A Quadratic Synchronization Rule for Distributed Deep Learning

Xinran Gu*,1 Kaifeng Lyu*,2 Sanjeev Arora2 Jingzhao Zhang1 Longbo Huang1

1Tsinghua University 2Princeton 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)
  
  $H \sim \eta^{-2}$ ($\eta$: learning rate)

  Improve comm. efficiency & test acc. simultaneously!

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

  $\bar{g}(0)$
  communication round $0$
  $\bar{g}(t+1)$
  communication round $t+1$
  $\bar{g}(t+2)$
  communication round $t+2$
  $\bar{g}(T)$
  next global iteration

  worker 1
  $H$ steps of SGD/AdamW
  $\bar{g}(t+1)$
  $H$ steps of SGD/AdamW
  $\bar{g}(T)$
  next global iteration

  worker $k$
  $H$ steps of SGD/AdamW
  $\bar{g}(t+1)$
  $H$ steps of SGD/AdamW
  $\bar{g}(T)$
  next global iteration

  worker $K$
  $H$ steps of SGD/AdamW
  $\bar{g}(t+1)$
  $H$ steps of SGD/AdamW
  $\bar{g}(T)$
  next global iteration

  average parameters

  global iteration

Generalization Benefits of Local SGD

- Local steps improve generalization (Lin et al., 2020)
  - Run #1: Parallel SGD ($H = 1$)
  - 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

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 $\Gamma$
  - Analyze dynamics of (Local) SGD near $\Gamma$

  Fast and slow dynamics in SGD
  
  - Fast Dynamics (short term)
    - Diffuse locally near a minimizer ($O(\eta^{-1})$ steps)
    - a tiny shift $\propto \text{cov}(\text{noise}) = 1/\delta t$ (from 3rd Taylor expansion)
  
  - Slow Dynamics (long term)
    - “Center” of the diffusion shifts ($O(\eta^{-2})$ steps)

  SDE approximations for different scalings of $H$

  **Theorem (informal).** For $O(\eta^{-1})$ steps, Local SGD with different scalings of $H$ can be approximated by the following SDEs on $\Gamma$:

  1. $H = \beta/\eta$ (Gu et al., 2023)
    
    $\mathrm{d} q(t) = P_{\eta}(\beta) \left( \frac{1}{\beta} \sum_{i=1}^{N} \xi_i \frac{\nabla q(t)}{\nabla q(t)} \right) \mathrm{d} t - \frac{k-1}{2} \nabla q(t) \nabla q(t) \mathrm{d} t - \frac{\xi_i}{\beta} \nabla q(t) \nabla q(t) \mathrm{d} t$

    - Same as SGD (Li et al., 2022)
    - Unique drift term of Local SGD $\Delta q(t)$ increases with $H$, goes to 0 as $H \to 0$ and goes to $E_{\lambda}(\xi)$ as $H \to \infty$

  2. $H = (\alpha/\gamma)^2$ (our new result)
    
    $\mathrm{d} q(t) = P_{\eta}(\beta) \left( \frac{1}{\beta} \sum_{i=1}^{N} \xi_i \frac{\nabla q(t)}{\nabla q(t)} \right) \mathrm{d} t - \frac{k-1}{2} \nabla q(t) \nabla q(t) \mathrm{d} t - \frac{\xi_i}{\beta} \nabla q(t) \nabla q(t) \mathrm{d} t$

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

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

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

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