# On the Convergence of Local SGD on Identical and Heterogeneous Data

**Ahmed Khaled**

FLOW: Federated Learning One World Seminar
May 13th, 2020

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

## This Talk is Based on

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**Tighter Theory for Local SGD on Identical and Heterogeneous Data**
To appear in Artificial Intelligence and Statistics (AISTATS) 2020

## Earlier Workshop Papers

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**First Analysis of Local GD on Heterogeneous Data**
NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik
**Better Communication Complexity for Local SGD**
NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality

# Collaborators

Peter Richtárik

Professor of Computer Science

KAUST

Konstantin Mishchenko

CS PhD candidate

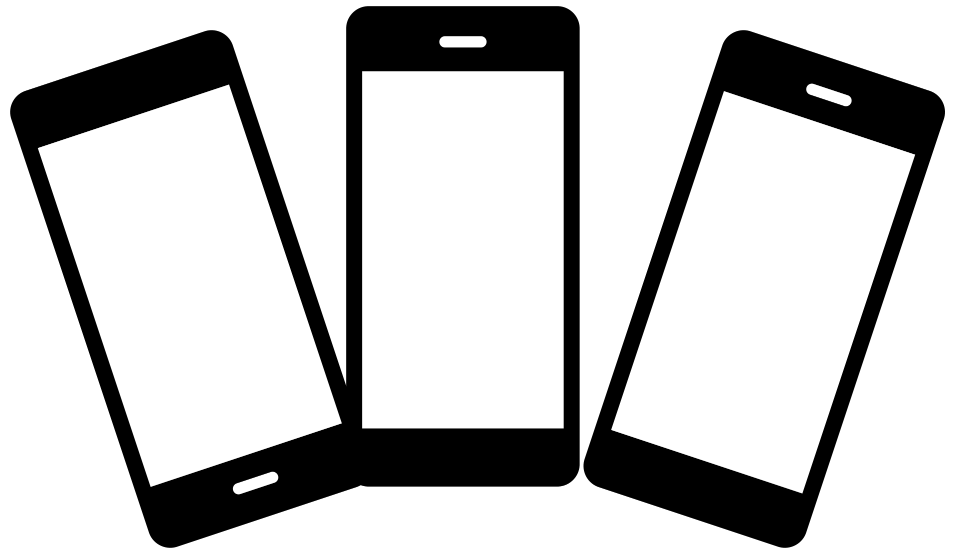
KAUST

# Plan

- **Introduction (20 mins)**
  - Problem Definition
  - Mini-batch SGD vs Local SGD
  - Goals and Contributions
- **Theory (20 mins)**
  - Heterogeneous Data
  - Identical Data

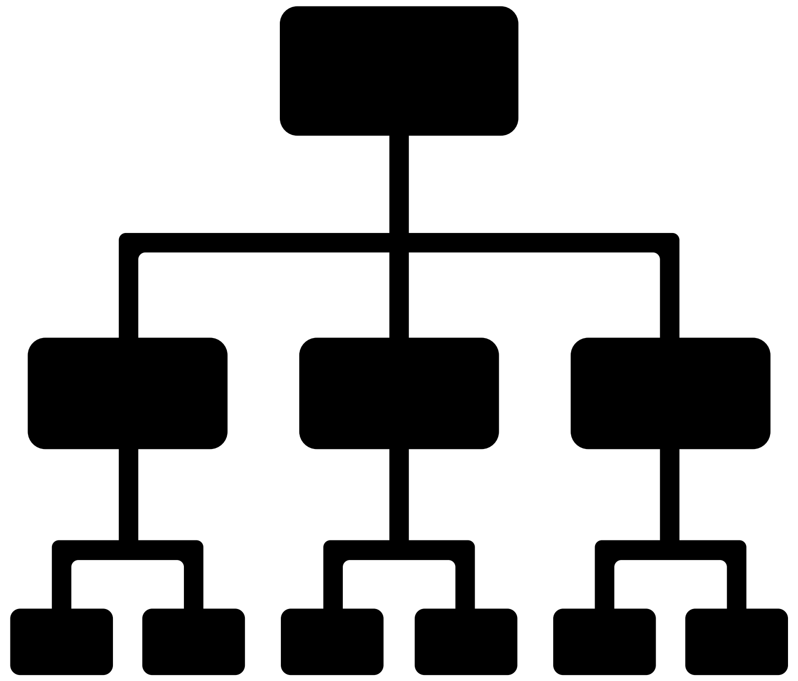
# Introduction

# Federated Learning



- A distributed machine learning setting where data is **distributed over many clients** with potentially unreliable connections.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, Dave Bacon
**Federated Learning: Strategies for Improving Communication Efficiency**
NIPS Workshop on Private Multi-Party Machine Learning, 2016



- **Many applications:** mobile text prediction, medical research, and many more!

- Federated Learning poses highly **interdisciplinary** problems: **optimization,** privacy, security, information theory, statistics, and many other fields intersect.

# Problem Definition

Smooth and μ-convex (for $\mu \geq 0$)

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)] \right\}$$

Model dimension

We can query stochastic gradients

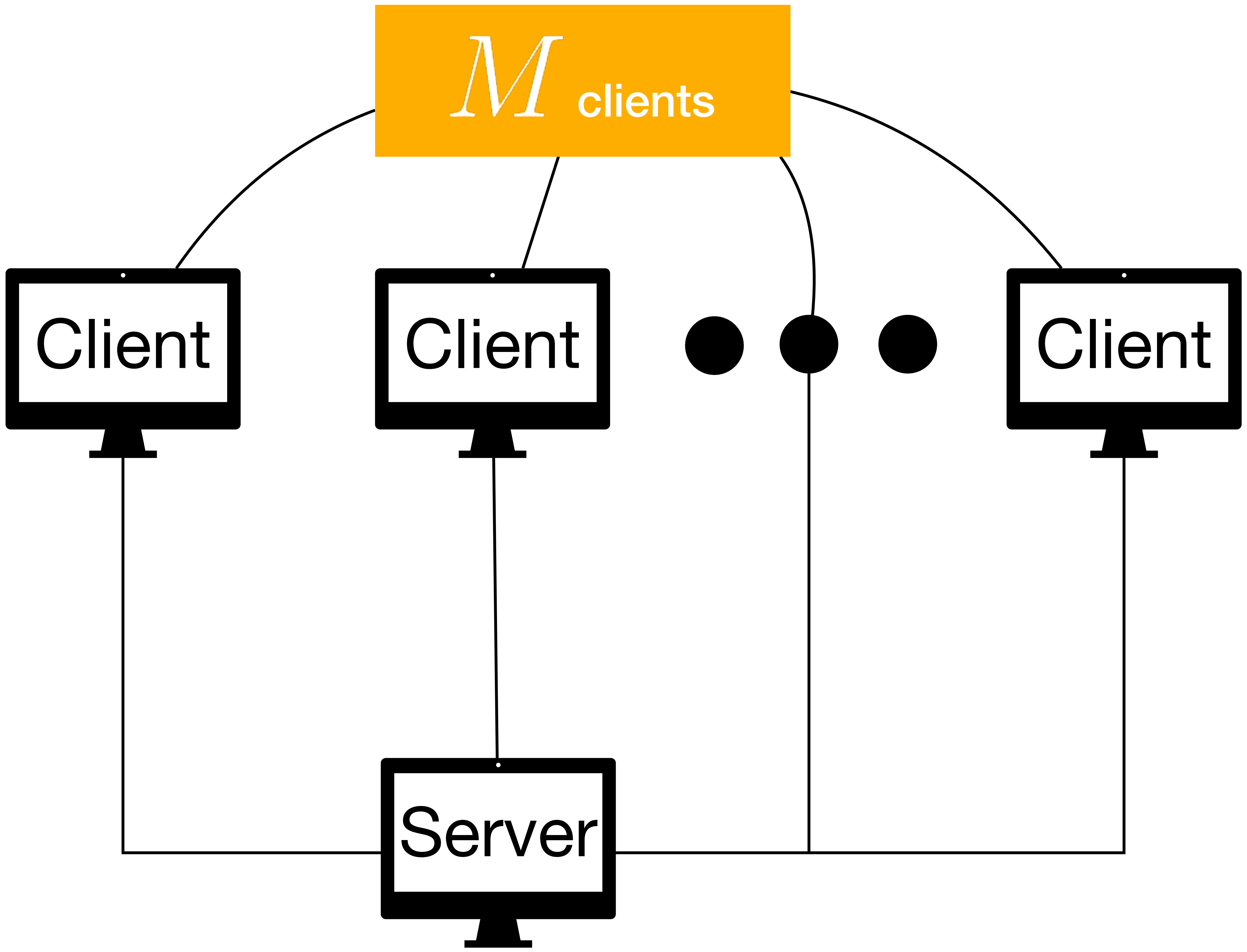
Example: **Finite-sum minimization**

Ubiquitous in machine learning

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

# Distributed Setting

- We desire a scalable, parallel optimization method.
- In typical parameter server applications, **Mini-batch Stochastic Gradient Descent (Mini-batch SGD)** is the popular baseline algorithm.
- **Communication is the bottleneck.**
- There are **two regimes...**



# Data Regime 1 : Heterogenous Data

- Each client has access to
  - **its own optimization objective**
  - **its own dataset**
- Each local objective **is also written as a stochastic expectation.**
- Arises in Federated Learning applications because the data is inherently distributed, can not be centralized due to **privacy protection.**


Has access to

$$f_1(x) = \mathbb{E}_{\xi \sim \mathcal{D}_1}\{f_1(x;\xi)\}$$


Has access to

$$f_2(x) = \mathbb{E}_{\xi \sim \mathcal{D}_2}\{f_2(x;\xi)\}$$

# Data Regime 2 : Identical Data

- Each client has access to **the same dataset.**
- The clients may draw different samples from the dataset, or **have different sampling distributions.**
- Arises in the parameter server framework.
- Can be **insightful into the usefulness of local steps.**


Each has access to

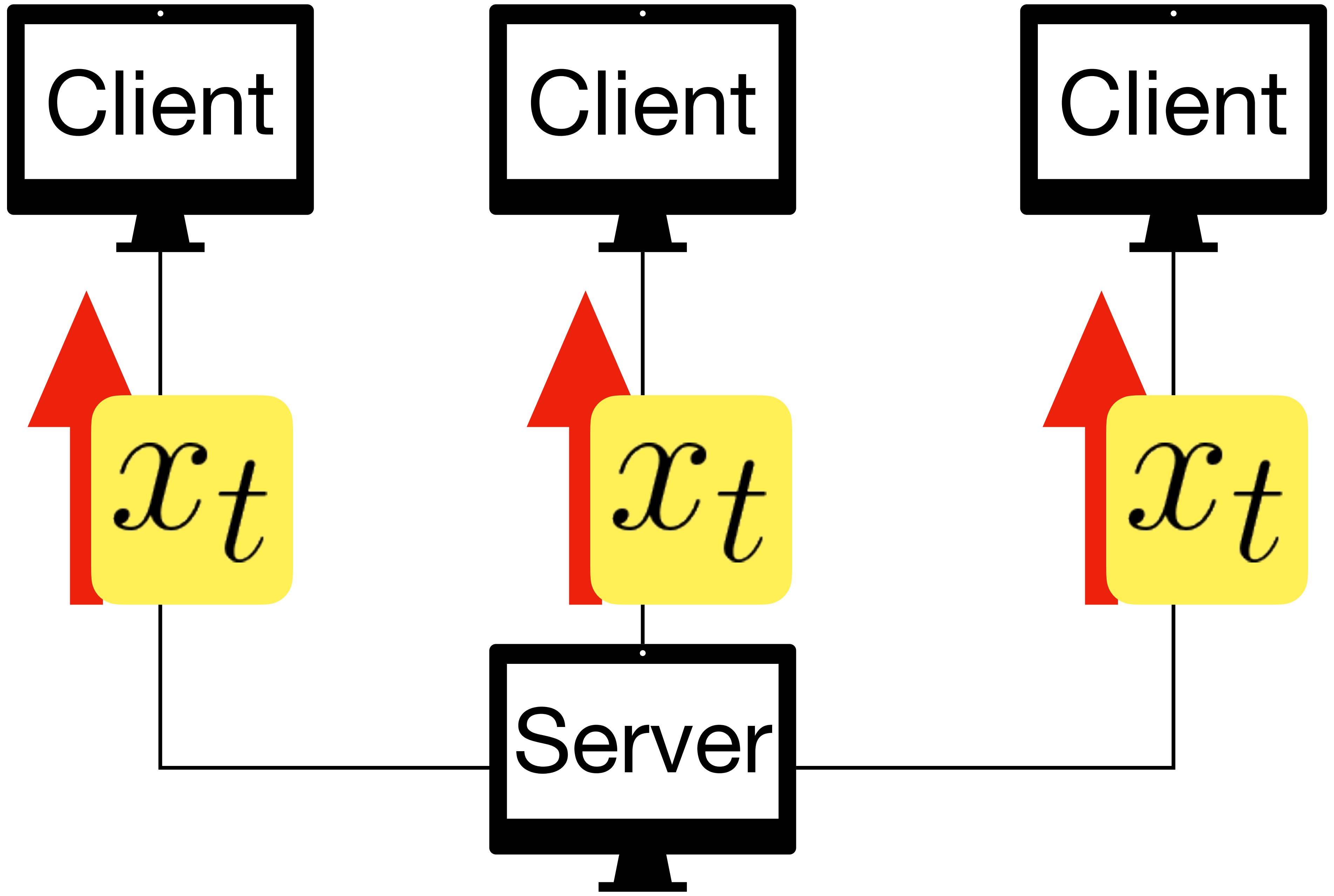
$$f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} \{ f(x; \xi) \}$$

# Mini-batch SGD

Sample a local stochastic gradient $g^1(x_t; \xi^1)$

Sample a local stochastic gradient $g^2(x_t; \xi^2)$

Sample a local stochastic gradient $g^3(x_t; \xi^3)$




**If the data is identical:**

$$\mathbb{E}\left[g^m\left(x_t; \xi^m\right)\right] = \nabla f\left(x_t\right)$$

**If the data is heterogeneous:**

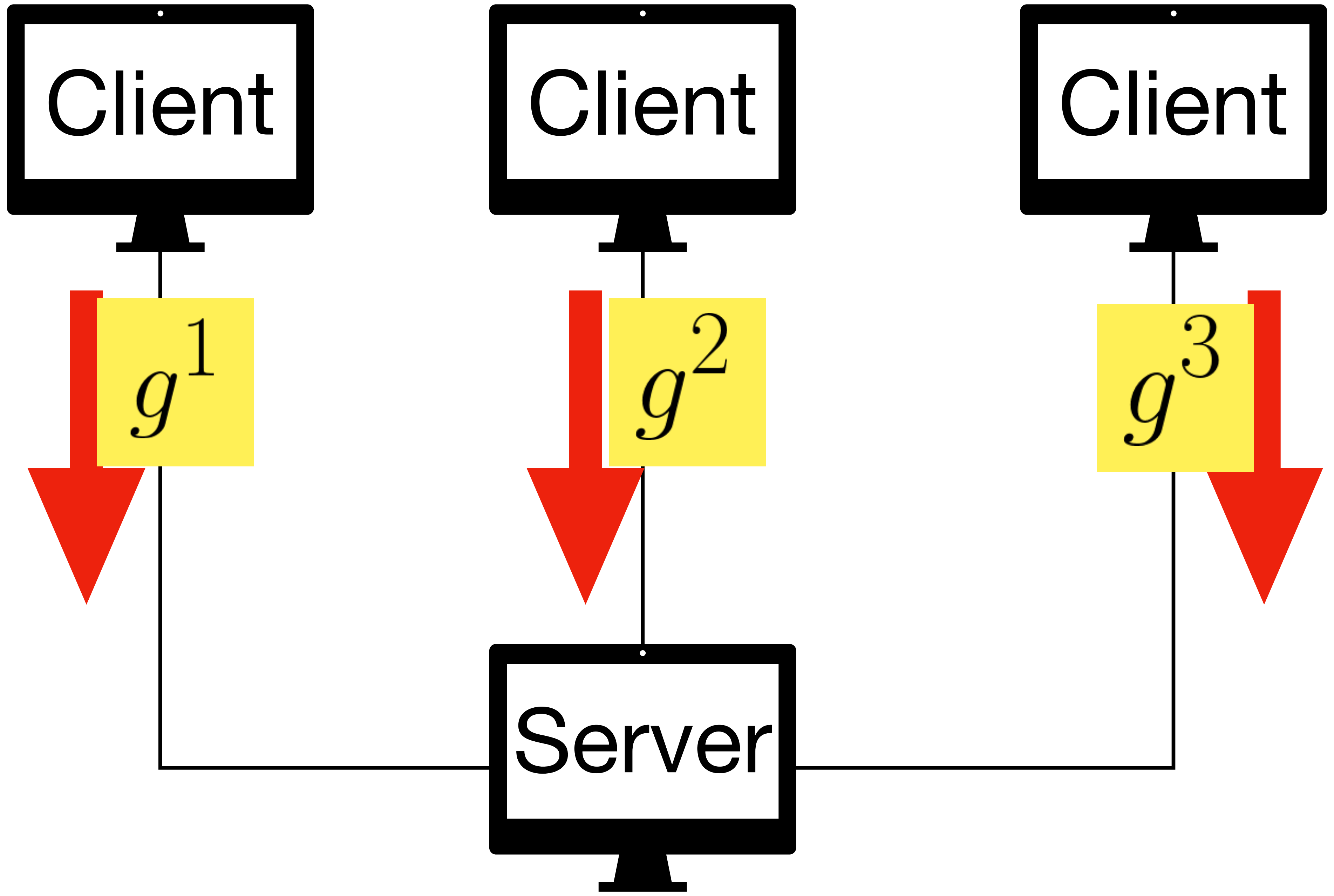
$$\mathbb{E}\left[g^m\left(x_t; \xi^m\right)\right] = \nabla f_m\left(x_t\right)$$

# Mini-batch SGD

Sample a local stochastic gradient $g^1(x_t; \xi^1)$

Sample a local stochastic gradient $g^2(x_t; \xi^2)$

Sample a local stochastic gradient $g^3(x_t; \xi^3)$




**The server then performs aggregation and averaging:**

$$x_{t+1} = x_t - \frac{\gamma}{M} \sum_{m=1}^{M} g^m(x_t; \xi^m)$$

**where** $\gamma > 0$ **is a stepsize**

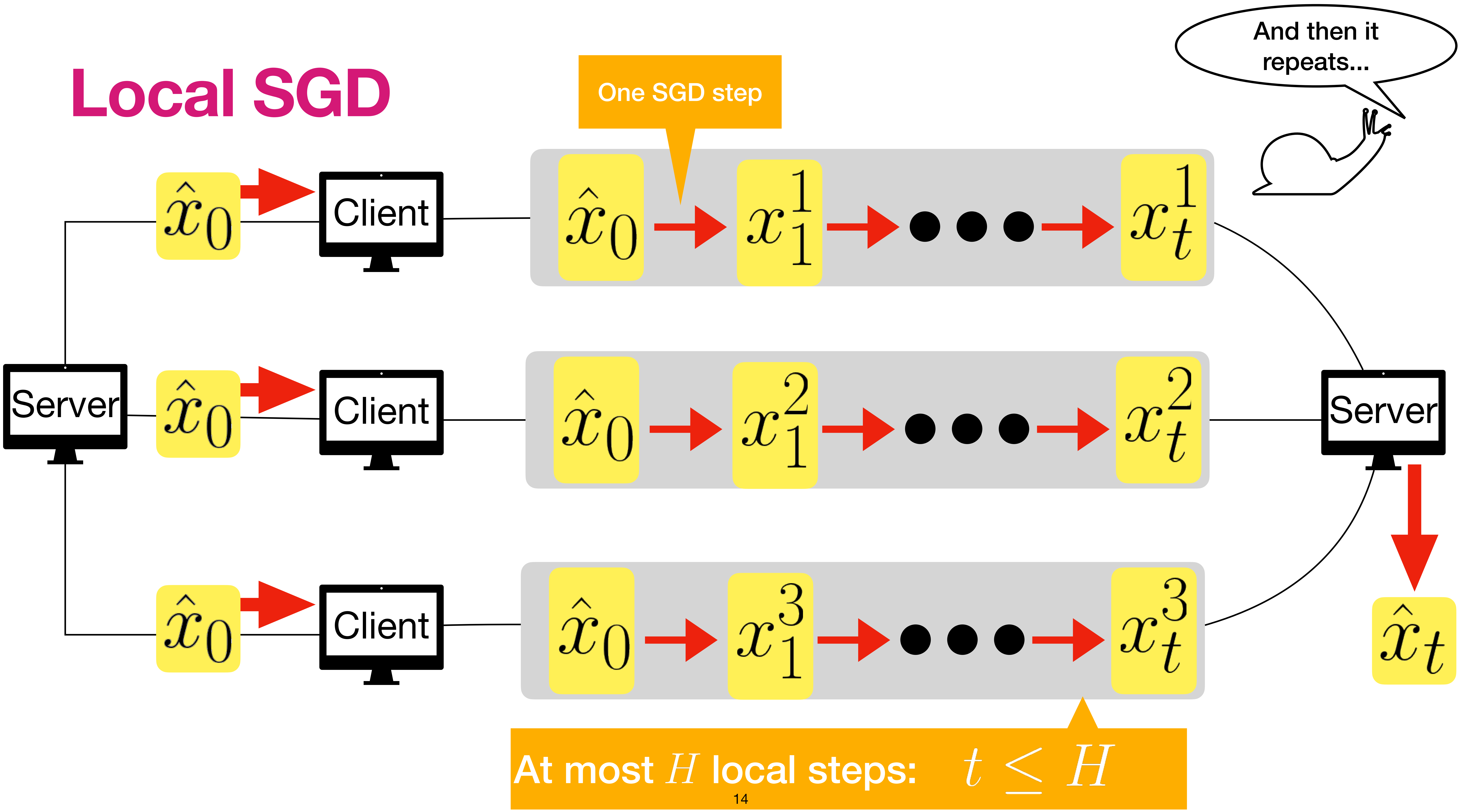
# Local SGD

- Equivalently, we can write the parallel mini-batch SGD update as follows:

$$x_{t+1} = \frac{1}{M} \sum_{m=1}^{M} \left(x_t - \gamma g^m(x_t; \xi^m)\right)$$

- Observation: we take **one "local step"** and follow it by averaging.
- What about **multiple local steps?**

# Local SGD



One SGD step
And then it repeats...
Server
$\hat{x}_0$
Client
$\hat{x}_0$
$x_1^1$
$x_t^1$
$x_1^2$
$x_t^2$
$x_1^3$
$x_t^3$
Server
$\hat{x}_t$
At most $H$ local steps: $t \leq H$

# Our Contributions

# Goal

**Can we achieve the same training error**

**as Mini-batch SGD**

**but with less communication?**

# Our Contributions

- **Heterogeneous data regime:**
  - We critically examine data similarity assumptions and show they **do not hold for even very simple functions.**
  - We obtain the **first convergence guarantee** for Local SGD on **arbitrarily heterogeneous local losses.** The guarantee shows Local SGD **is communication-efficient** at least in some settings.
- **Identical data regime:**
  - We show that even **more dramatic communications savings are possible** for convex and strongly-convex objectives.
  - In particular, we show that for strongly convex objectives the number of communications **can be a constant independent of the total number of iterations!**

# Notation

Total Number of iterations $T$

Synchronization Interval $H$

Number of Communication Steps $C \geq T/H$

Number of Nodes $M$

# Theory for Heterogeneous Data

# Setting and assumptions

Local loss function

- We assume the existence of **at least one minimizer**

$$x_* \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \right\}$$

- Each function is **convex.**
- The results can be extended to the **strongly convex case.**

# Related work

Early work on asymptotic convergence

PDF

Olvi L. Mangasarian.
**Parallel Gradient Distribution in Unconstrained Optimization.**
SIAM Journal on Control and Optimization, 33(6):1916–1925, 1995.

Also consider quantization!

PDF

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi
**Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification and Local Computations.**
In Advances in Neural Information Processing Systems 32 p. 14668–14679, 2019.

PDF

Hao Yu, Sen Yang, and Shenghuo Zhu.
**Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning.**
Proceedings of the AAAI Conference on Artificial Intelligence, 33:5693–5700, 2019.

**However, the last two use the "bounded gradients" assumption...**

# Related work

PDF

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. C. Makaya, Ting He, Kevin Chan.
**When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning**.
arXiv preprint arXiv:1804.05271, 2018.

Also consider quantization!

PDF

Peng Jiang and Gagan Agrawal
**A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication.**
In Advances in Neural Information Processing Systems 31 p. 2525–2536, 2018.

Show that communication savings are possible, however they use a ***bounded dissimilarity*** assumption...

# More related work

Consider FedAvg (with sampling)

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang
**On the Convergence of FedAvg on Non-IID Data.**
Eighth International Conference on Learning Representations (ICLR), 2020.

Farzin Haddadpour and Mehrdad Mahdavi
**On the Convergence of Local Descent Methods in Federated Learning**
arXiv preprint arXiv:1910.14425, 2019.

Obtain results for non-convex objectives under a bounded diversity assumption

**More later (and in the paper)...**

# Assumptions on similarity: bounded dissimilarity

- A common assumption to obtain convergence rates is **bounded dissimilarity**:

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x) - \nabla f(x)\|^2 \leq \sigma^2$$

$$\text{for all } x \in \mathbb{R}^d$$

# Assumptions on similarity: bounded dissimilarity

- The bounded dissimilarity condition may not be satisfied **for 1-dimensional quadratics:**

$$f_m(x) \overset{\text{def}}{=} \frac{a_m}{2} x^2$$

$$\frac{1}{M}\sum_{m=1}^{M} \|\nabla f_m(x) - \nabla f(x)\|^2$$

$$= \left(\frac{1}{M}\sum_{m=1}^{M}\left(a_m - \frac{1}{M}\sum_{j=1}^{M} a_j\right)^2\right) \cdot x^2$$

Can be arbitrarily large

# Assumptions on similarity: bounded gradients

- The bounded gradients assumption is also in common usage:

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x)\|^2 \leq G^2 \qquad \text{for all } x \in \mathbb{R}^d$$

- Problem 1: **special case** of bounded dissimilarity without the benefit of characterizing similarity..

$$\frac{1}{M}\sum_{m=1}^{M}||\nabla f_m(x) - \nabla f(x)||^2 = \frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x)\|^2 - \|\nabla f(x)\|^2 \leq G^2$$

# Assumptions on similarity: bounded gradients

- The bounded gradients assumption is also in common usage:

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x)\|^2 \leq G^2$$

- Problem 2: **contradicts** global strong convexity.

L. Nguyen, P. Ha Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, & M. Takáč.
**SGD and Hogwild! Convergence Without the Bounded Gradients Assumption**.
Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:3750-3758, 2018.

# Assumptions on similarity: bounded gradients

- The bounded gradients assumption is also in common usage:

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(x)\|^2 \leq G^2$$

- Problem 3: **questionable** applicability to practice.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien.
**Reducing Noise in GAN Training with Variance Reduced Extragradient**.
In Advances in Neural Information Processing Systems 32, p. 391–401, 2019.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky.
**Revisiting Stochastic Extragradient**.
To appear in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

There are no results that apply to arbitrarily heterogeneous data

# The alternative

- Our theory is built upon the **variance at the optimum**

$$\sigma_{\mathrm{dif}}^2 \stackrel{\mathrm{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi \sim \mathcal{D}_m} \left[ \left\| \nabla f_m \left( x_*; \xi \right) \right\|^2 \right]$$

- Naturally relates the difference between the functions **at a single point**.

- Zhang and Li show that when this quantity is zero, we get linear convergence for strongly convex objectives with any $H$.

PDF Chi Zhang and Qianxiao Li.
**Distributed Optimization for Over-Parameterized Learning.**
arXiv preprint arXiv:1906.06205, 2019

# Main Theorem (Heterogeneous Data)

For any sufficiently small step size

$$\gamma \leq \min\left\{\frac{1}{4L}, \frac{1}{8L(H-1)}\right\}$$

Initial distance to the optimum

# nodes

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{4\|r_0\|^2}{\gamma T} + \frac{20\gamma\sigma_{\mathrm{dif}}^2}{M} + 16\gamma^2 L(H-1)^2\sigma_{\mathrm{dif}}^2.$$

Average iterate

Smoothness constant

Synchronization interval

# Main Theorem (Heterogeneous Data)

For any sufficiently small step size

$$\gamma \le \min\left\{\frac{1}{4L}, \frac{1}{8L(H-1)}\right\}$$

Same as Mini-batch SGD (up to constants)

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \boxed{\frac{4\|r_0\|^2}{\gamma T} + \frac{20\gamma\sigma_{\text{dif}}^2}{M}}$$
$$+ \boxed{16\gamma^2 L(H-1)^2\sigma_{\text{dif}}^2.}$$

An error term controlled by the synchronization interval $H$

# Communication Complexity

Desired accuracy

- If we properly chose the stepsize...

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leq \varepsilon$$

- **Communication Complexity:** iterations to guarantee

Smoothness constant

$$C = \Omega\left(\frac{\|r_0\|^2}{\varepsilon}\max\left\{L, \frac{\sigma_{\text{dif}}^2}{HM\varepsilon}, \frac{\sqrt{L}\sigma_{\text{dif}}}{\sqrt{\varepsilon}}\right\}\right)$$

Initial distance to the optimum

Synchronization Interval

# clients

# Communication Complexity

- For a small enough desired accuracy:

$$C = \Omega\left(\frac{\|r_0\|^2\sqrt{L}\sigma_{\mathrm{dif}}}{\varepsilon^{3/2}}\right) \qquad C = \Omega\left(\frac{\|r_0\|^2\sigma_{\mathrm{dif}}^2}{\varepsilon^2 M}\right)$$

Local SGD

Minibatch SGD

**We get a reduction in the number of communications as a function of the accuracy even for arbitrarily heterogeneous data!**

# Optimal Synchronization Interval

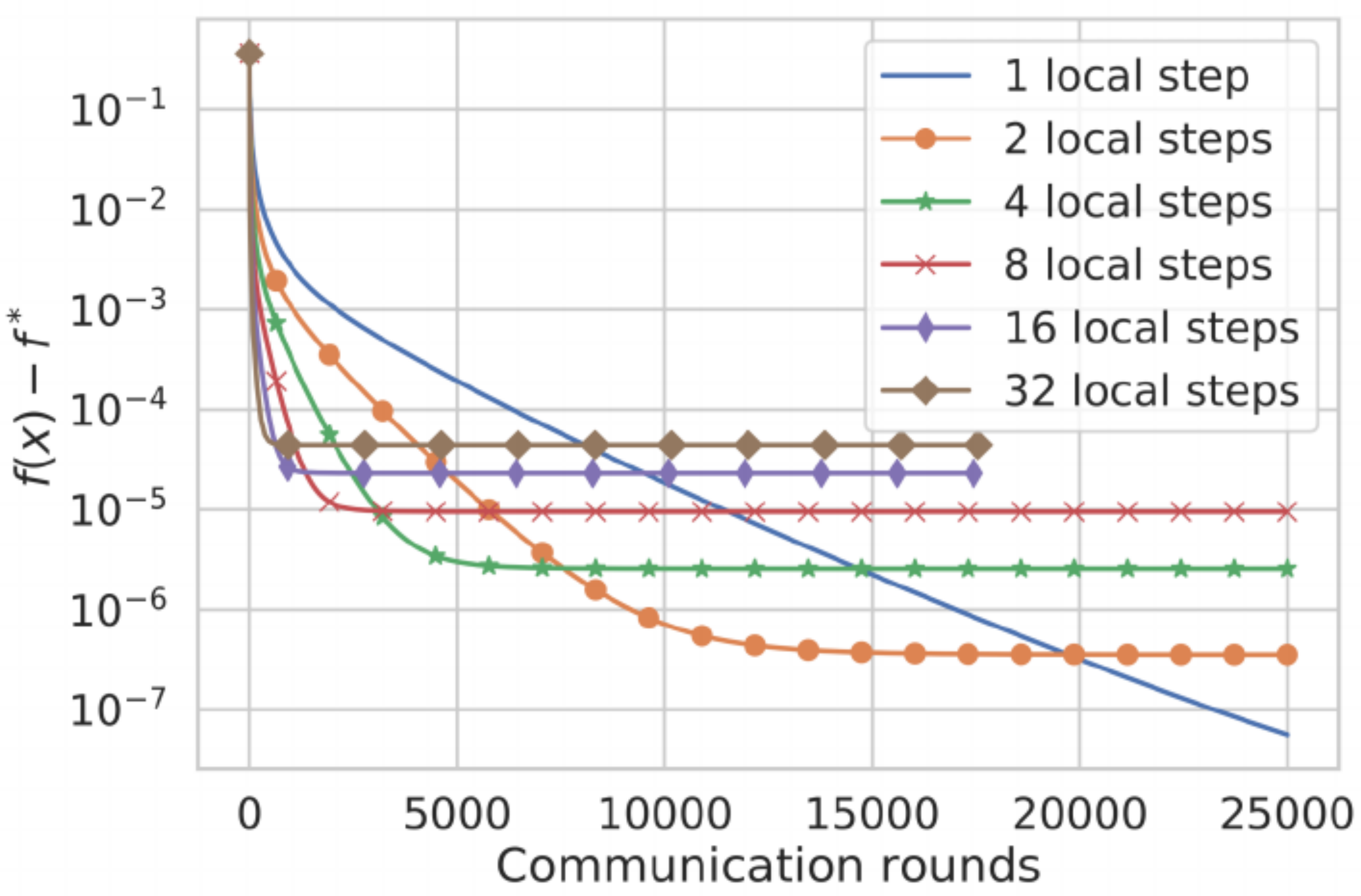
- **We show that that the optimal *H* for attaining the same rate as Minibatch SGD is**

$$H = 1 + \left\lfloor T^{1/4} M^{-3/4} \right\rfloor$$

- And the corresponding communication complexity then is

$$C = \Omega\left(\min\{T, T^{3/4} M^{-3/4}\}\right)$$

# Experimental Results



# Theory for Identical Data

# Setting


Each has access to

$$f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} \{ f(x; \xi) \}$$


- We assume that each $f(x;\xi)$ is almost surely **convex and smooth**. All clients share the same objective. **Will present results for $\mu$-strongly convex as well.**
- The measure of variance is also the **variance at the optimum:**

$$\sigma_{\mathrm{opt}}^2 \stackrel{\mathrm{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi \sim \mathcal{D}_m} \left[ \| \nabla f(x_*; \xi) \|^2 \right]$$

# Background 1

- Stich (2019) analyzes Local SGD with identical data.
- For strongly convex objectives, the communication complexity to reach the same error as Minibatch SGD is:

$$C = \Omega\left(\sqrt{\kappa M T}\right)$$

The condition number
$\kappa \stackrel{\text{def}}{=} L/\mu$

# clients

Sebastian U. Stich
**Local SGD Converges Fast and Communicates Little.**
In the Seventh International Conference on Learning Representations ICLR, 2019.

# Background 2

- One-shot averaging is running SGD on each node and communicating only once at the end.

- This communication complexity tells us that one-shot averaging is not convergent. **But it should be. Why?**

$$C = \Omega\left(\sqrt{\kappa M T}\right)$$

Grows with the total number of iterations

- There are **no results** for minimizing convex (but not strongly convex) objectives.

# Theorem (Identical Data, Strong Convexity)

- With an appropriately chosen constant stepsize:

The condition number

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\boxed{\frac{\|r_0\|^2}{T^2} + \frac{\sigma_{\text{opt}}^2}{\mu^2 M T}} + \boxed{\frac{\sigma_{\text{opt}}^2 \kappa (H-1)}{\mu^2 T^2}}\right)$$

Same as Minibatch SGD

Strong convexity constant

Error Term

# Interpreting the Result

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{\|r_0\|^2}{T^2} + \frac{\sigma_{\text{opt}}^2}{\mu^2 M T} + \frac{\sigma_{\text{opt}}^2 \kappa (H-1)}{\mu^2 T^2}\right)$$

- **Optimal synchronization interval** $H = 1 + \lfloor T/(\kappa M) \rfloor$
  - Reaches the same convergence error as Mini-batch SGD (up to absolute constants) but with a **communication complexity** of

$$\tilde{\Omega}(\min(T, \kappa M))$$

**Number of communications can be constant!**

# Interpreting the Result

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{\|x_0 - x_*\|^2}{T^2} + \frac{\sigma^2}{\mu^2 MT} + \frac{\kappa\sigma^2(H-1)}{\mu^2 T^2}\right)$$

- **One-shot averaging**
  - Put $H = T + 1$, then we obtain a convergence rate of $\tilde{\mathcal{O}}\left(\frac{\sigma^2\kappa}{\mu^2 T}\right)$
  - An improvement, but applying Jensen's inequality yields $\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 T}\right)$

**There is room for improvement!**

# Theorem (Identical Data, Convexity)

- For the (non-strongly) convex case, we get a similar result

$$\mathbb{E}\left[f\left(\bar{x}_T\right) - f\left(x_*\right)\right] \leqslant \frac{2}{\gamma T}\left\|x_0 - x_*\right\|^2 + \frac{2\gamma\sigma^2}{M} + 4\gamma^2 L\sigma^2(H-1)$$

- Same guarantee as the heterogenous case, but with a **linear instead of quadratic** dependence on the synchronization interval.
- Translates to **more communications savings**

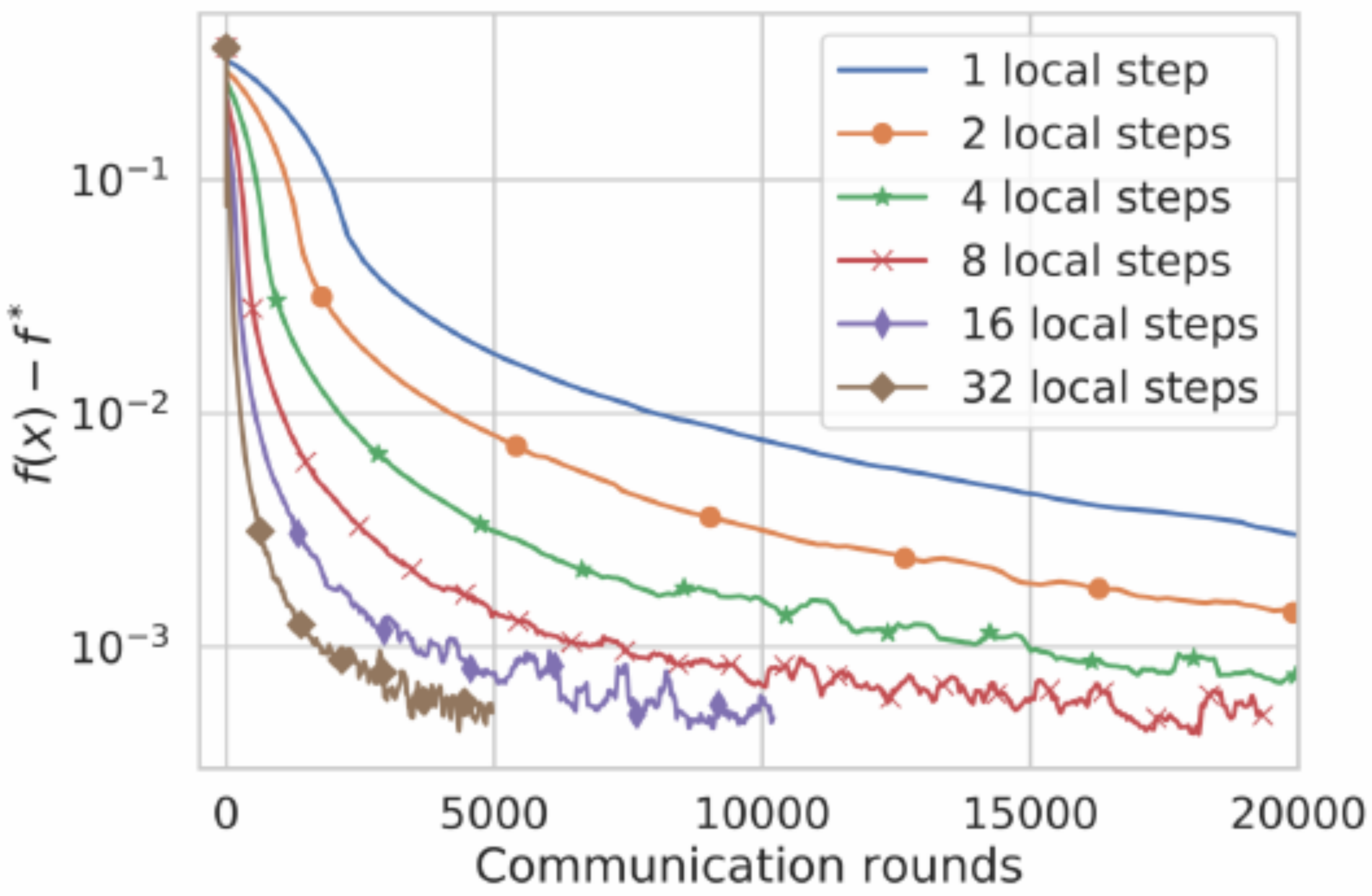
# Concurrent Work

- Similar results for identical data **were obtained in concurrent work** of Stich and Karimireddy who use a different proof technique.

Sebastian U. Stich and Sai Praneeth Karimireddy
**The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Communication**.
arXiv preprint arXiv:1909.05350, 2019.

- More discussion is given in the paper.

# Experimental Results



# Open Questions 1

- Can we get **better convergence results** for Local SGD or Federated Averaging compared to Minibatch SGD?
  - For general convex objectives and identical data:

    Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro.
    **Is Local SGD Better than Minibatch SGD?**
    arXiv preprint arXiv:2002.07839, 2020.

  - For heterogeneous data, client sampling and using two stepsizes:

    Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri and Sashank J. Reddi, Sebastian U. Stich and, Ananda Theertha Suresh.
    **SCAFFOLD: Stochastic Controlled Averaging for Federated Learning.**
    arXiv preprint arXiv:1910.06378, 2019.

# Open Questions 2

- Do local methods give benefits **other than optimization?**
  - Meta Learning point of view:

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan
**Improving Federated Learning Personalization via Model Agnostic Meta Learning**
arXiv preprint arXiv:1909.12488, 2020.

  - Another perspective on personalization:

Filip Hanzely and Peter Richtárik
**Federated Learning of a Mixture of Global and Local Models**
arXiv preprint arXiv:2002.05516, 2020.

# Questions?

# Thank you!

# On Non-convex Objectives

- Even for the single-machine finite-sum optimization problem, convergence bounds in the non-convex setting often rely on restrictive assumptions.
- Often results **rely on bounded variance or gradient dissimilarity assumptions.**
- The relation of these assumptions to each other **is not clear.**
- We consider this and obtain a **more general result** in our new paper:

PDF

Ahmed Khaled and Peter Richtárik.

**Better Theory for SGD in the Nonconvex World**

arXiv preprint arXiv:2002.03329, 2020

# References


- Li, Andersen, Park, Smola, Ahmed, Josifovski, Long, Shekita, and Su - Scaling distributed machine learning with the parameter server, OSDI'14: Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation, p.583-598, 2014.

- Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning. arXiv preprint arXiv:1804.05271, 2018.

- Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 33:5693–5700, 2019.

- Hao Yu, Rong Jin, and Sen Yang. On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization. In Proceedings of the 36th International Conference on Machine Learning, 2019.

# References


- Sebastian U. Stich. Local SGD Converges Fast and Communicates Little. In International Conference on Learning Representations, 2019.
- Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification and Local Computa- tions. In Advances in Neural Information Processing Systems 32, pages 14668–14679, 2019.
- L. Nguyen, P.H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, & M. Takáč. SGD and Hogwild! Convergence Without the Bounded Gradients Assumption. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:3750-3758, 2018.
- Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. In Advances in Neural Information Processing Systems 32, pages 391–401. Curran Associates, Inc., 2019.

# References


- Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting Stochastic Extragradient. To appear in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

- Chi Zhang and Qianxiao Li. Distributed Optimization for Over-Parameterized Learning. arXiv preprint arXiv:1906.06205, 2019.

- Peng Jiang and Gagan Agrawal. A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication. In Advances in Neural Information Processing Systems 31, pages 2525–2536. 2018.

- Jianyu Wang and Gauri Joshi. Cooperative SGD: A Unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. arXiv preprint arXiv:1808.07576, 2018.

# References


- Blake Woodworth and Kumar Kshitij Patel and Sebastian U. Stich and Zhen Dai and Brian Bullins and H. Brendan McMahan and Ohad Shamir and Nathan Srebro. Is Local SGD Better than Minibatch SGD? arXiv preprint arXiv: 2002.07839, 2020.

- Sai Praneeth Karimireddy and Satyen Kale and Mehryar Mohri and Sashank J. Reddi and Sebastian U. Stich and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. arXiv preprint arXiv: 1910.06378, 2019.

- Sebastian U. Stich and Sai Praneeth Karimireddy. The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Communication. arXiv preprint arXiv:1909.05350, 2019.

# Acknowledgments


- PDF icon made by Pixel Buddha on www.flaticon.com.