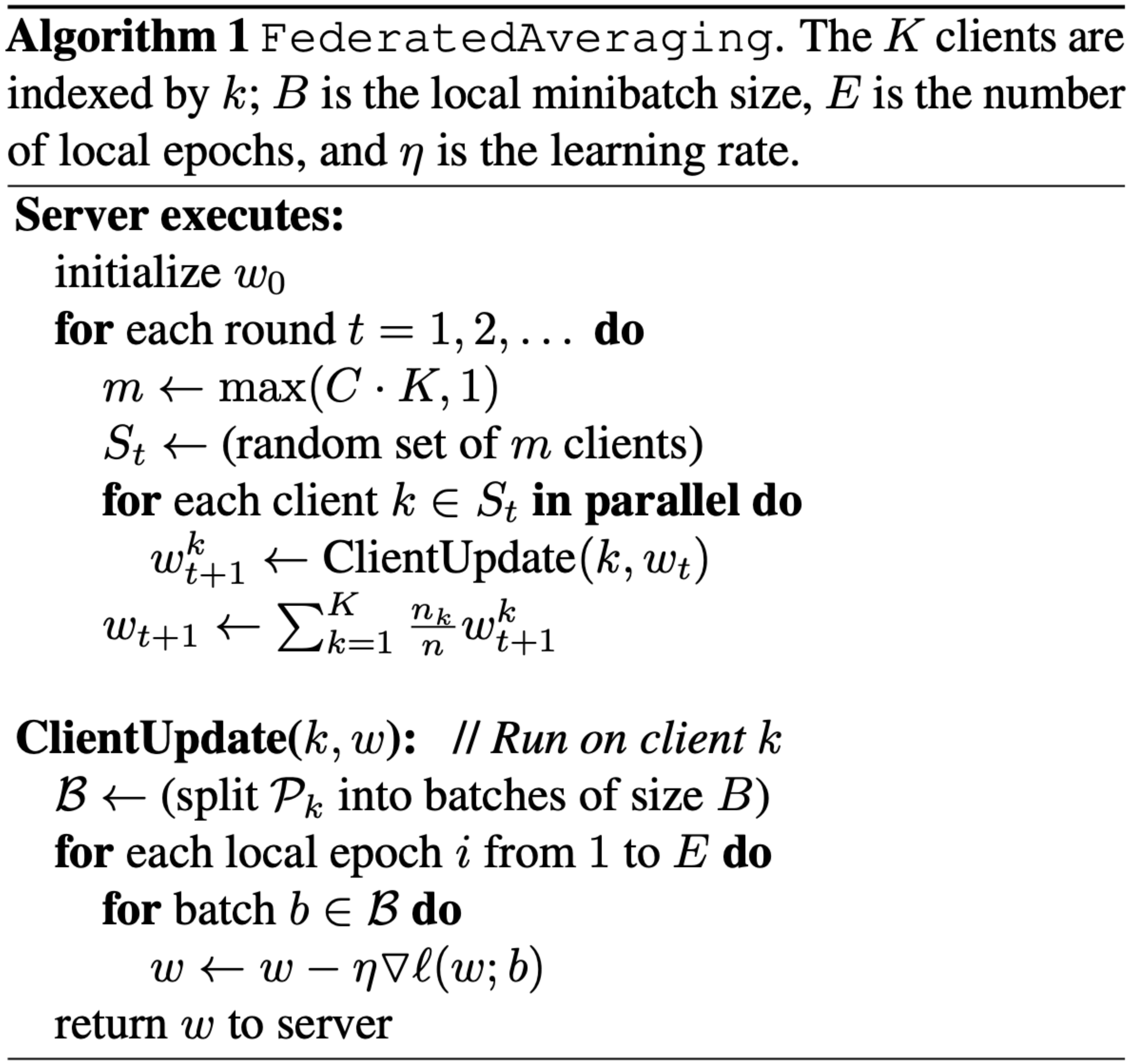


Dataset	Devices	Samples	Samples/device	
			mean	stdev
MNIST	1,000	69,035	69	106
FEMNIST	200	18,345	92	159
Shakespeare	143	517,106	3,616	6,808
Sent140	772	40,783	53	32

- **Systems heterogeneity simulation**
  - Fix a global number of epochs E, and force some devices to perform fewer updates than E epochs. In particular, for varying heterogeneous setting, assign x (chosen uniformly random between $[1,E]$) number of epochs to 0%, 50, and 90% of selected devices.

Backup 3

- The original FedAvg algorithm



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate($k, w$):   // Run on client $k$
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

Backup 4

- Complete theorem

Assume the functions F_k are non-convex, L-Lipschitz smooth, and there exists $L\_ > 0$, such that $\nabla^2 F_k \succeq -L\_\mathbf{I}$, with $\bar{\mu} = \mu - L\_ > 0$. Suppose that w^t is not a stationary solution and the local functions F_k are B-dissimilar, i.e., $B(w^t) \leq B$. If μ, K, and γ_k^t are chosen such that

$$\rho^t = \left(\frac{1}{\mu} - \frac{\gamma^t B}{\mu} - \frac{B(1+\gamma^t)\sqrt{2}}{\bar{\mu}\sqrt{K}} - \frac{LB(1+\gamma^t)}{\bar{\mu}\mu} - \frac{L(1+\gamma^t)^2 B^2}{2\bar{\mu}^2} - \frac{LB^2(1+\gamma^t)^2}{\bar{\mu}^2 K}\left(2\sqrt{2K}+2\right) \right) > 0,$$

then at the iteration t of FedProx, we have the following expected decrease in the global objective:

$$\mathbb{E}_{S_t}[f(w^{t+1})] \leq f(w^t) - \rho^t \|\nabla f(w^t)\|^2,$$

where S_t is the set of K devices chosen at iteration t and $\gamma^t = \max_{k \in S_t} \gamma_k^t$.