A Sharper Analysis of SCAFFOLD on Quadratics

Paul Mangold Eric Moulines

CMAP, CNRS, École Polytechnique, Institut Polytechnique de Paris

FLTA 2025

Handling heterogeneity in federated learning

- Challenge: heterogeneity between clients

  $\Longrightarrow$ local drift and bias.

- FedAvg[1]: suffers under heterogeneity.
- Scaffold[2]: uses **control variates** to correct drift.

  $\Longrightarrow$ Yet, theory of Scaffold remains incomplete:

  - what is its correct contraction rate?
  - does it converge for any fixed number of local steps?

In this work, we answer both questions, positively!

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

[2] Sai Praneeth Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: *ICML*. 2020.

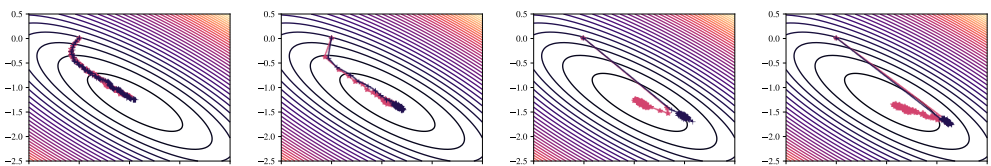
Goal of This Work

Goal: Provide a **refined theoretical analysis** of SCAFFOLD for quadratic objectives.

Contributions:

- We establish convergence for *any number of local steps*
  ...by decomposing dynamics along the **eigenspaces of the Hessian**.
- We identify optimal scaling of the number of local updates:

$$H_\gamma \propto \frac{1}{\gamma\sqrt{\mu L}}.$$



Blue: FedAvg, Pink: Scaffold. Left to right: $1, 10, 100, 1000$ local steps

Problem Setup

We minimize the quadratic FL objective:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{N} \sum_{c=1}^{N} \left(\tfrac{1}{2} \theta^\top A_c \theta - b_c^\top \theta \right),$$

with $A_c \succ 0$ and heterogeneous across clients.

Problem Setup

We minimize the quadratic FL objective:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{N} \sum_{c=1}^{N} \left(\tfrac{1}{2} \theta^\top A_c \theta - b_c^\top \theta \right),$$

with $A_c \succ 0$ and heterogeneous across clients.

Scaffold updates (deterministic, but our results also cover stochastic case):

$$\theta_{t,h+1}^c = \theta_{t,h}^c - \gamma (A_c \theta_{t,h}^c - b_c + \xi_t^c)$$

$$\theta_{t+1} = \frac{1}{N} \sum_{c=1}^{N} \theta_{t,H}^c \qquad \xi_{t+1}^c = \xi_t^c + \frac{1}{\gamma H} (\theta_{t,H}^c - \theta_{t+1})$$

- Control variates ξ_c estimate gradient at optimum.
- Bias correction mitigates client drift.

Key Analysis Idea: Spectral Decomposition

Main Result: assume $\mu I \preccurlyeq A_c \preccurlyeq LI$, and $\gamma \leq 1/L$. For any number of local steps $H > 0$,

$$\mathbb{E}[\|X_{t+1} - X_*\|_\Lambda^2] \leq \rho_{\gamma,H}\, \mathbb{E}[\|X_t - X_*\|_\Lambda^2],$$

where $X_t = (\theta_t, \xi_t^1, \ldots, \xi_t^N)$ and $X_* = (\theta_*, \xi_*^1, \ldots, \xi_*^N)$

$$\rho_{\gamma,H} = \max\left\{(1-\gamma\mu)^H,\, 1 - \frac{1-1/e}{\gamma L H}\right\} < 1.$$

Proof idea:

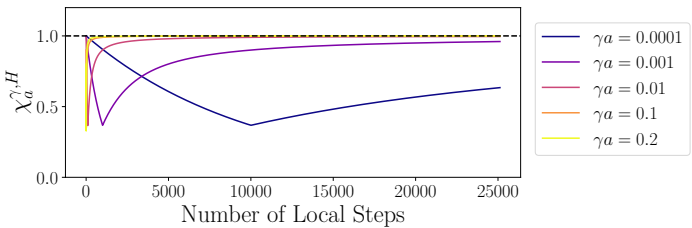
- Write each local Hessian: $A_c = U_c D_c U_c^\top$.
- Decompose the dynamics along eigen-directions of A_c.
- This per-eigenvalue analysis of contraction!

$\Longrightarrow$ **Scaffold remains convergent for all H!**

Interpretation of the Contraction Rate

For each eigenvalue a of A_c:

- Two regimes:
  1. $\gamma aH \leq 1$: exponential decay $(1-\gamma a)^H \leq (1-\gamma\mu)^H$ $\rightarrow$ **faster with larger H**
  2. $\gamma aH > 1$: algebraic decay $1 - \frac{1-1/e}{\gamma aH} \leq 1 - \frac{1-1/e}{\gamma LH}$ $\rightarrow$ **slower with larger H**
- Transition when $\gamma aH = 1$.



Contraction rate vs number of local steps.

Acceleration and Optimal Local Steps

Balancing the two regimes requires setting:

$$(1-\gamma\mu)^H = 1 - \frac{1-1/e}{\gamma L H} \Rightarrow \boldsymbol{H_\gamma = \left\lceil \sqrt{\frac{2(1-1/e)}{\gamma^2 L \mu}} \right\rceil}$$

[3] Konstantin Mishchenko et al. "Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally!" In: *ICML*. 2022.

[4] Zhengmian Hu and Heng Huang. "Tighter analysis for proxskip". In: *ICML*. 2023.

Acceleration and Optimal Local Steps

Balancing the two regimes requires setting:

$$(1-\gamma\mu)^H = 1 - \frac{1-1/e}{\gamma L H} \Rightarrow \boldsymbol{H_\gamma = \left\lceil \sqrt{\frac{2(1-1/e)}{\gamma^2 L \mu}} \right\rceil}$$

- Optimal rate: $\rho_{\text{opt}} = 1 - \sqrt{\frac{2(1-1/e)\mu}{L}}$.
- Communication complexity: $T = O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$,

  i.e., $\sqrt{L/\mu}$ improvement over FedAvg.

Recovers both the rates of ProxSkip and its refined analyses... but without stochastic communication scheme.[3,4]

[3] Konstantin Mishchenko et al. "Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally!" In: *ICML*. 2022.

[4] Zhengmian Hu and Heng Huang. "Tighter analysis for proxskip". In: *ICML*. 2023.

Extension to Stochastic Setting

SCAFFOLD with noise converges to a unique stationary distribution. Its variance Σ_X satisfies:

$$\Sigma_X = A_{\gamma,H} \Sigma_X A_{\gamma,H}^\top + B_{\gamma,H}.$$

Extension to Stochastic Setting

SCAFFOLD with noise converges to a unique stationary distribution. Its variance Σ_X satisfies:

$$\Sigma_X = A_{\gamma,H} \Sigma_X A_{\gamma,H}^\top + B_{\gamma,H}.$$

Result on variance: for homogeneous quadratics, $\mathrm{tr}(\Sigma_\theta) \lesssim \frac{\gamma \sigma^2}{N}$

$\Longrightarrow$ Linear speed-up with number of clients N.

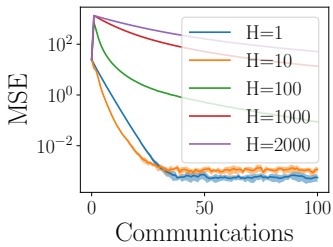
Conjectures:

1. this result still holds in heterogeneous settings
2. this result still holds in non-linear (strongly-convex) case

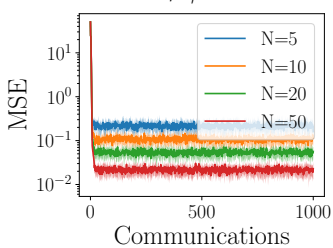
Numerical Results

- SCAFFOLD converges for all H if $\gamma \leq 1/L$.
- Accelerated convergence up to optimal H_γ.
- Variance decreases linearly with number of clients.

$N = 10$, $\gamma = 1$:



$H = 10$, $\gamma = 1$:



Discussion and Perspectives

- New spectral framework explains convergence of SCAFFOLD for any number of local steps.
- Establishes acceleration (and speed-up in homogeneous cases).
- Open directions:
  - Influence of heterogeneity on convergence? Does Scaffold match the lower bounds?
  - Establishing speed-up beyond homogeneous functions.
  - Going beyond quadratic functions.

Questions?