ICLR, May 6–9, New Orleans

Local SGD Converges Fast and Communicates Little

sebastian.stich@epfl.ch

Summary: Theoretical Analysis of (Communication Efficient) Local SGD

Stochastic Optimization Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\Big[f(\mathbf{x}) := \mathbb{E}_\xi F(\mathbf{x},\xi)\Big] \qquad \left(f(\mathbf{x}) \overset{\text{Example}}{=} \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})\right)$$

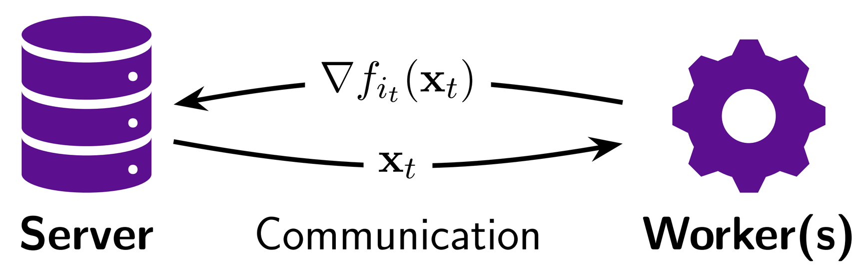
Assumptions:

- access to gradient oracles, $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^d$, s.t. $\forall \mathbf{x}\in\mathbb{R}^n$:
  $$\mathbb{E}\,\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x}), \quad \mathbb{E}\,\|\mathbf{g}\|^2 \leq G^2, \quad \operatorname{Var}\mathbf{g} \leq \sigma^2$$
- $f\colon \mathbb{R}^n \to \mathbb{R}$ μ-strongly convex, L-smooth, $\kappa := \frac{L}{\mu}$

Notation:

- B mini-batch size
- H steps of local SGD between communication rounds
- T iterations
- W parallel workers

← **Download the Paper**



Frequent communication between worker nodes (e.g. GPU's) is a **major bottleneck** for distributed training of DL models.

Local SGD (aka parallel SGD) enables models to be different on the worker nodes for a few iterations, that is, some all-to-all **communication rounds can be skipped**.

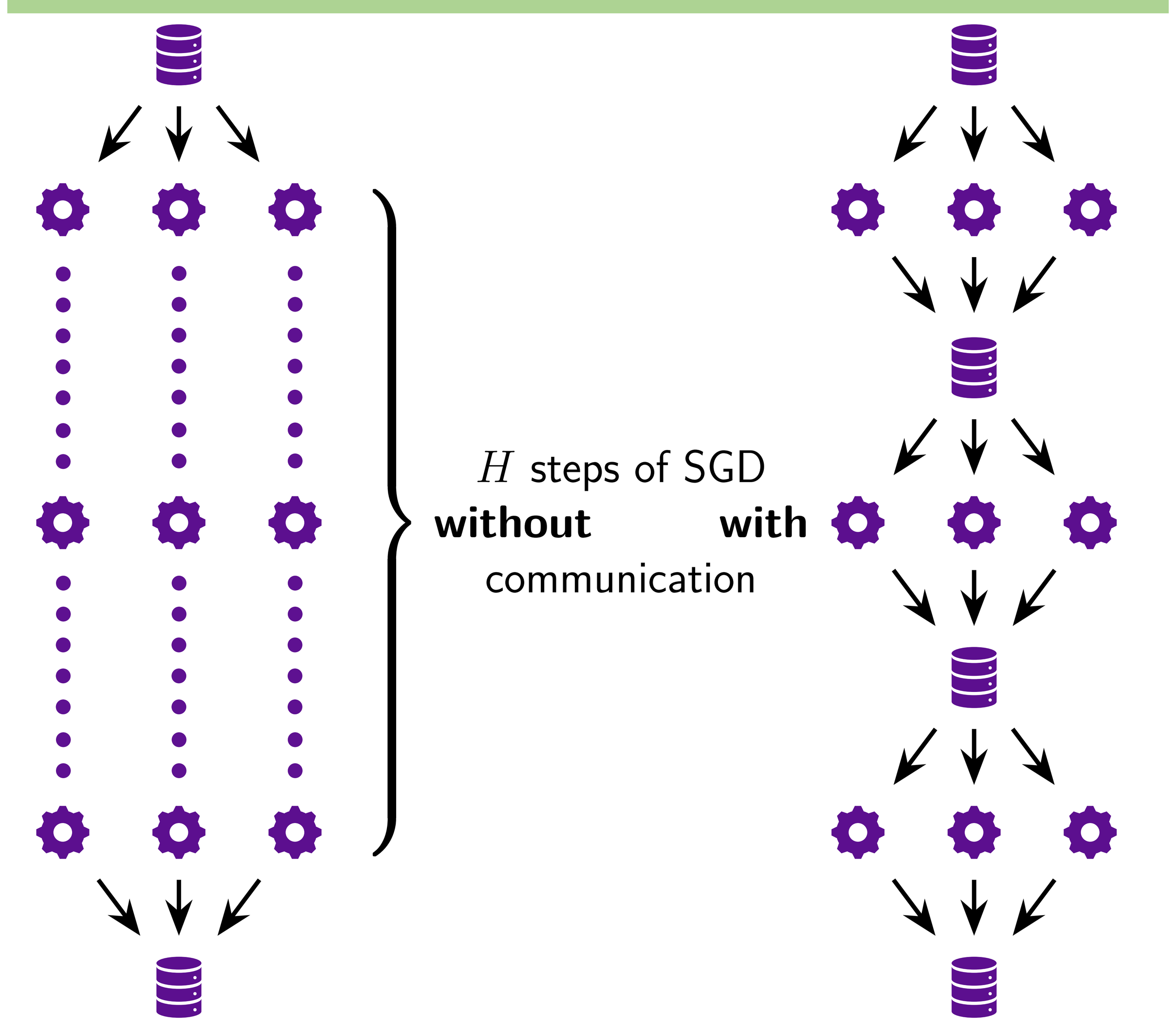
Local SGD could be an alternative to large-batch training.

We show that in the convex setting

Local SGD is as good as mini-batch SGD

while requiring fewer communication rounds.

- our technique offers a promising direction to extend the analysis to the non-convex setting in future work

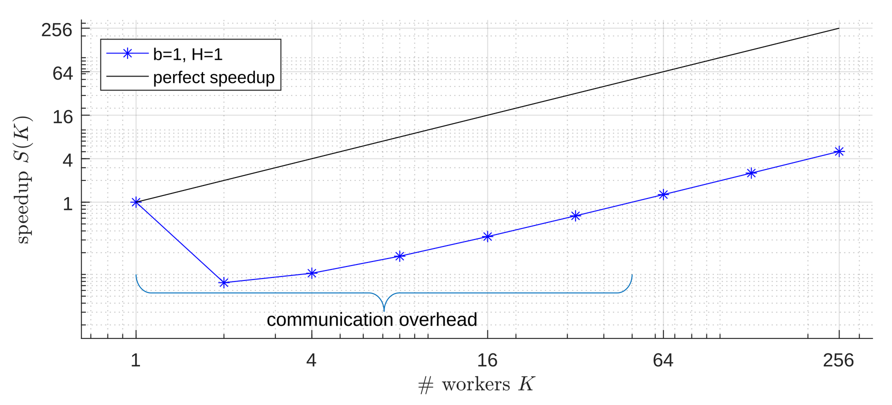


Local SGD communicates $H\times$ **less** than mini-batch SGD

Details

Illustration: Impact of high communication cost

An algorithm converging as $O\left(\frac{1}{BTW}\right)$ achieves linear speedup in terms of batch size B and number of workers W **in terms of iterations**. However, the speedup also depends on the *communication cost*.

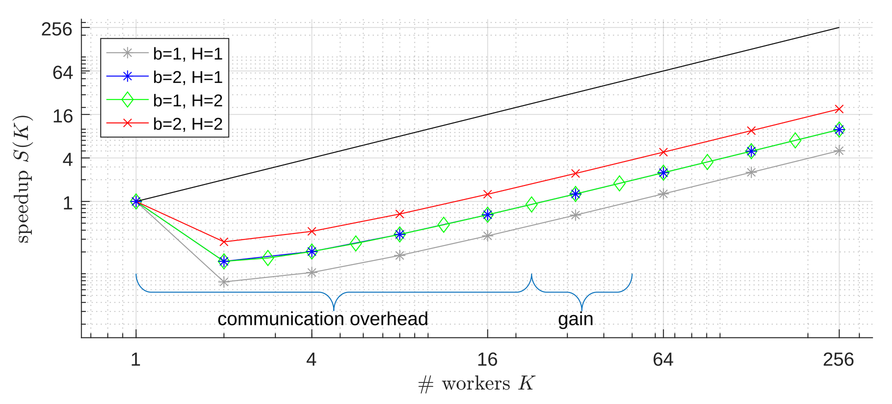


time-wise speedup, with communication cost

(assuming communication is 25× slower than computation)

This suggests two strategies:

- **increase batch size** B (mini-batch SGD)
- **increase local steps** H (local SGD)



Algorithm: Local SGD (for batch size $B = 1$)

1: Initialize variables $\mathbf{x}_0^w = \mathbf{x}_0$ on every worker $w \in [W]$
2: **for** t **in** $0 \dots T-1$ **do**
3: **parallel for** $w \in [W]$ **do**
4: Sample $\mathbf{g}_t^w = \mathbf{g}(\mathbf{x}_t^w)$
5: **if** $H \mid t+1$ **then**
6: $\mathbf{x}_{t+1}^w \leftarrow \frac{1}{W}\sum_{w=1}^W(\mathbf{x}_t^w - \eta_t \mathbf{g}_t^w)$ ▷ global synchronization
7: **else**
8: $\mathbf{x}_{t+1}^w \leftarrow \mathbf{x}_t^w - \eta_t \mathbf{g}_t^w$ ▷ local update
9: **end if**
10: **end parallel for**
11: **end for**

Special cases:

- $H = 1$: Mini-batch SGD. Communication in every round.
- $H = T$: One-shot averaging. Only one communication round at the end.

Baseline result:

Previous analyses did not show a speedup in W, the number of workers (expect for special cases).

$$f(\bar{\mathbf{x}}_T) - f^\star = O\left(\frac{\sigma^2}{\mu BT}\right) \qquad \text{(no dependence on } W\text{)}$$

Theorem:

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth, μ-strongly convex, and step-sizes $\eta_t := \frac{4}{\mu(a+t)}$ for $a \geq \max\{H, 16\kappa\}$. (technical conditions) Then

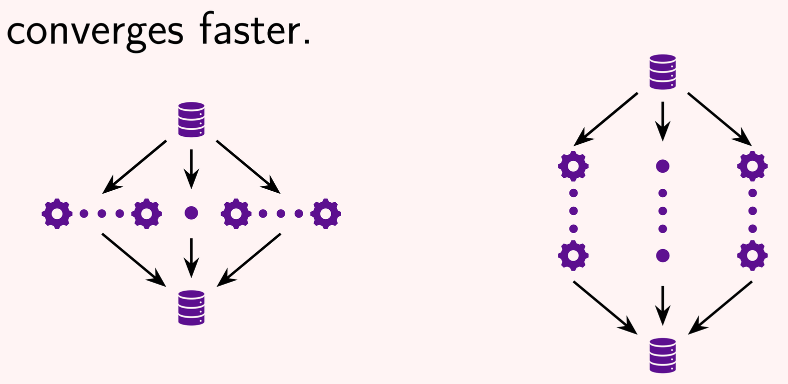
$$f(\bar{\mathbf{x}}_T) - f^\star = O\left(\frac{\sigma^2}{\mu BTW} + \frac{\kappa H^2 G^2}{\mu T^2}\right) \qquad \text{(simplified)}$$

for weighted average $\bar{\mathbf{x}}_T := \frac{1}{WS_T}\sum_{w=1}^W\sum_{t=0}^{T-1}\lambda_t \mathbf{x}_t^w$, with weights $\lambda_t = (a+t)^2$, $S_t := \sum_{t=0}^{T-1}\lambda_t$.

- for $H \leq \sqrt{\frac{T}{\kappa BW}}$ we recover the convergence rate of mini-batch SGD, i.e. linear speedup in batch size B and in number of workers W
- choosing $H = \sqrt{\frac{T}{\kappa BW}}$ reduces the communication rounds by a factor $O\Big(\sqrt{\frac{T}{\kappa BW}}\Big)$ compared to mini-batch SGD
- when the number of steps T is unknown, one could use an adaptive strategy (e.g. 'doubling trick') to successively increase the number of local steps (more communication steps for small t, less as t grows)

Discussion & Open Problems

- the result is not optimized for extreme settings of H, W, L, σ or G. For instance, we do not recover the convergence rate of SGD for $H = T$.
- the assumptions on the gradient oracle (e.g. bounded gradient assumption, unbiased on every worker) can potentially be relaxed
- the proof technique mainly leverages smoothness, allowing for extension of the results to the non-convex setting
- **huge-batch** SGD (i.e. SGD with mini-batch size BHW) converges under these assumptions with rate $O\left(\frac{\sigma^2}{\mu BTW}\right)$. This rate is strictly better than our established upper bound for local SGD. However, it is conjectured than local SGD converges faster.



two algorithms with the same computation and communication cost, which one is faster?

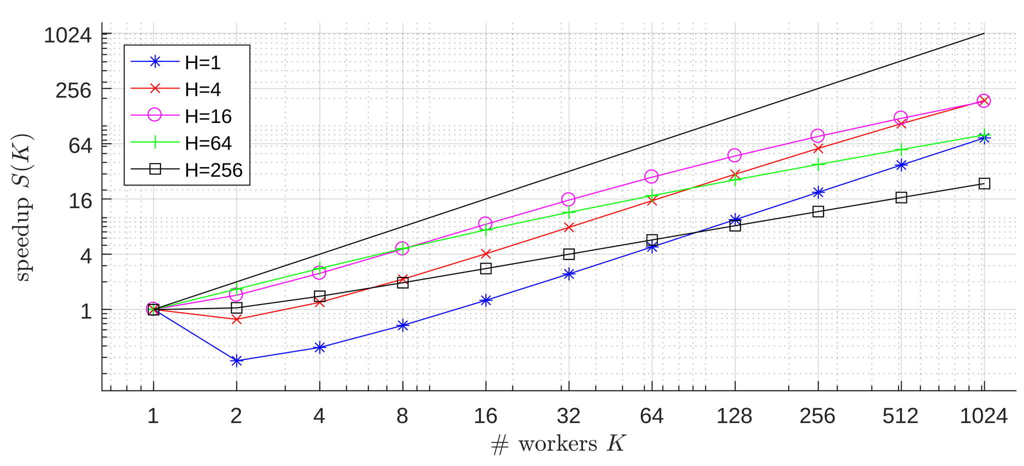
- recent work showed limitations of huge batch training. Local SGD could be promising direction (as the local mini-batches are considerably smaller). However, the current analysis does not resolve this.

Experiments

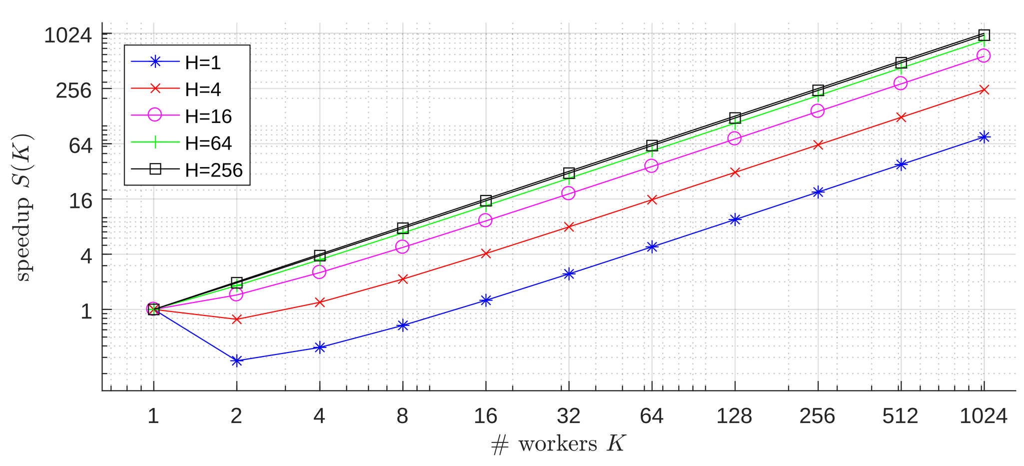
Logistic regression: for w8a dataset ($d = 300$, $n = 49749$).

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + e^{-b_i \mathbf{a}_i^\top \mathbf{x}}\right) + \frac{1}{2n}\|\mathbf{x}\|^2$$

theoretical speedup of local SGD for different H and number of workers W

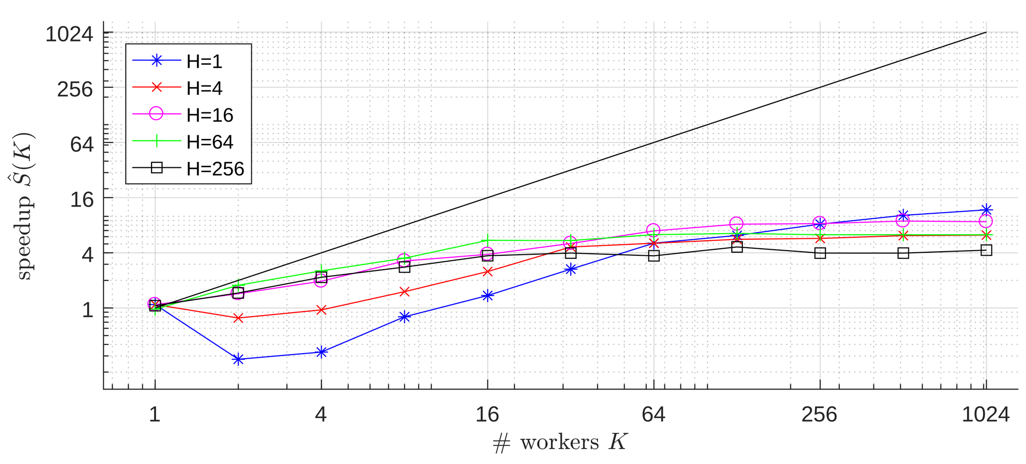


($\epsilon > 0$, T small)

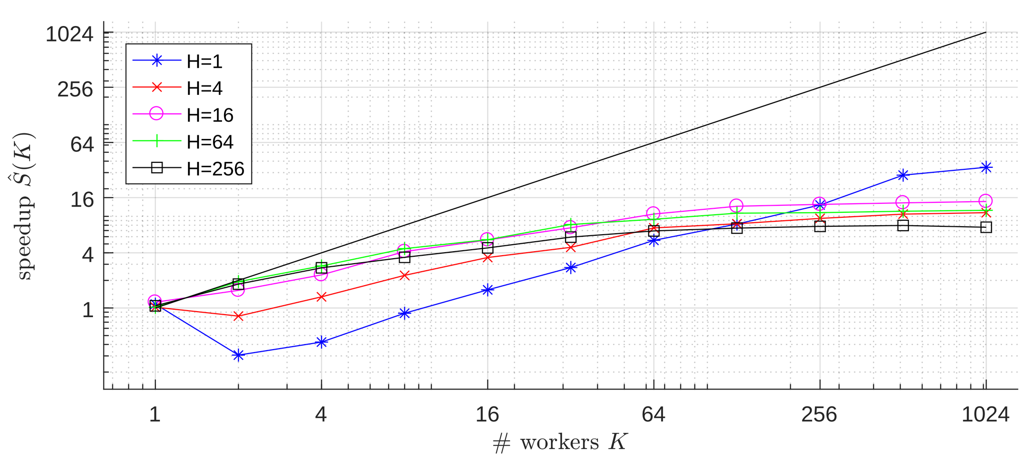


($\epsilon \to 0$, $T \to \infty$)

measured speedup of local SGD, $B = 4$



($\epsilon = 0.005$)



($\epsilon = 0.0001$)