Acceleration Tradeoff between Momentum and Asynchrony in Distributed Nonconvex Stochastic Optimization

Tianyi Liu*, Shiyang Li$^\diamond$, Jianping Shi*, Enlu Zhou*, Tuo Zhao*

*Georgia Tech $^\diamond$Harbin Institute of Technology *Sensetime

Background

Consider an empirical risk minimization problem,

$$\min_\theta \mathcal{F}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i, \theta)),$$

- n observations: $\{(x_i, y_i)\}_{i=1}^n$
- ℓ: loss function
- f: decision function associated with θ.

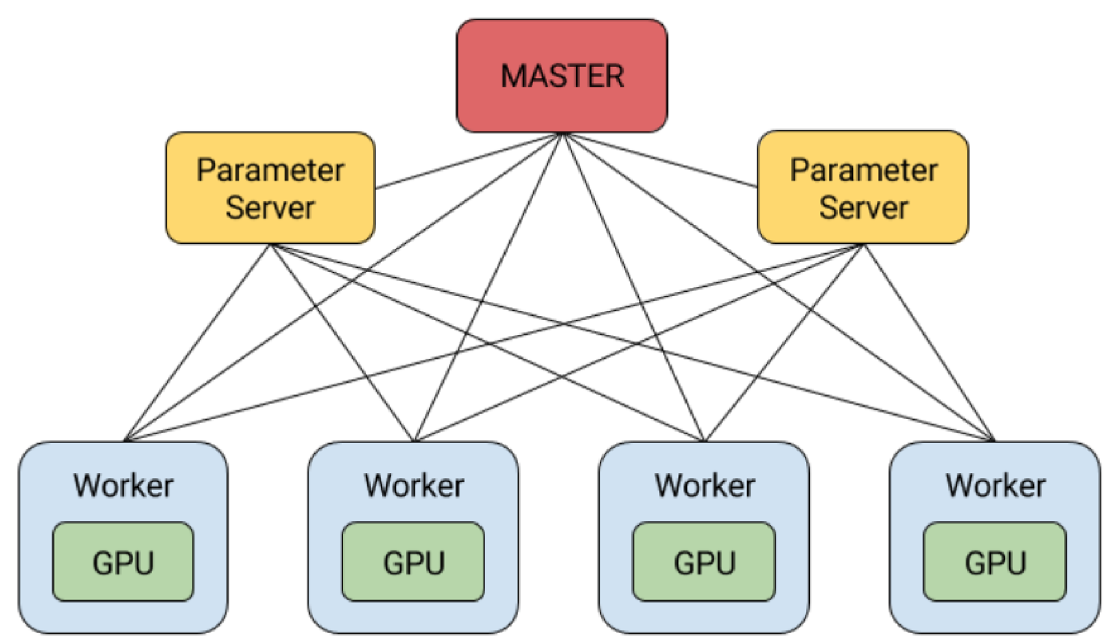
A Popular Algorithm: *Momentum SGD:*

$$\theta^{(k+1)} = \theta^{(k)} - \eta\nabla\ell(y_i, f(x_i, \theta^{(k)})) + \mu(\theta^{(k)} - \theta^{(k-1)}).$$

Drawback: **Slow on a single machine for large data!**

- ImageNet-1k: 10^6 training images, 224×224
- ResNet-50: 25.6×10^6 parameters
- 10 days on one GPU for 90-epoch training!

Solution: **Parameter Server Framework**



- Synchronous: Wait for the slowest worker. **Low parallel efficiency!**
- Asynchronous: Stale stochastic gradients:

$$\theta^{(k+1)} = \theta^{(k)} - \eta\nabla\ell(y_i, f(x_i, \theta^{(k-\tau_k)})) + \mu(\theta^{(k)} - \theta^{(k-1)}),$$

where τ_k denotes the delay.

Does there exist a Tradeoff between Asynchrony τ and momentum μ?

Streaming PCA

- A simple, but nontrivial example:

$$\max_v v^\top \mathbb{E}_{X\sim\mathcal{D}}[XX^\top]v \quad \text{s.t.} \quad v^\top v = 1,$$

where $X_k \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, $\mathbb{E}X_k = 0$, $\forall k > 0$.

- Async-MSGD for Streaming PCA:

$$v_{k+1} = v_k + \mu(v_k - v_{k-1}) + \eta\underbrace{(I - v_{k-\tau_k}v_{k-\tau_k}^\top)X_kX_k^\top v_{k-\tau_k}}_{\text{Approximate Manifold Gradient}},$$

where $\mu \in [0, 1)$ and τ_k denotes the delay.

- Assumption (Eigen-gap): $\Sigma = E[XX^\top]$ is positive definite with eigenvalues

$$\lambda_1 > \lambda_2 \geq ... \geq \lambda_d > 0$$

and associated normalized eigenvector v^1, v^2, ..., v^d.

- Assumption (Boundedness):

$$\mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma, \|X\| \leq C_d,$$

where C_d is a constant (possibly dependent on d).

Intuition

Discrete: $\dfrac{v_{k+1} - v_k}{\eta} = \mu\dfrac{v_k - v_{k-1}}{\eta} + \nabla F(v_k, X_k)$

weak $\Downarrow$ $\eta \to 0$

Continuous: $\dot{V} = \mu\dot{V} + \nabla\mathcal{F}(V)$

Discrete/Stochastic $\to$ Continuous/Deterministic.
Similar to the Law of Large Number, not reliable!

Consider the normalized error $\{u_n^{\eta,\tau} = \frac{v_n^{\eta,\tau} - v_i}{\sqrt{\eta}}\}$:

Discrete: $u_{k+1} - u_k = \mu(u_k - u_{k-1}) + \sqrt{\eta}\nabla F(v_k)$

weak $\Downarrow$ $\eta \to 0$

Continuous: $dU_t = \mu dU_t + \nabla^2\mathcal{F}(v_i)dt + \Sigma dB_t$

Randomness Returns.
Similar to the Central Limit Theorem!

Proof Technique: Fixed-State-Chain [2,3]

Convergence Theory

Theorem 1 (Global Convergence)**.** *Suppose for any $i > 0$, $v_{-i} = v_0 = v_1 \in \mathbb{S}$. When the delay satisfies:*

$$\tau_k \asymp (1-\mu)^2/(\lambda_1\eta^{1-\gamma}), \ \forall k > 0,$$

for some $\gamma \in (0, 1]$, we have $V^\eta(\cdot) \Rightarrow V(\cdot)$ in the weak sense as $\eta \to 0$, where $V(\cdot)$ satisfies the following ODE:

$$\dot{V} = \frac{1}{1-\mu}[\Sigma V - V^\top\Sigma V V], \quad V(0) = v_0.$$

Theorem 2 (Local Behavior)**.** *Condition on the event that $h_k^\eta - e_j \asymp \sqrt{\eta}$ for $k = 1, 2....$ Then for $i \neq j$, if for any k, the delay satisfies:*

$$\tau_k \asymp \frac{(1-\mu)^2}{(\lambda_1 + C_d)\eta^{\frac{1}{2}-\gamma}}, \ \forall k > 0,$$

for some $\gamma \in (0, 0.5]$, $\{U^{\eta,i}(\cdot)\}$ converges weakly to a solution of

$$dU = \frac{\lambda_i - \lambda_j}{1-\mu}U dt + \frac{\alpha_{i,j}}{1-\mu}dB_t.$$

O-U process, converges when $\lambda_i < \lambda_j$.

Remark 3. *Extension to unbounded random delay: Moment conditions, e.g.,*

$$\mathbb{E}(\tau_k) \asymp \frac{(1-\mu)^2}{(\lambda_1 + C_d)\eta^{\frac{1}{2}-\gamma}}, \ \forall k > 0,$$

for some $\gamma \in (0, 0.5]$.

Conclusions

Asynchrony does not yield Momentum.
They are in Conflicts!

Given η used in MSGD and τ such that for some $\gamma \in (0, 0.5]$,

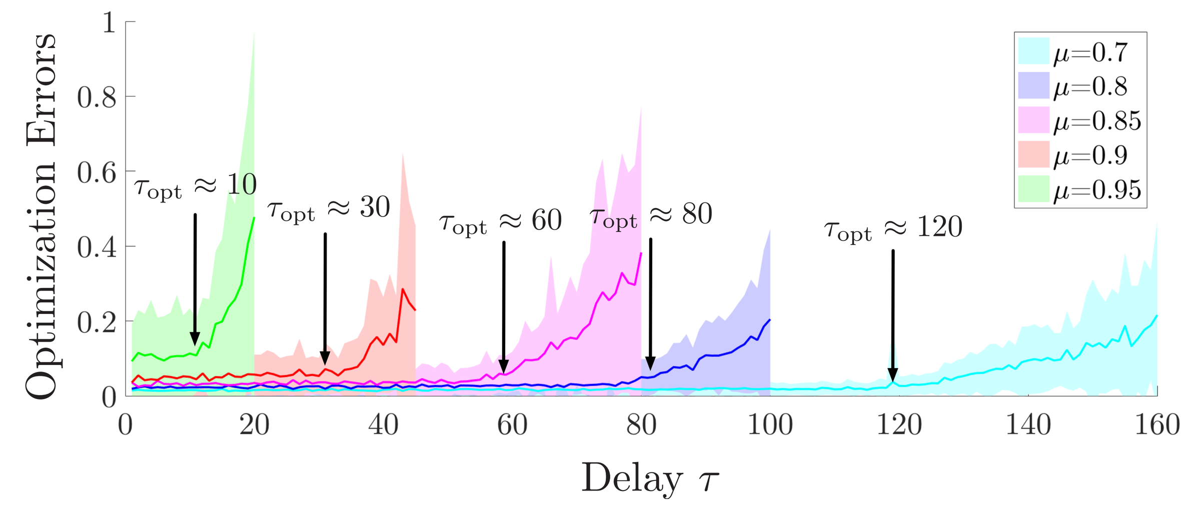
$$\tau \asymp (1-\mu)^2/(\lambda_1 + C_d)\eta^{\frac{1}{2}-\gamma},$$

the effective iteration complexity of Async-MSGD enjoys a linear acceleration, i.e.,

$$N_{\text{async}} \asymp \frac{(\lambda_1 + C_d)\phi^{\frac{1}{2}+\gamma}}{(1-\mu)(\lambda_1 - \lambda_2)^{\frac{3}{2}+\gamma}\epsilon^{\frac{1}{2}+\gamma}}.$$

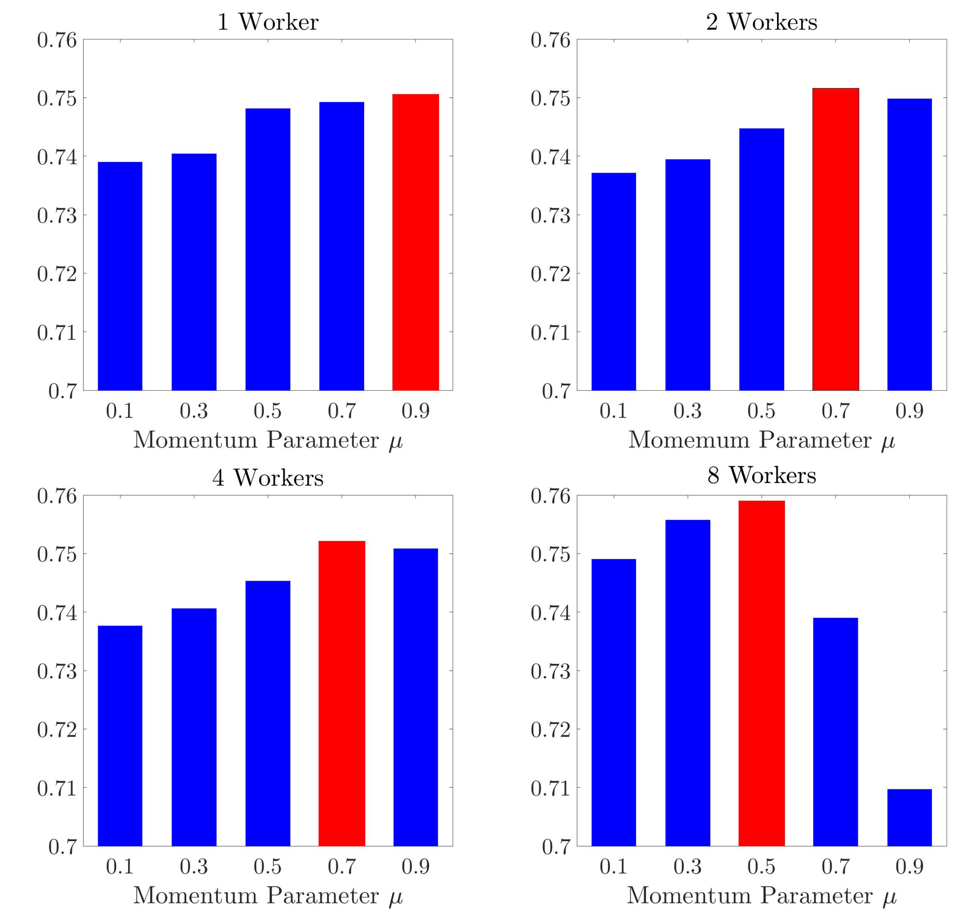
Experiments

- Streaming PCA: $d = 4$ and $\Sigma = \text{diag}\{4, 3, 2, 1\}$.



Optimal Delay Decreases!

- Deep Neural Network: WideResNet (36.6×10^6 parameters), CIFAR-100 (60×10^3 images for 100-class classification).



The average validation accuracies.
Optimal Momentum Decreases!

References

[1] Liu, et al. Towards Understanding Acceleration Tradeoff between Momentum and Asynchrony in Distributed Nonconvex Stochastic Optimization. Annual Conference on Neural Information Processing Systems (NIPS), 2018.

[2] Kushner and Yin. Stochastic approximation and recursive algorithms and applications, Stochastic modelling and applied probability, vol. 35.

[3] Liu, et al. Toward Deeper Understanding of Nonconvex Stochastic Optimization with Momentum using Diffusion Approximations. arXiv.1802.05155.