Taming Heterogeneity in Federated Linear Stochastic Approximation and Federated Learning

Paul Mangold
CMAP, École polytechnique, France

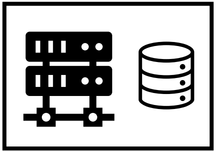
Joint Work with E. Moulines (Polytechnique), S. Samsonov (HSE Russia), S. Labbi (Polytechnique), I. Levin (HSE Russia), R. Alami (TII, UAE), A. Naumov (HSE Russia)

November 4, 2024
ARGO Seminar

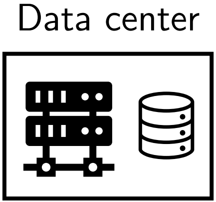
Background on Federated Learning

Data Collection

Data center



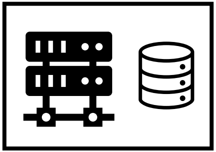
Data Collection



vs.

Data Collection

Data center

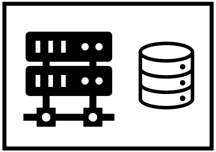


vs.

Data collection *by users*

Data Collection

Data center



vs.

Data collection *by users*

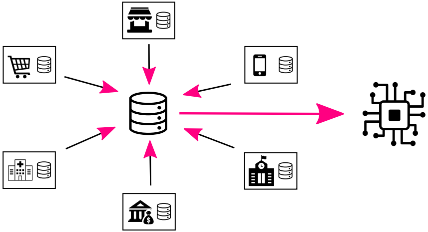
→ **how to use all this data?**

Centralizing in a data center is difficult

Centralizing data is often impossible

- *Privacy*:
  → data may be sensitive (e.g. health records, geolocation)
- *Volume of data*:
  → data may be large (e.g. cameras of self-driving car)
- *Time*:
  → it may be needed to take decisions quickly (e.g. reinforcement learning)

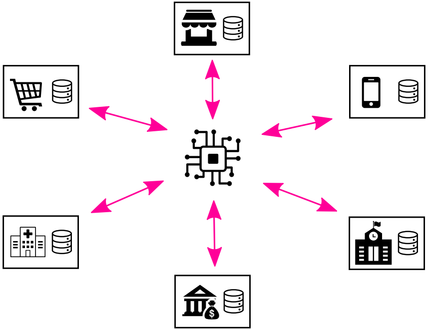
Classical vs Federated Learning



A single optimization problem

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x,y \sim D} \left[\ell(\theta; x, y) \right]$$

Classical vs Federated Learning



Multiple sub-problems

$$\min_{\theta \in \mathbb{R}^d} \sum_{c=1}^{N} \mathbb{E}_{x^c, y^c \sim \mathcal{D}^c} \left[\ell(\theta; x^c, y^c) \right]$$

$\rightarrow$ but only *one shared solution*

Best Scenario: Homogeneous Data

N local sub-problems

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^1, y^1 \sim \mathcal{D}^1} \left[\ell(\theta; x^1, y^1) \right] \to \theta_\star^1$$

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^2, y^2 \sim \mathcal{D}^2} \left[\ell(\theta; x^2, y^2) \right] \to \theta_\star^2$$

$$\vdots$$

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^N, y^N \sim \mathcal{D}^N} \left[\ell(\theta; x^N, y^N) \right] \to \theta_\star^N$$

Best Scenario: Homogeneous Data

N local sub-problems

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^1, y^1 \sim \mathcal{D}^1} \left[\ell(\theta; x^1, y^1) \right] \to \theta_\star^1$$

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^2, y^2 \sim \mathcal{D}^2} \left[\ell(\theta; x^2, y^2) \right] \to \theta_\star^2$$

$$\vdots$$

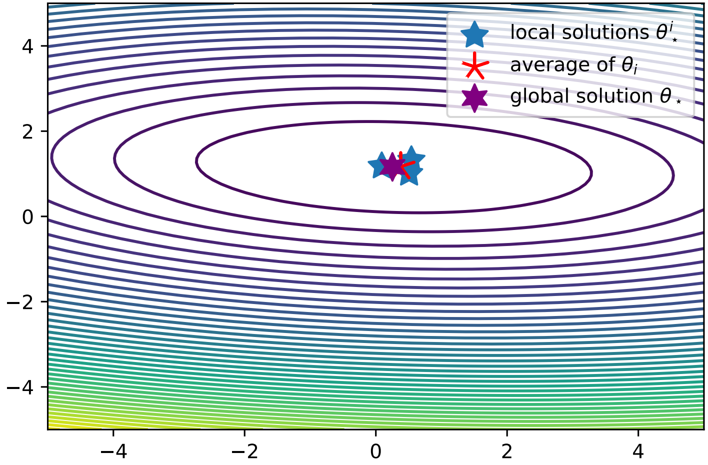
$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{x^N, y^N \sim \mathcal{D}^N} \left[\ell(\theta; x^N, y^N) \right] \to \theta_\star^N$$

Estimate global solution

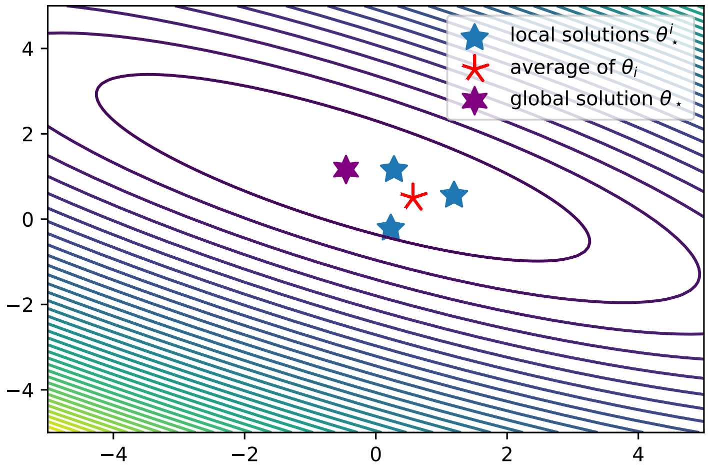
$$\theta_\star = \frac{1}{N} \sum_{c=1}^{N} \theta_\star^c$$

OK if $\mathcal{D}_1 = \mathcal{D}_2 = \cdots = \mathcal{D}_N$

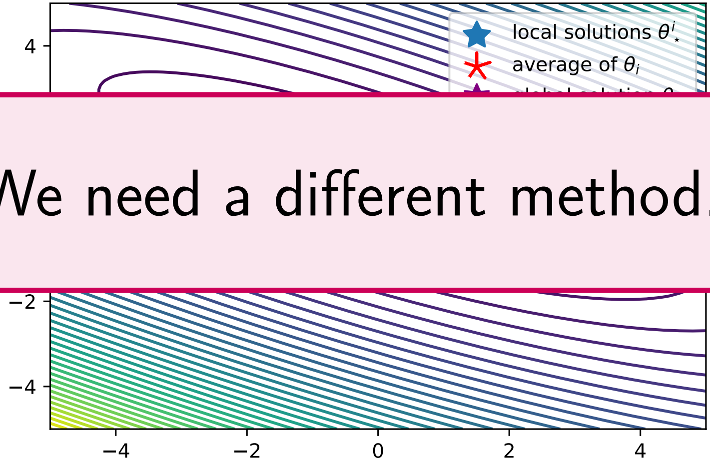
Best Scenario: Homogeneous Data



Failure: Heterogeneous Data



Failure: Heterogeneous Data



We need a different method...

Federated Optimization

$$\theta_\star \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{c=1}^{N} f^c(\theta) \quad , \qquad \text{where } f^c(\theta) = \mathbb{E}_{x^c, y^c \sim \mathcal{D}^c}\Big[\ell(\theta; x^c, y^c)\Big]$$

[1] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *AISTATS*. PMLR. 2017, pp. 1273–1282.

Federated Optimization

$$\theta_\star \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{c=1}^{N} f^c(\theta) \;, \qquad \text{where } f^c(\theta) = \mathbb{E}_{x^c, y^c \sim \mathcal{D}^c}\left[\ell(\theta; x^c, y^c)\right]$$

Federated Averaging (or local (S)GD)[1]

- For each $t = 0 \ldots$:
  - Set $\theta_{t,0}^c = \theta_t$
  - For each agent c, do H gradient updates:

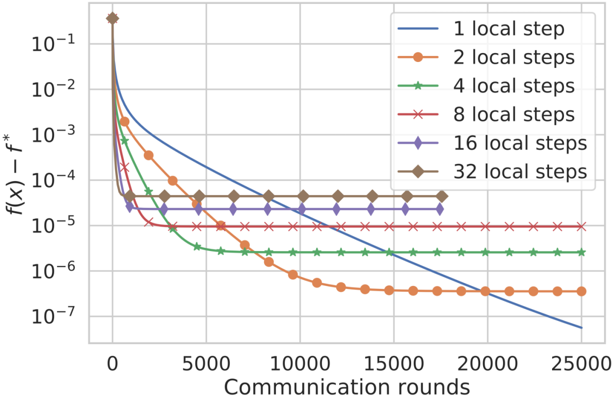
$$\theta_{t,h+1}^c = \theta_{t,h}^c - \eta \nabla f^c(\theta_{t,h}^c)$$

- Aggregate models: $\theta_{t+1} = \frac{1}{N} \sum_{c=1}^{N} \theta_{t,H}^c$

[1] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *AISTATS*. PMLR. 2017, pp. 1273–1282.

Communication and Sample Complexity

Local Training vs. Precision



(Figure from Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. "Tighter Theory for Local SGD on Identical and Heterogeneous Data". In: *AISTATS*. 2020, pp. 4519–4529)

Beyond Federated Optimization: Federated TD and LSA

Some problems do not fit this framework...

Example: TD Learning with linear approximation (I)

In Federated TD learning, N agent use a shared policy π in N different environments:

$$S_0^c = s, A_k^c \sim \pi(\cdot|S_k^c), \text{ and } S_{k+1}^c \sim P_{\mathrm{MDP}}^c(\cdot|S_k^c, A_k^c)$$

Some problems do not fit this framework...

Example: TD Learning with linear approximation (I)

In Federated TD learning, N agent use a shared policy π in N different environments:

$$S_0^c = s, A_k^c \sim \pi(\cdot|S_k^c), \text{ and } S_{k+1}^c \sim P_{\text{MDP}}^c(\cdot|S_k^c, A_k^c)$$

Goal: estimate its value in each environment, for $s \in \mathcal{S}$,

$$V^{c,\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r^c(S_k^c, A_k^c)\right]$$

where r^c is a reward obtained by agent c

Some problems do not fit this framework...

Example: TD Learning with linear approximation (II)

Idea: build a *shared estimate* of all values

$$V^{c,\pi}(s) \approx \theta^\top \varphi(s)$$

using $\theta \in \mathbb{R}^d$ and embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$

Some problems do not fit this framework...

Example: TD Learning with linear approximation (II)

Idea: build a *shared estimate* of all values

$$V^{c,\pi}(s) \approx \theta^\top \varphi(s)$$

using $\theta \in \mathbb{R}^d$ and embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$

Is this meaningful to use a shared estimate? Yes, because:

- If agents are homogeneous, it reduces sample complexity
- If agents are heterogeneous, it may reduce bias of local data

Linear Stochastic Approximation

Special case: only one agent

TD (with linear approx.) can be seen as solving a linear system

$$\bar{A}\theta_\star = \bar{b}$$

where $\bar{A}$ and $\bar{b}$ are known through stochastic estimates $A(Z)$, $b(Z)$ for a sequence of random variables Z

... variance of $A(Z)$ and $b(Z)$ are typically very large

... and $\bar{A}$ is not symmetric

Linear Stochastic Approximation

Special case: only one agent

TD (with linear approx.) can be seen as solving a linear system

$$\bar{A}\theta_\star = \bar{b}$$

where $\bar{A}$ and $\bar{b}$ are known through stochastic estimates $A(Z)$, $b(Z)$ for a sequence of random variables Z

... variance of $A(Z)$ and $b(Z)$ are typically very large

... and $\bar{A}$ is not symmetric

Note: It is inefficient to cast it as a minimization problem with loss $\|\bar{A}\theta_\star - \bar{b}\|^2$
$\rightarrow$ This requires a different method, with a different analysis

Algorithm for LSA

Initialize $\theta_0 \in \mathbb{R}^d$

for $t = 0$ to $T - 1$ **do**

Observe Z_t and update:

$$\theta_t = \theta_{t-1} - \eta(A(Z_t)\theta_{t-1} - b(Z_t))$$

end for

Context, analysis of TD (I)[2]

Initialize $\theta_0 \in \mathbb{R}^d$
for $t = 0$ to $T - 1$ **do**
    Observe $Z_{t,h}^c$ and update: $\theta_t = \theta_{t-1} - \eta(A(Z_t)\theta_{t-1} - b(Z_t))$
end for

[2] Sergey Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *COLT*. PMLR. 2024, pp. 4511–4547.

Context, analysis of TD (I)[2]

Initialize $\theta_0 \in \mathbb{R}^d$
for $t = 0$ to $T - 1$ **do**
  Observe $Z^c_{t,h}$ and update: $\theta_t = \theta_{t-1} - \eta(A(Z_t)\theta_{t-1} - b(Z_t))$
end for

Stochastic Expansion

We may write: $\theta_t - \theta_\star = (\mathrm{Id} - \eta A(Z_t))(\theta_{t-1} - \theta_\star) - \eta\varepsilon(Z_t)$

[2] Sergey Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *COLT*. PMLR. 2024, pp. 4511–4547.

Context, analysis of TD (I)[2]

Initialize $\theta_0 \in \mathbb{R}^d$
for $t = 0$ to $T-1$ **do**
  Observe $Z^c_{t,h}$ and update: $\theta_t = \theta_{t-1} - \eta(A(Z_t)\theta_{t-1} - b(Z_t))$
end for

Stochastic Expansion

We may write: $\theta_t - \theta_\star = (\mathrm{Id} - \eta A(Z_t))(\theta_{t-1} - \theta_\star) - \eta\varepsilon(Z_t)$

Assumptions

- Oracle: i.i.d sequence Z_t's such that $\mathbb{E}[A(Z_t)] = \bar{A}$, and $\mathbb{E}[b(Z_t)] = \bar{b}$
- Exponential stability: $\mathbb{E}[\| \prod_{t=\ell}^{k}(\mathrm{Id} - \eta A(Z_t))\|^2] \leq (1-\eta a)^{k-\ell}$ for some $a > 0$
- Noise $\varepsilon(Z) = (A(Z) - \bar{A})\theta_\star + (b(Z) - \bar{b})$ has finite variance $\sigma_\star^2$

[2] Sergey Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *COLT*. PMLR. 2024, pp. 4511–4547.

Context, analysis of TD (II)[3]

Stochastic Expansion

$$\theta_T - \theta_\star = \Gamma_{1:T}(\theta_0 - \theta_\star) + \eta \sum_{t=1}^{T} \Gamma_{t+1:T}\varepsilon(Z_t)$$

Where $\Gamma_{t:t'}$ "accumulates the updates" from t to t':

$$\Gamma_{t:t'} = (\mathrm{Id} - \eta A(Z_{t'}))(\mathrm{Id} - \eta A(Z_{t'-1}))\cdots(\mathrm{Id} - \eta A(Z_t))$$

[3] Sergey Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *COLT*. PMLR. 2024, pp. 4511–4547.

Context, analysis of TD (III)[4]

Stochastic Expansion

$$\theta_T - \theta_\star = \Gamma_{1:T}(\theta_0 - \theta_\star) + \eta \sum_{t=1}^{T} \Gamma_{t+1:T}\varepsilon(Z_t)$$

Using $\mathbb{E}[\|\Gamma_{t:t'} u\|^2] \leq (1-\eta a)^{t'-t+1}\|u\|^2$ to bound each term:

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \leq (1-\eta a)^T \|\theta_0 - \theta_\star\|^2 + \frac{\eta \sigma_\star^2}{a}$$

[4] Sergey Samsonov et al. "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *COLT*. PMLR. 2024, pp. 4511–4547.

Federated LSA

Take $\bar{A}^c, \bar{b}^c$ such that $\bar{A}^c \theta_\star^c = \bar{b}^c$ for $c = 1..N$

Federated LSA

Take $\bar{A}^c, \bar{b}^c$ such that $\bar{A}^c \theta_\star^c = \bar{b}^c$ for $c = 1..N$

Goal: solve collaboratively

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

Federated LSA

Take $\bar{A}^c, \bar{b}^c$ such that $\bar{A}^c \theta_\star^c = \bar{b}^c$ for $c = 1..N$

Goal: solve collaboratively

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

Assumptions

- $\theta_\star$ and $\theta_\star^c$ are unique, and $\bar{A}^c$ and $\bar{b}^c$ are split among N agents
- Oracle: i.i.d sequence Z_t^c's such that $\mathbb{E}[A(Z_t^c)] = \bar{A}^c$, and $\mathbb{E}[b(Z_t^c)] = \bar{b}^c$
- Exponential stability: $\mathbb{E}[\| \prod_{t=\ell}^{k} (\mathrm{Id} - \eta A^c(Z_t^c)) \|^2] \leq (1 - \eta a)^{k-\ell}$ for $a > 0$
- Noise $\varepsilon^c(Z) = (A^c(Z) - \bar{A}^c)\theta_\star^c + (b^c(Z) - \bar{b}^c)$ has variance bounded by $\sigma_\star^2$

Solving Federated LSA

Paul Mangold et al. "SCAFFLSA: Taming Heterogeneity in Federated Linear Stochastic Approximation and TD Learning". In: *NeurIPS* (2024)

FedLSA Algorithm

for $t = 0$ to $T - 1$ **do**

Initialize $\theta_{t,0} = \theta_t$

for each agent $c = 1..N$ **do**

for $h = 1$ to H **do**

Observe $Z^c_{t,h}$ and perform local update:

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

end for

end for

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

end for

Analysis of FedLSA

Stochastic Expansion (over one communication round)

$$\theta_t - \theta_\star = \frac{1}{N}\sum_{c=1}^{N} \Gamma^c_{t,1:H}(\theta_{t-1} - \theta_\star) + \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \Gamma^c_{t,1:H})(\theta^c_\star - \theta_\star)$$
$$+ \frac{\eta}{N}\sum_{c=1}^{N}\sum_{h=1}^{H} \Gamma^c_{t,h+1:H}\varepsilon^c(Z^c_t)$$

Where $\Gamma^c_{t,h:h'}$ "accumulates local updates", round t, from h to h',

$$\Gamma^c_{t,h:h'} = (\mathrm{Id} - \eta A^c(Z^c_{t,h'}))(\mathrm{Id} - \eta A^c(Z^c_{t,h'-1}))\cdots(\mathrm{Id} - \eta A^c(Z^c_{t,h}))$$

Analysis of FedLSA

We can characterize the bias of FedLSA:

$$\theta_{\infty}^{\mathrm{bias}} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \bar{\Gamma}_{t,1:H})^{-1}(\mathrm{Id} - (\mathrm{Id} - \eta \bar{A}^{c})^{H})\{\theta_{\star}^{c} - \theta_{\star}\}$$

where $\bar{\Gamma}_{t,1:H} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \eta \bar{A}^{c})^{H}$

Analysis of FedLSA

We can characterize the bias of FedLSA:

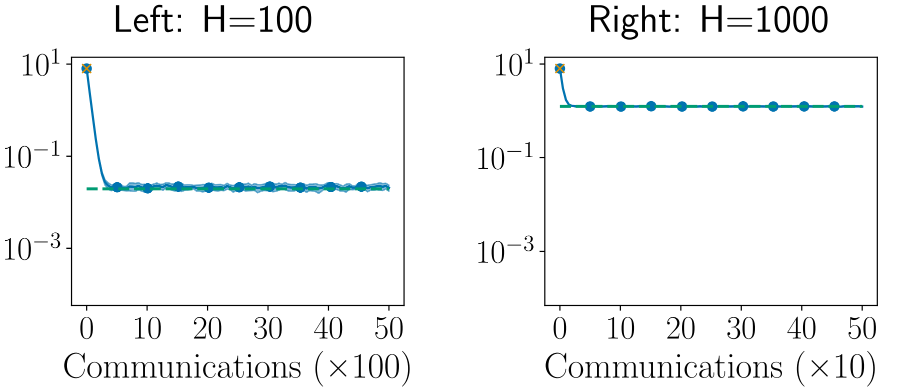
$$\theta_\infty^{\text{bias}} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \bar{\Gamma}_{t,1:H})^{-1}(\mathrm{Id} - (\mathrm{Id} - \eta \bar{A}^c)^H)\{\theta_\star^c - \theta_\star\}$$

where $\bar{\Gamma}_{t,1:H} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \eta \bar{A}^c)^H$

And give a convergence rate

$$\mathbb{E}\left[\|\theta_t - \theta_\infty^{bias} - \theta_\star\|^2\right] = O\left((1-\eta a)^{Ht}\|\theta_0 - \theta_\star\|^2 + \frac{\eta\sigma_\star^2}{Na}\right)$$

Numerical Illustration ($N = 100$ agents)



Blue line: FedLSA's mean squared error
Green line: FedLSA's bias as predicted by our theory

Problem: heterogeneity requires lots of communications

To achieve $\mathbb{E}\left[\|\theta_T - \theta_\star\|^2\right] \leq \epsilon^2$, we need

- $\frac{\eta\sigma_\star^2}{Na} \leq \epsilon^2$ $\quad\rightarrow \eta = \frac{Na\epsilon^2}{\sigma_\star^2}$
- $\|\theta_T^{\text{bias}}\|^2 \leq \epsilon^2$ $\quad\rightarrow H = \frac{\sigma_\star^2}{N\epsilon\Delta_{\text{het}}}$
- $(1-\eta a)^{HT}\|\theta_0 - \theta_\star\|^2 \leq \epsilon^2$ $\quad\rightarrow T = \frac{\Delta_{\text{het}}}{a^2\epsilon}\log\frac{\|\theta_0-\theta_\star\|}{\epsilon}$

where $\Delta_{\text{het}} = \frac{1}{N}\sum_{c=1}^{N}\|\theta_\star - \theta_\star^c\|$

Solution: Control variates (SCAFFLSA)[5]

for $t = 0$ to $T - 1$ **do**
  Initialize $\theta_{t,0} = \theta_t$
  for each agent $c = 1..N$ **do**
    for $h = 1$ to H **do**
      Observe $Z^c_{t,h}$ and perform local update:
$$\theta_{t,h} = \theta^c_{t,h-1} - \eta(A^c(Z^c_{t,h})\theta^c_{t,h-1} - b^c(Z^c_{t,h}) - \xi_t)$$
    end for
  end for
  Aggregate local updates $\theta_{t+1} = \frac{1}{N}\sum_{c=1}^{N} \theta^c_{t,H}$
  Update control variate $\xi_{t+1} = \xi_t - \frac{1}{\eta H}(\theta_{t+1} - \theta^c_{t,H})$
end for

[5] Based on Sai Praneeth Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: *ICML*. PMLR. 2020, pp. 5132–5143

Theoretical analysis

We prove, assuming $H \leq \frac{a}{\eta \max_c \|\bar{A}^c\|^2}$

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \left(1 - \frac{\eta a H}{2}\right)^T \psi_0 + \frac{\eta \sigma_\star^2}{N a}$$

with $\psi_0 = \|\theta_0 - \theta_\star\|^2 + \frac{\eta^2 H^2}{N} \sum_{c=1}^N \|\bar{A}^c(\theta_\star^c - \theta_\star)\|^2$

Theoretical analysis

We prove, assuming $H \leq \frac{a}{\eta \max_c \|\bar{A}^c\|^2}$

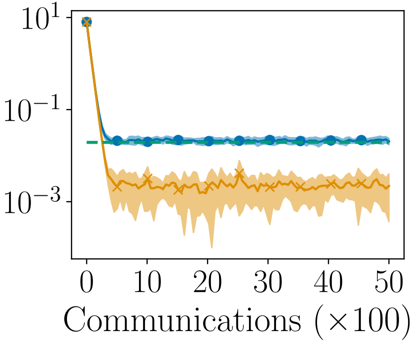
$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \left(1 - \frac{\eta a H}{2}\right)^T \psi_0 + \frac{\eta \sigma_\star^2}{\boldsymbol{Na}}$$

with $\psi_0 = \|\theta_0 - \theta_\star\|^2 + \frac{\eta^2 H^2}{N} \sum_{c=1}^{N} \|\bar{A}^c(\theta_\star^c - \theta_\star)\|^2$

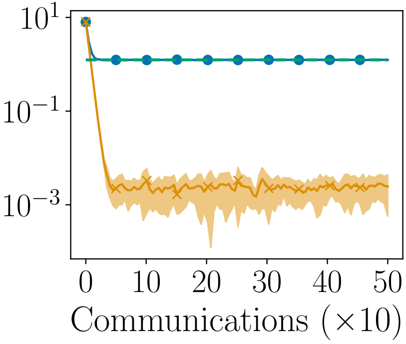
Note on analysis
Direct analysis “à la LSA” does not work. We need a “Lyapunov” analysis, and to carefully study covariances of control variates to obtain linear speed-up.

Numerical Illustration ($N = 100$ agents)

Left: H=100



Right: H=1000



Blue line: FedLSA's mean squared error

Orange line: SCAFFLSA's mean squared error

Communication Complexity

To achieve $\mathbb{E}\left[\|\theta_T - \theta_\star\|^2\right] \leq \epsilon^2$, we need

- $\frac{\eta\sigma_\star^2}{Na} \leq \epsilon^2$ $\rightarrow \eta = \frac{Na\epsilon^2}{\sigma_\star^2}$
- $H \leq \frac{a}{\eta \max_c \|\bar{A}^c\|^2}$ $\rightarrow H = \frac{\sigma_\star^2}{N\epsilon^2 \max_c \|\bar{A}^c\|^2}$
- $(1 - \frac{\eta a H}{2})^T \psi_0 \leq \epsilon^2$ $\rightarrow T = \frac{2 \max_c \|\bar{A}^c\|^2}{a^2} \log \frac{\psi_0}{\epsilon}$

$\rightarrow H \propto 1/N\epsilon^2$ rather than $1/N\epsilon$, and T independent on ϵ

What about the Non-Linear Case?

Back to FedAvg

$$\theta_\star \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{c=1}^{N} f^c(\theta) \ , \qquad \text{where } f^c(\theta) = \mathbb{E}_{x^c, y^c \sim \mathcal{D}^c}\left[\ell(\theta; x^c, y^c)\right]$$

Federated Averaging (or local (S)GD)[6]

- For each $t = 0 \ldots$:
  - Set $\theta^c_{t,0} = \theta_t$
  - For each agent c, do H gradient updates:

$$\theta^c_{t,h+1} = \theta^c_{t,h} - \eta \nabla f^c(\theta^c_{t,h})$$

- Aggregate models: $\theta_{t+1} = \frac{1}{N} \sum_{c=1}^{N} \theta^c_{t,H}$

[6] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *AISTATS*. PMLR. 2017, pp. 1273–1282.

Back to FedAvg

$$\theta_\star \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{c=1}^{N} f^c(\theta) \ , \qquad \text{where } f^c(\theta) = \mathbb{E}_{x^c, y^c \sim \mathcal{D}^c}\left[\ell(\theta; x^c, y^c)\right]$$

F

What can we say about the bias?

$$\theta^c_{t,h+1} = \theta^c_{t,h} - \eta \nabla f^c(\theta^c_{t,h})$$

- Aggregate models: $\theta_{t+1} = \frac{1}{N} \sum_{c=1}^{N} \theta^c_{t,H}$

[6] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *AISTATS*. PMLR. 2017, pp. 1273–1282.

For Quadratics ($f^c(\theta) = (1/2)\theta^\top \bar{A}^c \theta + \bar{b}^c \theta$)

The bias is the same as FedLSA

$$\theta_\infty^{\text{bias}} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \bar{\Gamma}_{t,1:H})^{-1}(\mathrm{Id} - (\mathrm{Id} - \eta \bar{A}^c)^H)\{\theta_\star^c - \theta_\star\}$$

where $\bar{\Gamma}_{t,1:H} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \eta \bar{A}^c)^H$

For Quadratics $(f^c(\theta) = (1/2)\theta^\top \bar{A}^c\theta + \bar{b}^c\theta)$

The bias is the same as FedLSA

$$\theta_\infty^{\text{bias}} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \bar{\Gamma}_{t,1:H})^{-1}(\mathrm{Id} - (\mathrm{Id} - \eta\bar{A}^c)^H)\{\theta_\star^c - \theta_\star\}$$

where $\bar{\Gamma}_{t,1:H} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id} - \eta\bar{A}^c)^H$

And we can give first order expansion:

$$\theta_\infty^{\text{bias}} = \frac{\eta(H-1)}{2N}\sum_{c=1}^{N}\nabla^2 f^c(\theta_\star)^{-1}(\nabla^2 f^c(\theta_\star) - \nabla^2 f(\theta_\star))\nabla f^c(\theta_\star) + O(\eta^2 H^2)$$

In the General Case

(Strongly convex and smooth functions f^c)

Bias is in *two* parts!

$$\theta_\infty^{\text{bias}} = \frac{\eta(H-1)}{2N}\sum_{c=1}^{N} \nabla^2 f^c(\theta_\star)^{-1}(\nabla^2 f^c(\theta_\star) - \nabla^2 f^c(\theta_\star))\nabla f^c(\theta_\star)$$
$$+ \frac{\eta}{2N}\nabla^2 f^c(\theta_\star)^{-1}\nabla^3 f(\theta_\star)\mathbf{A}\mathcal{C}(\theta_\star) + O(\eta^{3/2}H + \eta^2 H^2)$$

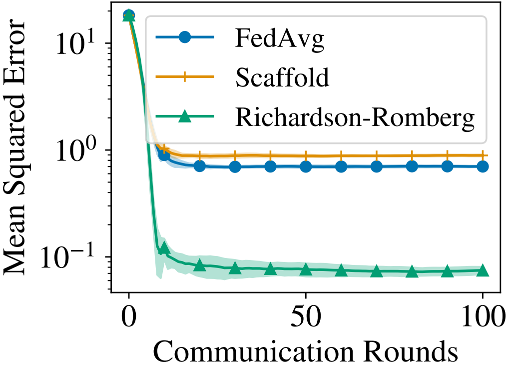
where:

- $\mathbf{A} = (\mathrm{Id} \otimes \nabla^2 f(\theta_\star) + \nabla^2 f(\theta_\star) \otimes \mathrm{Id})^{-1}$
- $\mathcal{C}(\theta_\star)$ is the gradient's covariance at $\theta_\star$

A new Federated Method?

(A bit of teasing on Richardson-Romberg)

Running FedAvg with step sizes η and 2η, we can correct the bias:



$\rightarrow$ it seems Scaffold cannot correct bias due to stochasticity!

Conclusion and Perspectives

Summary:

- We studied FedLSA's communication complexity
- We extended control variates methods to FedLSA
- We showed that both methods have linear speed-up (up to bias)
- We proved first-order expansion of FedAvg's bias

Perspectives:

- SCAFFLSA's analysis is good for small step-size: what about larger steps?
- Direct analysis of SCAFFLSA "à la FedLSA"?
- Removing hyperparameters?
- Asynchronous federated learning?

Thank you!

Questions?

See the papers:

P. Mangold, S. Samsonov, S. Labbi, I. Levin, R. Alami, A. Naumov, and E. Moulines. "SCAFFLSA: Taming Heterogeneity in Federated Linear Stochastic Approximation and TD Learning". In: *NeurIPS* (2024)

On FedAvg and Richardson-Romberg (with E. Moulines, A. Durmus, A. Dieuleveut and S. Samsonov): soon on arXiv!