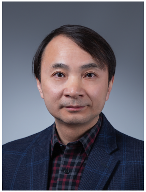
Complete Asymptotic Analysis for Projected Stochastic Approximation and Debiased Variants

Xiang Li

Joint work with

Jiadong Liang

Yuze Han



Zhihua Zhang

Peking University

Allerton Conf.

Sep. 27, 2023

Outline

Outline

Federated Learning (FL)

- FL collaboratively trains a global model from data held by remote devices without data sharing [McMahan et al., 2017].
- Assume K devices with weight p_k and objective function

$$\tilde{f}_k(x) := \mathbb{E}_{\xi_k \sim \mathcal{D}_k} \tilde{f}(x, \xi_k).$$

- The central server tries to $\min_x \sum_{k=1}^K p_k \tilde{f}_k(x)$.
- Equivalent to the global consensus problem:

$$\min_{x_1, \cdots, x_K} \sum_{k=1}^K p_k \tilde{f}_k(x_k) \quad \text{such that} \quad x_1 = \cdots = x_K.$$

What if we concatenate all local x_k's and $\tilde{f}_k$'s?

Linearly Constrained Problem

- We consider

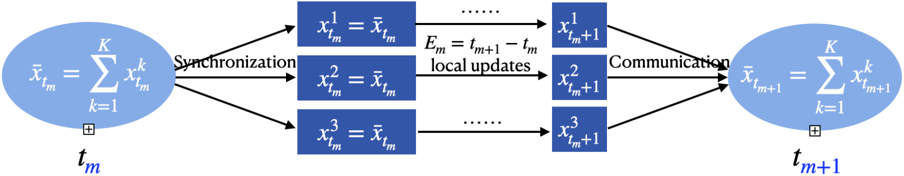
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \mathbb{E}_{\zeta\sim\mathcal{D}} f(\boldsymbol{x},\zeta) \ \text{ subject to } \boldsymbol{A}^\top \boldsymbol{x} = \boldsymbol{0}. \tag{1}$$

- Reduced to FL if $\boldsymbol{x} = (x_1^\top, \cdots, x_K^\top)^\top, \zeta = (\xi_1^\top, \cdots, \xi_K^\top)^\top \in \mathbb{R}^{Kd}$ and

$$f(\boldsymbol{x},\zeta) = \sum_{k=1}^{K} p_k \tilde{f}(x_k,\xi_k), \boldsymbol{A} = (\mathbf{I}, -\mathbf{I}, \mathbf{0}; \mathbf{0}, \mathbf{I}, -\mathbf{I}, \mathbf{0}; \cdots) \in \mathbb{R}^{(K-1)d\times Kd}.$$

- A simple and general formulation, if f and $\boldsymbol{A}$ are general.
- Data heterogeneity: $\operatorname{argmin} f(\boldsymbol{x}) \neq \operatorname{argmin}_{\boldsymbol{A}^\top\boldsymbol{x}=\boldsymbol{0}} f(\boldsymbol{x}) =: \boldsymbol{x}^\star$.
- Communication/synchronization $\Leftrightarrow$ projection.

Local SGD



$$x_{t+1}^k = \begin{cases} x_t^k - \eta_m g_t^k & \text{if } t_m < t+1 < t_{m+1} \\ \sum_{k=1}^{K} p_k \left[x_t^k - \eta_m g_t^k \right] & \text{if } t+1 = t_{m+1} \end{cases}$$

Figure: Illustration of Local SGD.

What's the counterpart algorithm of Local SGD in solving the general (1)?

Loopless Projected Stochastic Approximation (LPSA)

- Key idea: lower projection frequency to improve projection efficiency.
- The algorithm: for each iteration $n \geq 1$,
  1. Sample data $\zeta_n \sim \mathcal{D}$ and $\omega_n \sim \mathrm{Bernoulli}(p_n)$.
  2. $\boldsymbol{x}_{n+\frac{1}{2}} = \boldsymbol{x}_n - \eta_n \nabla f(\boldsymbol{x}_n, \zeta_n)$.
  3. $\boldsymbol{x}_{n+1} = \mathcal{P}_{\boldsymbol{A}^\perp} \boldsymbol{x}_{n+\frac{1}{2}}$ if $\omega_n = 1$ else $= \boldsymbol{x}_{n+\frac{1}{2}}$.
- A loopless version of Local SGD [Stich, 2019].

Main Question

How different p_n's changes the asymptotic behavior of LPSA?

Why this Question Matter?

Main Question

How $p_n \propto \eta_n^{\beta}$ changes the asymptotic behavior of LPSA?

- Popularity of lazy communication/local updates.
- Limited theoretical understanding (typically MSE).
- Provide insights for future algorithm design.

Outline

Convergence Result

Decompose $\boldsymbol{x}_n = \boldsymbol{u}_n + \boldsymbol{v}_n$ where $\boldsymbol{u}_n = \mathcal{P}_{\boldsymbol{A}^\perp}(\boldsymbol{x}_n)$ and $\boldsymbol{v}_n = \mathcal{P}_{\boldsymbol{A}}(\boldsymbol{x}_n)$.

Theorem (Convergence)

Under some standard assumptions [Liang et al., 2022] and $\eta_n \propto n^{-\alpha}$ and $p_n = \min\{\eta_n^{\beta}, 1\}$ with $\alpha \in (0, 1]$ and $\beta \in [0, 1)$,

$$\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2 = \mathcal{O}(n^{-\alpha \min\{1,2(1-\beta)\}}) \text{ and } \mathbb{E}\left\|\boldsymbol{v}_n\right\|^2 = \mathcal{O}(n^{-2\alpha(1-\beta)}).$$

- With frequent projection $\beta \le 0.5$, $\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2 = \mathcal{O}(\eta_n)$.
- With occasional projection $\beta \ge 0.5$, $\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2 = \mathcal{O}(\eta_n^{2(1-\beta)})$.

Asymptotic Behaviors

Theorem (Asymptotic behaviors)

Under some standard assumptions, we can find a PSD matrix $\tilde{\Sigma}$ and a vector $\boldsymbol{\mu}$ such that

- *Frequent projection* $\beta \in [0, 1/2)$: $\frac{1}{\eta_n^{1/2}}(\boldsymbol{u}_n - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(0, \tilde{\Sigma})$.
- *Occasional projection* $\beta \in (1/2, 1)$: $\frac{1}{\eta_n^{1-\beta}}(\boldsymbol{u}_n - \boldsymbol{x}^\star) \xrightarrow{L_2} \boldsymbol{\mu}$.
- *(New) Moderate projection* $\beta = 1/2$: $\frac{1}{\eta_n^{1/2}}(\boldsymbol{u}_n - \boldsymbol{x}^\star) \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \tilde{\Sigma})$.

Summary

β	$\mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2$	Asym. dist. of $\boldsymbol{u}_n - \boldsymbol{x}^\star$	Behavior
$[0, 1/2)$	$\mathcal{O}(\eta_n)$	$\sqrt{\eta_n} \cdot \mathcal{N}(0, \tilde{\Sigma})$	Var dominate
$1/2$	$\mathcal{O}(\eta_n)$	$\sqrt{\eta_n} \cdot \mathcal{N}(\boldsymbol{\mu}, \tilde{\Sigma})$	Bias $\approx$ Var
$(1/2, 1)$	$\mathcal{O}(\eta_n^{2(1-\beta)})$	$\eta_n^{1-\beta}\|\boldsymbol{\mu}\|$	Bias dominate

Outline

Where the Bias Comes from

$$\begin{aligned} \boldsymbol{u}_{n+1} &= \boldsymbol{u}_n - \eta_n \mathcal{P}_{\boldsymbol{A}^\perp} \nabla f(\boldsymbol{x}_n) + \eta_n \xi_n^{(1)}, \\ &\approx \boldsymbol{u}_n - \eta_n \mathcal{P}_{\boldsymbol{A}^\perp} \nabla^2 f(\boldsymbol{x}^\star)(\boldsymbol{x}_n - \boldsymbol{x}^\star) + \eta_n \xi_n^{(1)}. \end{aligned}$$

Let $\Delta_n = \mathbb{E}\left\|\boldsymbol{u}_n - \boldsymbol{x}^\star\right\|^2$. Then,

$$\Delta_{n+1} \lesssim (1 - c\eta_n)\Delta_n - 2\eta_n \mathbb{E}\langle \boldsymbol{u}_n - \boldsymbol{x}^\star, \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{v}_n \rangle + \eta_n^2.$$

Lemma

$$\left|\mathbb{E}\left\langle \boldsymbol{u}_n - \boldsymbol{x}^\star, \nabla^2 f(\boldsymbol{x}^\star)(\boldsymbol{v}_n - \mathbb{E}\boldsymbol{v}_n)\right\rangle\right| = o(\eta_n^{2(1-\beta)}).$$

$$\Delta_{n+1} \lesssim (1 - c\eta_n)\Delta_n - 2\eta_n \mathbb{E}\langle \boldsymbol{u}_n - \boldsymbol{x}^\star, \nabla^2 f(\boldsymbol{x}^\star)\mathbb{E}\boldsymbol{v}_n \rangle + o(\eta_n^{3-2\beta}).$$

Remove the Bias

If the gradient is evaluated at $\boldsymbol{x}_n - \mathbb{E}\boldsymbol{v}_n$ rather than $\boldsymbol{x}_n$, i.e.,

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n - \eta_n \mathcal{P}_{\boldsymbol{A}^\perp} \nabla f(\boldsymbol{x}_n - \mathbb{E}\boldsymbol{v}_n) + \eta_n \xi_n^{(1)},$$

we then have

$$\Delta_{n+1} \precsim (1 - c\eta_n)\Delta_n + o(\eta_n^{3-2\beta}).$$

How to Approximate $\mathbb{E}\boldsymbol{v}_n$

Lemma

$$\frac{\boldsymbol{v}_n}{\eta_n^{1-\beta}} \xrightarrow{d} -\frac{\nabla f(\boldsymbol{x}^\star)}{\|\nabla f(\boldsymbol{x}^\star)\|} \cdot \mathcal{E}\left(\|\nabla f(\boldsymbol{x}^\star)\|\right),$$

where $\mathcal{E}(\theta)$ represents the exponential distribution with expectation θ.

$$\mathbb{E}\boldsymbol{v}_n \approx -\eta_n^{1-\beta}\nabla f(\boldsymbol{x}^\star).$$

However, $\boldsymbol{x}^\star$ is unknown in practice......

Solution

Replace $\nabla f(\boldsymbol{x}^\star)$ by $\nabla f(\boldsymbol{x}_n, \zeta_n')$.

Debiased LPSA (DLPSA)

- Key idea: evaluate gradient at $\boldsymbol{x}_n + \eta_n^{1-\beta} \nabla f(\boldsymbol{x}_n, \zeta_n')$ instead of $\boldsymbol{x}_n$.

For each iteration $n \geq 1$

1. Sample $\zeta_n, \zeta_n' \sim \mathcal{D}$ and $\omega_n \sim \mathrm{Bernoulli}(p_n)$ independently.
2. $\boldsymbol{x}_{n+\frac{1}{2}} = \boldsymbol{x}_n - \eta_n \nabla f\left(\boldsymbol{x}_n + \eta_n^{1-\beta} \nabla f(\boldsymbol{x}_n, \zeta_n'), \zeta_n\right)$.
3. $\boldsymbol{x}_{n+1} = \mathcal{P}_{\boldsymbol{A}^\perp} \boldsymbol{x}_{n+\frac{1}{2}}$ if $\omega_n = 1$ else $= \boldsymbol{x}_{n+\frac{1}{2}}$.

Theorem (Convergence for DLPSA)

Under some assumptions,

$$\mathbb{E} \left\| \boldsymbol{u}_n - \boldsymbol{x}^\star \right\|^2 = \mathcal{O}\left(n^{-\alpha \min\{1, 3(1-\beta)\}} \right).$$

Corollary

The projection complexity is $\mathcal{O}(\varepsilon^{-1/2})$ for LPSA and $\mathcal{O}(\varepsilon^{-1/3})$ for DLPSA.

Illustration

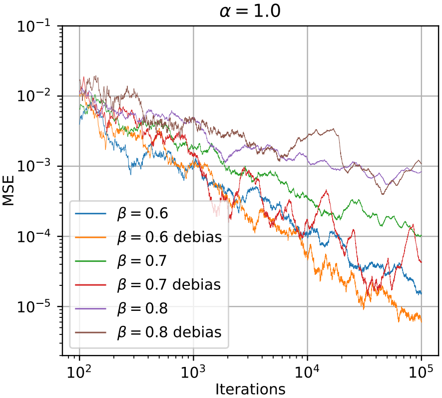


Figure: Comparison between LPSA and DLPSA in averaged MSEs over 10 repetitions.

Outline

Conclusion

- Introduce LPSA to solve linearly constrained problems.
- Characterize the full phase transition of asymptotic behaviors when varying projection frequency by tuning β.
- Develope debiased LPSA to improve projection efficiency.

References

Jiadong Liang, Yuze Han, Xiang Li, and Zhihua Zhang. Asymptotic behaviors of projected stochastic approximation: A jump diffusion perspective. In *Advances in Neural Information Processing Systems*, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Sebastian Urban Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.