ICLR

# A Quadratic Synchronization Rule for Distributed Deep Learning


Xinran Gu*[,1] Kaifeng Lyu*[,4] Sanjeev Arora[4] Jingzhao Zhang[1,2,3] Longbo Huang[1]

[1]Tsinghua University, [2]Shanghai Qizhi Institute, [3]Shanghai AI Laboratory, [4]Princeton University


## Overview

Local gradient methods, e.g., Local SGD, improve the communication efficiency of data parallel training by letting workers communicate only every $H$ steps.

**Key question: how to set the synchronization period H?**

**Why hard?**

- Optimization theory: larger $H \Rightarrow$ slower convergence, communication & convergence tradeoff (Stich, 2018; Yu-Yang-Zhu, 2018)
- But for **modern neural nets**
  - same train loss $\not\Rightarrow$ same test loss
  - In some cases, increase $H \Rightarrow$ higher test acc. (Lin et al., 2020)

**Main contribution: a theory-grounded strategy to set $H$!**

**Quadratic Synchronization Rule (QSR)**

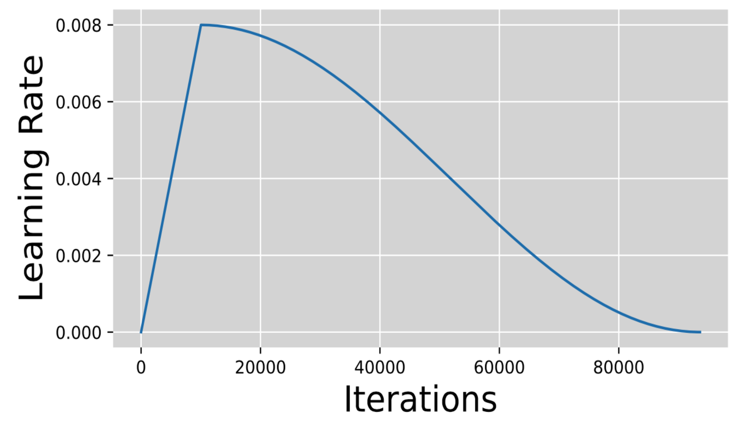
$$H^{(s)} = \max\{H_{\text{base}}, \lfloor(\alpha/\eta_t)^2\rfloor\}$$

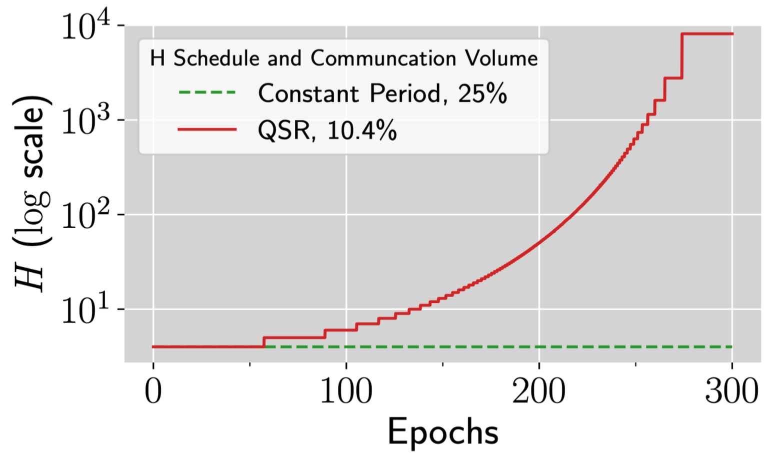
- improve generalization by quadratically scaling $H$ as LR decays
- save communication simultaneously

| | time | val. acc. |
|---|---|---|
| data parallel | 26.7h | 79.86% |
| QSR | **20.2h** ↓ | **80.98%** ↑ |

Setting: Local AdamW, 300 epochs, ViT-B, ImageNet

**save 7h, improve 1%**




## Background: Local Gradient Methods

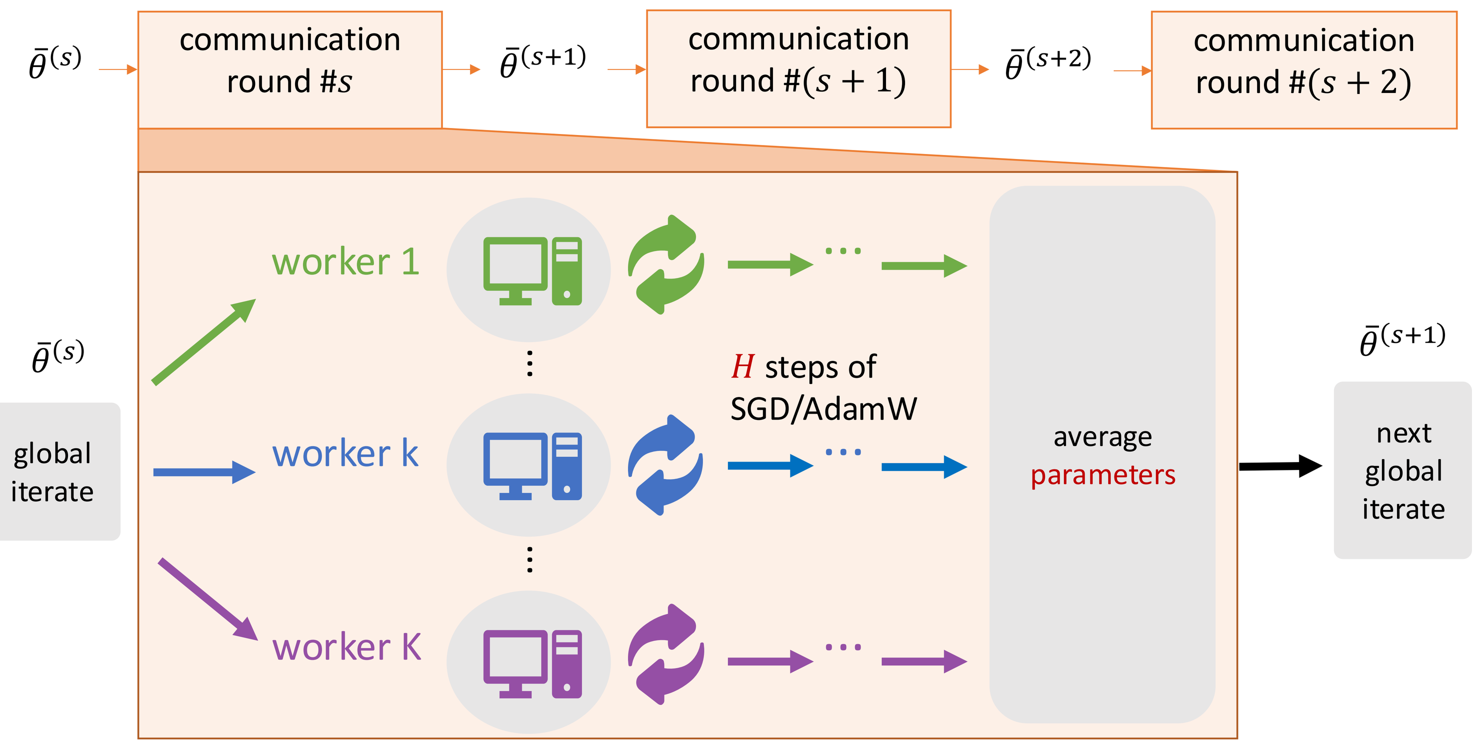
**Data parallel training**

- Distribute gradient computation on $B$ samples to $K$ workers
- Each iteration, each worker: 1. compute gradients on $B/K$ samples; 2. average gradients via All-Reduce; 3. update using the averaged gradient & optimizer OPT

Issue: high comm. cost due to frequent synchronization

**Local gradient methods**

- Worker locally updates own replica with OPT
- Average model params. every $H$ steps



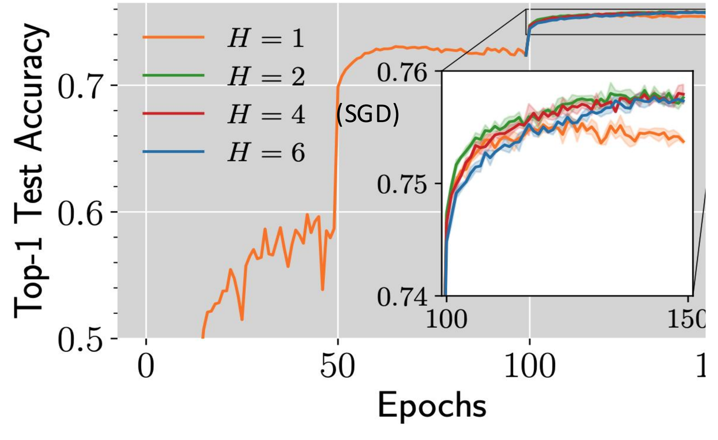

## Controversy on the Generalization Benefit of Local SGD

**Local steps improve generalization** (Lin et al., 2020, Fig. (a))

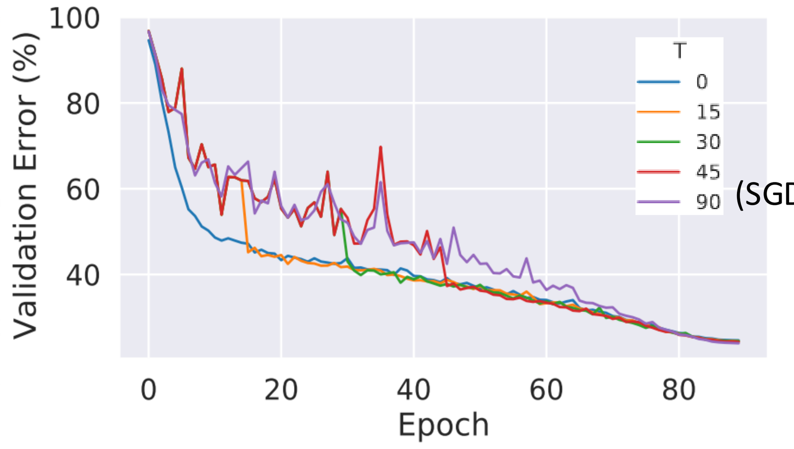
- Run #1: Parallel SGD ($\equiv$ Local SGD with $H=1$)
- Run #2: #1 + switch to Local SGD with $H>1$ at some epoch $t_0$ (Post-local SGD)
- Result: test acc. #2 > #1

**The improvement seems only short term** (Ortiz et al., 2021, Fig. (b))

- For cos LR decay, the generalization benefit appears only shortly after switching




(a) Constant LR after switching




(b) Cosine LR decay (Ortiz et al., 2021)

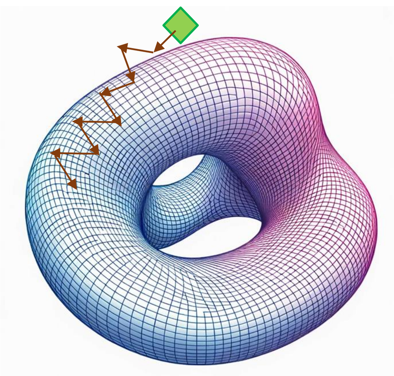
ImageNet, ResNet-50

Hint: the generalization benefit has something to do with LR

## Theory: Why does Local SGD Generalize Better?

**Setup** (Follow Blanc et al., 2020; Damian et al., 2021; Li et al., 2022)

Assume (1) a minimizer manifold $\Gamma$; (2) a small LR $\eta$; (3) Analyze dynamics of (Local) SGD near $\Gamma$

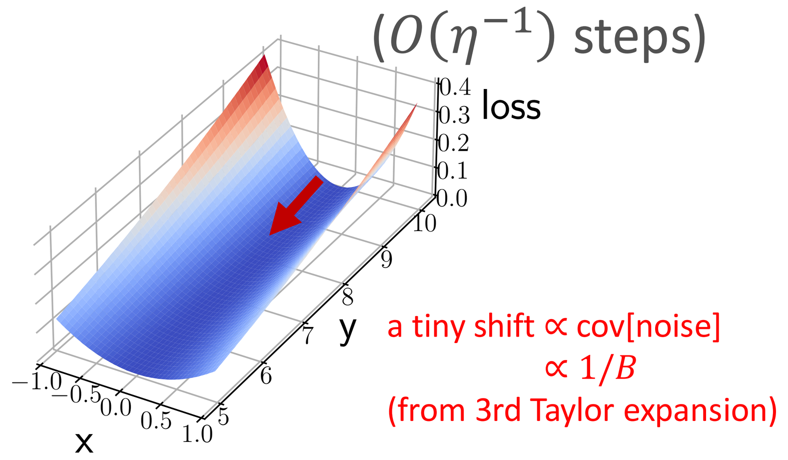


**Fast and slow dynamics in SGD**

(Blanc et al., 2020; Damian et al., 2021; Li et al., 2022)

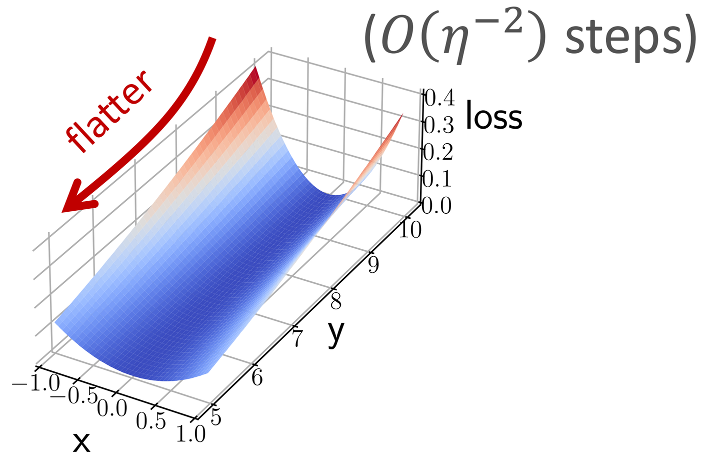
**Fast Dynamics (short term)**

Diffuse locally near a minimizer ($O(\eta^{-1})$ steps)

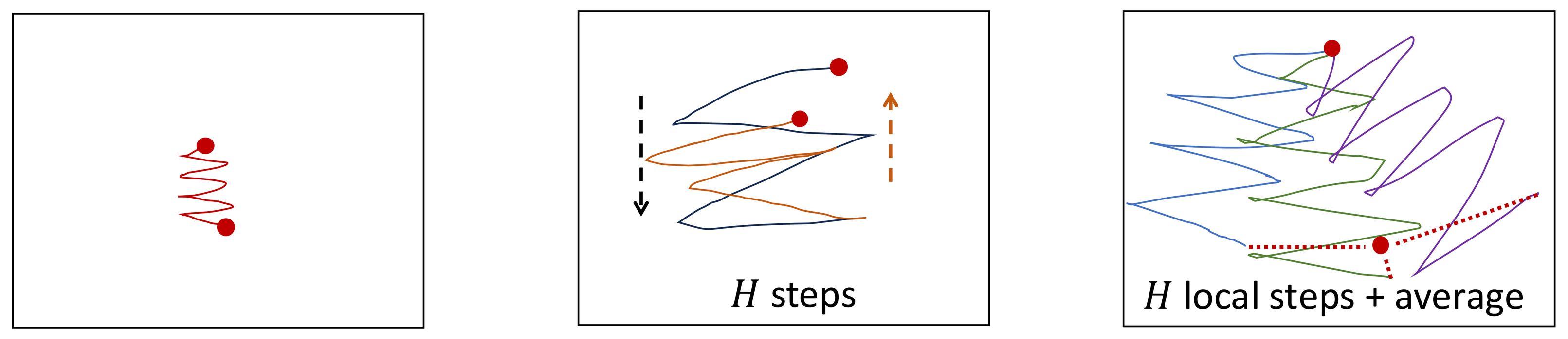



**Slow Dynamics (long term)**

"Center" of the diffusion shifts ($O(\eta^{-2})$ steps)




**Local SGD drifts faster to flatter minima**




- SGD, batch size $B$: drift slowly
- SGD, batch size $B/K$: drift fast in expectation, but go back and forth (large var.)
- Local SGD, $B_{\text{loc}} = B/K$: drift fast in expectation, averaging reduces var.

**SDE approximations for different scalings of $H$**

**Theorem (informal).** For $O(\eta^{-2})$ steps, Local SGD can be approximated by the following SDEs on $\Gamma$:

1. $H = \beta/\eta$ (Gu et al., 2023)

$$d\zeta(t) = P_\zeta\Big(\tfrac{1}{\sqrt{B}}\Sigma_\parallel^{1/2}(\zeta)dW_t - \tfrac{1}{2B}\nabla^3\mathcal{L}(\zeta)[\hat\Sigma_\diamond(\zeta)]dt - \tfrac{K-1}{2B}\nabla^3\mathcal{L}(\zeta)[\hat\Psi(\zeta)]dt\Big)$$

Same as SGD (Li et al., 2022) — Unique drift term of Local SGD

- larger ⇒ stronger implicit bias
- increases with $H$; → 0 as $H\eta \to 0$; → $\hat\Sigma_\diamond(\zeta)$ as $H\eta\to\infty$

**Remark:** (1) $H$ should be at least $\eta^{-1}$ to see the benefit (2) stronger implicit bias for larger $H$ (3) but also higher approximation error for larger $H$ (valid for $o(\eta^{-2})$, fails for $\omega(\eta^{-2})$)

2. $H = (\alpha/\eta)^2$ (our new result)

$$d\zeta(t) = P_\zeta\Big(\tfrac{1}{\sqrt{B}}\Sigma_\parallel^{1/2}(\zeta)dW(t) - \tfrac{K}{2B}\nabla^3\mathcal{L}(\zeta)[\hat\Sigma_\diamond(\zeta)]dt\Big)$$

$K$ times of SGD; Local SGD with $H = \beta/\eta$ when $\beta\to\infty$

What about $H \sim \eta^{-2}$

**$H\sim\eta^{-1}$ to see the benefit, $H\sim\eta^{-2}$ to maximize it!**

(also tried $H\sim\eta^{-3}$, worse than QSR)

## Our Roadmap

- **Goal**: find the $H$ schedule to maximize test acc.
- **Theory**: understanding how the generalization benefit arises
- **Practical guidance**: QSR ($H\sim\eta^{-2}$)
- **Empirical validation**: Local SGD & AdamW