Local SGD with Periodic Averaging: Tighter Analysis and Adaptive Synchronization

Farzin Haddadpour   Mohammad Mahdi Kamani   Mehrdad Mahdavi   Viveck Cadambe

Pennsylvania State University

Background

- **Goal**: $\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) \triangleq \Sigma_i f_i(\mathbf{x})$
- **Main tool**: SGD $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\tilde{\mathbf{g}}(\mathbf{x}^{(t)}, \xi^{(t)})$, $\tilde{\mathbf{g}}(\mathbf{x}^{(t)}, \xi^{(t)})$: gradient observed over mini-batch $\xi^{(t)}$.
- **Challenge**: Higher computational cost

☑ Solution: Parallelization of SGD

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\frac{1}{p}\Sigma_{j=1}^{p}\mathbf{g}(\mathbf{x}^{(t)}, \xi_j^{(t)})$$

ξ_j: Mini-batch sampled from worker j, p: # workers

- **Challenge**: Communication cost is bottleneck due to *data exchange per iterations* and *communication rounds (# iterations)*.

☑ Solution: Reduce communication rounds by **Local SGD with periodic averaging**.

- **Question**: How many communication rounds do we need?

Local SGD

➤Update rule:

$$\mathbf{x}_j^{(t+1)} = \begin{cases} \frac{1}{p}\Sigma_{j=1}^{p}\left(\mathbf{x}_j^{(t)} - \tilde{\mathbf{g}}_j^{(t)}\right) & \text{if } \tau | T, \\ \mathbf{x}_j^{(t)} - \tilde{\mathbf{g}}_j^{(t)} & \text{Otherwise,} \end{cases}$$

$\mathbf{x}_j$: Model at worker j, $\tilde{\mathbf{g}}_j$: observed gradient over random mini-batch by worker j.

Assumptions

- Unbiased estimation: $\mathbb{E}[\tilde{\mathbf{g}}_j] = \mathbf{g}_j$
- Bounded variance:

$$\mathbb{E}_{\xi_j}\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}_j\|^2\right] \leq C_1\|\mathbf{g}_j\|^2 + \frac{\sigma^2}{B}$$

- L-smoothness :

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

- μ-Polyak-Łojasiewicz (PL) :

$$\frac{1}{2}\|\nabla F(\mathbf{x})\|_2^2 \geq \mu(F(\mathbf{x}) - F(\mathbf{x}^*)), \ \forall \mathbf{x} \in \mathbb{R}^d$$

☞ PL condition is a generalization of strong convexity, meaning, μ-strong convexity implies μ-PL condition.

Local SGD: Convergence analysis

Table: Comparison of different local-SGD with periodic averaging based algorithms.

Strategy	Convergence Rate	Communication Rounds (T/τ)	Extra Assumption	Setting
Yu et al. (2019)	$\mathcal{O}\left(\frac{G^2}{\sqrt{pT}}\right)$	$\mathcal{O}\left(p^{\frac{3}{4}}T^{\frac{3}{4}}\right)$	Bounded Gradients	Non-convex
Wang and Joshi (2018)	$\mathcal{O}\left(\frac{1}{\sqrt{pT}}\right)$	$\mathcal{O}\left(p^{\frac{3}{2}}T^{\frac{1}{2}}\right)$	No	Non-convex
Stich (2019)	$\mathcal{O}\left(\frac{G^2}{pT}\right)$	$\mathcal{O}\left(p^{\frac{1}{2}}T^{\frac{1}{2}}\right)$	Bounded Gradients	Strongly Convex
This Paper	$\boldsymbol{\mathcal{O}\left(\frac{1}{pT}\right)}$	$\boldsymbol{\mathcal{O}\left(p^{\frac{1}{3}}T^{\frac{1}{3}}\right)}$	**No**	**Non-convex under PL Condition**

Questions

➤ Can we further improve the number of communication rounds from $\mathcal{O}\left((pT)^{\frac{1}{3}}\right)$?

➡ Answer: Yes, using Adaptive Synchronization.

➤ What is the motivation behind adaptive synchronization?

➡ Answer: Getting closer to the optimal point ➪ local solutions are getting closer to each other ➪ less number of communication rounds is needed.

➤ How to capture this in the convergence analysis?

➡ Answer: $\tau = \mathcal{O}\left(T^{\frac{2}{3}}/p^{\frac{1}{3}}\left[F(\bar{\mathbf{x}}^{(0)}) - F^*\right]^{\frac{1}{3}}\right)$

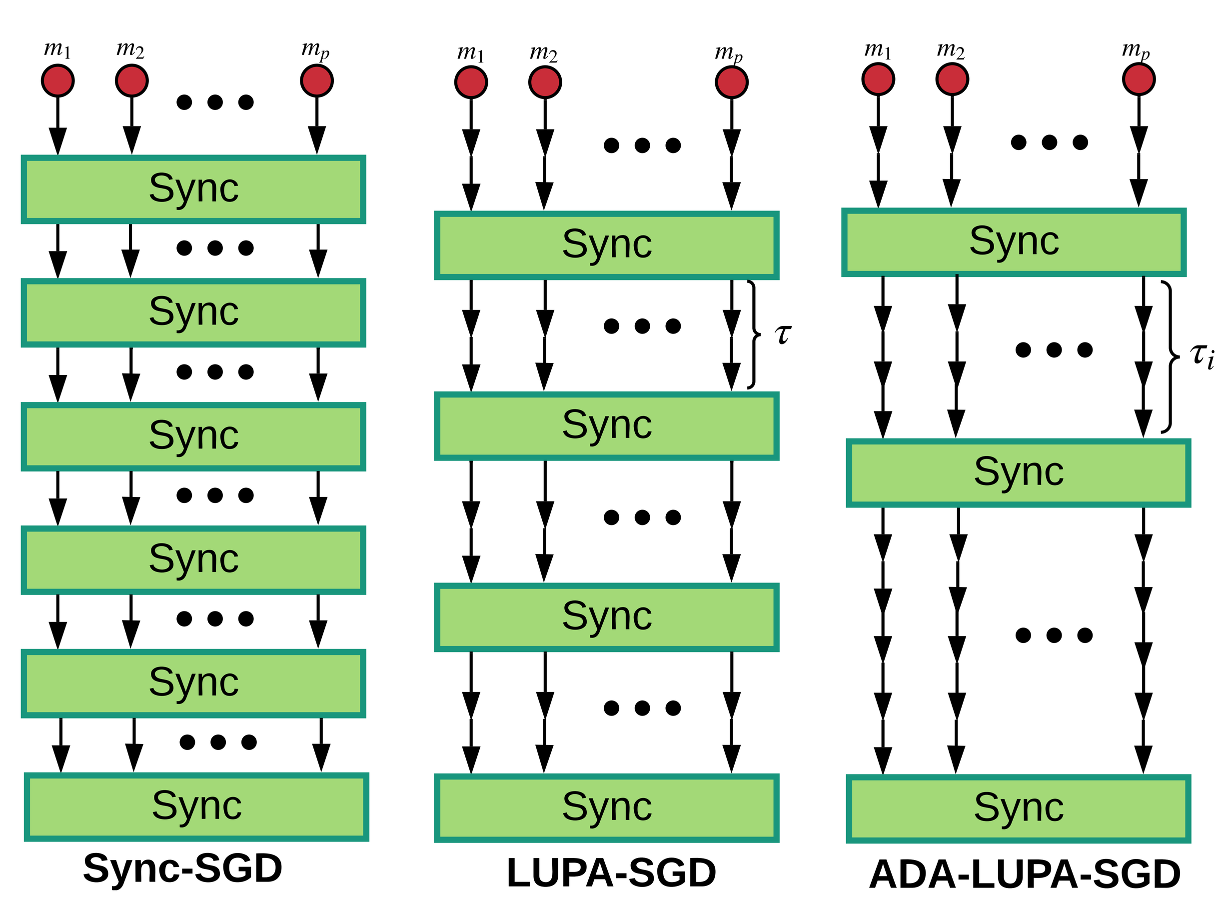
➤ A suggestion for adaptive τ?

➡ Answer: $\tau_i = \left\lceil\left(\frac{F(\bar{\mathbf{x}}^{(0)})}{F(\bar{\mathbf{x}}^{(i\tau_0)})}\right)^{\frac{1}{3}}\right\rceil\tau_0$, for ith communication round, starting with τ_0.

➤ Convergence rate with this adaptive τ?

➡ Answer: The same rate if τ_i satisfies: i) $\Sigma_{i=1}^{E}\tau_i = T$, ii) $\Sigma_{i=1}^{E}\tau_i(\tau_i - 1) = \mathcal{O}(T^2)$, iii) $(\max_{1\leq i\leq E}\tau_i)^3 = \mathcal{O}\left(\frac{T^2}{pB}\right)$, with E is the number of communication rounds.

Synchronous, Local, and Adaptive Local SGD



LUPA-SGD vs. Sync-SGD

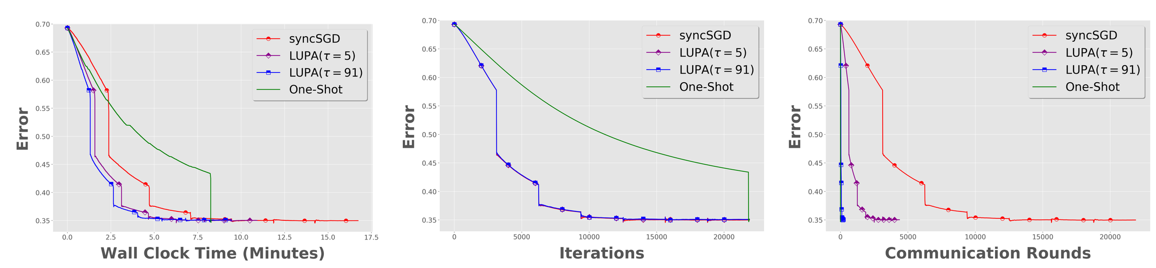


Figure: Comparison of the convergence rate of Sync-SGD with LUPA-SGD with $\tau = 5$ (Stich, 2019), $\tau = 91$ (ours) and one-shot (with only one communication round).

Number of Machines

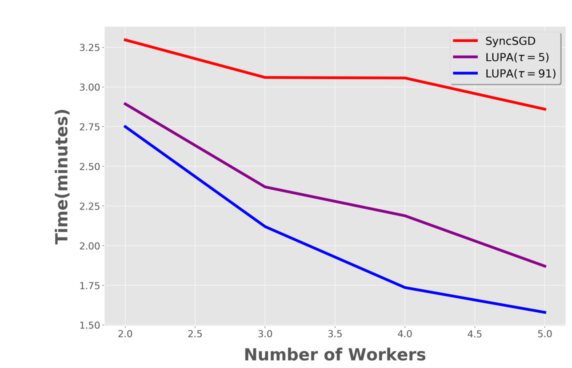


Figure: Changing the number of machines and calculate time to reach certain level of error rate ($\epsilon = 0.35$).

Adaptive Synchronization

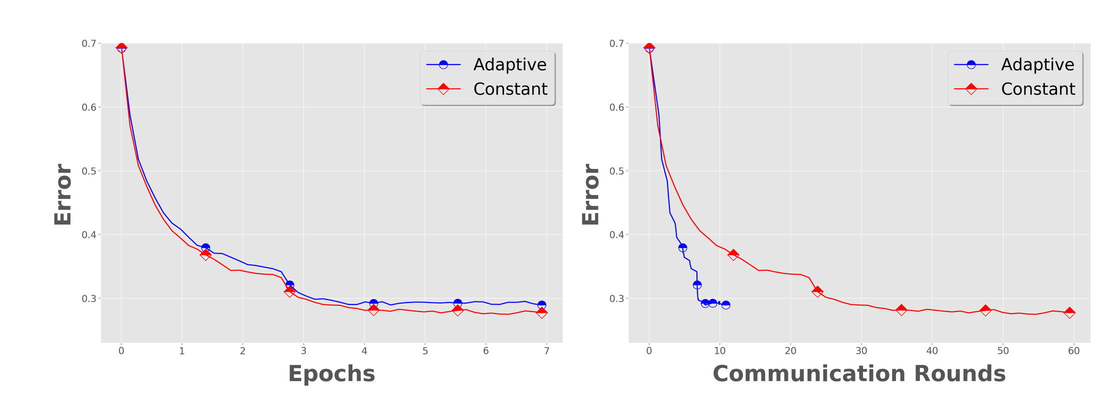


Figure: Comparison of the convergence rate of LUPA-SGD with ADA-LUPA-SGD. $\tau = 91$ for LUPA-SGD, and $\tau_0 = 91$ and $\tau_i = (1 + i\alpha)\tau_0$, with $\alpha = 1.09$ for ADA-LUPA-SGD to have 10 rounds of communication.

References

Sebastian Urban Stich. Local sgd converges fast and communicates little. In *ICLR 2019 ICLR 2019 International Conference on Learning Representations*, number CONF, 2019.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.