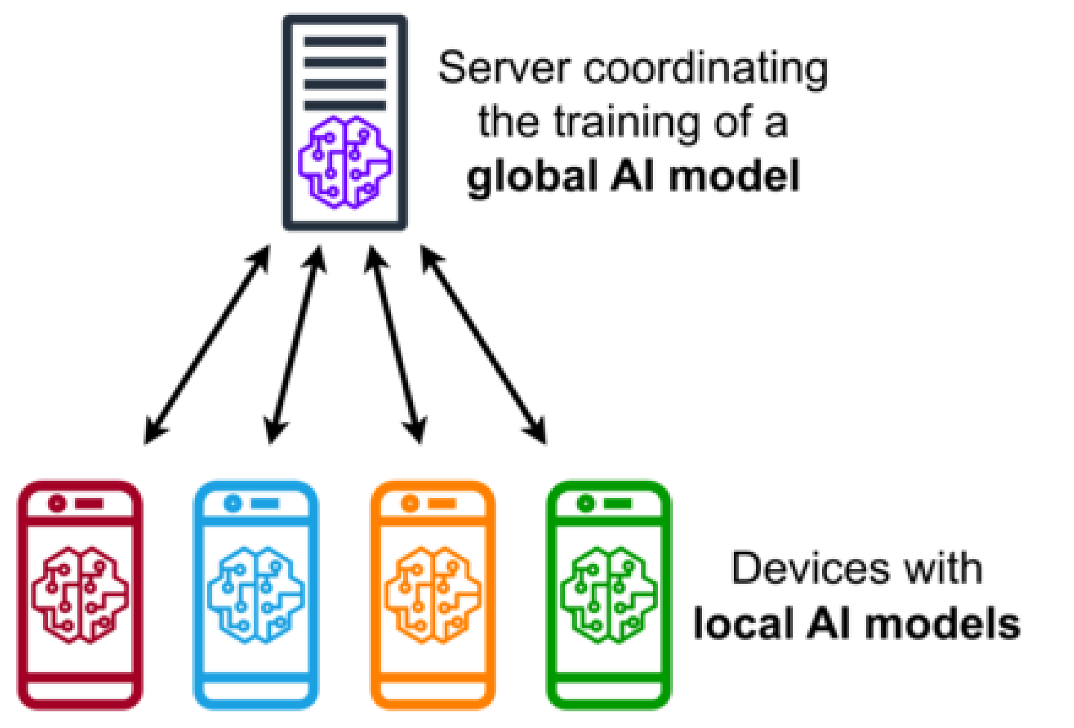
Local SGD converges **faster** for **quadratic-like** objectives and requires **less** communication.

Motivation and Challenges

- Larger models need data and tasks to be shared across many (M) devices.
- Devices calculate local stochastic gradients and transmit them to a central server.
- Transmitting large amounts of data is costly.
- Our goal: **Reduce** the number of **communication** rounds.

We denote the concept above as "Federated Learning"



Federated Learning process

Local SGD

- The most popular Federated Learning method is called **Local SGD**.
- It performs multiple local SGD steps between communications.
- Problem: if we reduce the number of communications and increase the number of local steps (H), the performance **degrades**.

Woodworth et al., 2020 noted the following: For **quadratic** objectives, Local SGD convergence rate **is not affected** by the number of local steps, making it **highly efficient** for such problems.

But what happens when we diverge from pure quadratic setting?

Thus, our aim was to establish better communication complexity rates for objectives somehow **close to the quadratic form.**

In order to measure the proximity of an objective F to the quadratic form, we decompose F into the sum: $F = Q + R$, where Q is a convex quadratic function, and R is some convex residue.

Then we introduce was *quadraticity* parameter $\varepsilon := \frac{L_R}{L} \leq 1$.

- For quadratic objectives, where F is equal to Q, **the value of ε is zero**
- For quadratic-like objectives, i.e. cases where Q is somewhat larger than R, ε is **small**.

Quadraticity concept allows us to improve over the previous lower bounds for Local SGD

Breaking existing bounds

Under the assumption of *uniformly bounded variance*, when $E\left\|\nabla F(x) - \nabla F(x,z)\right\|^2 \leq \sigma^2$:

Case $\mu = 0$, $E[F(x_T) - F(x_*)] =$

$$O\left(\frac{LD^2}{T} + \frac{\sigma D}{\sqrt{MT}} + \left(\frac{HL\sigma^2 D^4}{T^2}\right)^{1/3}\right) \quad \text{[Woodworth et al., 2020]}$$

$$\downarrow$$

$$O\left(\frac{LD^2}{T} + \frac{\sigma D}{\sqrt{MT}} + \left(\frac{\varepsilon HL\sigma^2 D^4}{T^2}\right)^{1/3}\right) \quad \text{[This work]}$$

If we denote $\lambda = \mu_Q + \mu_R$ we can also get an estimate for the case $\lambda > 0$:

$$\tilde{O}\left(\text{exp.decay} + \frac{\sigma^2}{\mu MT} + \frac{HL\sigma^2}{\mu^2 T^2}\right) \quad \text{[Woodworth et al., 2020]}$$

$$\downarrow$$

$$\tilde{O}\left(\text{exp.decay} + \frac{\sigma^2}{\lambda MT} + \frac{\varepsilon HL\sigma^2}{\lambda^2 T^2}\right) \quad \text{[This work]}$$

Abandoning restrictive assumption

If we replace uniformly bounded variance assumption with more **general** one, i.e.

$$E\left\|\nabla F(x) - \nabla F(x,z)\right\|^2 \leq \sigma^2 + \rho\left\|\nabla F(x)\right\|^2$$

the acceleration given by quadraticity **persists**.

Case $\lambda > 0$:

$$E[F(x_T) - F(x_*)] = O\left(\text{exp.decay} + \frac{\sigma^2}{\lambda MT} + \frac{\rho HL^2\sigma^2}{\lambda^3 MT^3} + \frac{\varepsilon HL\sigma^2}{\lambda^2 T^2}\right)$$

Variance reduction term

Represents the impact of $\rho\left\|\nabla F(x)\right\|^2$

Represents the ***drift*** caused by rare communication

In all the estimates above, the last term represents ***drift*** that appears due to many local steps (or rare communication, which is equivalent)

So, when it is multiplied by the ε factor it shows that **the influence of rare communications weakens for quadratic functions.**

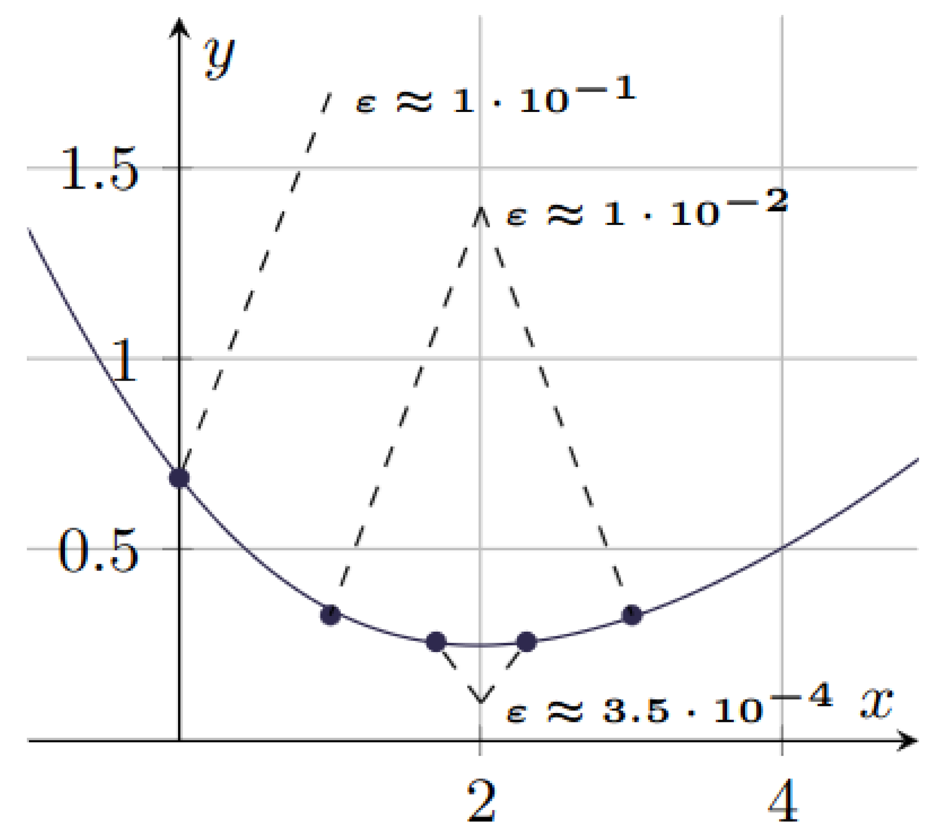
Notation

The following symbols and definitions are used throughout this work:

Symbol	Definition
M	Number of devices
H	Number of local SGD steps
T	Total number of iterations on a given device
D	Initial distance to the optimum, $\|x_0 - x_*\|$
μ	Strong convexity constant
L	Lipschitz gradient constant

Discussion

An important observation about quadraticity is that for functions with a **Lipschitz Hessian**, ε decreases rapidly, as illustrated in the graph below.



Decrease of ε for LogLoss with l_2 regularization

References

Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., & Srebro, N. (2020). Is local sgd better than minibatch sgd?

Local SGD for Near-Quadratic Problems: Improving Convergence under Unconstrained Noise Conditions

Andrey Sadchikov, Savelii Chezhegov, Aleksandr Beznosikov and Alexander Gasnikov
Moscow Institute of Physics and Technology, Moscow, Russia