SCAFFLSA: Taming Heterogeneity in Federated Stochastic Approximation

To appear in NeurIPS-2024

Paul Mangold [1], Sergey Samsonov [2], Safwan Labbi [1], Ilya Levin [2], Reda Alami [3], Alexey Naumov [2], Eric Moulines [1]
[1]Ecole Polytechnique, France [2]HSE University, Russia [3]Technology Innovation Institute, UAE

Federated Linear Stochastic Optimization

Take N matrices $\bar{\mathbf{A}}^c$'s and vectors $\bar{\mathbf{b}}^c$'s
Goal: solve collaboratively

$$\Big(\frac{1}{N}\sum_{c=1}^{N}\bar{\mathbf{A}}^c\Big)\theta_\star = \frac{1}{N}\sum_{c=1}^{N}\bar{\mathbf{b}}^c$$

assuming $\theta_\star$ is unique, and $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$ are split among N agents, with stochastic oracles $\mathbf{A}^c(Z^c_{t,h})$ and $\mathbf{b}^c(Z^c_{t,h})$.
Oracles are unbiased with bounded variance, and for $\eta < \eta_\infty$, there exists $a > 0$ such that

$$\mathbb{E}[\|\mathsf{Id} - \eta\mathbf{A}^c(Z^c_{t,h})\|^2] \leq 1 - \eta a$$

Applications: federated TD learning, linear regression...

Existing Method: FedLSA Algorithm

Input: $\eta > 0$, $\theta_0 \in \mathbb{R}^d$, $T, N, H > 0$
for $t = 0$ to $T-1$ **do**
  Initialize $\theta_{t,0} = \theta_t$
  for $c = 1$ to N **do**
    for $h = 1$ to H **do**
      Observe $Z^c_{t,h}$ and perform local update:

$$\theta_{t,h} = \theta^c_{t,h-1} - \eta(\mathbf{A}^c(Z^c_{t,h})\theta^c_{t,h-1} - \mathbf{b}^c(Z^c_{t,h}))$$

  Aggregate local updates $\theta_{t+1} = \frac{1}{N}\sum_{c=1}^{N}\theta^c_{t,H}$

We propose a new analysis of FedLSA, inspired by Wang et al., 2022 and Samsonov et al., 2024

We show that local training with control variates in federated LSA accelerates while preserving the linear speed-up

Parameter setting required to reach $\mathbb{E}\left[\|\theta_T - \theta_\star\|^2\right] \leq \epsilon^2$ for different algorithms/analyses

	Algorithm	Communication T	Local updates H	Sample complexity TH
	FedLSA [Existing analysis Doan, 2020]	$\mathcal{O}\left(\frac{N^2}{a^2\epsilon^2}\log\frac{1}{\epsilon}\right)$	1	$\mathcal{O}\left(\frac{N^2}{a^2\epsilon^2}\log\frac{1}{\epsilon}\right)$
new results	FedLSA [Our analysis]	$\mathcal{O}\left(\frac{1}{a^2\epsilon}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{N\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{Na^2\epsilon^2}\log\frac{1}{\epsilon}\right)$
new results	Scaffnew [Extended to LSA]	$\mathcal{O}\left(\frac{1}{a\epsilon}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{a\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{a^2\epsilon^2}\log\frac{1}{\epsilon}\right)$
new results	Scafflsa [Our analysis]	$\mathcal{O}\left(\frac{1}{a^2}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{N\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{Na^2\epsilon^2}\log\frac{1}{\epsilon}\right)$

Proposed Method: SCAFFLSA algorithm

[Inspired by Mishchenko et al., 2022's ProxSkip!]

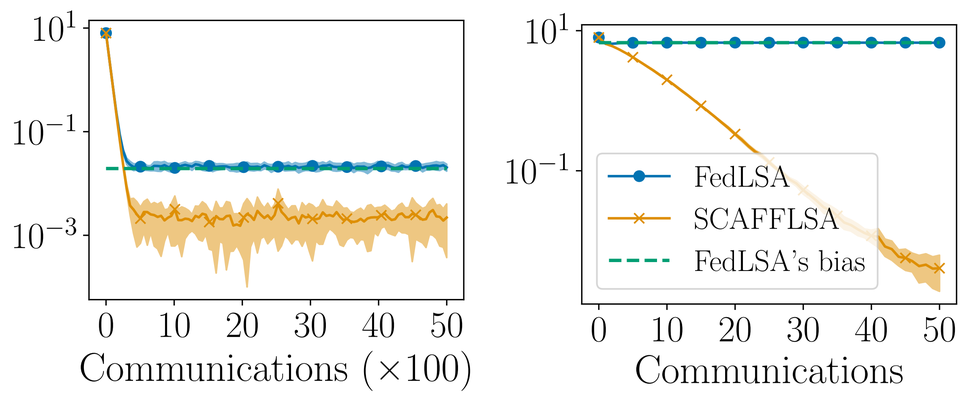
Input: $\eta > 0$, $\theta_0, \xi_0 \in \mathbb{R}^d$, $T, N, H > 0$
for $t = 0$ to $T-1$ **do**
  Initialize $\theta_{t,0} = \theta_t$
  for $c = 1$ to N **do**
    for $h = 1$ to H **do**
      Observe $Z^c_{t,h}$ and perform local update:

$$\theta_{t,h} = \theta^c_{t,h-1} - \eta(\mathbf{A}^c(Z^c_{t,h})\theta^c_{t,h-1} - \mathbf{b}^c(Z^c_{t,h}) - \xi^c_t)$$

  Aggregate local updates $\theta_{t+1} = \frac{1}{N}\sum_{c=1}^{N}\theta^c_{t,H}$
  Update control variates: $\xi^c_{t+1} = \xi^c_t + \frac{1}{\eta H}(\theta_{t+1} - \theta^c_{t,H})$

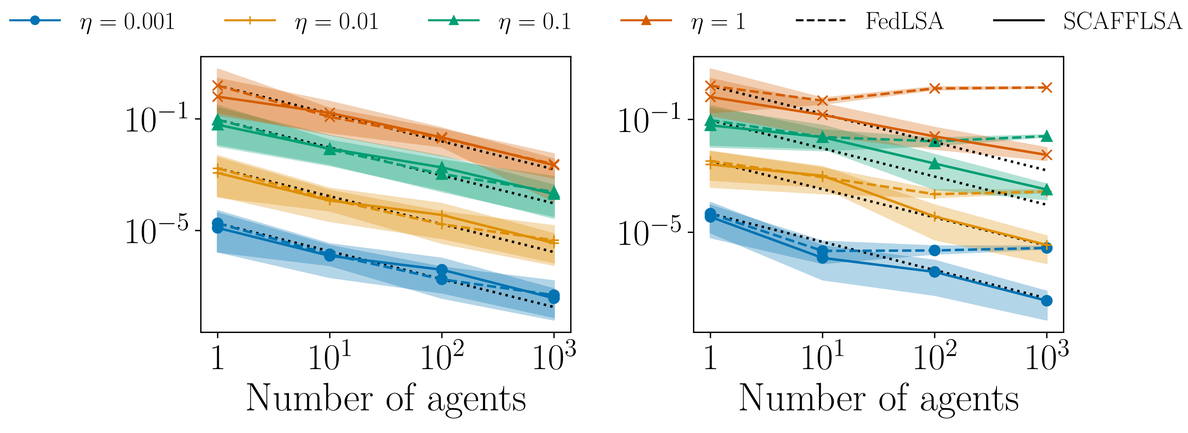
Numerical Study [on heterogeneous Garnet]

SCAFFLSA does not have bias when H increases!



(a) Heterogeneous, $N = 100, H = 100$ (b) Heterogeneous, $N = 100, H = 10000$

Both algorithms have linear speed-up, FedLSA is biased...



(a) Heterogeneous, $H = 1$ (b) Heterogeneous, $H = 100$

References

[1] J. Wang et al. "On the unreasonable effectiveness of federated averaging with heterogeneous data". In: *arXiv* (2022).
[2] S. Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 4511–4547.
[3] T. T. Doan. "Local stochastic approximation: A unified view of federated learning and distributed multi-task reinforcement learning algorithms". In: *arXiv* (2020).
[4] K. Mishchenko et al. "Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally!" In: *ICML*. PMLR. 2022, pp. 15750–15769.