A Sharper Analysis of Scaffold on Quadratics

Paul Mangold
CMAP, CNRS, École Polytechnique,
Institut Polytechnique de Paris, 91120 Palaiseau, France

paul.mangold@polytechnique.edu

Eric Moulines
CMAP, CNRS, École Polytechnique,
Institut Polytechnique de Paris, 91120 Palaiseau, France
MBZUAI, UAE,
eric.moulines@polytechnique.edu

***Abstract*—Heterogeneity across client datasets poses a challenge in federated learning (FL), impairing the convergence and performance of classical distributed optimization methods. The SCAFFOLD algorithm has emerged as a prominent approach to mitigate heterogeneity; despite many efforts, its theoretical properties remain incompletely understood. In this paper, we present a refined analysis of SCAFFOLD for quadratic objectives. Our results establish convergence guarantees valid for an arbitrary number of local steps. A distinctive aspect of our analysis is the spectral decomposition of the Hessian matrix governing the quadratic optimization problem, revealing that the convergence dynamics of SCAFFOLD are determined by its behavior across individual eigenspaces. We complement our theoretical contributions with illustrative numerical examples, elucidating empirical observations previously unexplained by existing analyses. Our findings deepen the understanding of variance-reduction techniques in federated learning and provide insights for future algorithmic design.**

***Index Terms*—federated learning, optimization, local training, heterogeneity**

I. INTRODUCTION

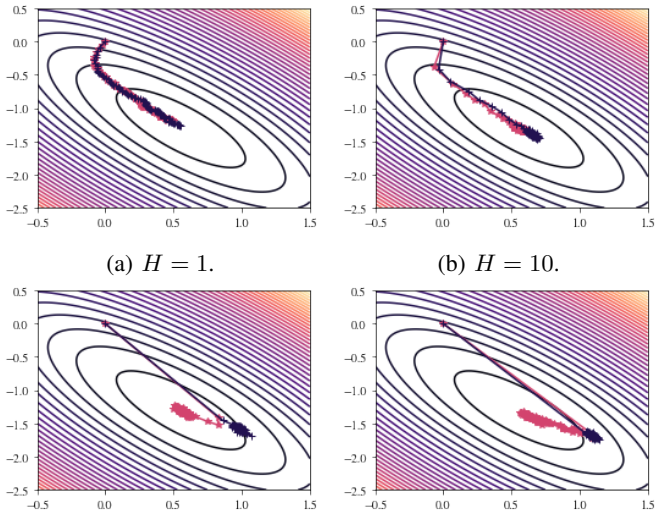
In federated learning (FL), multiple clients train a shared machine learning model without exchanging their raw data. Clients engage in distributed optimization coordinated by a central server, with periodic synchronization. A significant challenge arises from the divergence of local data distributions, commonly referred to as *statistical heterogeneity*.

To reduce the communication overhead inherent to federated learning, clients often perform multiple local updates before communicating with the server. However, under statistical heterogeneity, such local steps may induce large bias, hindering convergence and precision of the result. Various approaches address heterogeneity-induced biases, most notably the SCAFFOLD algorithm, which uses control variates to correct for client-level discrepancies.

Despite its empirical success, the theoretical foundations of SCAFFOLD remain incomplete. The original analysis of SCAFFOLD covers both convex and non-convex objectives [1]–[3]. However, these theoretical guarantees require the number of local steps to be bounded by conservative problem-dependent quantities, which are generally unknown in practice. Consequently, practical deployments of SCAFFOLD often operate outside the theoretical conditions guaranteeing convergence.

To address some of these limitations, a variant of SCAFFOLD, termed SCAFFNEW, was introduced by [4]. In SCAFFNEW, clients decide whether to communicate at each iteration, leading to a variable number of local steps between synchronizations. Although SCAFFNEW provides convergence guarantees for arbitrary communication probabilities [4], [5], its analysis suffers from two significant drawbacks. First, the analysis technique employed fails to track the iterates' variance, thereby precluding a theoretical analysis of the variance of the global iterates. Second, SCAFFNEW's stochastic communication scheme departs from common practice and may compromise performance in practical deployments by introducing additional randomness.

Our work aims to bridge these theoretical gaps, providing a sharper analysis of the original SCAFFOLD algorithm and delivering precise guarantees. We focus on the simpler quadratic setting, which allows the derivation of very precise results, offering new insights on the SCAFFOLD algorithm.



(a) $H = 1$. (b) $H = 10$. (c) $H = 100$. (d) $H = 1000$.

Fig. 1: Iterates of FEDAVG and SCAFFOLD on a 2D quadratic problem with two clients, with different numbers of local steps H. As H increases, FEDAVG becomes increasingly biased, whereas SCAFFOLD still converges. More crucially, SCAFFOLD converges even for very large numbers of local steps: this behavior is not covered by existing analyses. We give a theoretical explanation for this in Section III.

Our contributions are the following:

- We show that SCAFFOLD converges with any number of local steps, and derive a precise convergence rate.
- Our analysis precisely analyzes the properties of SCAFFOLD in each eigenspace of the problem's Hessian matrix, (i) determining the number of local steps that one can perform in each direction without harming convergence, and (ii) showing that the optimal number of local updates scales in $\frac{1}{\gamma\sqrt{\mu L}}$, recovering both ideas from [4] and [5].
- We illustrate the phenomena that arise from our theory in multiple, simple, numerical problems, shedding light on the properties of SCAFFOLD.

This study lays the foundation for developing more sophisticated analyses of SCAFFOLD in broader contexts and may ultimately inform the design of even more efficient SCAFFOLD variants. We begin by describing our experimental setup and presenting initial numerical insights in Section II. Section III introduces our novel theoretical analysis, followed by comprehensive numerical validation of the identified phenomena in Section IV. In Section V, we discuss the implications of our findings and the new research directions they reveal.

Notations. For $(u_t)_{t\geq 0}$, $(v_t)_{t\geq 0}$, we denote $u_t \lesssim v_t$ if there exists a constant $c > 0$ such that $u_t \leq c \cdot v_t$ for all $t \geq 0$. For a vector $u \in \mathbb{R}^d$ and $1 \leq j \leq d$, we denote u_j the j-th coordinate of u.

II. Problem Setup

Federated Learning. We consider the following quadratic federated optimization problem, where each client $c \in \{1, \dots, N\}$, holds a symmetric, positive definite matrix $\mathbf{A}_c \in \mathbb{R}^{d\times d}$, and a vector $\mathbf{b}_c \in \mathbb{R}^d$.

$$\theta^\star = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{N}\sum_{c=1}^{N}\left(\tfrac{1}{2}\theta^\top \mathbf{A}_c\theta - \mathbf{b}_c^\top\theta\right) \ , \qquad (1)$$

where we assume that clients have access to a stochastic gradient oracle of the local objective functions of the form $\theta \mapsto \mathbf{A}_c\theta - \mathbf{b}_c + \varepsilon_c$, where ε_c is of expectation zero and has bounded variance. To solve (1), clients typically perform multiple local gradient updates, which are periodically averaged by the central server. This procedure, introduced by [6], is called Federated Averaging (FEDAVG). However, FEDAVG is sensitive to client heterogeneity, which induces *client drift* towards local solutions.

SCAFFOLD. To overcome this issue, [1] proposed the SCAFFOLD algorithm, which uses control variates to suppress heterogeneity bias. In Scaffold, each client performs $H > 0$ local updates of the form, e.g. for $h \in \{0, \dots, H\}$,

$$\theta_c^{t,h+1} = \theta_c^{t,h} - \gamma(\mathbf{A}_c\theta_c^{t,h} - \mathbf{b}_c + \varepsilon_c^{t,h+1} + \xi_c^t) \ ,$$

where, for $t, h \geq 0$, $c \in \{1, \dots, N\}$ $\{\varepsilon_c^{t,h}\} \in \mathbb{R}^d$ is a collection of i.i.d random variables such that $\mathbb{E}[\varepsilon_c^{t,h}] = 0$ for all $\theta \in \mathbb{R}^d$ and

$$\mathbb{E}[\|\varepsilon_c^{t,h}\|^2] \leq \sigma^2 \ , \text{ for all } t, h \geq 0, \text{and } c \in \{1, \dots, N\} \ . \quad (2)$$

Algorithm 1 SCAFFOLD

Input: initial $\theta^0 \in \mathbb{R}^d$ and $\xi_1^0, \dots, \xi_N^0 \in \mathbb{R}^d$, step size $\gamma > 0$, number of rounds $T > 0$, number of clients $N > 0$, number of local steps $H > 0$

for $t = 0$ to $T - 1$ **do**
  for $c = 1$ to N **do**
    Initialize $\theta_c^{t,0} = \theta^t$
    for $h = 0$ to $H - 1$ **do**
      Set $\theta_c^{t,h+1} = \theta_c^{t,h} - \gamma(\mathbf{A}_c\theta_c^{t,h} - \mathbf{b}_c + \varepsilon_c^{t,h+1} + \xi_c^t)$
    end for
  end for
  Update: $\theta^{t+1} = \frac{1}{N}\sum_{c=1}^{N}\theta_c^{t,H}$
  Update: $\xi_c^{t+1} = \xi_c^t + \frac{1}{\gamma H}(\theta_c^{t,H} - \theta^{t+1})$
end for
Return: θ^T

The local iterates are then aggregated by the server as

$$\theta^{t+1} = \tfrac{1}{N}\textstyle\sum_{c=1}^{N}\theta_c^{t,H} \ .$$

Upon receiving the updated global parameter, each client updates its local control variate ξ_c^t according to

$$\xi_c^{t+1} = \xi_c^t + \tfrac{1}{\gamma H}(\theta_c^{t,H} - \theta^{t+1}) \ .$$

These control variates are crucial in the algorithm as they slowly learn the local drift. Specifically, they approximate the gradient at the optimal solution, with the ideal control variates being $\xi_c^\star = -(\mathbf{A}_c\theta^\star - \mathbf{b}_c)$, where $\theta^\star$ denotes the global optimum. Existing convergence analyses for SCAFFOLD (see, e.g., [1], [3]) establish convergence bounds of the following form, with some upper bound on the number of local steps,

$$\mathbb{E}[\|\theta^t - \theta^\star\|^2] \lesssim (1-\gamma\mu)^{Ht}\psi_0 + \tfrac{\gamma\sigma^2}{\mu} \ ,$$

where $\psi_0 = \|\theta^0 - \theta^\star\|^2 + \frac{\gamma^2 H^2}{N}\sum_{c=1}^{N}\|\xi_c^0 - \xi_c^\star\|^2$, and μ represents the strong convexity parameter (i.e., the smallest eigenvalue among all local Hessians $\mathbf{A}_c$).

III. Sharp Analysis for Quadratics

We analyze SCAFFOLD for quadratic objective functions, where each client $c \in \{1, \dots, N\}$ has a local function of the form $f_c(\theta) = \mathbf{A}_c\theta + \mathbf{b}_c$, with $\mathbf{b}_c \in \mathbb{R}^d$ and positive definite matrices $\mathbf{A}_c$ with eigendecomposition $\mathbf{A}_c = \mathbf{U}_c\mathbf{D}_c\mathbf{U}_c^\top$. Here, $\mathbf{U}_c$ is an orthonormal matrix (i.e., $\mathbf{U}_c^\top\mathbf{U}_c = \mathrm{Id}$) and $\mathbf{D}_c = \mathrm{diag}(a_{c,1}, \dots, a_{c,d})$ with $a_{c,i} > 0$ for all i.

SCAFFOLD as a Markov chain. The sequence of iterates generated by SCAFFOLD naturally forms a Markov process, which corresponds to an iterated random function system. The Markov property follows from the algorithm's structure: given the current state $\mathrm{X}^t = (\theta^t, \xi_1^t, \dots, \xi_N^t)$, the distribution of the next state X^{t+1} depends only on X^t and is independent of the entire history prior to time t. More formally, the global parameters and control variates of SCAFFOLD evolve as a time-homogeneous Markov chain on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{X} = \mathbb{R}^d \times (\mathbb{R}^d)^N$ is the joint space of

global and local control variates. We denote the corresponding transition kernel by $\mathcal{K}_{\gamma,H}$ and its t-fold composition by $\mathcal{K}^t_{\gamma,H}$. Starting from an initial probability distribution ν on $\mathcal{X}$, the law of the state at time t is given by $\nu\,\mathcal{K}^t_{\gamma,H}$.

Our subsequent analysis establishes the existence and uniqueness of a stationary distribution toward which the law of the iterates converge. This convergence result relies fundamentally on demonstrating that the transition kernel $\mathcal{K}_{\gamma,H}$ is contractive, a property that we will establish through careful analysis of the algorithm's dynamics.

Convergence to stationarity. We recall that the local updates, with $\mathbf{A}_c = \mathbf{U}_c\mathbf{D}_c\mathbf{U}_c^\top$, take the form

$$\theta_c^{t,h+1} = \theta_c^{t,h} - \gamma\mathbf{A}_c(\theta_c^{t,h} - \theta^\star) - \gamma(\xi_c^t - \xi_c^\star) - \gamma\varepsilon_c^{t,h+1} \quad .$$

After H local iterations, the deviation from the optimum can be expressed as

$$\theta_c^{t,H} - \theta^\star = \Gamma_c^H(\theta^t - \theta^\star) - \gamma\textstyle\sum_{h=0}^{H-1}\Gamma_c^h(\xi_c^t - \xi_c^\star) - \gamma\mathcal{E}_c^t \ ,$$

where $\Gamma_c = \mathrm{Id} - \gamma\mathbf{A}_c$ is the local contraction matrix and

$$\mathcal{E}_c^t = \textstyle\sum_{h=1}^{H}\Gamma_c^{H-h}\varepsilon_c^{t,h} \tag{3}$$

is the noise accumulated over the local updates.

Using the matrix identity $\sum_{h=0}^{H-1}B^h = (\mathrm{Id} - B)^{-1}(\mathrm{Id} - B^H)$ when B is invertible, we can rewrite

$$\theta_c^{t,H} - \theta^\star = \Gamma_c^H(\theta^t - \theta^\star) - \mathbf{A}_c^{-1}(\mathrm{Id} - \Gamma_c^H)(\xi_c^t - \xi_c^\star) - \gamma\mathcal{E}_c^t \quad . \tag{4}$$

To analyze the algorithm's convergence, we use the following weighted distance metric on $\mathbb{R}^{(N+1)d}$, originally introduced in [3] and inspired by the Lyapunov functions from [1], [4]

$$\|\mathrm{X}^t\|_\Lambda^2 \triangleq \|\theta^t\|^2 + \tfrac{\gamma^2H^2}{N}\textstyle\sum_{c=1}^N\|\xi_c^t\|^2 \quad .$$

This metric weights the global parameter error and the control variate errors based on their respective roles in the algorithm's dynamics. The scaling factor $\frac{\gamma^2H^2}{N}$ for the control variate terms reflects their influence on the convergence rate and ensures that the metric captures the algorithm's inherent structure.

We now establish our central convergence result, which provides a novel contraction analysis for SCAFFOLD applied to quadratic objective functions.

Lemma 1. *Let $(\mathrm{X}^t)_{t\geq0}$ and $(\widetilde{\mathrm{X}}^t)_{t\geq0}$ be two sequences of* SCAFFOLD *iterates with the same noise sequence $(\varepsilon_c^{t,h})_{t,h\geq0,1\leq c\leq N}$ but different initializations. Assume that local Hessians satisfy $\mu\mathrm{Id} \preccurlyeq \mathbf{A}_c \preccurlyeq L\mathrm{Id}$ for some constants $0 < \mu \leq L$. If the step size satisfies $\gamma \leq 1/L$, then*

$$\mathbb{E}[\|\mathrm{X}^{t+1} - \widetilde{\mathrm{X}}^{t+1}\|_\Lambda^2] \leq \rho_{\gamma,H}\cdot\mathbb{E}[\|\mathrm{X}^t - \widetilde{\mathrm{X}}^t\|_\Lambda^2],$$

where the contraction factor is given by

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

Proof. The key insight is to exploit the control variate updates' structure and the cancellation properties arising from SCAFFOLD's design. We decompose the proof in multiple steps.

Step 1: Global parameter update analysis. Since the control variates satisfy $\sum_{c=1}^N(\xi_c^t - \xi_c^\star) = 0$ (a fundamental property of SCAFFOLD), we can write

$$\|\theta^{t+1} - \widetilde{\theta}^{t+1}\|^2 = \left\|\tfrac{1}{N}\textstyle\sum_{c=1}^N(\theta_c^{t,H} - \widetilde{\theta}_c^{t,H} + \gamma H(\xi_c^t - \widetilde{\xi}_c^t))\right\|^2.$$

Expanding the squared norm and using [7, Lemma D.3] gives

$$\begin{aligned}\|\theta^{t+1} - \widetilde{\theta}^{t+1}\|^2 = \tfrac{1}{N^2}\textstyle\sum_{c=1}^N\|\theta_c^{t,H} - \widetilde{\theta}_c^{t,H} + \gamma H(\xi_c^t - \widetilde{\xi}_c^t)\|^2 \\ - \tfrac{\gamma^2H^2}{N^2}\textstyle\sum_{c=1}^N\|\xi_c^{t+1} - \widetilde{\xi}_c^{t+1}\|^2 \quad .\end{aligned}$$

Using the definition of the Λ-norm, we get

$$\|\mathrm{X}^{t+1} - \widetilde{\mathrm{X}}^{t+1}\|_\Lambda^2 = \tfrac{1}{N}\textstyle\sum_{c=1}^N\|\theta_c^{t,H} - \widetilde{\theta}_c^{t,H} + \gamma H(\xi_c^t - \widetilde{\xi}_c^t)\|^2 \quad .$$

From (4) and noting that noise terms are identical (hence cancel), we have

$$\begin{aligned}\theta_c^{t,H} - \widetilde{\theta}_c^{t,H} + \gamma H(\xi_c^t - \widetilde{\xi}_c^t) = \Gamma_c^H(\theta^t - \widetilde{\theta}^t) \\ + \gamma H\left(\mathrm{Id} - (\gamma H\mathbf{A}_c)^{-1}(\mathrm{Id} - \Gamma_c^H)\right)(\xi_c^t - \widetilde{\xi}_c^t) \quad .\end{aligned} \tag{5}$$

Step 2: Eigendecomposition analysis. Using the spectral decomposition $\mathbf{A}_c = \mathbf{U}_c\mathbf{D}_c\mathbf{U}_c^\top$ where $\mathrm{F}_c = \mathrm{Id} - \gamma\mathbf{D}_c$, we can rewrite (5) as

$$\begin{aligned}\theta_c^{t,H} - \widetilde{\theta}_c^{t,H} + \gamma H(\xi_c^t - \widetilde{\xi}_c^t) = \mathbf{U}_c\mathrm{F}_c^H\mathbf{U}_c^\top(\theta^t - \widetilde{\theta}^t) \\ + \gamma H\mathbf{U}_c\left(\mathrm{Id} - \mathbf{D}_c^{-1}(\mathrm{Id} - \mathrm{F}_c^H)\right)\mathbf{U}_c^\top(\xi_c^t - \widetilde{\xi}_c^t).\end{aligned} \tag{6}$$

Step 3: Coordinate-wise analysis via eigendecomposition. Multiplying (6) by $\mathbf{U}_c^\top$ from the left, we obtain

$$\begin{aligned}\mathbf{U}_c^\top(\theta_c^{t,H} - \widetilde{\theta}_c^{t,H}) + \gamma H\mathbf{U}_c^\top(\xi_c^t - \widetilde{\xi}_c^t) = \mathrm{F}_c^H\mathbf{U}_c^\top(\theta^t - \widetilde{\theta}^t) \\ + \gamma H\left(\mathrm{Id} - \mathbf{D}_c^{-1}(\mathrm{Id} - \mathrm{F}_c^H)\right)\mathbf{U}_c^\top(\xi_c^t - \widetilde{\xi}_c^t).\end{aligned} \tag{7}$$

To simplify the analysis, we introduce the coordinate representations of the error in the eigenbasis of $\mathbf{A}_c$

$$\begin{aligned}\Delta\theta_c^{t,H} = \mathbf{U}_c^\top(\theta_c^{t,H} - \widetilde{\theta}_c^{t,H}) \ , \\ \Delta\theta_c^t = \mathbf{U}_c^\top(\theta^t - \widetilde{\theta}^t) \ , \quad \Delta\xi_c^t = \mathbf{U}_c^\top(\xi_c^t - \widetilde{\xi}_c^t) \ .\end{aligned}$$

Since $\mathbf{D}_c = \mathrm{diag}(a_{c,1},\dots,a_{c,d})$ and $\mathrm{F}_c = \mathrm{Id} - \gamma\mathbf{D}_c = \mathrm{diag}(1-\gamma a_{c,1},\dots,1-\gamma a_{c,d})$, we can analyze (7) coordinate by coordinate. For each $j \in \{1,\dots,d\}$, we have

$$\begin{aligned}\Delta\theta_{c,j}^{t,H} + \gamma H\Delta\xi_{c,j}^t = (1-\gamma a_{c,j})^H\Delta\theta_{c,j}^t \\ + \gamma H\left(1 - \tfrac{1-(1-\gamma a_{c,j})^H}{\gamma H a_{c,j}}\right)\Delta\xi_{c,j}^t.\end{aligned}$$

To bound the squared magnitude of each coordinate, we apply Young's inequality. For any $\beta > 0$, we have

$$\begin{aligned}&|\Delta\theta_{c,j}^{t,H} + \gamma H\Delta\xi_{c,j}^t|^2 \\ &\leq (1+\beta)(1-\gamma a_{c,j})^{2H}|\Delta\theta_{c,j}^t|^2 \\ &\quad + \left(1 + \tfrac{1}{\beta}\right)\left(1 - \tfrac{1-(1-\gamma a_{c,j})^H}{\gamma H a_{c,j}}\right)^2\gamma^2H^2|\Delta\xi_{c,j}^t|^2.\end{aligned} \tag{8}$$

We optimize over the parameter $\beta > 0$. For $u, v \in \mathbb{R}_+$, taking $\beta = v/u$, $\max((1+\beta)u^2, (1+1/\beta)v^2) \leq (u+v)\max(u,v)$. Using this inequality with $u = (1-\gamma a_{c,j})^H$ and $v = \gamma H\left|1 - \frac{1-(1-\gamma a_{c,j})^H}{\gamma H a_{c,j}}\right|$, we obtain the coordinate-

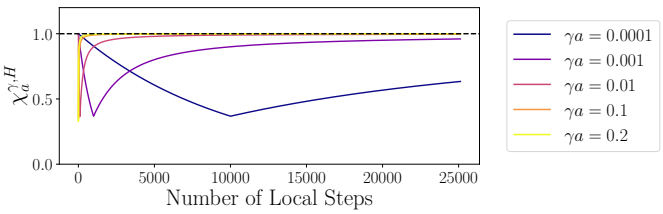


Fig. 2: Contraction rate $\chi_a^{\gamma,H}$ as a function of the number of communications H for different values of γa.

wise contraction bound

$$|\Delta\theta_{c,j}^{t,H}+\gamma H\Delta\xi_{c,j}^t|^2 \leq \chi_{a_{c,j}}^{\gamma,H}\left\{|\Delta\theta_{c,j}^t|^2+\gamma^2H^2|\Delta\xi_{c,j}^t|^2\right\}, \quad (9)$$

where for $a \in \mathbb{R}$ and $h \geq 0$, we define

$$\chi_a^{\gamma,h} = \max\left\{(1-\gamma a)^h, 1-\frac{1-(1-\gamma a)^h}{\gamma a h}\right\} \ ,$$

where we used $(1-\gamma a)^h + (1-\frac{1-(1-\gamma a)^h}{\gamma a h}) \leq 1$ for $\gamma a \leq 1$.

Step 4: Study of $\chi_a^{\gamma,h}$. The contraction factor is determined by whichever term dominates in the maximum (see Figure 2). The first term $(1-\gamma a)^h$ dominates when

$$(1-\gamma a)^h \geq 1-\frac{1-(1-\gamma a)^h}{\gamma a h},$$

which can be rearranged as $(\gamma a h-1)(1-\gamma a)^h \geq \gamma a h-1$, leading two different regimes based on the value of $\gamma a h$.

- **First case: $\gamma a h \leq 1$.** When $\gamma a h \leq 1$, the inequality $(\gamma a h-1)(1-\gamma a)^h \geq \gamma a h-1$ automatically satisfied since both sides are non-positive and $(1-\gamma a)^h \leq 1$. Therefore, the exponential decay term dominates, yielding

  $$\chi_a^{\gamma,h} \leq (1-\gamma a)^h.$$

  This represents the classical exponential decay regime.
- **Second case: $\gamma a h \geq 1$.** When $\gamma a h \geq 1$, the control variate correction term becomes dominant, and we have

  $$\chi_a^{\gamma,h} = 1+\frac{(1-\gamma a)^h-1}{\gamma a h}.$$

  To bound this expression, we use the fact that for $\gamma a h \geq 1$, the exponential term satisfies $(1-\gamma a)^h \leq \mathrm{e}^{-\gamma a h} \leq \mathrm{e}^{-1} = 1/e$. Substituting this bound:

  $$\chi_a^{\gamma,h} \leq 1+\frac{1/e-1}{\gamma a h} = 1-\frac{1-1/e}{\gamma a h}.$$

  This regime demonstrates the benefit of SCAFFOLD's control variates: even with larger effective step sizes, the algorithm maintains a contraction rate that decay like $1/(\gamma a h)$.

The transition between these two regimes occurs at $\gamma a h = 1$, representing the boundary between exponential and algebraic decay rates in the contraction analysis.

Step 5: Conclusion. Since $\mathbf{U}_c$ is an orthogonal matrix, we have that $\|\mathbf{U}_c^\top w\|^2 = \|w\|^2$ for any vector $w \in \mathbb{R}^d$, which gives the following identity

$$\begin{aligned}\|\theta_c^{t,H}-\widetilde{\theta}_c^{t,H}+\gamma H(\xi_c^t-\widetilde{\xi}_c^t)\|^2 &= \|\mathbf{U}_c^\top(\theta_c^{t,H}-\widetilde{\theta}_c^{t,H}+\gamma H(\xi_c^t-\widetilde{\xi}_c^t))\|^2 \\ &= \|\Delta\theta_c^t-\gamma H\Delta\xi_c^t\|^2 \ .\end{aligned}$$

Using (9) and summing the coordinate-wise contraction on all coordinates, we have

$$\begin{aligned}&\|\theta_c^H-\theta^\star+\gamma H(\xi_c-\xi_c^\star)\|^2 \\ &\quad\leq \max_{1\leq j\leq d}\chi_{a_{c,j}}^{\gamma,H}\left\{\|\theta-\theta^\star\|^2+\gamma^2H^2\|\xi_c-\xi_c^\star\|^2\right\} \ .\end{aligned}$$

Now, we divide the coordinates in two sets, the ones such that $\gamma a_{c,j} h \leq 1$ and the ones where $\gamma a_{c,j} h > 1$. We refer to these sets as $\mathcal{S}_{\leq 1}$ and $\mathcal{S}_{\geq 1}$ respectively. Now we directly obtain

$$\begin{aligned}\max_{j\in\mathcal{S}_{\leq 1}}\chi_{a_{c,j}}^{\gamma,H} &\leq \max_{j\in\mathcal{S}_{\leq 1}}(1-\gamma a_{c,j})^H \leq (1-\gamma\mu)^H \ , \\ \max_{j\in\mathcal{S}_{\geq 1}}\chi_{a_{c,j}}^{\gamma,H} &\leq \max_{j\in\mathcal{S}_{\geq 1}}1-\tfrac{1-1/e}{\gamma a_{c,j}H} \leq 1-\tfrac{1-1/e}{\gamma LH} \ .\end{aligned}$$

Consequently, we have that

$$\max_{1\leq j\leq d}\chi_{a_{c,j}}^{\gamma,H} \leq \max\left\{(1-\gamma\mu)^H, 1-\tfrac{1-1/e}{\gamma LH}\right\} \ ,$$

and the result follows after averaging over $c = 1\dots N$. □

This lemma has two important consequences.

Consequence 1: Geometric convergence to stationarity. The contraction property establishes convergence of SCAFFOLD with noisy gradients for any step size $\gamma \leq 1/L$ and any number of local steps $H \geq 1$. To quantify this convergence, we employ the Λ-weighted Wasserstein distance between probability distributions ξ and ξ', defined as

$$\mathbf{W}_\Lambda(\xi,\xi') = \inf_{\rho\in\mathcal{C}(\xi,\xi')}\int\|x-x'\|_\Lambda^2\,\rho(\mathrm{d}(x,x')),$$

where $\mathcal{C}(\xi,\xi')$ denotes the set of all couplings between ξ and ξ'; see [8, Chapter 20]. This metric naturally captures the convergence behavior in the algorithm's intrinsic geometry defined by the Λ-norm.

Theorem 1. *Consider the* SCAFFOLD *algorithm applied to quadratic objectives with step size $\gamma \leq 1/L$ and $H \geq 1$ local steps. Then the Markov chain $(\mathrm{X}^t)_{t\geq 0}$ converges geometrically to a unique stationary distribution $\pi_{\gamma,H}$. Specifically, for any initial distribution ξ and time $t \in \mathbb{N}$,*

$$\mathbf{W}_\Lambda(\xi\mathcal{K}_{\gamma,H}^t,\pi_{\gamma,H}) \leq \rho_{\gamma,H}^t\,\mathbf{W}_\Lambda(\xi,\pi_{\gamma,H}),$$

where $\rho_{\gamma,H} < 1$ is the contraction factor from Lemma 1.

Proof. The result follows from the general theory of uniform convergence in the Wasserstein distance. For a detailed treatment of this approach, see [8, Theorem 20.3.4]. □

Optimal balancing and accelerated rates. The contraction rate is minimized when the two competing terms in the maximum are balanced:

$$(1-\gamma\mu)^H = 1-\frac{1-1/\mathrm{e}}{\gamma LH}.$$

To analyze this tradeoff, we consider the case $\gamma\mu H \leq 1$, which gives the bound $(1-\gamma\mu)^H \geq 1-\gamma\mu H$ (Bernoulli inequality). For the exponential decay term to dominate, we need

$$(1-\gamma\mu)^H \leq 1 - \frac{1-1/\mathrm{e}}{\gamma LH}.$$

A sufficient condition for this dominance is $1-\gamma\mu H/2 \leq 1-(1-1/\mathrm{e})/(\gamma LH)$, which simplifies to

$$H^2\gamma^2 L\mu \geq 2(1-1/\mathrm{e}). \tag{10}$$

To achieve optimal contraction, we choose the minimal H_γ satisfying (10) with equality:

$$H_\gamma = \left\lceil \sqrt{\frac{2(1-1/\mathrm{e})}{\gamma^2 L\mu}} \right\rceil = \left\lceil \frac{\sqrt{2(1-1/\mathrm{e})}}{\gamma\sqrt{L\mu}} \right\rceil \quad . \tag{11}$$

Note that $\gamma\mu H_\gamma \leq 1$, showing that this choice is compatible with the small stepsize limit. The optimal contraction rate is

$$\rho_{\text{opt}} = 1 - \sqrt{\frac{2(1-1/\mathrm{e})\mu}{L}}. \tag{12}$$

If $\gamma\mu H \geq 1$, there is no tradeoff, since $(1-\gamma\mu)^H < 1-(1-1/\mathrm{e})/(\gamma LH)$. The optimal contraction is obtained with $H_\gamma = 1/(\gamma\mu)$ and is given by $1-\mu/L$, which is always worse than ρ_{opt}. In practice, this means that it is never advantageous to choose $\gamma\mu H \geq 1$.

We stress that here we discuss only the rate of convergence to stationarity, but it should be noted that the stationary distribution itself depends on γ and H. In practice, the choice of the step size γ is governed by the behavior of the stationary distribution $\pi_{\gamma,H}$. In particular, controlling the variance at stationarity may require taking a very small step size, a setting that is well captured by our analysis.

Consequence 2: Acceleration in the deterministic case. The second consequence is that SCAFFOLD achieves acceleration compared to standard federated averaging by enabling more local steps through its control variate mechanism.

Theorem 2 (Deterministic convergence and acceleration)**.** *When gradients are deterministic (i.e., $\varepsilon_c^{t,h} = 0$ for all $t, h \geq 0$ and $c \in \{1,\dots,N\}$), the global parameter error satisfies*

$$\|\theta^t - \theta^\star\|^2 \leq \rho_{\gamma,H}^t \|\mathrm{X}^0 - \mathrm{X}^\star\|_\Lambda^2,$$

where $\rho_{\gamma,H} < 1$ is the contraction factor from Lemma 1 and $\mathrm{X}^\star = [(\theta^\star)^\top, (\xi_1^\star)^\top, \dots, (\xi_N^\star)^\top]^\top$.

Optimal parameter selection for acceleration: *By choosing $\gamma = 1/L$ and balancing the two terms in the maximum, the optimal number of local steps is H_γ yielding to the contraction rate ρ_{opt} defined in (12), respectively.*

Communication complexity: *To achieve $\|\theta^t - \theta^\star\|^2 \leq \epsilon^2$ for $\epsilon > 0$, we need*

$$t \geq \frac{\log(1/\epsilon^2)}{\log(1/\rho_{\text{opt}})} = O\left(\sqrt{\frac{L}{\mu}}\log(1/\epsilon)\right)$$

communication rounds, which represents a $\sqrt{\kappa}$ improvement over the $\kappa = L/\mu$ dependence of federated averaging.

Proof. Follows directly from Lemma 1 by setting $\widetilde{\mathrm{X}}^t = \mathrm{X}^\star$, which is a fixed point in the deterministic setting. □

This analysis demonstrates that SCAFFOLD achieves accelerated convergence rates when using fixed-length local update blocks. When choosing the maximal step size $\gamma = 1/L$, our result recovers convergence rates similar to those established in [4] for the stochastic communication setting.

Comparison with standard rates. This represents a significant improvement over standard federated averaging, which typically achieves rates of the form $1 - O(\mu/L)$. Our result demonstrates a $\sqrt{\kappa}$ acceleration where $\kappa = L/\mu$ is the condition number, transforming the linear dependence on μ/L into a square-root dependence.

Variance and convergence rate. To obtain finite-time bound for SCAFFOLD, we decompose the error $\mathbb{E}[\|\theta^t - \theta^\star\|^2]$ between a transient error (which measures the distance between the iterates and the stationary distribution) and the quadratic error under stationarity. We define $(\mathrm{X}^t)_{t\geq 0}$ the iterates of SCAFFOLD started at X^0, and define $(\widetilde{\mathrm{X}}^t)_{t\geq 0}$ the hypothetical stationary version of SCAFFOLD generated by sampling $\widetilde{\mathrm{X}}^0 \sim \pi_{\gamma,H}$. The error can thus be decomposed as

$$\mathbb{E}[\|\theta^t - \theta^\star\|^2] \leq 2\mathbb{E}[\|\theta^t - \widetilde{\theta}^t\|^2] + 2\mathbb{E}[\|\widetilde{\theta}^t - \theta^\star\|^2] \quad . \tag{13}$$

By Lemma 1, the first term decays exponentially fast with the contraction rate $\rho_{\gamma,H} < 1$, ensuring that the algorithm quickly approaches its stationary regime regardless of initialization.

The second term represents the steady-state mean-square error, which quantifies the fluctuations of SCAFFOLD's stationary distribution around the optimum. This term captures the tradeoff between convergence speed and steady-state accuracy. Unlike deterministic optimization, the algorithm does not converge to the optimum but oscillates with a variance determined by the algorithm parameters and noise variance. Following the approach in [3], the covariance matrix of the state vector X^t in the stationary regime is given by a discrete-time Lyapunov equation: denoting $\mathbf{\Sigma}^{\mathbf{X}}$ the stationary covariance matrix of X,

$$\mathbf{\Sigma}^{\mathbf{X}} = \mathrm{A}_{\gamma,H}\,\mathbf{\Sigma}^{\mathbf{X}}\,\mathrm{A}_{\gamma,H}^\top + \mathrm{B}_{\gamma,H},$$

where $\mathrm{A}_{\gamma,H}$ is the deterministic part of the transition operator and $\mathrm{B}_{\gamma,H}$ captures the noise injection at each iteration. Contrary to [3], the analysis is performed without imposing additional constraints on H and holds as long as $\gamma \leq 1/L$. Since the state vector has the block structure $\mathrm{X}^\top = [\theta^\top, \xi_1^\top, \dots, \xi_N^\top] \in \mathbb{R}^{(N+1)d}$, the corresponding covariance matrix Σ^{X} admits a natural block decomposition:

$$\mathbf{\Sigma}^{\mathbf{X}} = \begin{bmatrix} \mathbf{\Sigma}^{\theta} & \mathbf{\Sigma}_1^{\theta,\xi} & \cdots & \mathbf{\Sigma}_N^{\theta,\xi} \\ \mathbf{\Sigma}_1^{\xi,\theta} & \mathbf{\Sigma}_{1,1}^{\xi} & \cdots & \mathbf{\Sigma}_{1,N}^{\xi} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_N^{\xi,\theta} & \mathbf{\Sigma}_{N,1}^{\xi} & \cdots & \mathbf{\Sigma}_{N,N}^{\xi} \end{bmatrix},$$

where each block has dimension $d \times d$.

- **Diagonal blocks**: $\mathbf{\Sigma}^{\theta} \in \mathbb{R}^{d\times d}$ represents the covariance of the global parameter around the optimum, while $\Sigma_{c,c}^{\xi} \in$

$\mathbb{R}^{d\times d}$ captures the variability of the c-th client's control variate around its optimal value $\xi_c^\star$.

- **Cross-covariance blocks**: $\mathbf{\Sigma}_c^{\theta,\xi}, \mathbf{\Sigma}_c^{\xi,\theta} \in \mathbb{R}^{d\times d}$ quantifies the statistical dependence between the global parameter and the c-th control variate, reflecting how fluctuations in local control variates propagate to the global model.
- **Inter-client blocks**: $\mathbf{\Sigma}_{c,c'}^{\xi} \in \mathbb{R}^{d\times d}$ for $c \neq c'$ measures the cross-correlation between control variates of different clients, which typically arises from their shared dependence on the global parameter evolution.

This block structure is crucial for understanding the steady-state behavior of SCAFFOLD, showing how noise and control variate errors propagate. The off-diagonal blocks capture the coupling effects that distinguish SCAFFOLD from simpler FedAvg where local updates are conditionally independent given the local parameter. To express the covariance, we follow [3] and give a system of equations on the covariance matrices. We define the following matrices, for $c\{1,\dots,N\}$,

$$\Upsilon_c = \Gamma_c^H \ , \quad \bar{\Upsilon} = \tfrac{1}{N}\textstyle\sum_{c=1}^N \Upsilon_c \ , \quad \widetilde{\Upsilon}_c = \Upsilon_c - \bar{\Upsilon} \ ,$$
$$\mathrm{B}_c = \mathrm{Id} - (\gamma H \mathbf{A}_c)^{-1}(\mathrm{Id} - \Upsilon_c) \ ,$$
$$\bar{\mathbf{\Sigma}}_\varepsilon^c = \mathbb{E}[\mathcal{E}_c \mathcal{E}_c^\top] \ , \quad \bar{\mathbf{\Sigma}}_\varepsilon = \tfrac{1}{N}\textstyle\sum_{c=1}^N \bar{\mathbf{\Sigma}}_\varepsilon^c \ ,$$

where $\mathcal{E}_c$ is defined in (3). Rewriting the expression from [3] in our notations, we obtain

$$\begin{aligned}\mathbf{\Sigma}^\theta &= \bar{\Upsilon}\mathbf{\Sigma}^\theta\bar{\Upsilon} + \tfrac{\gamma H}{N}\textstyle\sum_{i=1}^N(\bar{\Upsilon}\mathbf{\Sigma}_i^{\theta,\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_i^{\xi,\theta}\bar{\Upsilon}) \\ &\quad + \tfrac{\gamma^2H^2}{N^2}\textstyle\sum_{i=1}^N\sum_{i'=1}^N \mathrm{B}_i\mathbf{\Sigma}_{i,i'}^{\xi}\mathrm{B}_{i'} + \tfrac{\gamma^2}{N}\bar{\mathbf{\Sigma}}_\varepsilon \ ,\end{aligned}$$

$$\begin{aligned}\gamma H\mathbf{\Sigma}_c^{\theta,\xi} &= \bar{\Upsilon}\mathbf{\Sigma}^\theta\widetilde{\Upsilon}_c + \gamma H\bar{\Upsilon}\mathbf{\Sigma}_c^{\theta,\xi}\mathrm{B}_c \\ &\quad - \tfrac{\gamma H}{N}\textstyle\sum_{i=1}^N\{\bar{\Upsilon}\mathbf{\Sigma}_i^{\theta,\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_i^{\xi,\theta}\widetilde{\Upsilon}_c\} + \tfrac{\gamma^2H^2}{N}\textstyle\sum_{i=1}^N\mathrm{B}_i\mathbf{\Sigma}_{i,c}^{\xi}\mathrm{B}_c \\ &\quad - \tfrac{\gamma^2H^2}{N^2}\textstyle\sum_{i=1}^N\sum_{i'=1}^N \mathrm{B}_i\mathbf{\Sigma}_{i,i'}^{\xi}\mathrm{B}_{i'} + \tfrac{\gamma^2}{N}(\bar{\mathbf{\Sigma}}_\varepsilon^c - \bar{\mathbf{\Sigma}}_\varepsilon) \ ,\end{aligned}$$

$$\begin{aligned}\gamma^2H^2\mathbf{\Sigma}_{c,c}^{\xi} &= \widetilde{\Upsilon}_c\mathbf{\Sigma}^\theta\widetilde{\Upsilon}_c + \gamma H\widetilde{\Upsilon}_c\mathbf{\Sigma}_c^{\theta,\xi}\mathrm{B}_c + \gamma H\mathrm{B}_c\mathbf{\Sigma}_c^{\xi,\theta}\widetilde{\Upsilon}_c \\ &\quad - \tfrac{\gamma H}{N}\textstyle\sum_{i=1}^N\{\widetilde{\Upsilon}_c\mathbf{\Sigma}_i^{\theta,\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_i^{\xi,\theta}\widetilde{\Upsilon}_c\} \\ &\quad + \gamma^2H^2\mathrm{B}_c\mathbf{\Sigma}_{c,c}^{\xi}\mathrm{B}_c - \tfrac{\gamma^2H^2}{N}\textstyle\sum_{i=1}^N\{\mathrm{B}_c\mathbf{\Sigma}_{c,i}^{\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_{i,c}^{\xi}\mathrm{B}_c\} \\ &\quad + \tfrac{\gamma^2H^2}{N^2}\textstyle\sum_{i=1}^N\sum_{i'=1}^N \mathrm{B}_i\mathbf{\Sigma}_{i,i'}^{\xi}\mathrm{B}_{i'} + \big(1-\tfrac{2}{N}\big)\gamma^2\bar{\mathbf{\Sigma}}_\varepsilon^c + \tfrac{\gamma^2}{N}\bar{\mathbf{\Sigma}}_\varepsilon \ ,\end{aligned}$$

$$\begin{aligned}\gamma^2H^2\mathbf{\Sigma}_{c,c'}^{\xi} &= \widetilde{\Upsilon}_c\mathbf{\Sigma}^\theta\widetilde{\Upsilon}_{c'} + \gamma H\widetilde{\Upsilon}_c\mathbf{\Sigma}_{c'}^{\theta,\xi}\mathrm{B}_{c'} + \gamma H\mathrm{B}_c\mathbf{\Sigma}_{c'}^{\xi,\theta}\widetilde{\Upsilon}_{c'} \\ &\quad - \tfrac{\gamma H}{N}\textstyle\sum_{i=1}^N\{\widetilde{\Upsilon}_c\mathbf{\Sigma}_i^{\theta,\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_i^{\xi,\theta}\widetilde{\Upsilon}_{c'}\} \\ &\quad + \gamma^2H^2\mathrm{B}_c\mathbf{\Sigma}_{c,c'}^{\xi}\mathrm{B}_{c'} - \tfrac{\gamma^2H^2}{N}\textstyle\sum_{i=1}^N\{\mathrm{B}_c\mathbf{\Sigma}_{c,i}^{\xi}\mathrm{B}_i + \mathrm{B}_i\mathbf{\Sigma}_{i,c'}^{\xi}\mathrm{B}_{c'}\} \\ &\quad + \tfrac{\gamma^2H^2}{N^2}\textstyle\sum_{i=1}^N\sum_{i'=1}^N \mathrm{B}_i\mathbf{\Sigma}_{i,i'}^{\xi}\mathrm{B}_{i'} + \tfrac{\gamma^2}{N}\bar{\mathbf{\Sigma}}_\varepsilon - \tfrac{\gamma^2}{N}\bar{\mathbf{\Sigma}}_\varepsilon^c - \tfrac{\gamma^2}{N}\bar{\mathbf{\Sigma}}_\varepsilon^{c'} \ .\end{aligned}$$

Adapting [3]'s Theorem 5.8 to the quadratic case, we obtain the following theorem.

Theorem 3 (Adapted from Theorem 5.8 in [3])**.** *Assume that* $\mu\mathrm{Id} \preccurlyeq \mathbf{A}_c \preccurlyeq L\mathrm{Id}$ *for all* $c \in \{1,\dots,N\}$*, with constants* $\mu, L > 0$*. Define the inhomogeneity factor as*

$$\zeta_2^2 = \frac{1}{N}\sum_{c=1}^N \|\mathbf{A}_c - \bar{\mathbf{A}}\|^2,$$

where $\bar{\mathbf{A}} = \frac{1}{N}\sum_{c=1}^N \mathbf{A}_c$*. Under the conditions* $\gamma H\zeta_2 \lesssim \mu$ *and* $\gamma HL \lesssim 1$*, it holds that* $\|\mathbf{\Sigma}^\theta\| \lesssim \frac{\gamma\sigma^2}{N}$.

Proof. The result follows from Theorem 5.8 of [3] applied to quadratic functions, where the third derivative vanishes. □

This theorem yields a straightforward consequence for the special case where (i) all Hessian matrices are diagonal, and (ii) the local problems are perfectly homogeneous.

Corollary 1. *Assume that* $\mathbf{A}_c = \mathrm{diag}(a_1,\dots,a_d)$*, with common diagonal entries* $a_1,\dots,a_d > 0$ *shared across all clients. Let* $\mu = \min_j a_j$ *and* $L = \max_j a_j$*. Then, we have*

$$\mathrm{tr}\,\mathbf{\Sigma}^\theta \leq \frac{\gamma\sigma^2}{N} \ .$$

Proof. When the Hessian matrices are diagonal, each coordinate evolves independently. In this scenario, SCAFFOLD can be viewed as solving d independent one-dimensional problems, each characterized by coordinate-specific constants μ_j and L_j. Under perfect homogeneity, we have $\mu_j = L_j$ for each coordinate j, implying $\zeta_2 = 0$. The result then directly follows by applying Theorem 5.8 to each coordinate separately. □

We argue that, although the above corollary is stated under restrictive assumptions, its conclusions likely extend to more general settings, relaxing both (i) the homogeneity assumption and (ii) the diagonal Hessian condition. This motivates the following conjectures:

Conjecture 1: SCAFFOLD *has linear speed-up in the number of agents* for any number of local updates, i.e. $\mathrm{tr}\,\mathbf{\Sigma}^\theta \lesssim \frac{\gamma\sigma^2}{N}$ up to a multiplicative constant that may depend on H.

Conjecture 2: SCAFFOLD has linear speed-up even when the Hessian matrices are not diagonal.

Under these two conjectures, we have the following finite-sample bound for SCAFFOLD, provided that $\gamma \leq 1/L$,

$$\mathbb{E}[\|\theta^T - \theta^\star\|^2] \lesssim \rho_{\text{opt}}^T\|\mathrm{X}^0 - \mathrm{X}^\star\|_\Lambda^2 + \tfrac{\gamma\sigma^2}{N\mu} \ , \qquad (14)$$

where

$$\|\mathrm{X}^0 - \mathrm{X}^\star\|_\Lambda^2 = \|\theta^0 - \theta^\star\|^2 + \tfrac{\gamma^2H^2}{N}\textstyle\sum_{c=1}^N\|\xi_c^0 - \xi_c^\star\|^2 \ ,$$

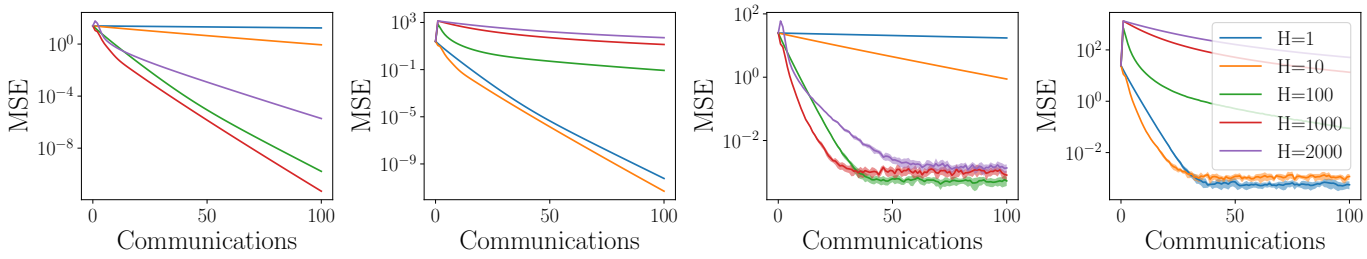
and $\rho_{\text{opt}} = \max\{(1-\gamma\mu)^H, 1 - \frac{1-1/e}{\gamma LH}\}$ as defined in (12). This results allows to derive the corresponding sample and communication complexity. Indeed, to obtain $\mathbb{E}[\|\theta^t - \theta^\star\|^2] \leq \epsilon^2$, for $\epsilon > 0$, Equation (14) requires that

$$\gamma \leq \min\Big(\frac{1}{L}, \frac{N\mu\epsilon^2}{\sigma^2}\Big) \ , \quad T \geq \frac{2\log(\|\mathrm{X}^0 - \mathrm{X}^\star\|/\epsilon)}{\log(1/\rho_{\text{opt}})} \ .$$

Setting $H = \left\lceil \frac{\sqrt{2(1-1/e)}}{\gamma\sqrt{L\mu}} \right\rceil$ as in (11), we have that SCAFFOLD reaches mean squared error ϵ^2 after

$$T = O\Big(\sqrt{\frac{L}{\mu}}\log\Big(\frac{1}{\epsilon}\Big)\Big) \text{ communications} \ ,$$

$$TH = O\Big(\max\Big(\frac{L}{\mu}, \frac{\sigma^2}{N\mu^2\epsilon^2}\Big)\log\Big(\frac{1}{\epsilon}\Big)\Big) \text{ updates} \ .$$



(a) $\gamma = 0.01$, deterministic (b) $\gamma = 1$, deterministic (c) $\gamma = 0.01$, stochastic (d) $\gamma = 1$, stochastic

Fig. 3: Mean Squared Error of SCAFFOLD's global iterates as a function of the number of communications, with different step sizes and numbers of local training steps. In the stochastic regime, we draw $\varepsilon_c^{t,h} \sim \mathcal{N}(0; 0.01 \cdot \mathrm{Id})$ and average the results over 10 independent runs of the algorithm.

Assuming that our Conjectures hold, this shows that SCAFFOLD may simultaneously achieve accelerated communication, together with linear speed-up.

IV. NUMERICAL ILLUSTRATIONS

In this section, we study numerically the behavior of the SCAFFOLD algorithm on quadratic functions. To this end, we generate quadratic problems with various numbers of clients, and run the algorithm for a diverse number of steps. Each agent is assigned a matrix

$$\mathbf{A}_c = \mathbf{U}_c \mathbf{D}_c \mathbf{U}_c^\top \ , \tag{15}$$

where $\mathbf{U}_c$ is a random orthonormal matrix generated by applying the QR decomposition to a random matrix whose entries are drawn independently from a centered Gaussian distribution, and $\mathbf{b}_c$ is a random vector with entries similarly drawn from a centered Gaussian distribution. We emphasize that the primary purpose of these experiments is *to illustrate the phenomena described by our theory*. Consequently, we focus on simple scenarios, as our theory ensures that the observed behavior generalizes to more complex settings. We now describe our main numerical observations.

SCAFFOLD converges in all settings. In the first experiment, we consider a problem with $N = 10$ clients and $d = 20$ features. The Hessian matrices are generated according to (15), where each diagonal entry of $\mathbf{D}_c$ is drawn at random among 20 values spaced logarithmically between 0.01 and 1. We run SCAFFOLD with step sizes $\gamma \in \{0.01, 1.0\}$ and a number of local updates $H \in \{1, 10, 100, 1000, 2000\}$. We consider two scenarios: a deterministic gradient setting (i.e., $\varepsilon_c^{t,h} = 0$), and a stochastic gradient setting with noise $\varepsilon_c^{t,h} \sim \mathcal{N}(0,\ 0.1 \cdot \mathrm{Id})$. The mean squared error is reported in Figure 3 as a function of the number of gradients computed by each client. The results validate the two main insights derived from our theory

- SCAFFOLD *converges for any number of local steps*, provided the step size satisfies $\gamma \leq 1/L$. This convergence holds in both deterministic and stochastic gradient settings. Furthermore, the convergence speed increases when the number of local steps increases, up to a given threshold, where the speed starts to decrease.
- SCAFFOLD *exhibits faster convergence up to an optimal number of local updates* H. From our theoretical analysis, the optimal number of steps, defined in (11), is given by

$$H_\gamma = \left\lceil \frac{\sqrt{2(1-1/\mathrm{e})}}{\gamma\sqrt{L\mu}} \right\rceil ,$$

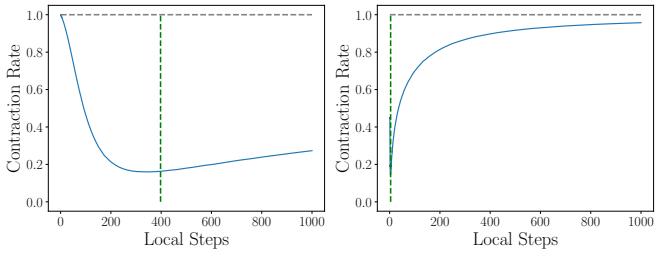
which, in our experimental setting where $\mu = 0.01$ and $L = 1$, yields approximately $H_\gamma \approx 8/\gamma$.

Empirical validation of SCAFFOLD's convergence rate. To further validate our theoretical results, we empirically reproduce the theoretical convergence rate depicted in Figure 2. Specifically, we generate a new problem instance following the procedure described above, and compute the empirical convergence rate as the ratio

$$\frac{\|\mathrm{X}^{t+1} - \mathrm{X}^\star\|_\Lambda^2}{\|\mathrm{X}^{t} - \mathrm{X}^\star\|_\Lambda^2}$$

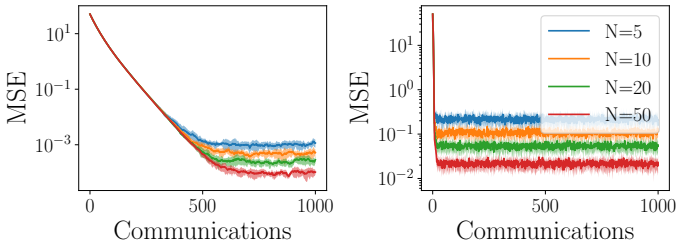
for multiple random initializations of θ^t and ξ_c^t drawn from a Gaussian distribution in $\mathbb{R}^d$. The results of this experiment are shown in Figure 4, and confirm that the convergence speed of SCAFFOLD increases, up to a given threshold. We report as a green dashed line the threshold predicted from our theory, and show that it perfectly matches the best convergence rate observed in practice. Moreover, the shape of the curves exactly matches the one predicted by our theory, reported in Figure 2.

SCAFFOLD has linear speed-up. We now study the linear speed-up phenomenon. To this end, we generate variants of the problems used in Figure 3, where we replicate the clients multiple times to obtain $N \in \{5, 10, 20, 50\}$ clients, with step size $\gamma \in \{0.01, 1\}$. This ensures that the global minimum remains the same for all problems, allowing for fair comparison of the results. In Figure 5, we show that the variance in the stationary regimes of SCAFFOLD decreases as the number of clients increases. As a consequence, one could increase the step size and still obtain accurate results. This confirms the result of Corollary 1, and is aligned with our **Conjectures 1 and 2**, where we argue that the linear speed-up holds for any number of local steps, and even when Hessian matrices are non-diagonal.



(a) $\gamma = 0.01$ (b) $\gamma = 1.0$

Fig. 4: Convergence rate, for different step sizes, as a function of the number of local updates. The green dashed line represents the optimal number of local updates as predicted by our theory. The grey dashed line corresponds to 1: for any number of local updates, SCAFFOLD converges.



(a) $\gamma = 0.01$, $H = 10$ (b) $\gamma = 1$, $H = 10$

Fig. 5: Illustration of the linear speed-up. We run SCAFFOLD on replicates of the same problem, with different numbers of clients. As the number of clients grow, the variance in the stationary regime decreases.

V. Discussion and Perspectives

We presented a novel theoretical analysis of the SCAFFOLD algorithm for quadratic objective functions. Our goal was to precisely characterize its convergence rate and demonstrate that SCAFFOLD (i) converges for any choice of local update steps, and (ii) accelerates significantly under an optimal choice of local updates. Our approach introduced a new technique to study convergence, based on a decomposition of the iterates' errors along the characteristic subspaces defined by the Hessian matrices of the local functions. Importantly, our work is the first to show that SCAFFOLD with stochastic gradients converges to a stationary distribution for an arbitrary number of local updates, under the mild stability condition on the step size $\gamma \leq 1/L$, where L denotes the gradient Lipschitz constant, which is classical in smooth optimization. Our theoretical findings are supported by numerical experiments, confirming that SCAFFOLD converges, albeit sometimes slowly, and consistently exhibits linear speed-up with respect to the number of participating clients.

Importantly, our results show that SCAFFOLD can achieve accelerated rates for both small and large step sizes. This unifies and extends prior analyses of SCAFFOLD and related methods, such as [4], [5], to the deterministic communication regime. However, despite the sharpness of our convergence rates, they do not reflect the influence of data heterogeneity, contrasting with existing communication lower bounds [9], which decrease with heterogeneity. Whether SCAFFOLD matches these bounds remains an open and intriguing question.

Motivated by our results, we proposed two conjectures regarding the general behavior of SCAFFOLD: (i) the algorithm achieves linear speed-up irrespective of the number of local updates, and (ii) linear speed-up persists even when the Hessians are not diagonal. We believe the refined analysis we developed in this paper may provide valuable tools for future studies on SCAFFOLD, and on federated optimization methods more broadly. Extending these insights beyond quadratic functions is particularly promising and constitutes an exciting direction for future research.

Acknowledgment

The work of P. Mangold has been supported by Technology Innovation Institute (TII). The work of E. Moulines has been partly funded by the European Union (ERC-2022-SYG-OCEAN- 101071601).

References

[1] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*, PMLR, 2020, pp. 5132–5143.

[2] R. Luo, S. U. Stich, S. Horváth, and M. Takáč, "Revisiting localsgd and scaffold: Improved rates and missing analysis," *arXiv preprint arXiv:2501.04443*, 2025.

[3] P. Mangold, A. Durmus, A. Dieuleveut, and E. Moulines, "Scaffold with stochastic gradients: New analysis with linear speed-up," *arXiv preprint arXiv:2503.07594*, 2025.

[4] K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtárik, "Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally!" In *International Conference on Machine Learning*, PMLR, 2022, pp. 15 750–15 769.

[5] Z. Hu and H. Huang, "Tighter analysis for proxskip," in *International Conference on Machine Learning*, PMLR, 2023, pp. 13 469–13 496.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[7] P. Mangold, S. Samsonov, S. Labbi, *et al.*, "Scaffsa: Taming heterogeneity in federated linear stochastic approximation and td learning," *arXiv preprint arXiv:2402.04114*, 2024.

[8] R. Douc, E. Moulines, P. Priouret, *et al.*, *Markov chains*. Springer, 2018.

[9] Y. Arjevani and O. Shamir, "Communication complexity of distributed convex learning and optimization," *Advances in neural information processing systems*, vol. 28, 2015.