SGD and Hogwild! Convergence Without the Bounded Gradients Assumption

Lam M. Nguyen[1,2] · Phuong Ha Nguyen[3] · Marten van Dijk[3]
Peter Richtárik[4] · Katya Scheinberg[1] · Martin Takáč[1]
[1]Lehigh University · [2]IBM Research · [3]University of Connecticut · [4]KAUST

The Problem and Assumptions

The Problem:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \mathbb{E}[f(w;\xi)] \right\}$$

– ξ is a random variable obeying some distribution

Assumptions:

- $F : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex
  $\exists \mu > 0$ such that $\forall w, w' \in \mathbb{R}^d$:
  $F(w) \geq F(w') + \langle \nabla F(w'), (w - w') \rangle + \frac{\mu}{2}\|w - w'\|^2$
- $f(w;\xi)$ is L-smooth for every realization of ξ
  $\exists L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$:
  $\|\nabla f(w;\xi) - \nabla f(w';\xi)\| \leq L\|w - w'\|$
- we can compute unbiased gradient $\mathbb{E}[\nabla f(w_t;\xi_t)] = \nabla F(w_t)$

The SGD Algorithm

1: **Input:** $\{\eta_t\}_{t=0}^{\infty}$ such that $\sum_t \eta_t = \infty$
2: choose $w_0 \in \mathbb{R}^d$
3: **for** $t = 0, 1, \ldots$ **do**
4: sample ξ_t
5: compute $\nabla f(w_t;\xi_t)$
6: update $w_{t+1} = w_t - \eta_t \nabla f(w_t;\xi_t)$
7: **end for**

Example:

- $F(w) = \frac{1}{2}(\underbrace{\tfrac{1}{2}w^2}_{f_1(w)} + \underbrace{w}_{f_2(w)})$ is smooth and SC
- with probability $(1/2)^t$ we will have $w_{t+1} = w_0 - \sum_{i=0}^{t} \eta_t$
  SGD can go arbitrary far with non-zero probability

Bounded Gradient Assumption

Common Assumption in SGD analysis

- $\exists G < \infty$ such that $\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq G, \ \forall w$

Clash with Strong Convexity Assumption

$$2\mu(F(w) - F^*) \leq \|\nabla F(w)\|^2 = \|\mathbb{E}[\nabla f(w;\xi)]\|^2 \leq \mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq G < \infty$$

Alternative Bound on Second Moment

- **$f(w;\xi)$ is convex:**

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq 4L[F(w) - F^*] + N,$$

- **$f(w;\xi)$ is nonconvex:**

$$\mathbb{E}[\|\nabla f(w;\xi)\|^2] \leq 4L\kappa[F(w) - F^*] + N,$$

where $\kappa = \frac{L}{\mu}$ and

$$N = 2\,\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$$

Convergence Rate of SGD

- **$f(w;\xi)$ is convex:**
  Let $\eta_t = \frac{2}{4L + \mu t} \leq \eta_0 = \frac{1}{2L}$. Then

$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{16N}{\mu} \cdot \frac{1}{4L + \mu(t - T)}$$

  for $t \geq T = \frac{4L}{\mu} \max\{\frac{L\mu}{N}\|w_0 - w_*\|^2 - 1, 0\}$
- **$f(w;\xi)$ is nonconvex:**
  Let $\eta_t = \frac{2}{4L\kappa + \mu t} \leq \eta_0 = \frac{1}{2L\kappa}$. Then

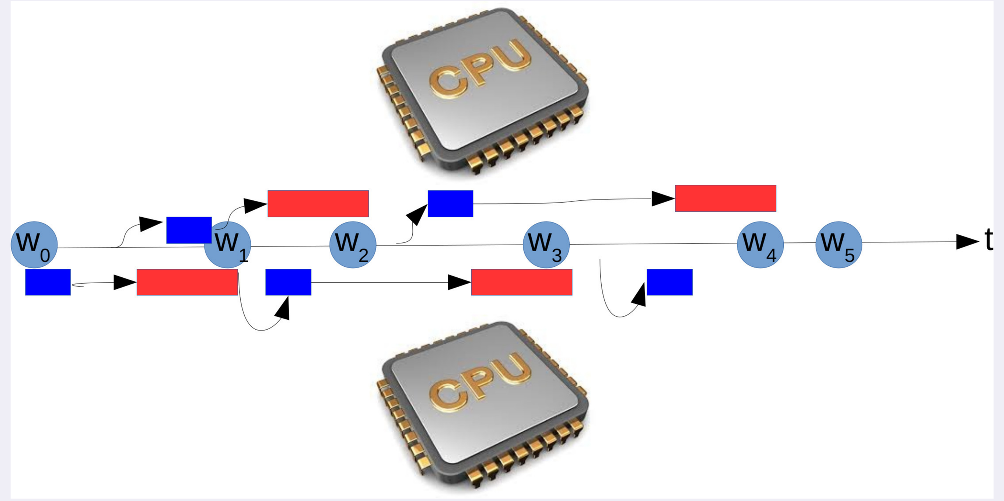
$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{16N}{\mu} \cdot \frac{1}{4L\kappa + \mu(t - T)}$$

  for $t \geq T = \frac{4L\kappa}{\mu} \max\{\frac{L\kappa\mu}{N}\|w_0 - w_*\|^2 - 1, 0\}$

HogWild!

- w_t - state of the shared memory after the t-th update is fully written
- $\hat{w}_t$ - state of the shared memory read which is used to produce w_t

$$w_t = w_{t-1} - \eta_t \nabla f(\hat{w}_t;\xi_t)$$



Convergence Rate of HogWild!

– τ - the maximum **delay** between "read" and "write"

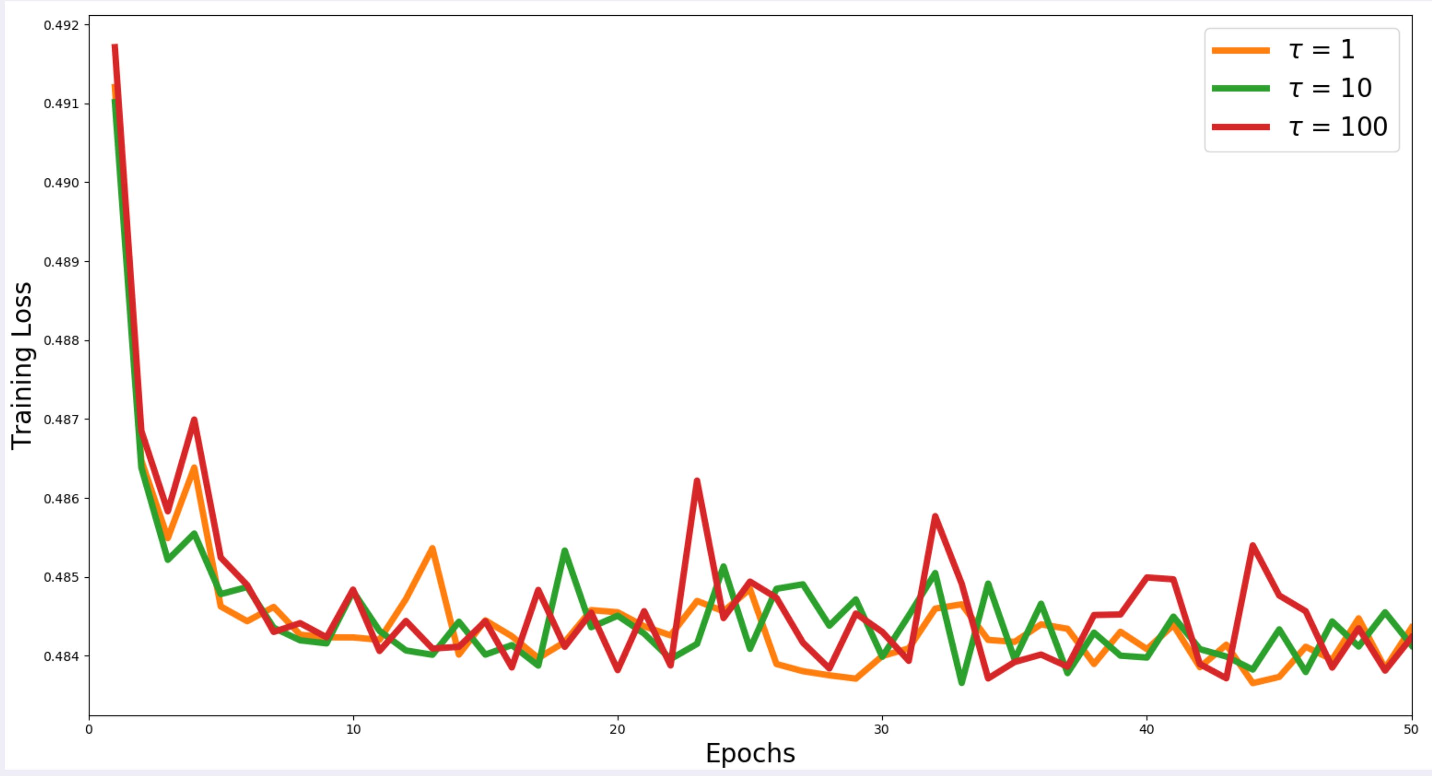
Theorem: Let $\eta_t = \frac{4}{\mu t + E}$, $E = \max\{16L, 2\tau\mu\}$ then $\mathbb{E}[\|\hat{w}_t - w_*\|^2]$ and $\mathbb{E}[\|w_t - w_*\|^2]$ are at most

$$\frac{64N}{\mu} \frac{t}{(\mu(t-1) + E)^2} + O\left(\frac{\ln t}{t^2}\right)$$

Note: In the paper, we also analyze **Lazy Hogwild!** (when only portion of gradient is applied)

Numerical Experiments

– Logistic regression
– covtype dataset



References

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal Optimization methods for large-scale machine learning, 2016
[2] Eric Moulines and Francis R. Bach Non-asymptotic analysis of stochastic approximation algorithms for machine learning, NIPS 2011.
[3] Benjamin Recht, Christopher Re, Stephen Wright and Feng Niu A Lock-Free Approach to Parallelizing Stochastic Gradient Descent, NIPS 2011.
[4] H. Mania, X. Pan, D. Papailiopoulos, B. Recht K. Ramchandran and M.I. Jordan Perturbed Iterate Analysis for Asynch. Stoch. Opt., 2015.
[5] Remi Leblond, Fabian Pedregosa and Simon Lacoste-Julien Improved asynchronous parallel optimization analysis for stochastic incremental methods, 2018.