# Local SGD for non-i.i.d. data

Konstantin Mishchenko
Work done together with
Ahmed Khaled and Peter Richtárik

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

# Problem

$$\min_x \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

**Convex**

# Problem

$$\min_x \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

**Convex**

**In practice, usually a neural network**

# Problem

$$\min_x \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

$$f_m(x) = \mathbb{E}_\xi f_m(x; \xi)$$

# Local SGD

$$\min_{x} \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

$$x_{t+1}^{m} = \begin{cases} \hat{x}_{t+1}, \text{ if } t \bmod H = 0 \\ x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m), \text{ otherwise} \end{cases}$$

# Local SGD

$$\min_x \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

$$x_{t+1}^m = \begin{cases} \frac{1}{M} \sum_{j=1}^{M} (x_t^j - \gamma \nabla f_j(x_t^j; \xi_t^j)), \text{if } t \bmod H = 0 \\ x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m), \text{ otherwise} \end{cases}$$

# Local SGD

$$\min_x \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

$$x_{t+1}^m = \begin{cases} \hat{x}_{t+1}, \text{if } t \bmod \boxed{H} = 0 \\ x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m), \ \text{otherwise} \end{cases}$$

$$H = 1 \longrightarrow \text{minibatch SGD}$$

$$H = T \longrightarrow \text{one-shot averaging}$$

# Local GD

$$x_{t+1}^{m} = x_{t}^{m} - \gamma \nabla f_{m}(x_{t}^{m})$$

# The Variance of Local GD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m)$$

$$\sigma_f^2 \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \|\nabla f_m(x_*)\|^2$$

# Analysis difficulties in local GD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m)$$

$$\hat{x}_t \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} x_t^m$$

# Analysis difficulties in local GD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m)$$

$$\hat{x}_t \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} x_t^m$$

$$g_t \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(x_t^m)$$

$$V_t \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \|x_t^m - \hat{x}_t\|^2$$

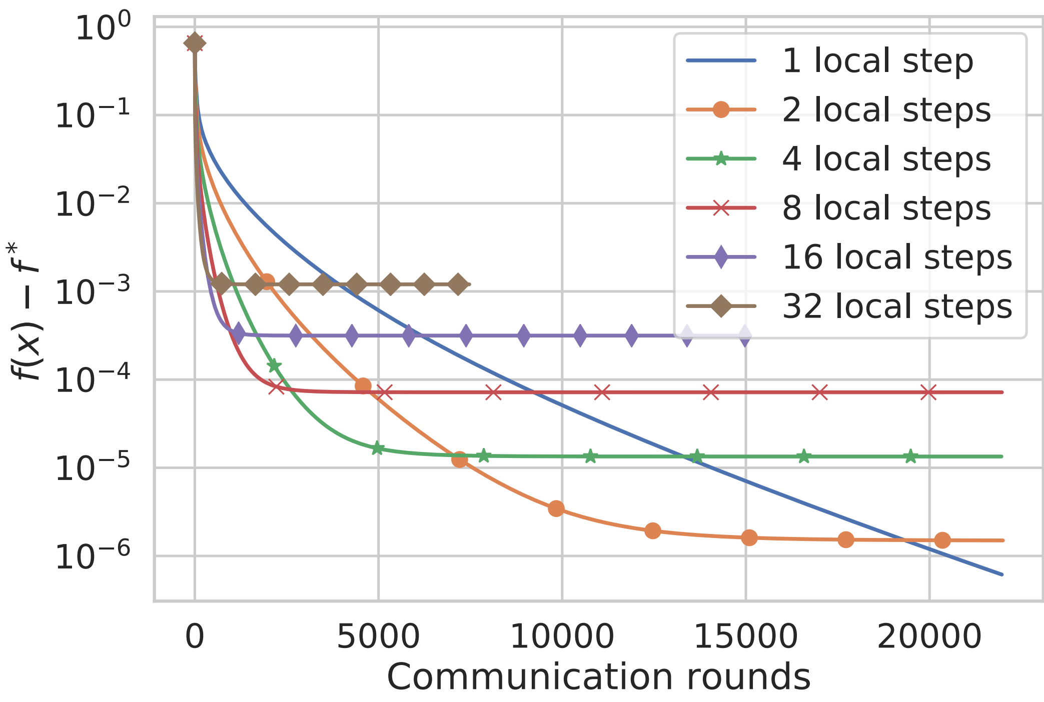
# Theorem

Choose $H$ such that $H \leq \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{4L\sqrt{T}} \leq \frac{1}{4HL}$, and hence

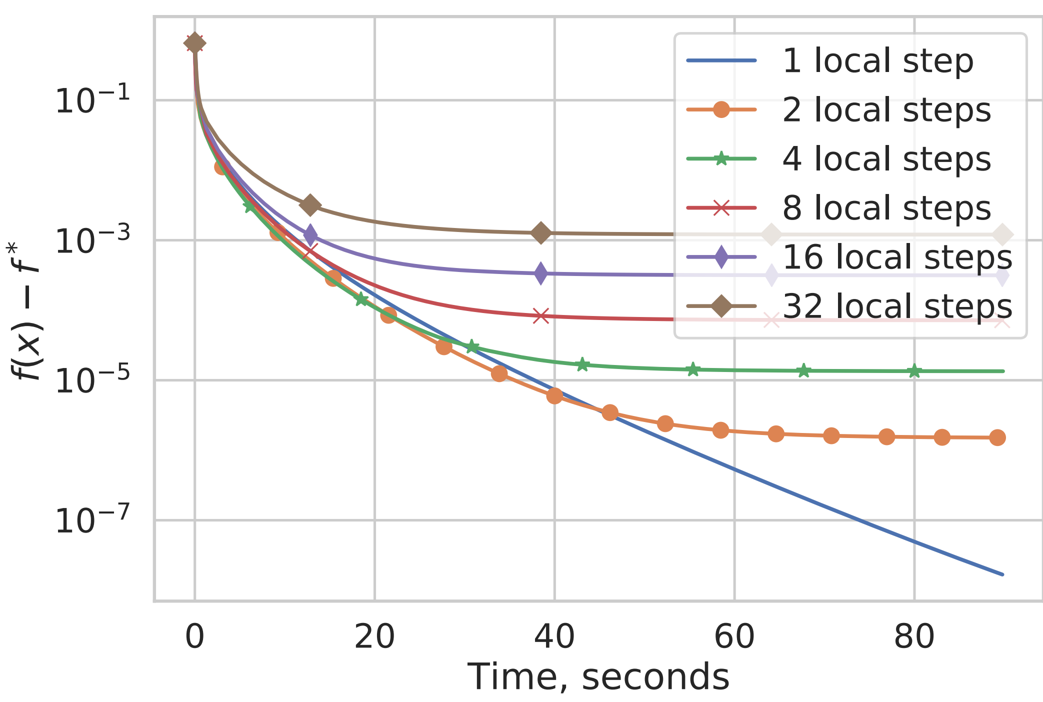
$$f(\hat{x}_T) - f(x_*) \leq \frac{8L\|x_0 - x_*\|^2}{\sqrt{MT}} + \frac{3M\sigma_f^2 H^2}{2LT}.$$

To get a convergence rate of $1/\sqrt{MT}$ we can choose $H = O(T^{1/4}M^{-3/4})$, which implies a total number of $\Omega(T^{3/4}M^{3/4})$ communication steps. If a rate of $1/\sqrt{T}$ is desired instead, we can choose a larger $H = O(T^{1/4})$.

# Plots

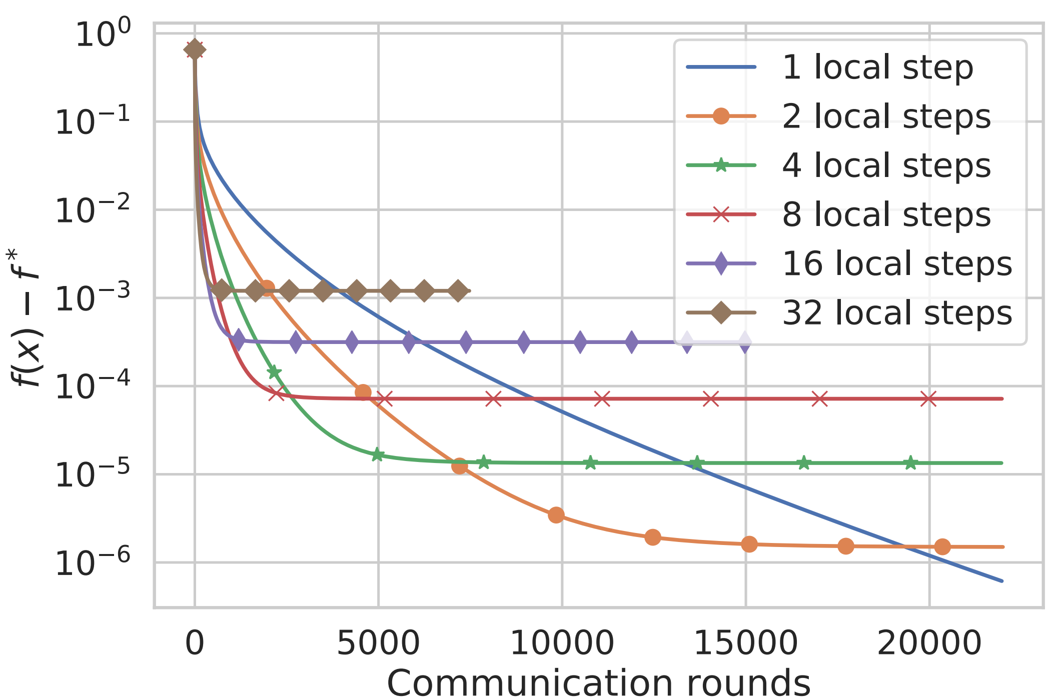


$10^0$
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
$10^{-5}$
$10^{-6}$
$f(x) - f^*$
0
5000
10000
15000
20000
Communication rounds
1 local step
2 local steps
4 local steps
8 local steps
16 local steps
32 local steps

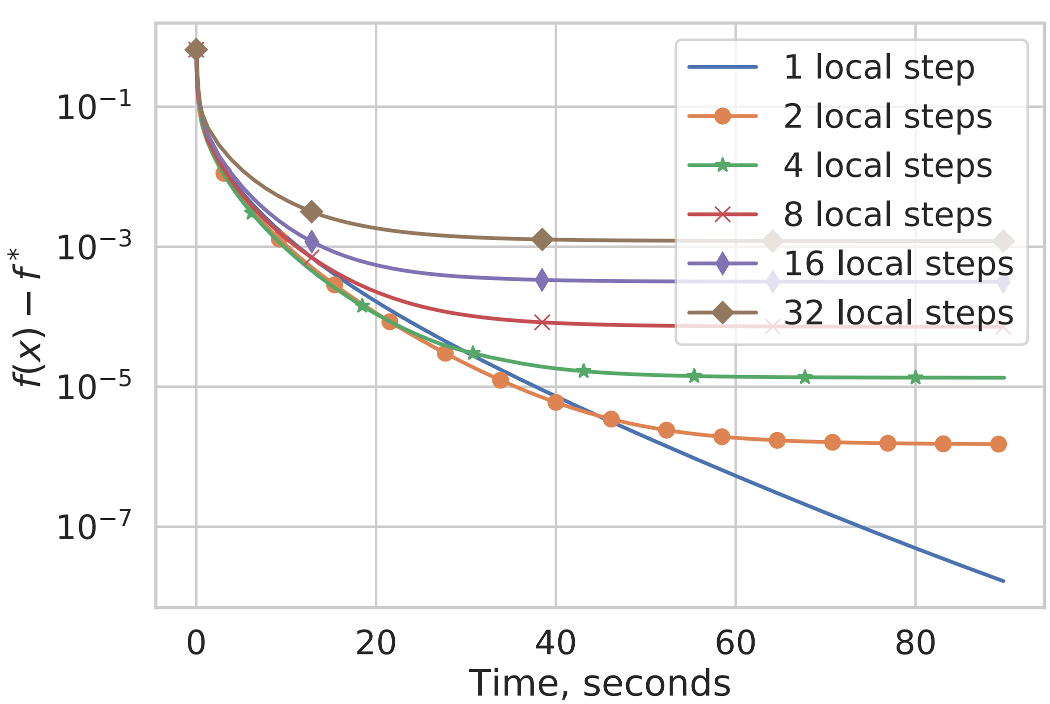



$10^{-1}$
$10^{-3}$
$10^{-5}$
$10^{-7}$
$f(x) - f^*$
0
20
40
60
80
Time, seconds
1 local step
2 local steps
4 local steps
8 local steps
16 local steps
32 local steps

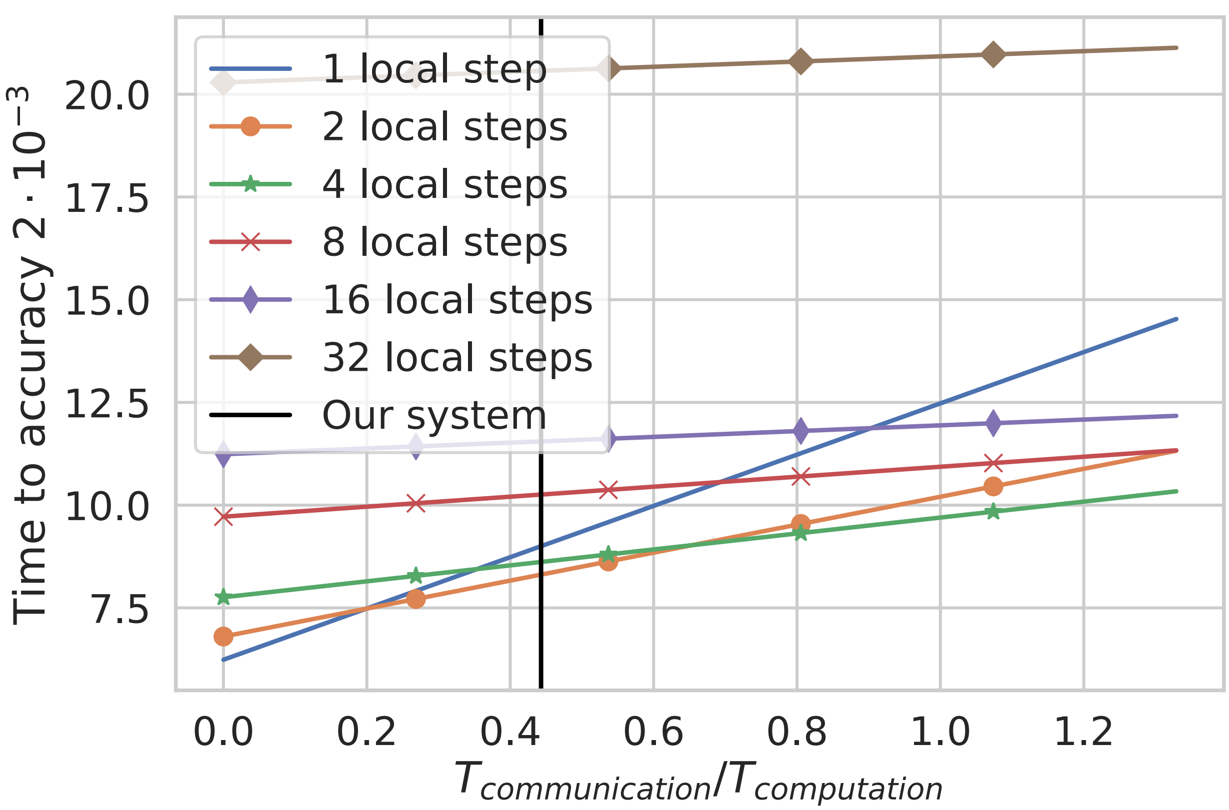
# Plots



$10^{0}$
$10^{-1}$
$10^{-2}$
$10^{-3}$
$10^{-4}$
$10^{-5}$
$10^{-6}$
$f(x) - f^*$
0
5000
10000
15000
20000
Communication rounds
1 local step
2 local steps
4 local steps
8 local steps
16 local steps
32 local steps

$10^{-1}$
$10^{-3}$
$10^{-5}$
$10^{-7}$
$f(x) - f^*$
0
20
40
60
80
Time, seconds
1 local step
2 local steps
4 local steps
8 local steps
16 local steps
32 local steps

Time to accuracy $2 \cdot 10^{-3}$
20.0
17.5
15.0
12.5
10.0
7.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$T_{communication}/T_{computation}$
1 local step
2 local steps
4 local steps
8 local steps
16 local steps
32 local steps
Our system

# Local SGD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m)$$

$$\mathbb{E}_\xi \|\nabla f_m(x;\xi) - \nabla f_m(x)\|^2 \leq \sigma^2$$

$$\mathbb{E}_\xi \|\nabla f_m(x;\xi) - \nabla f_m(x)\|^2 \leq 4 L D_{f_m}(x, x_*) + 2\sigma^2$$

# Local SGD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m)$$

$$\mathbb{E}_\xi \|\nabla f_m(x;\xi) - \nabla f_m(x)\|^2 \le 4 L D_{f_m}(x, x_*) + 2\sigma^2$$

$$\sigma_{\mathrm{dif}} \overset{\mathrm{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_\xi \|\nabla f_m(x_*, \xi)\|^2$$

# Local SGD

$$x_{t+1}^m = x_t^m - \gamma \nabla f_m(x_t^m; \xi_t^m)$$

$$\mathbb{E}_\xi \|\nabla f_m(x;\xi) - \nabla f_m(x)\|^2 \le 4L D_{f_m}(x, x_*) + 2\sigma^2$$

$$\sigma_{\text{dif}} \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_\xi \|\nabla f_m(x_*, \xi)\|^2$$

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

# Theorem

Choose $H$ such that $H \leq \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{8L\sqrt{T}} \leq \frac{1}{8HL}$ and $\mathbb{E} f(\hat{x}_T) - f(x_*) \leq \frac{32L\|\hat{x}_0 - x_*\|^2}{\sqrt{MT}} + \frac{5\sigma_{\mathrm{dif}}^2}{2L\sqrt{MT}} + \frac{\sigma_{\mathrm{dif}}^2 M(H-1)^2}{4LT}$.

# Theorem

Choose $H$ such that $H \le \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{8L\sqrt{T}} \le \frac{1}{8HL}$ and $\mathbb{E} f(\hat{x}_T) - f(x_*) \le \frac{32L\|\hat{x}_0 - x_*\|^2}{\sqrt{MT}} + \frac{5\sigma_{\mathrm{dif}}^2}{2L\sqrt{MT}} + \frac{\sigma_{\mathrm{dif}}^2 M(H-1)^2}{4LT}$.

$$\text{Optimal } H \text{ is } H = 1 + \boxed{\lfloor T^{1/4} M^{-3/2} \rfloor}$$

# Theorem

Choose $H$ such that $H \leq \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{8L\sqrt{T}} \leq \frac{1}{8HL}$ and

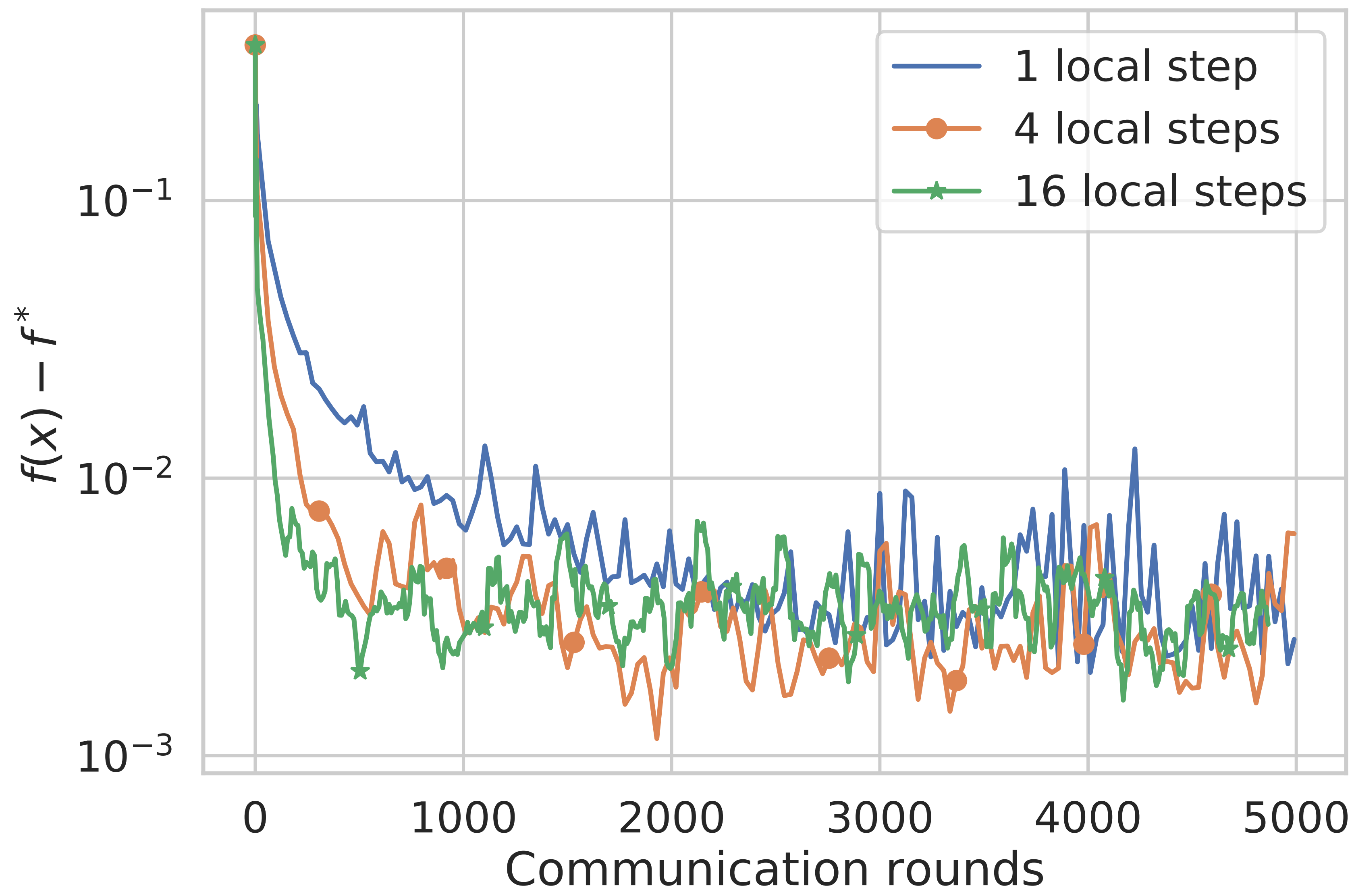
$$\mathbb{E} f(\hat{x}_T) - f(x_*) \leq \frac{32L\|\hat{x}_0 - x_*\|^2}{\sqrt{MT}} + \frac{5\sigma_{\mathrm{dif}}^2}{2L\sqrt{MT}} + \frac{\sigma_{\mathrm{dif}}^2 M(H-1)^2}{4LT}.$$

Optimal $H$ is $H = 1 + \lfloor T^{1/4} M^{-3/2} \rfloor$

Improves to $H = 1 + \lfloor T^{1/2} M^{-3/2} \rfloor$

if $\mathbb{E}\|\nabla f_m(x;\xi) - \nabla f_m(x)\|^2 \leq \sigma^2$

# Plot



1 local step
4 local steps
16 local steps
$10^{-1}$
$10^{-2}$
$10^{-3}$
$f(x) - f^*$
0
1000
2000
3000
4000
5000
Communication rounds

# Open questions

## Meta-Learning

We can learn an ”improvable” model

# Open questions

## Meta-Learning

We can learn an "improvable" model

$$\min_x \frac{1}{m} \sum_{m=1}^{M} f_m(x - \gamma \nabla f_m(x))$$

# Reference

**Better Communication Complexity**
**for Local SGD**
**arXiv:1909.04746**

**First Analysis of Local GD on Heterogeneous Data**
**arXiv:1909.04715**

**NeurIPS workshop on Federated Learning**
**http://federated-learning.org/fl-neurips-2019/**