A Quadratic Synchronization Rule for Distributed Deep Learning

Xinran Gu*[,1] Kaifeng Lyu*[,2] Sanjeev Arora[2] Jingzhao Zhang[1] Longbo Huang[1]
[1]Tsinghua University [2]Princeton University

Abstract

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

- **How to set the synchronization period H?**

- Optimization: communication & convergence tradeoff
- Generalization: proper $H \Rightarrow$ higher test acc.(Lin et al., 2020)

- We propose a theory-grounded strategy to set H

Quadratic Synchronization Rule (QSR)

$\boldsymbol{H \sim \eta^{-2}}$ (η: learning rate)

Improve comm. efficiency & test acc. simultaneously!

	time	val. acc.
data parallel	26.7h	79.86%
QSR	**20.2h**	**80.98%**
	save 7h, improve 1%	

Setting: 300 epoch on ViT-B, ImageNet

Background: Local Gradient Methods

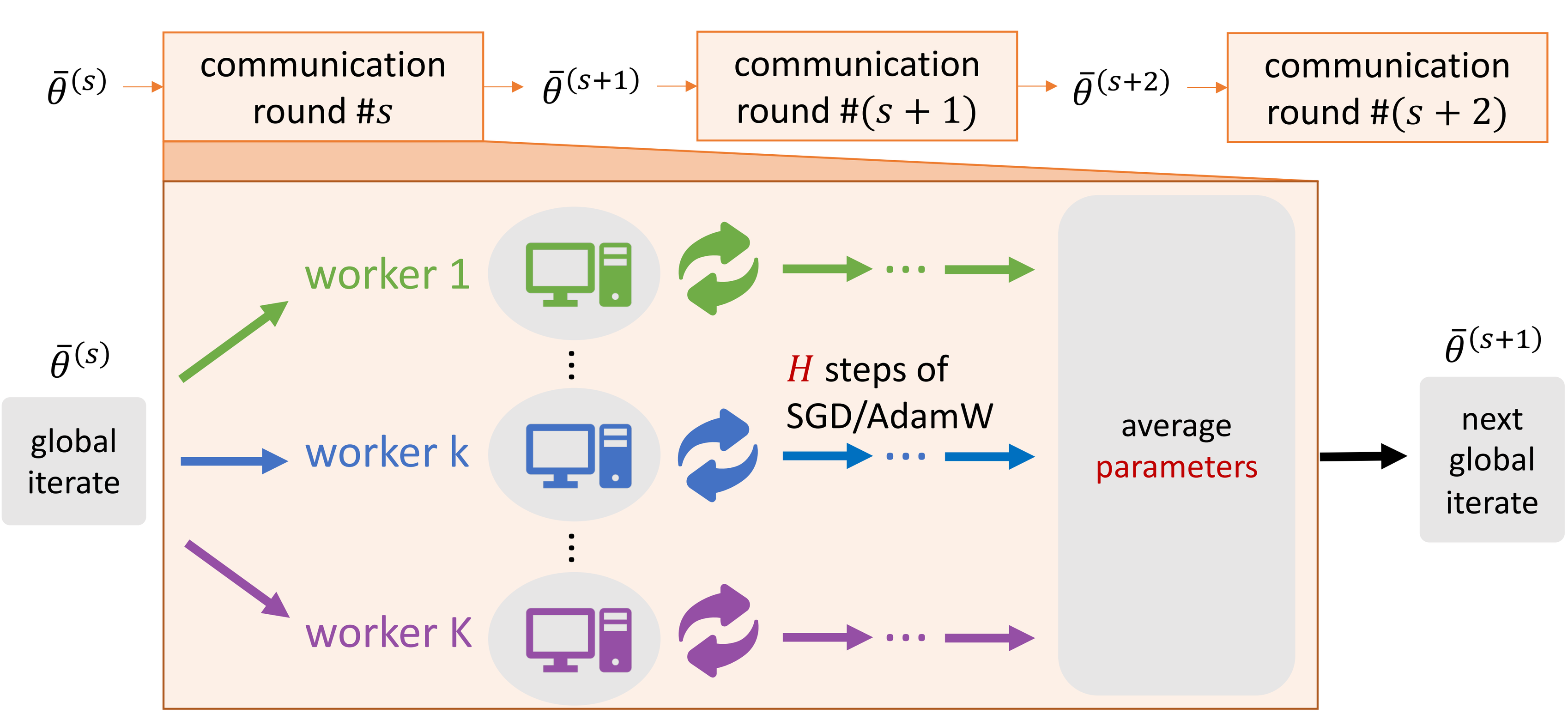
- **Data parallel approach**

- 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: frequent sync. $\Rightarrow$ high comm. cost

- **Local gradient methods**

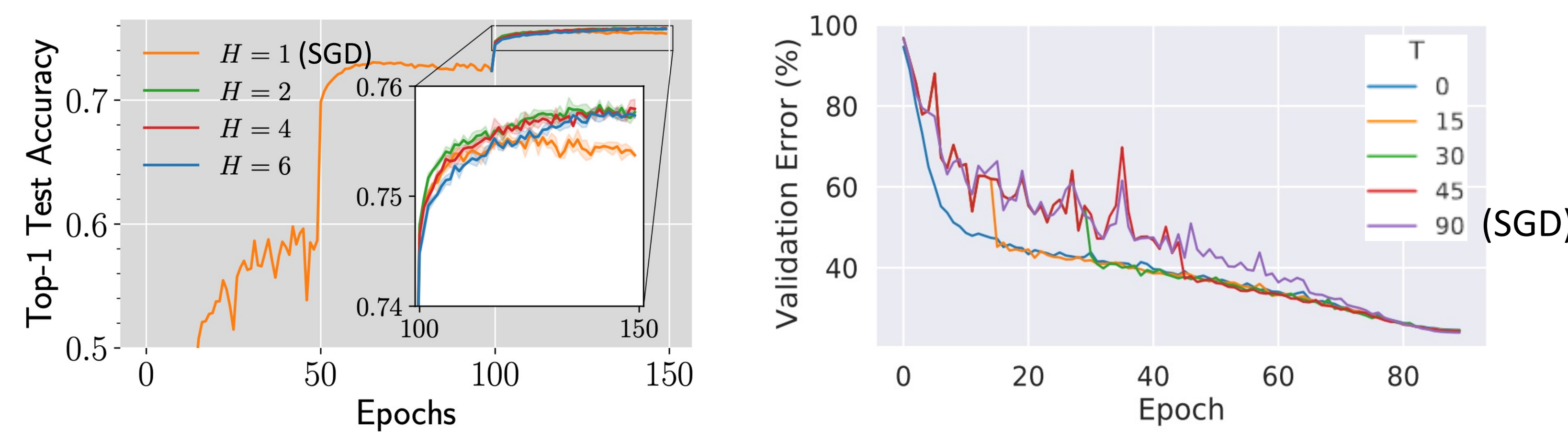
- Each worker locally updates its own replica with OPT
- Average model parameters every H steps



Generalization Benefits of Local SGD

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

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



(a) Constant LR after switching (b) Cosine LR decay (Ortiz et al., 2021)

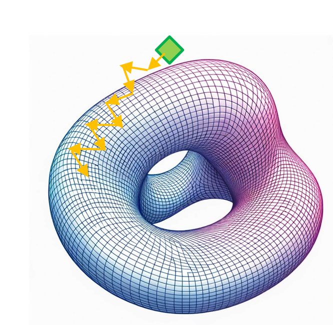
ImageNet, ResNet-50

Issue: short-term generalization benefits on cos decay (Ortiz et al., 2021)

Theory: Why does Local SGD Generalize Better?

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

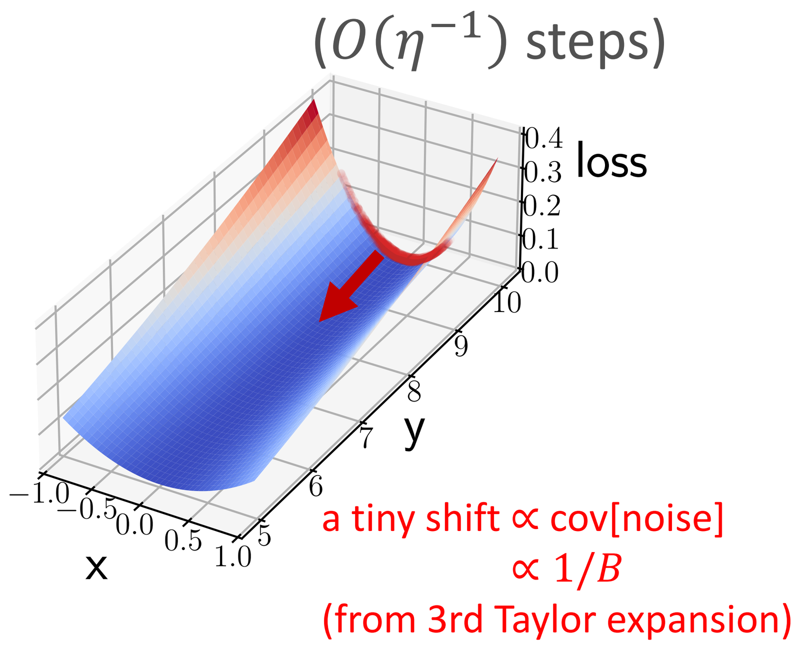
- Assume a minimizer manifold Γ
- Assume a smaller LR η
- Analyze dynamics of (Local) SGD near Γ



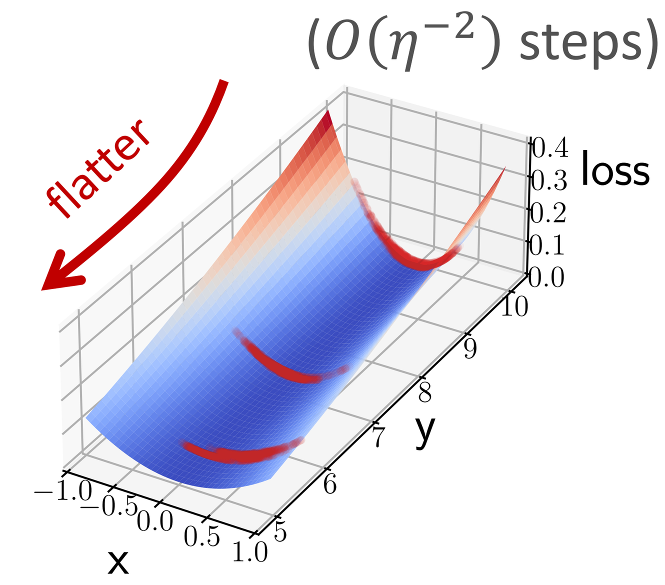
- **Fast and slow dynamics in SGD**

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

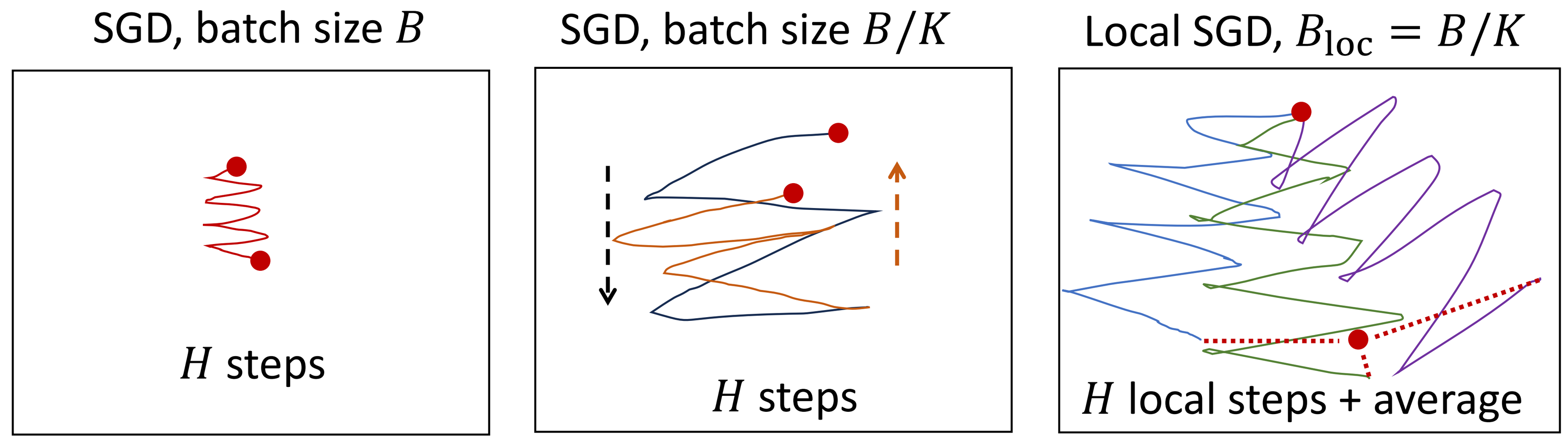
Fast Dynamics (short term)
Diffuse locally near a minimizer
($O(\eta^{-1})$ steps)



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



- **Local SGD drifts faster to flatter minima**



drift slowly | drift fast in expectation, but go back and forth (large var.) | drift fast in expectation, averaging reduces var.

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

Theorem (informal). For $O(\eta^{-2})$ steps, Local SGD with different scalings of H can be approximated by the following SDEs on Γ:

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

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\underbrace{\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}_t - \tfrac{1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{Same as SGD (Li et al., 2022)}} \underbrace{- \tfrac{K-1}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Psi}}(\boldsymbol{\zeta})]\mathrm{d}t}_{\text{Unique drift term of Local SGD}} \Big)$$

- $\widehat{\Psi}(\zeta)$ increases with H, goes to 0 as $H\eta \to 0$ and goes to $\hat{\Sigma}_{\diamond}(\zeta)$ as $H\eta \to \infty$

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

$$\mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta}}\Big(\tfrac{1}{\sqrt{B}}\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta})\mathrm{d}\boldsymbol{W}(t) \underbrace{- \tfrac{K}{2B}\nabla^3\mathcal{L}(\boldsymbol{\zeta})[\widehat{\boldsymbol{\Sigma}}_{\Diamond}(\boldsymbol{\zeta})]\mathrm{d}t}_{K\text{ times of SGD; Local SGD with } H = \beta/\eta \text{ when } \beta \to \infty} \Big)$$

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

Cannot find valid SDE approximation on the manifold for more aggressive scalings.