Communication trade-offs for synchronized distributed SGD (Local-SGD) with large step size

Kumar Kshitij Patel[1,3] **and Aymeric Dieuleveut**[1,2]

[1] MLO Team, EPFL, Lausanne; [2]CMAP, Polytechnique, France; [3] TTIC, Chicago

Local SGD

- **Stochastic gradient descent**: popular method, very important in ML.
- **Large steps size and averaging** achieve optimal performance for smooth and strongly convex function
- **Distributed setting**: very important and popular today.

Goal: minimize F smooth.
Setting:

- P machines, each of them running SGD.
- C the number of communication steps.
- between two communication rounds (*phase*) $t \in [C]$, for any worker $p \in [P]$, we perform N^t *local steps* of SGD.

Algorithm:

- **Initialisation:** All machines initially start from the same point: for any $p \in [P]$, $\boldsymbol{w}_{p,0}^1 = \boldsymbol{w}_0$.
- $\boldsymbol{w}_{p,k}^t$ the model proposed by worker p, at phase t, after k local iteration
- **Local-iterations:** for any $p \in [P], t \in [C], k \in [N^t]$:

$$\boldsymbol{w}_{p,k}^t = \boldsymbol{w}_{p,k-1}^t - \eta_k^t g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t). \tag{1}$$

- **Aggregation steps:** averaging the final local iterates of a phase:$t \in [C]$, $\hat{\boldsymbol{w}}^t = \frac{1}{P}\sum_{p=1}^P \boldsymbol{w}_{p,N^t}^t$.
- **Restart point:** every worker $p \in [P]$ restarts from the averaged model: $\boldsymbol{w}_{p,0}^{t+1} := \hat{\boldsymbol{w}}^t$.
- **Output:** Polyak-Ruppert (PR) averaged iterate:

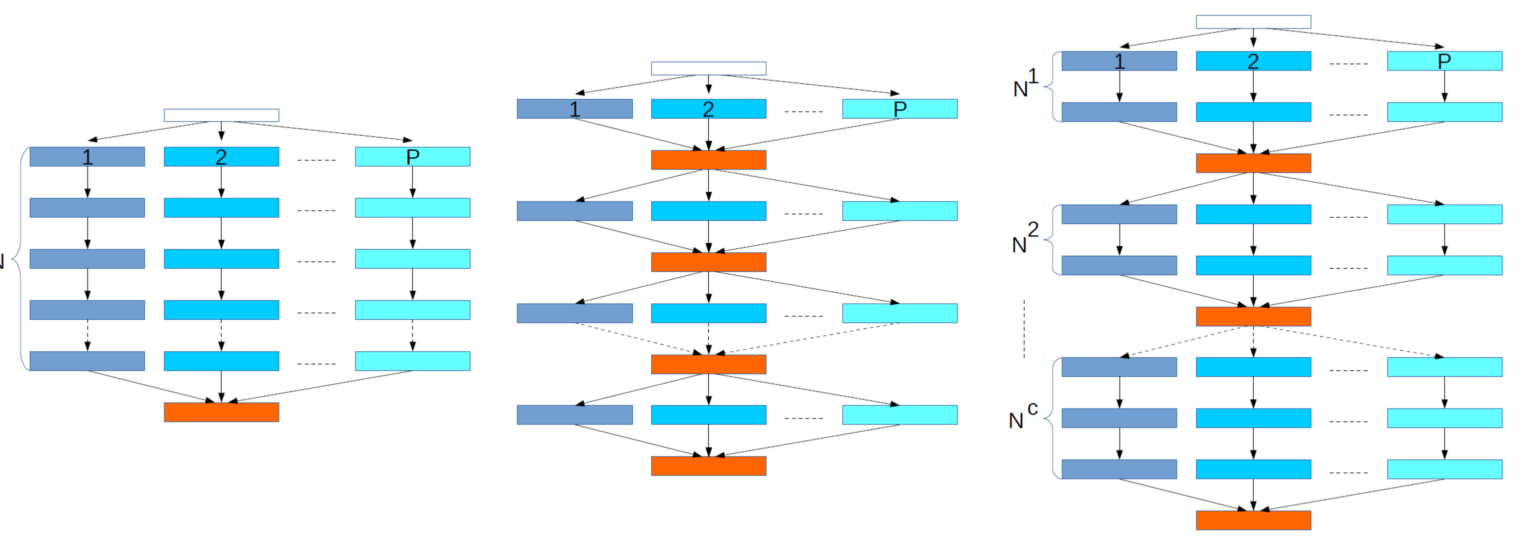
$$\overline{\overline{\boldsymbol{w}}}^C = \frac{1}{\sum_{t=1}^C N^t}\sum_{t=1}^C N^t \overline{\boldsymbol{w}}^t = \frac{1}{P\sum_{t=1}^C N^t}\sum_{t=1}^C\sum_{p=1}^P\sum_{k=1}^{N^t} \boldsymbol{w}_{p,k}^t,$$

with $\overline{\boldsymbol{w}}^t = \frac{1}{PN^t}\sum_{k=1}^{N^t}\sum_{p=1}^P \boldsymbol{w}_{p,k}^t$.

2 special "extreme" cases: MBA and OSA

Comparison setting: fixed total number of gradients T, with P workers.

- *One-Shot Averaging (OSA)*: $C = 1$ communication, and $(N^1) = T/P$
- *P-mini-batch averaging (MBA)*, $C = T/P$ communication rounds, and $(N^t)_{t \in [C]} = (1, \dots, 1)$.



Local-SGD is seen as a compromise between OSA and MBA.

Assumptions

A1: Strong convexity F is $\mu > 0$-strongly-convex
A2: Smoothness and regularity F is F is L-smooth $+ \mathcal{C}^3$, with uniformly bounded derivatives: $\sup_{\boldsymbol{w}\in\mathbb{R}^d} \left|\left|\left| F^{(3)}(\boldsymbol{w}) \right|\right|\right| < M$.
Q1: Quadratic function F is the quadratic function.
A3: Oracle on the gradient For any $(t,k) \in [C] \times [N^t]$ and $\boldsymbol{w} \in \mathbb{R}^d$, $\mathbb{E}[g_{p,k+1}^t(\boldsymbol{w}_{p,k}^t)|\boldsymbol{w}_{p,k}^t] = F'(\boldsymbol{w}_{p,k}^t)$.
For any *fixed* $\boldsymbol{w}$ the functions $(g_{p,k}^t)_{(t,k)}(\boldsymbol{w})$ are i.i.d. .
A4: Uniformly bounded variance (Additive noise) The variance of the error, $\mathbb{E}[\left\|g_{p,k}^t(\boldsymbol{w}_{p,k}^t) - F'(\boldsymbol{w}_{p,k}^t)\right\|^2] \le \sigma_\infty^2$.
A5: Cocoercivity of the random gradients For any p,t,k, $g_{p,k}^t$ is almost surely L-co-coercive: for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, $L\langle g_{p,k}^t(\boldsymbol{w}_1) - g_{p,k}^t(\boldsymbol{w}_2)\boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \ge \left\|g_{p,k}^t(\boldsymbol{w}_1) - g_{p,k}^t(\boldsymbol{w}_2)\right\|^2$.
A6: Finite variance at $\boldsymbol{w}^\star$ $\exists \sigma \ge 0$, s.t. for any t,k,p, $\mathbb{E}[\left\|g_{p,k}^t(\boldsymbol{w}^\star)\right\|^4] \le \sigma^4$.
Learning rate. $2\eta_k^t L \le 1$. 2 settings: *finite horizon* (FH) and *on-line*.

Related Work

- Local SGD [7]: small learning rate $(1/(\mu t))$, μ un-known in practice.
- Experimental [8].
- Parallel SGD (non asymptotic) [2].
- Proof technique [5, 1, 4] (non distributed)

Sketch of the proof [6]

"Decomposition":

$$\eta_k^t F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star) = \boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}_{p,k}^t - \eta_k^t[g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - F'(\boldsymbol{w}_{p,k-1}^t)] - \eta_k^t[F'(\boldsymbol{w}_{p,k-1}^t) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star)]. \tag{2}$$

Gives:

$$F''(\boldsymbol{w}^\star)\left(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star\right) = \frac{1}{P\sum_{t=1}^C N^t}\sum_{t=1}^C\sum_{p=1}^P\sum_{k=1}^{N^t}\left(\frac{\boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}_{p,k}^t}{\eta_k^t} - [g_{p,k}^t(\boldsymbol{w}_{p,k-1}^t) - F'(\boldsymbol{w}_{p,k-1}^t)] - [F'(\boldsymbol{w}_{p,k-1}^t) - F''(\boldsymbol{w}^\star)(\boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star)]\right). \tag{3}$$

3 terms: Initial Conditions, Noise, Residual
Noise and Residual depend on $\left\|\boldsymbol{w}_{p,k-1}^t - \boldsymbol{w}^\star\right\| \leftarrow$ control this quantity.

Convergence: non asymptotic comparison of MBA and OSA

Define

$$Q_{bias} = 1 + \frac{M^2\eta}{\mu}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2 + \frac{L^2\eta}{\mu P}, \quad Q_{1,var}(X) = \frac{L^2\eta}{\mu} + \frac{P}{X\eta\mu}, \quad Q_{2,var}(X) = \frac{M^2XP\eta^2\sigma^2}{\mu^2}.$$

Proposition 1. ***Mini-batch Averaging*** *Assume* ***A1,2,3,5,6*** *for any $t \in [C]$,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|^2\right] \le (1-\eta\mu)^t\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{2\sigma^2\eta}{P}\frac{1-(1-\eta\mu)^t}{\mu}, \tag{4}$$

$$\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|^2\right] \precsim \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{\eta^2C^2}Q_{bias} + \frac{\sigma^2}{T}\left(1 + \frac{Q_{1,var}(C)}{P} + \frac{Q_{2,var}(C)}{P^2}\right). \tag{5}$$

Proposition 2. ***One-shot Averaging*** *Assume* ***A1,2,3,5,6*** *for $p \in [P], t = 1, k \in [N]$,*

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^1 - \boldsymbol{w}^\star\right\|^2\right] \le (1-\eta\mu)^k\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + 2\sigma^2\eta\frac{1-(1-\eta\mu)^k}{\mu}, \tag{6}$$

$$\mathbb{E}\left[\left\|F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)\right\|^2\right] \precsim \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^\star\right\|^2}{\eta^2N^2}Q_{bias} + \frac{\sigma^2}{T}\left(1 + Q_{1,var}(N) + Q_{2,var}(N)\right). \tag{7}$$

Comments

1. **Identical asymptotic behavior** fixed P: initial condition ("bias") + variance decomposition.

- For the "local-process": Eqs. (4),(6): bias remains the same, but that the variance of the local process is reduced by a factor P.
- For the averaged process: Eqs. (5), (7) bias term is the same, and for $\eta = X^{-\alpha}$, $0.5 < \alpha < 1$, $X \in \{N, C\}$, the speed at which the *variance* is forgotten is the same (σ^2T^{-1} as $T \to \infty$).

"the noise is the noise and SGD doesn't care"

2. **Higher order terms matter**

- With $Q_{var}(N) = Q_{var}(C)$ the remaining terms are respectively P or P^2 times smaller for mini-batch.
- explanation of why mini-batch SGD outperforms one shot averaging in practice.
- Necessity of non asymptotic analysis

3. **Interpretation,** $P, T \to \infty$**.** Remaining terms are not always negligible. MBA could outperform OSA by a factor as large as P.
4. **Convergence in function values?** with $F(\overline{\overline{\boldsymbol{w}}}^C) - F(\boldsymbol{w}^\star) \le L\mu^{-2}||F''(\boldsymbol{w}^\star)(\overline{\overline{\boldsymbol{w}}}^C - \boldsymbol{w}^\star)||^2$

- sub-optimal dependence in μ
- but classical-proofs ($\eta_t \propto 1/(\mu t)$) do not get optimal asymptotic behavior of OSA $\to$ if the extreme are not tight, meaningless comparison.

4. **Comparing upper bounds:** caution !
5. **Rate in online setting better** but tradeoffs are the same.

Local SGD - "simple" assumptions - intuition

Proposition 3 (Local-SGD: Quadratic Functions with Bounded Noise)**.** *Under Assumptions Q1, A3, A4, we have the following bound for Local-SGD: for any $p \in [P], t \in [C], k \in [N^t]$,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{w}^\star\right\|^2\right] \le (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \frac{\sigma_\infty^2\eta}{P}\frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{\mu}$$

$$\mathbb{E}\left[\left\|\boldsymbol{w}_{p,k}^t - \boldsymbol{w}^\star\right\|^2\right] \le (1-\eta\mu)^{\boldsymbol{N}_1^{t-1}+k}\left\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\right\|^2 + \sigma_\infty^2\eta\left(\underbrace{\frac{1-(1-\eta\mu)^{\boldsymbol{N}_1^{t-1}}}{P\mu}}_{\textit{long term reduced var.}} + \underbrace{\frac{1-(1-\eta\mu)^k}{\mu}}_{\textit{local iteration var.}}\right).$$

Comments:

- proof: introduce a *ghost* sequence [3].
- "just after" communication, $\hat{w}^t \to$ same bound as mini-batch case
- Local iterates $\boldsymbol{w}_{p,k}^t \to$ variance composed of a "long term" reduced variance, $\to \frac{\sigma_\infty^2\eta}{P\mu}$ and extra variance $\eta\sigma_\infty^2\frac{1-(1-\eta\mu)^k}{\mu}$, increasing within the phase, and is upper bounded by $\sigma_\infty^2\eta^2k$.

Optimality of Local-SGD, "simple" assumptions

Corollary 1. *If for all $t \in [C]$, $N^t \le (\mu\eta P)^{-1}$, then*

- *the second order moment of $\boldsymbol{w}_{p,k}^t$ admits the same upper bound as the mini-batch iterate $\hat{\boldsymbol{w}}_{MB}^{\boldsymbol{N}_1^{t-1}+k}$ (Equation (4)) up to a constant factor of 2.*
- *As a consequence, Equation (5) is still valid, and Local-SGD performs "optimally".*

Interpretation

- if the algorithm communicates often enough, the convergence of the Polyak Ruppert iterate $\overline{\overline{\boldsymbol{w}}}^C$ is as good as in the mini-batch case, thus it is "optimal".
- more communication steps are necessary when more machines are used.
- **Example** With constant number of local steps $N^t = N$, and learning rate $\eta = c(NC)^{-1/2}$ in order to obtain an optimal $O(\sigma^2T^{-1})$ parallel variance[a] rate, local-SGD communicates $O(\sqrt{NC}/(P\mu))$ times less as compared to mini-batch averaging.

[a]in online setting, the same example would hold, resulting in a $O(\frac{\sigma^2}{T})$ convergence **rate** (not only variance).

Optimality of Local-SGD, general assumptions

Proposition 4. *Under either of the following sets of assumptions, the convergence of the Polyak Ruppert iterate $\overline{\overline{\boldsymbol{w}}}^C$ is as good as in the mini-batch case, up to a constant:*

- *Assume Q1, A3, A5, A6, and for any $t \in [C]$, $N^t \le (\mu\eta P)^{-1}$ and $\mu\eta^2\boldsymbol{N}_1^t = O(1)$.*
- *Assume A1,A2, A3, A4, and for any $t \in [C]$, $N^t \le \inf\left((\eta PM\mathbb{E}[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|])^{-1}, (\mu\eta P)^{-1}\right)$.*

Interpretation.

- Optimal rates if the communications happen often enough.
- Corresponds to practice [8]. But hard to use in practice.
- The first set of assumption is valid for LSR, the second for LR.
- In the second case, the maximal number of local steps is smaller than before, by a factor μ^{-1}, but the allowed maximal number of local steps can increase along with the epochs, as $\mathbb{E}[\left\|\hat{\boldsymbol{w}}^t - \boldsymbol{w}^\star\right\|]$ is typically decaying.

References

[1] F. Bach and E. Moulines. Non-asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *NIPS*, NIPS'11, USA, 2011.
[2] A. B. Godichon and S. Saadane. On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms. *ArXiv e-prints*, Oct. 2017.
[3] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *ArXiv e-prints*, July 2015.
[4] D. Needell, R. Ward, and N. Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In *NIPS*. 2014.
[5] B. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. 30:838–855, 07 1992.
[6] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.
[7] S. U. Stich. Local SGD Converges Fast and Communicates Little. *ICLR 2019*, 2019.
[8] J. Zhang, C. De Sa, I. Mitliagkas, and C. Ré. Parallel SGD: When does averaging help? *ArXiv e-prints*, June 2016.