Gaussian Approximation and Multiplier Bootstrap for Federated Linear Stochastic Approximation

Ilya Levin[*1] **Maksim Shuklin**[*1] **Eric Moulines**[2] **Paul Mangold**[3] **Sergey Samsonov**[1]

[1]HSE University, Russia
[2]Mohamed bin Zayed University of Artificial Intelligence, UAE
[3]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France

Abstract

In this paper, we establish Berry-Esseen-type bounds for federated linear stochastic approximation (LSA). Our results provide the first federated Gaussian approximations for LSA that explicitly capture communication-computation trade-offs and heterogeneity-aware error terms, quantifying the effects of local step size, number of local updates, and heterogeneity on convergence rates. We present results for both (i) constant step size regime and (ii) decreasing step size with an increasing number of local iterations, recovering the recent rates of Bonnerjee et al. [2026] as a special case. As a primary application of our results, we develop an online multiplier bootstrap procedure for inference on the last iterate, which avoids explicit estimation of the asymptotic covariance matrix, and obtain non-asymptotic validity guarantees for this procedure.

1 INTRODUCTION

Federated learning [McMahan et al., 2017, Kairouz and McMahan, 2021] enables collaborative training of machine learning models across distributed data sources. In this context, the Federated Averaging (FedAvg) algorithm [McMahan et al., 2017] is considered the most fundamental algorithm in federated learning. It reduces communication complexity by using local updates with periodic aggregation [Stich, 2019, Khaled et al., 2020, Woodworth et al., 2020]. Most of the federated learning literature adopts an optimization perspective, establishing convergence rates and communication complexity for variants of the FedAvg optimization scheme [Khaled et al., 2020, Karimireddy et al., 2020, Li et al., 2020, Mishchenko et al., 2022, Ogier du Terrail et al., 2022]. At the same time, an important question concerns inference procedures that aim to provide confidence intervals for parameters of interest or their functionals. The most challenging aspect of theoretical grounding in such algorithms is the Gaussian approximation (GAR) for the underlying model parameters. While there is a growing number of results studying the rates of Gaussian approximation for SGD [Shao and Zhang, 2022, Sheshukova et al., 2026] and other stochastic approximation algorithms [Samsonov et al., 2024a, Wu et al., 2024, 2025, Butyrin et al., 2025], the number of related contributions to the analysis of federated learning algorithms remains limited.

To our knowledge, the few existing works on statistical inference for FedAvg [Li et al., 2022, Gu and Chen, 2024] establish only *asymptotic* convergence to a limiting Gaussian distribution. A notable exception is the recent work of Bonnerjee et al. [2026], which establishes GAR rates for federated averaging with a *constant* number of local updates. Notably, they apply their rates to an inference procedure based on the multiplier bootstrap approach [Fang et al., 2018] and conjecture, without proof, a rate of approximation for this procedure for the Polyak-Ruppert averaged iterates. This remains an important gap in the federated literature, which lacks non-asymptotic methods with formal guarantees for quantifying uncertainty.

In this paper, we aim to close this gap and develop GAR results for federated averaging with an *increasing number of local iterations*, along with an inference procedure based on the multiplier bootstrap with non-asymptotic validity. Our main contributions are as follows:

- We present new high-order moment bounds for the error of the FedLSA algorithm with decreasing step size and increasing local iterations, capturing the effects of heterogeneity bias and noisy updates. To our knowledge, these are the first moment bounds of order p with $p \geq 2$ available in the literature.
- We present the first Gaussian approximation bound for the last iterate of FedLSA (see Theorem 3). Our analysis establishes a rate of order up to $t^{-1/2}$ in terms of convex dis-

[*]These authors contributed equally to this work.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3487–3545.

tance between probability distributions (see Section 5 for precise definitions). Our bounds are *fully non-asymptotic* and capture the interplay between the number of local iterations and the choice of step size.

- We present an inference procedure to approximate the error of the last iterate of FedLSA using the multiplier bootstrap [Fang et al., 2018]. Specifically, our bounds show that this distribution can be approximated at a rate of up to $t^{-1/2}$ in convex distance, bypassing the Gaussian approximation step involving the asymptotic covariance matrix of the last iterate, Σ_∞; see Section 6. This underscores the distinction between our proposed procedure and various plug-in or batch-mean techniques for constructing confidence intervals, which seek to directly approximate Σ_∞. Therefore, we provide *the first theoretically grounded method for assessing uncertainty in federated learning*, confirming the hypothesis of Bonnerjee et al. [2026].

Notations and definitions. For a matrix $A \in \mathbb{R}^{d\times d}$, we denote its operator norm by $\|A\|$. Let N be the number of agents; we use the notation $\mathbb{E}_c[a_c] = N^{-1}\sum_{c=1}^{N} a_c$ for the average over different agents. For a symmetric positive-definite matrix $Q = Q^\top \succ 0$, $Q \in \mathbb{R}^{d\times d}$, and $x \in \mathbb{R}^d$, we define the corresponding norm $\|x\|_Q = \sqrt{x^\top Q x}$, and the respective matrix Q-norm of a matrix $B \in \mathbb{R}^{d\times d}$ by $\|B\|_Q = \sup_{x\neq 0} \|Bx\|_Q/\|x\|_Q$. For sequences a_n and b_n, we write $a_n \lesssim_{\log_n} b_n$ if there exist $c, \alpha > 0$ (not depending on n) such that $a_n \leq c(1+\log n)^\alpha b_n$. In this text, the following abbreviations are used: "w.r.t." stands for "with respect to," "i.i.d." for "independent and identically distributed," and "GAR" for "Gaussian Approximation."

2 RELATED WORK

Federated stochastic approximation. Most of the federated learning literature has aimed to establish convergence rates for federated averaging under various assumptions [McMahan et al., 2017, Khaled et al., 2020, Wang et al., 2024b, Mangold et al., 2025a, Glasgow et al., 2022, Patel et al., 2023, Stich, 2019, Gorbunov et al., 2021], along with mechanisms to mitigate the impact of heterogeneity using control variates [Karimireddy et al., 2020, Mishchenko et al., 2022, Mangold et al., 2025b, Luo et al., 2025], local proximal updates [Li et al., 2020, Grudzień et al., 2023], or modified objectives [Wang et al., 2021, Mitra et al., 2021].

Finite-sample properties of federated SA were first studied by Doan [2020], Wai [2020], Liu and Olshevsky [2023] without local training. Khodadadian et al. [2022], Wang et al. [2024a] subsequently examined federated SA with local updates. More recently, Mangold et al. [2024], Zhu et al. [2026] analyzed federated linear SA with local training and bias correction. In this paper, we extend this line of work by addressing the decreasing-step-size regime alongside an increasing number of local iterations, and by obtaining p-moment bounds.

Gaussian approximation for SGD and SA. Classical asymptotic normality results for stochastic approximation and SGD algorithms date back to foundational papers [Polyak and Juditsky, 1992, Kushner and Yin, 2003, Benveniste et al., 2012]. Gaussian approximation and Berry-Esseen type bounds for nonlinear statistics of i.i.d. random variables were extensively studied in Shorack [2000], Shao and Zhang [2022]. These methods were applied to Polyak–Ruppert averaged linear stochastic approximation in Samsonov et al. [2024a], Wu et al. [2024], and to SGD in Sheshukova et al. [2026], along with non-asymptotic multiplier bootstrap guarantees. The case of stochastic approximation with Markovian noise was recently studied in Wu et al. [2025], Samsonov et al. [2026], Liu [2025]. In the decentralized setting, Gaussian approximation results are much more limited: Bonnerjee et al. [2026] established Gaussian approximation results for local SGD, including a Berry-Esseen bound for Polyak–Ruppert averaged SA and outlined the corresponding results for the last iterate, both in the regime of a *constant* number of local training steps. In contrast, our results track the heterogeneity-aware error terms that arise in the more general regime of an increasing number of local training steps.

Statistical inference. Classical methods for constructing confidence sets for optimal parameters are based on direct estimation of the asymptotic covariance matrix using plug-in or batch-mean methods Chen et al. [2020, 2021], Chang et al. [2026]. For federated learning, Li et al. [2022] studied asymptotic versions of the functional CLT and their applications to statistical estimation via plug-in and random scaling methods. Gu and Chen [2024] developed similar inference based on batch-mean methods. Both papers remain asymptotic in their analysis. Alternative approaches use bootstrap and resampling schemes Fang et al. [2018], Zhong et al. [2023], with non-asymptotic analysis provided in Samsonov et al. [2024a, 2026]. I this paper, we aim to generalize the fully non-asymptotic results obtained in these papers for the setting of Federated LSA.

3 FEDERATED LSA SETTING

In this paper, we consider federated linear stochastic approximation (LSA) with N agents, who aim to collaboratively solve the following linear system:

$$\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}\,,\ \text{with}\ \bar{\mathbf{A}} = \frac{1}{N}\sum_{c=1}^{N}\bar{\mathbf{A}}^c,\ \bar{\mathbf{b}} = \frac{1}{N}\sum_{c=1}^{N}\bar{\mathbf{b}}^c, \quad (1)$$

where, for $c \in \{1,\ldots,N\}$, $\bar{\mathbf{A}}^c \in \mathbb{R}^{d\times d}$ and $\bar{\mathbf{b}}^c \in \mathbb{R}^d$ are distributed among multiple agents. Furthermore, we assume that the system matrices $\bar{\mathbf{A}}^c$ and vectors $\bar{\mathbf{b}}^c$ are unknown, and agent c only has access to estimates of $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$.

Stochastic estimation. To estimate $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$, agent c obtains independent samples $Z_1^c, \ldots, Z_T^c$ from a distribution π_c, and computes, for $t \in \{1, \ldots, T\}$, the estimates $\mathbf{A}^c(Z_t^c)$ and $\mathbf{b}^c(Z_t^c)$ such that

$$\mathbb{E}_{Z\sim\pi_c}[\mathbf{A}^c(Z)] = \bar{\mathbf{A}}^c\,, \quad \mathbb{E}_{Z\sim\pi_c}[\mathbf{b}^c(Z)] = \bar{\mathbf{b}}^c\,. \tag{2}$$

We introduce the notation for the errors of observations

$$\begin{gathered}\widetilde{\mathbf{A}}^c(z) = \mathbf{A}^c(z) - \bar{\mathbf{A}}^c\,, \quad \widetilde{\mathbf{b}}^c(z) = \mathbf{b}^c(z) - \bar{\mathbf{b}}^c\,, \\ \varepsilon^c(\theta, z) = \widetilde{\mathbf{A}}^c(z)\theta - \widetilde{\mathbf{b}}^c(z)\,.\end{gathered} \tag{3}$$

We assume that the $\mathbf{A}^c(\cdot)$ satisfy the following assumption, which is classical in the LSA literature [Samsonov et al., 2024b, Mangold et al., 2024].

A1 (p)**.** *For each agent $c \in [N]$, the matrix $-\bar{\mathbf{A}}^c$ is Hurwitz, and there exist constants $a > 0$ and $\eta_{\infty,p} > 0$ such that $\eta_{\infty,p} a \le \frac{1}{2}$ and for all $0 < \eta < \eta_{\infty,p}$ and $u \in \mathbb{R}^d$,*

$$\mathbb{E}^{1/p}\big[\|(\mathrm{I} - \eta\mathbf{A}^c(Z))u\|^p\big] \le (1 - \eta a)\|u\|\,,$$

for $Z \sim \pi_c$ with a distribution π_c over Z.

This assumption ensures the existence and uniqueness of $\theta_\star$ and $\theta_\star^c$, respectively defined as the solutions of $\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}$ and $\bar{\mathbf{A}}^c\theta_\star^c = \bar{\mathbf{b}}^c$. Given the existence of $\theta_\star^c$, we will sometimes omit the dependence on θ in $\varepsilon^c(\theta_\star^c, z)$, denoting

$$\varepsilon^c(z) = \varepsilon^c(\theta_\star^c, z) = \widetilde{\mathbf{A}}^c(z)\theta_\star^c - \widetilde{\mathbf{b}}^c(z)\,. \tag{4}$$

To handle stochasticity in the estimates of $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$, we introduce the following quantities:

$$\|\varepsilon\|_\infty = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\varepsilon^c(z)\|\,, \quad \mathrm{C}_{\mathbf{A}} = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\mathbf{A}^c(z)\|\,,$$

which measure observation noise and the stochastic part of the LSA error. Now, we can introduce assumptions about the underlying random process and the stability of the system.

A 2. *For each agent c, $(Z_k^c)_{k\in\mathbb{N}}$ are i.i.d. random variables with values in $(\mathsf{Z}, \mathcal{Z})$ and distribution π_c satisfying $\mathbb{E}_{\pi_c}[\mathbf{A}^c(Z_k^c)] = \bar{\mathbf{A}}^c$ and $\mathbb{E}_{\pi_c}[\mathbf{b}(Z_k^c)] = \bar{\mathbf{b}}^c$. Moreover, the uniform boundness condition holds*

$$\|\varepsilon\|_\infty = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\varepsilon^c(z)\| < \infty\,,$$

$$\mathrm{C}_{\mathbf{A}} = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\mathbf{A}^c(z)\| \vee \|\widetilde{\mathbf{A}}^c(z)\| < \infty\,.$$

Heterogeneity. For federated LSA, a crucial challenge is to ensure that we obtain a solution to the averaged system (1), even though $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$ differ across agents. To assess the impact of heterogeneity, we measure the differences between $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$ using the following two quantities.

$$\zeta_1^2 = \frac{1}{N}\sum_{c=1}^N \|\bar{\mathbf{A}}^c(\theta_\star^c - \theta_\star)\|^2\,,\ \zeta_2^2 = \frac{1}{N}\sum_{c=1}^N \|\bar{\mathbf{A}}^c - \bar{\mathbf{A}}\|^2\,, \tag{5}$$

where, ζ_1 measures the heterogeneity of the solutions, while ζ_2 measures the heterogeneity of the matrices $\bar{\mathbf{A}}^c$. Under the standard assumptions **A**1 and **A**2, these two quantities are always defined and bounded. They play a key role in analyzing the impact of heterogeneity.

Algorithm 1 FedLSA

Input: stepsizes $\{\eta_t\}_{t=1}^T$, initial model $\theta_0 \in \mathbb{R}^d$, number of rounds T, number of agents N, local steps H_t
for $t = 1$ to T **do**
  for $c = 1$ to N **do**
    Initialize current round $\theta_{t,0}^c \leftarrow \theta_{t-1}$
    for $h = 1$ to H_t **do**
      Receive data noise $Z_{t,h}^c$
      Compute: $g_{t,h}^c \leftarrow \mathbf{A}^c(Z_{t,h}^c)\theta_{t,h-1}^c - \mathbf{b}^c(Z_{t,h}^c)$
      Update local parameter: $\theta_{t,h}^c \leftarrow \theta_{t,h-1}^c - \eta_t g_{t,h}^c$
  Update global parameter $\theta_t \leftarrow \frac{1}{N}\sum_{c=1}^N \theta_{t,H_t}^c$
Return: final parameter θ_T

Federated LSA algorithm. To solve the federated LSA problem (1), we use the FedLSA algorithm Doan [2020], Mangold et al. [2024]. This method leverages local training on each agent to reduce communication overhead. At each communication round t, the server broadcasts the current global model θ_t to all agents. Each agent c then performs H_t local stochastic updates using its own data, starting from $\theta_{t,0}^c = \theta_{t-1}$ and, for $h = 1, \ldots, H_t$, performing the local update

$$\theta_{t,h}^c = \theta_{t,h-1}^c - \eta_t\big(\mathbf{A}^c(Z_{t,h}^c)\,\theta_{t,h-1}^c - \mathbf{b}^c(Z_{t,h}^c)\big)\,, \tag{6}$$

where η_t is the learning rate at round t. After completing the local updates, the server aggregates the agent models by averaging: $\theta_{t+1} = \frac{1}{N}\sum_{c=1}^N \theta_{t,H_t}^c$. The full procedure is detailed in Algorithm 1. In contrast to previous work, we consider a decreasing sequence of step sizes along with an increasing sequence of local step numbers.

4 MOMENT BOUNDS FOR FEDLSA

We first derive bounds on the moments of the error in FedLSA. First, we bound the second moment of the FedLSA error when using decreasing step sizes and increasing numbers of local iterations. This result will be used in Section 5 to obtain rates for Gaussian approximation in FedLSA. Next, we extend the analysis to high-order moment bounds that will be used in the analysis of the federated multiplier bootstrap procedure presented in Section 6.

4.1 ERROR DECOMPOSITION FOR FEDLSA

To derive our moment bounds, we extend the error decomposition framework from Mangold et al. [2024] to accommodate non-constant step sizes and numbers of local iterations. We define the following quantities:

$$\Gamma_{t,l:r}^c = \prod_{h=l}^r (\mathrm{I} - \eta_{t,h}\mathbf{A}^c(Z_{t,h}^c))\,,\ \Gamma_t^{\mathsf{avg}} = \frac{1}{N}\sum_{c=1}^N \Gamma_{t,1:H_t}^c\,.$$

For brevity, we will denote $\Gamma_t^c = \Gamma_{t,1:H_t}^c$. We also define the noisy part of these matrices as

$$\Delta_{s,t} = \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \mathbb{E}\big[\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\big]\,,$$

which captures the fluctuations of the matrix product Γ_i^{avg}. Using these notations, we obtain the following error decomposition for the last iterate:

$$\theta_t - \theta_\star = \tilde{\theta}_t^{(\mathrm{tr})} + \tilde{\theta}_t^{(\mathrm{bi,bi})} + \tilde{\theta}_t^{(\mathrm{fl,bi})} + \tilde{\theta}_t^{(\mathrm{fl})}\,, \tag{7}$$

where we introduce the following notations

$$\tilde{\theta}_t^{(\mathrm{tr})} = \prod_{s=1}^t \Gamma_s^{\mathsf{avg}} \{\theta_0 - \theta_\star\}\,, \tag{8}$$

$$\tilde{\theta}_t^{(\mathrm{bi,bi})} = \sum_{s=1}^t \mathbb{E}\Big[\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\Big] \rho_s^{\mathsf{avg}}\,, \tag{9}$$

$$\tilde{\theta}_t^{(\mathrm{fl,bi})} = \sum_{s=1}^t \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} \tau_s^{\mathsf{avg}} + \Delta_{s,t}\rho_s^{\mathsf{avg}}\,, \tag{10}$$

$$\tilde{\theta}_t^{(\mathrm{fl})} = \sum_{s=1}^t \big\{ \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} \cdot \varphi_s^{\mathsf{avg}} \big\}\,. \tag{11}$$

The first term, $\tilde{\theta}_t^{(\mathrm{tr})}$, characterizes the rate at which the initial error decays. The terms $\tilde{\theta}_t^{(\mathrm{bi,bi})}$ and $\tilde{\theta}_t^{(\mathrm{fl,bi})}$ capture the bias and fluctuations arising from statistical heterogeneity. In the special case of homogeneous agents (i.e., $\bar{\mathbf{A}}^c = \bar{\mathbf{A}}$ for all agents $c \in [N]$), these two terms vanish. Finally, $\tilde{\theta}_t^{(\mathrm{fl})}$ accounts for the fluctuations of θ_t around the solution $\theta_\star$.

4.2 MOMENT BOUNDS

We now establish bounds on the moments of the error. We emphasize that, although similar to Mangold et al. [2024]'s decomposition, the varying step size and number of local updates change the nature of the terms from decomposition Equation (7). Thus we emphasize that the bounds derived in this section do not follow from the ones obtained in Mangold et al. [2024]. We consider two choices of step size and local update schedules:

S1 (Constant). *For any $t > 0$, we set $\eta_t = \eta$ and $H_t = H$ for some $0 < \eta \le \eta_{\infty,p}$ and $H > 0$.*

S2 (Decreasing). *For any $t > 0$, we set $\eta_t = \eta(1+t)^{-\gamma_\eta}$ and $H_t = \lceil H(1+t)^{\gamma_H} \rceil$, for some $\eta \in (0,1)$, $H > 0$ and $\gamma_\eta \in [\frac{1}{2}, 1)$, $\gamma_H \ge 0$, such that $\gamma_\eta \ge \gamma_H$. Additionally, we assume that $\eta H < a^{-1}$ and we denote $\gamma = \gamma_\eta - \gamma_H$.*

The scheme S2 extends the classical polynomially decaying step size regime by allowing the number of local updates to increase over time. In federated settings, such scaling is natural: as the step size decreases, bias reduces, and performing more local updates per communication round can improve communication efficiency. The condition $\gamma_\eta \ge \gamma_H$ ensures that the overall contraction remains stable, while $\eta H < a^{-1}$ guarantees uniform stability of the local dynamics. To our knowledge, this joint regime of decaying step sizes and increasing number of local updates has not been previously analyzed for federated learning.

Second moment bounds. We first derive a general theorem on the squared error, applicable to any sequence of decreasing step sizes and increasing numbers of local iterations.

Theorem 1. *Assume A1(2), A2. Let $\eta_s \le \eta_{\infty,2}$, be decreasing, and H_s be increasing, for all $s \ge 0$. Then, it holds*

$$\begin{aligned}
&\mathbb{E}^{1/2}[\|\theta_t - \theta_\star\|^2] \le \prod_{s=1}^t (1-\eta_s a)^{H_s} \|\theta_0 - \theta_\star\| \\
&+ \zeta_1\zeta_2 \sum_{s=1}^t \eta_s^2 H_s^2 \exp\Big(-a \sum_{i=s+1}^t \eta_i H_i \Big) \\
&+ \tfrac{\bar{\sigma}_\varepsilon + 2\bar{\sigma}_{het}}{\sqrt{N}} \sqrt{\sum_{s=1}^t \eta_s^2 H_s \exp(-2a\sum_{r=1}^t \eta_r H_r))} \\
&+ \sqrt{\tfrac{4\bar{\sigma}_{\mathbf{A}}^2 \zeta_1^2\zeta_2^2}{Nae}} \sqrt{\sum_{s=1}^t \eta_s^5 H_s^4 \exp\big(-a \sum_{r=s+1}^t \eta_r H_r \big)}\,,
\end{aligned}$$

where $\bar{\sigma}_\varepsilon, \bar{\sigma}_{het}$ and $\bar{\sigma}_{\mathbf{A}}$ are defined in (28).

The proof of Theorem 1 relies on bounding each term in the error decomposition (7). We provide a proof with explicit constants in Section A.1. When the step size and the number of local updates are constant over time, the above result simplifies and recovers an improved variant of Theorem A.6 in Mangold et al. [2024].

Corollary 1. *Assume A1(2), A2 and S1. Then, it holds*

$$\begin{aligned}
\mathbb{E}^{1/2}[\|\theta_t - \theta_\star\|^2] &\le (1-\eta a)^{Ht} \|\theta_0 - \theta_\star\| + \tfrac{\eta H \zeta_1\zeta_2}{a} \\
&+ \sqrt{\tfrac{\eta \bar{\sigma}_\varepsilon^2}{Na}} + \sqrt{\tfrac{\eta \bar{\sigma}_{het}^2}{Na}} + \sqrt{\tfrac{4\eta^4 H^3 \bar{\sigma}_{\mathbf{A}}^2 \zeta_1^2\zeta_2^2}{Na^2 e}}\,.
\end{aligned}$$

This bound shows that in order to obtain MSE of order $\mathbb{E}[\|\theta_t - \theta_\star\|^2] \le \epsilon^2$, one needs to choose the step size $\eta \asymp \epsilon^2$, and the product $\eta H \asymp \epsilon$. This choice is due to non-vanishing variance terms. In contrast, one can make the variance vanish by choosing appropriately decreasing step size. We thus establish the result on the last iterate moment bounds under the scheme S2.

Corollary 2. *Assume A1(2), A2 and S2. Then, for all $t \ge 1$, the following bound holds*

$$\begin{aligned}
&\mathbb{E}^{1/2}\big[\|\theta_t - \theta_\star\|^2\big] \lesssim \exp\big(-\tfrac{a\eta H}{1-\gamma}(t+1)^{1-\gamma} \big) \|\theta_0 - \theta_\star\| \\
&+ \tfrac{8\zeta_1\zeta_2\eta H}{a(t+1)^\gamma} + \sqrt{\tfrac{\eta}{Na}} \tfrac{\bar{\sigma}_\varepsilon + \bar{\sigma}_{het}}{(t+1)^{\gamma_\eta/2}} + \sqrt{\tfrac{\eta^4 H^3}{Nae}} \tfrac{\bar{\sigma}_{\mathbf{A}}\zeta_1\zeta_2}{(t+1)^{(4\gamma_\eta - 3\gamma_H)/2}}\,.
\end{aligned}$$

This corollary shows that, when using decaying step size, one can obtain convergence to arbitrary precision without starting with tiny step sizes. Specifically, this is the case as long as $\gamma > 0$, implying that step sizes must decrease faster than the number of local iteration increases.

Higher-order moment bounds. Similarly, we establish the result for higher-order moment bounds, which are crucial for further multiplier bootstrap analysis in Section 6.

Theorem 2. *Assume* **A**1*(p),* **A**2*. Let* $\eta_s \leq \eta_{\infty,2}$*, be decreasing, and* H_s *be increasing, for all* $s \geq 0$*. Then, it holds*

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|\theta_t - \theta_\star\|^p] \leq \exp\Big(-a \textstyle\sum_{s=1}^t \eta_s H_s\Big)\|\theta_0 - \theta_\star\| \\
&+ \sqrt{\tfrac{p^3\|\varepsilon\|_\infty^2}{N^2} \textstyle\sum_{s=1}^t \eta_s^2 H_s \exp\Big(-2a \displaystyle\sum_{i=s+1}^{t} \eta_i H_i\Big)} \\
&+ \zeta_1\zeta_2 \textstyle\sum_{s=1}^t \eta_s^2 H_s^2 \exp\Big(-a \sum_{i=s+1}^t \eta_i H_i\Big) \\
&+ 2\,\mathrm{C}_{\mathbf{A}}\, p^2 \zeta_* \sqrt{\textstyle\sum_{s=1}^t \eta_s^2 H_s \exp\Big(-2a \sum_{i=s}^t \eta_i H_i\Big)} \\
&+ \tfrac{\mathrm{C}_{\mathbf{A}}\, p^3 \zeta_1\zeta_2}{\sqrt{Nae}} \sqrt{\textstyle\sum_{s=1}^t \eta_s^5 H_s^4 \exp\Big(-a \sum_{i=s+1}^t \eta_i H_i\Big)}\,,
\end{aligned}
$$

where $\zeta_* = 1/N \sum_{c=1}^N \|\theta_\star^c - \theta_\star\|$.

This theorem is very similar to Theorem 1, up to an additional factor depending on p. Following the second-order case, we obtain the counterpart of Corollaries 1 and 2, which we state in Section A.3, together with a proof of Theorem 2.

5 GAUSSIAN APPROXIMATION

5.1 GAUSSIAN APPROXIMATION FRAMEWORK

In this section, we present our main result, establishing non-asymptotic Gaussian approximation rates for the last iterate of FedLSA. Before presenting particular results, we outline a general scheme for proving the normal approximation in Section 5.1.

The classical object studied in the literature on Gaussian approximation [Chen and Shao, 2007, Shao and Zhang, 2022] is the general vector-valued nonlinear statistic $T(X_1, \ldots, X_n) \in \mathbb{R}^d$, which can be represented as

$$T = W + D\,, \tag{12}$$

where W is a linear statistic of $X_1, \ldots, X_n$, and D has small moments. We consider the decomposition (12) and assume, without loss of generality, that $\mathbb{E}[WW^\top] = \mathrm{I}_d$. To measure the approximation quality, we focus on the convex distance d_C, defined for probability measures μ, ν on $\mathbb{R}^d$ as

$$\mathsf{d}_\mathsf{C}(\mu,\nu) = \sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mu(B) - \nu(B)|\,. \tag{13}$$

Estimating $\mathsf{d}_\mathsf{C}(T, \mathcal{N}(0, \mathrm{I}_d))$ can be reduced to the following two steps. First, one needs to estimate the convex distance $\mathsf{d}_\mathsf{C}(W, \mathcal{N}(0, \mathrm{I}_d))$ using the tools for normal approximation for linear statistics, such as Bentkus [2003], Shao and Zhang [2022]. In the second step, one proceeds with one of the following approaches:

1. Estimate moments $\mathbb{E}[\|D\|^p]$ for some $p \geq 1$ and then apply Proposition 5 (see also [Sheshukova et al., 2026, Proposition 1]);
2. Estimation of the perturbations of the term D within the randomized concentration approach Shao and Zhang [2022], requiring a bound of the fluctuations of the term D when one of the variables X_i, for some $i \in [n]$, is modified.

The first approach is technically simpler, yet it requires one to estimate moments of increasing order p, which produces some subtle logarithmic factors in the final convergence rates. The second one is technically more involved, yet it requires only estimating the second moment of the term D and provides the bounds that are sharper in terms of extra logarithmic factors.

5.2 LINEAR STATISTIC OF FEDLSA

Self-normalized statistic. To apply Gaussian approximation results for nonlinear statistics of independent random variables, we extract the leading linear component of the noise, a linear statistic of $\{Z_k\}_{k \geq 0}$, and apply Shao and Zhang [2022, Theorem 2.1]. We write the error as follows.

$$\eta_t^{-1/2}(\theta_t - \theta_\star) = M_t + D_t^{\mathrm{rm}}, \tag{14}$$

where M_t collects the linear statistics of $\{Z_{s,h}^c\}$ and D_t^{rm} is a higher-order remainder term. Following this decomposition, we define the covariance matrix

$$\Sigma_t := \mathbb{E}[M_t M_t^\top]\,, \tag{15}$$

which characterizes the effective variance of the last iterate. Under **S**2, Σ_t converges to the solution of the associated discrete-time Lyapunov equation Σ_∞, which determines the asymptotic covariance of FedLSA. We then analyze the self-normalized statistic

$$T_t := \eta_t^{-1/2} \Sigma_t^{-1/2}(\theta_t - \theta_\star), \tag{16}$$

where Σ_t is a covariance matrix representing the main linear part of non-linear statistics T_t, and is defined in (15). Under the stability conditions imposed in the previous section, T_t weakly converges to $\mathcal{N}(0, \mathrm{I}_d)$ as $t \to \infty$. Our objective is to quantify this convergence at finite time using the convex distance (13).

Linear statistic of FedLSA's error. We start from the decomposition

$$
\begin{aligned}
\theta_t - \theta_\star = &\textstyle\prod_{i=1}^t \Gamma_i^{\mathsf{avg}}(\theta_0 - \theta_\star) \\
&+ \textstyle\sum_{s=1}^t \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}(\Delta_{s,H_s}^{\mathsf{avg}} - \eta_s \mathcal{E}_{s,H_s}^{\mathsf{avg}}),
\end{aligned} \tag{17}
$$

where we recall that

$$
\begin{aligned}
\Delta_{t,h}^{\mathsf{avg}} &= N^{-1} \textstyle\sum_{c=1}^N (\mathrm{I} - \Gamma_{t,h}^c)(\theta_\star^c - \theta_\star), \\
\mathcal{E}_{t,h}^{\mathsf{avg}} &= N^{-1} \textstyle\sum_{c=1}^N \sum_{s=1}^h \Gamma_{t,s+1:h}^c \varepsilon^c(Z_{t,s,c}).
\end{aligned}
$$

The transient term $\prod_{i=1}^{t} \Gamma_i^{\mathsf{avg}}(\theta_0 - \theta_\star)$ in (17) is negligible under **A**1. The main stochastic contribution arises from the second term, which aggregates agent noise and heterogeneity-induced fluctuations. By Corollary 2, both components contribute at the same order up to heterogeneity constants ($\bar{\sigma}_\varepsilon^2$ and $\bar{\sigma}_{\text{het}}^2$ respectively) and must be handled jointly. To obtain the linear statistic, we center all stochastic contraction matrices around their expectation in (17), using the following matrices, defined for $t, l, r \geq 0$,

$$G_{t,l:r}^{c} = \mathbb{E}[\Gamma_{t,l:r}^{c}]\,, \text{ and } G_i^{\mathsf{avg}} = \mathbb{E}[\Gamma_i^{\mathsf{avg}}]\,. \tag{18}$$

Unrolling the resulting decomposition and identifying the leading term at each step gives the linear statistic

$$M_t = -N^{-1}\eta_t^{-1/2}\sum_{s=1}^{t}\eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\,(\mathrm{U}_s^{\text{noise}} + \mathrm{U}_s^{\text{het}})\,,$$

where $\mathrm{U}_s^{\text{noise}}$ accounts for the noise directly due to noisy updates, and $\mathrm{U}_s^{\text{het}}$ is a noise term which arises due to the heterogeneity of the local solutions,

$$\begin{aligned}\mathrm{U}_s^{\text{noise}} &:= \textstyle\sum_{c=1}^{N}\sum_{h=1}^{H_s} G_{s,h+1:H_s}^{c}\varepsilon^c(Z_{s,h,c})\,,\\ \mathrm{U}_s^{\text{het}} &:= \textstyle\sum_{c=1}^{N}\sum_{h=1}^{H_s} G_{s,1:h-1}^{c}\widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star^c - \theta_\star)\,.\end{aligned}$$

The remainder term is then given by

$$D_t^{\mathsf{rm}} = \eta_t^{-1/2}\sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}(\Delta_{s,H_s}^{\mathsf{avg}} - \eta_s\mathcal{E}_{s,H_s}^{\mathsf{avg}}) - M_t\,.$$

Then, we can rewrite our self-normalized statistic T_t as

$$\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star) = W_t + D_t\,, \tag{19}$$

where $W_t := \Sigma_t^{-1/2}M_t$ and $D_t = \Sigma_t^{-1/2}D_t^{\mathsf{rm}}$.

5.3 LAST ITERATE GAUSSIAN APPROXIMATION

We now present the main result of this section. Applying Shao and Zhang [2022, Theorem 2.1] to the decomposition (19) yields a non-asymptotic Gaussian approximation for $\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star)$, leading to the following result.

Theorem 3. *Under assumptions **A**1(2), **A**2 and **S**2, setting $Y \sim \mathcal{N}(0, \mathrm{I}_d)$, we obtain*

$$\begin{aligned}&\mathsf{d}_{\mathsf{C}}(\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star), Y)\\ &\lesssim N^{-1/2}(1+t)^{-\gamma_\eta/2} + N^{1/2}(1+t)^{-\gamma}\\ &+ (1+t)^{-(3\gamma-\gamma_\eta)/2}\,,\end{aligned}$$

where Σ_t is defined in (15).

Theorem 3 extends [Bonnerjee et al., 2026, Theorem 2.2] - which assume a constant number of local steps ($\gamma_H = 0$). Note that Theorem 3 implies, since convex distance is invariant under non-degenerate linear mappings, approximation rates for

$$\mathsf{d}_{\mathsf{C}}(\eta_t^{-1/2}(\theta_t - \theta_\star), \mathcal{N}(0, \Sigma_t))\,,$$

and not for

$$\mathsf{d}_{\mathsf{C}}(\eta_t^{-1/2}(\theta_t - \theta_\star), \mathcal{N}(0, \Sigma_\infty))\,,$$

where Σ_∞ is the limiting covariance matrix of the last iterate of federated LSA defined as the unique solution of the Lyapunov equation (see e.g. Fort [2015]):

$$\bar{\mathbf{A}}\Sigma_\infty + \Sigma_\infty\bar{\mathbf{A}}^\top = N^{-1}\Sigma_*^{\mathsf{avg}}\,. \tag{20}$$

Here Σ_*^{avg} is the covariance of noise at $\theta_\star$

$$\Sigma_*^{\mathsf{avg}} = N^{-1}\textstyle\sum_{c=1}^{N}\mathbb{E}[\varepsilon^c(\theta_\star, Z_c)(\varepsilon^c(\theta_\star, Z_c))^\top]\,.$$

Although the empirical covariance matrix Σ_t converges to Σ_∞ as $t \to \infty$, the rate of convergence may be slow and can therefore dominate the overall Gaussian approximation error. Similar effects were previously observed for Polyak-Ruppert averaged SGD in the single-agent setting in Shao and Zhang [2022] and Sheshukova et al. [2026]. Interestingly, the same phenomenon persists for the last iterate bounds. The next proposition quantifies this effect.

Proposition 1. *Under assumptions **A**1(2), **A**2, and **S**2, and choosing $\eta H \leq \beta_\infty$ for some $\beta_\infty > 0$, we have*

$$\|\Sigma_t - \Sigma_\infty\| \leq C_{\infty,1}(1+t)^{\gamma-1} + C_{\infty,2}(1+t)^{-\gamma}\,,$$

where the constants $C_{\infty,1}$ and $C_{\infty,2}$ are given in (99) and (100).

Proposition 1 reveals an intrinsic trade-off in the choice of the parameter $\gamma = \gamma_\eta - \gamma_H$. On the one hand, γ cannot be too small, as sufficiently fast decay is required to ensure moment control and Gaussian approximation rates. On the other hand, taking γ close to 1 deteriorates the covariance stabilization rate through the term

$$C_{\infty,1}(1+t)^{\gamma-1}\,. \tag{21}$$

Interestingly, we can show the lower bound demonstrating that both terms in the bound of Proposition 1 are essential.

Lower bound. Consider a 1-dimensional linear stochastic approximation with single agent ($N = 1$):

$$\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}, \qquad \bar{\mathbf{A}} = 1, \quad \bar{\mathbf{b}} = 0\,,$$

where the unbiased stochastic estimates are given by

$$\mathbf{A}(Z_j) = 1 + \xi_j, \qquad \mathbf{b}(Z_j) = \bar{\mathbf{b}} = 0,$$

with $\xi_j \sim \mathcal{N}(0,1)$. Note that in this setting there are no synchronization steps across agents (since $N = 1$), so $H_t = 1$, and the algorithm reduces to the usual 1-dimensional LSA.

Then, the FEDLSA algorithm with starting point $\theta_0 = 0$ can be written as the recurrence

$$\theta_t = \theta_{t-1} - \eta_t(1+\xi_i)\theta_{t-1}, \quad \theta_0 = 0. \tag{22}$$

We set the step sizes as $\eta_t = 1/(1+t)^{\gamma_\eta}$, where $\gamma_\eta \in (0,1)$. Note that in the single-agent case we have $\gamma_H = 0$, so that $\gamma = \gamma_\eta$. In this case, it is possible to compute the exact value of the limiting variance σ_∞^2 and to derive a lower bound for the convex distance:

Proposition 2. *Assume that $\eta_1 = 0$ and $\eta_\ell = (1+\ell)^{-\gamma}$ with $0 < \gamma < 1$. For 1-dimensional single agent* FEDLSA *with $\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}, \bar{\mathbf{A}} = 1, \bar{\mathbf{b}} = 0$ and $\mathbf{A}(Z_j) = 1+\xi_j, \mathbf{b}(Z_j) = \bar{\mathbf{b}}$ with $\xi_j \sim \mathcal{N}(0,1)$ it holds that*

$$\sigma_\infty^2 := \lim_{t\to\infty} \mathrm{Var}[\eta_t^{-1/2}(\theta_t - \theta_\star)] = \frac{1}{2}\,.$$

Moreover, for any $t \geq 4$, it can be shown that

$$\left| v_t - \frac{1}{2} \right| \geq \mathrm{C}_{\mathbf{v}}(t^{\gamma-1} + t^{-\gamma})\,, \tag{23}$$

for some positive constant $\mathrm{C}_{\mathbf{v}} > 0$ that depends only on γ.

We provide the proof of this lower bound in Section B.3. To the best of our knowledge, this is the first result that establishes a non-asymptotic lower bound for the last iterate of LSA in terms of the convex distance, thereby extending the lower bound results for Polyak–Ruppert averaging due to Sheshukova et al. [2026, Theorem 5].

Discussion. Theorem 3 demonstrates that increasing the number of local updates over time (i.e., $\gamma_H > 0$) reduces γ and therefore deteriorates the rate of normal approximation. While enlarging the number of local iterations may improve optimization accuracy or communication efficiency, it adversely affects distributional convergence. From the perspective of Gaussian approximation, this suggests that a constant number of local updates ($\gamma_H = 0$) is preferable. However, this choice can not always be implementable due to communication constraints.

6 MULTIPLIER BOOTSTRAP FOR FEDLSA

We now aim to perform statistical inference for the last-iterate estimator θ_t, and to construct confidence intervals for $\theta_\star$. Instead of explicitly estimating the asymptotic covariance matrix Σ_∞ and relying on plug-in Gaussian approximations [Li et al., 2022], we use an online multiplier bootstrap that directly approximates the sampling distribution of θ_t in a data-driven manner. To this end, we define a bootstrapped sequence, starting with the same initialization as the original FedLSA dynamics, but using randomly weighted local updates. Bootstrap iterates are defined by

$$\theta_{t,h}^{\mathsf{b},c} = \theta_{t,h-1}^{\mathsf{b},c} - \eta_t w_{t,h,c}\big(\mathbf{A}^c(Z_{t,h,c})\theta_{t,h-1}^{\mathsf{b},c} - \mathbf{b}^c(Z_{t,h,c})\big),$$

where the weights $w_{t,h,c}$ are i.i.d. random variables independent of the data $\{Z_{s,h}^c\}$ which satisfy

$$\mathbb{E}[w_{1,1,1}] = 1, \qquad \mathrm{Var}[w_{1,1,1}] = 1.$$

For ease of notation, we gather these weights in a set

$$\mathcal{W}^t = \{w_{s,h,c} : s \in [t],\, h \in [H_s],\, c \in [N]\}\,.$$

After each block of local updates, the global bootstrap iterate is updated as $\theta_t^{\mathsf{b}} = 1/N \sum_{c=1}^N \theta_{t,H_t}^{\mathsf{b},c}$.

Confidence intervals. Conditionally on the observed data, the distribution of $\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t)$ approximates the sampling distribution of $\eta_t^{-1/2}(\theta_t - \theta_\star)$. More precisely, denoting conditional probability and expectation given the data by

$$\mathbb{P}^b(\cdot) = \mathbb{P}(\cdot \mid \mathcal{Z}^t), \qquad \mathbb{E}^b(\cdot) = \mathbb{E}(\cdot \mid \mathcal{Z}^t)\,.$$

Our objective, which is the main result of this section, is to show that the bootstrap validity holds, that is, the quantity

$$\sup_{B\in\mathrm{Conv}(\mathbb{R}^d)} \Big| \mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) - \mathbb{P}(\eta_t^{-1/2}(\theta_t - \theta_\star) \in B) \Big| \tag{24}$$

converges to zero as $t \to \infty$, and to establish a rate for this convergence. This approximation result is essential, as it allows us to construct confidence intervals for the *true solution* $\theta_\star$ based only on the distribution of $\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t)$. For example, for a coordinate $j \in [d]$ and confidence level $(1-\alpha)$, with $\alpha \in (0,1)$, let $q_{\alpha/2}^b$ and $q_{1-\alpha/2}^b$ denote the empirical conditional quantiles of $\eta_t^{-1/2}(\theta_{t,j}^{\mathsf{b}} - \theta_{t,j})$. A $(1-\alpha)$ confidence interval for $\theta_\star$'s j-th coordinate is then

$$\big[\theta_{t,j} - \eta_t^{1/2} q_{1-\alpha/2}^b,\; \theta_{t,j} - \eta_t^{1/2} q_{\alpha/2}^b\big].$$

Similarly, simultaneous confidence intervals can be constructed using bootstrap quantiles of $\|\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t)\|$. This approach avoids explicit covariance estimation and remains fully online. To establish its validity, we first specify the linear statistics of the bootstrap iterates in Section 6.1, and then establish our main theorem in Section 6.2.

6.1 BOOTSTRAP DECOMPOSITION

We now aim to extract the linear statistic from $\theta_t^{\mathsf{b}} - \theta_t$. To this end, we introduce the bootstrap counterparts of matrix product, for $t, \ell, r \geq 0$, as

$$\Gamma_{t,\ell:r}^{\mathsf{b},c} = \textstyle\prod_{h=l}^{r}(\mathrm{I} - \eta_t w_{t,h,c}\mathbf{A}^c(Z_{t,h,c}))\,, \quad \Gamma_t^{\mathsf{b},c} = \Gamma_{t,1:H_t}^{\mathsf{b},c}\,,$$

together with $\Gamma_t^{\mathsf{b},\mathsf{avg}} = N^{-1}\sum_{c=1}^N \Gamma_t^{\mathsf{b},c}$, the averaged of these bootstrap matrix products. The main difference between these matrices and the $\Gamma_{t,\ell:r}$ from Section 4.1 is that they incorporate the weights $w_{t,h,c}$. Arguing as in the original process, we obtain the following decomposition of $\theta_t^{\mathsf{b}} - \theta_t$ using these bootstrap matrices

$$\begin{aligned}\theta_t^{\mathsf{b}} - \theta_t = &\textstyle\prod_{i=1}^t \Gamma_i^{\mathsf{b},\mathsf{avg}}(\theta_0^{\mathsf{b}} - \theta_0) \\ &- \textstyle\sum_{s=1}^t \eta_s \prod_{i=s+1}^t \Gamma_i^{\mathsf{b},\mathsf{avg}} \mathcal{E}_s^{\mathsf{b},\mathsf{avg}} ,\end{aligned}$$

where $\mathcal{E}_s^{\mathsf{b},\mathsf{avg}}$ is defined as

$$\mathcal{E}_s^{\mathsf{b},\mathsf{avg}} = N^{-1} \textstyle\sum_{c=1}^N \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{\mathsf{b},c} \tilde{\varepsilon}_{t,h}^c ,$$

with $\tilde{\varepsilon}_{t,h}^c = \varepsilon^c(Z_{t,h,c}) + \mathbf{A}^c(Z_{t,h,c})(\theta_{t,h-1}^c - \theta_\star^c)$. Note that here, in contrast with the decomposition (17), there is no additional heterogeneity term. This is expected, as we are considering the difference $\theta_t^{\mathsf{b}} - \theta_t$ and not the error $\theta_t - \theta_\star$. Extracting the leading linear statistics M_t^{b} in terms of bootstrap weights $w_{s,h,c}$ yields the following decomposition

$$\eta_t^{-1/2} (\Sigma_t^{\mathsf{b}})^{-1/2} (\theta_t^{\mathsf{b}} - \theta_t) = W_t^{\mathsf{b}} + D_t^{\mathsf{b}}, \tag{25}$$

where we define the bootstrap covariance matrix as

$$\Sigma_t^{\mathsf{b}} = \mathbb{E}^b \big[M_t^{\mathsf{b}} (M_t^{\mathsf{b}})^\top \big] , \tag{26}$$

together with $D_t^{\mathsf{b}} = \eta_t^{-1/2} (\Sigma_t^{\mathsf{b}})^{-1/2} (\theta_t^{\mathsf{b}} - \theta_t) - W_t^{\mathsf{b}}$, and $W_t^{\mathsf{b}} := (\Sigma_t^{\mathsf{b}})^{-1/2} M_t^{\mathsf{b}}$ where the linear term M_t^{b} given by

$$\begin{aligned} M_t^{\mathsf{b}} &= - N^{-1} \eta_t^{-1/2} \textstyle\sum_{s=1}^t \eta_s \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} \, \mathrm{U}_s^{\mathrm{noise},\mathsf{b}} , \\ \mathrm{U}_s^{\mathrm{noise},\mathsf{b}} &= \textstyle\sum_{c=1}^N \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^c \varepsilon^c(\theta_\star, Z_{s,h,c}) . \end{aligned}$$

Again, in comparison with results from Section 5.2, conditionally on $\mathcal{Z}^t$, no heterogeneity term appears, since we compare θ_t^{b} with θ_t.

6.2 BOOTSTRAP VALIDITY

We now establish the validity of the proposed multiplier bootstrap procedure for the last-iterate estimator in FedLSA. To conduct our analysis, we impose one additional technical assumption on the bootstrap weights $\mathcal{W}^t$, together with a condition on the minimal number of iterations.

A3. *The bootstrap weights $\mathcal{W}^t = \{w_{s,h,c} : s \in [t], h \in [H_s], c \in [N]\}$ satisfy $0 < W_{\min} < w_{1,1,1} < W_{\max} < +\infty$ almost surely for some $W_{\min}, W_{\max} > 0$. Additionally, we assume that $\eta < \eta_{\infty,p} / W_{\max}$.*

A4. *Number of observations t is large enough, that is, $t \geq t_0$. Precise expression for t_0 is given in Appendix, see* **A'***4.*

These assumptions are mostly technical. Examples of random variables that satisfy the assumption **A**3 are provided in, e.g., Sheshukova et al. [2026]; **A**4 ensures sufficient convergence of Σ_t to Σ_∞ and Σ_t^{b} to Σ_t, see Proposition 1. We now show that the conditional distribution of the bootstrap statistic provides an accurate Gaussian approximation to the sampling distribution of the normalized estimation error.

Theorem 4 (Bootstrap validity)**.** *Assume* ***A****1*$(\log(Nt))$*,* ***A****2,* ***S****2,* ***A****3,* ***A****4, and let $\eta H \leq \beta_\infty$. Then there exists an event Ω_0 with $\mathbb{P}(\Omega_0) \geq 1 - 4/t$ such that, on this event,*

$$\begin{aligned} \sup_{B \in \mathit{Conv}(\mathbb{R}^d)} \Big| &\mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) \\ &- \mathbb{P}(\eta_t^{-1/2}(\theta_t - \theta_\star) \in B) \Big| \\ \lesssim_{\log_t} \; & N^{-1/2} t^{-\gamma_\eta/2} + N^{1/2} t^{-\gamma} + t^{-(3\gamma - \gamma_\eta)/2} . \end{aligned}$$

Proof sketch. The proof relies on Gaussian approximations for both the original and bootstrap iterates, linked via a Gaussian comparison inequality. The main idea of the proof is illustrated with the diagram in Figure 1, following the general pipeline outlined in Spokoiny and Zhilova [2015], Samsonov et al. [2024a]:

$$\begin{tikzcd} \text{Real world: } \eta_t^{-1/2}(\theta_t - \theta_\star) \arrow[r, leftrightarrow, "\text{Theorem 3}"] & \mathcal{N}(0, \Sigma_t) \arrow[d, leftrightarrow, "\text{Gaussian comparison}"', "\text{Proposition 6}"] \\ \text{Bootstrap: } \eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \arrow[r, leftrightarrow, "\text{Theorem 5}"] & \mathcal{N}(0, \Sigma_t^{\mathsf{b}}) \end{tikzcd}$$

Figure 1: Illustration of the multiplier bootstrap bounds: real-world estimates are close to $\mathcal{N}(0, \Sigma_t)$, which is itself close to $\mathcal{N}(0, \Sigma_t^{\mathsf{b}})$.

We define random vectors $Y^{\mathsf{b}} \sim \mathcal{N}(0, \Sigma_t^{\mathsf{b}})$ under $\mathbb{P}^{\mathsf{b}}$-probability and $Y \sim \mathcal{N}(0, \Sigma_t)$. We define the convex distance under the conditional measure $\mathbb{P}^b$ as, for random vectors $X, Y \in \mathbb{R}^d$,

$$\mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(X, Y) = \sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \left| \mathbb{P}^{\mathsf{b}}(X \in B) - \mathbb{P}^{\mathsf{b}}(Y \in B) \right| .$$

Then the convex distance is split using the triangle inequality

$$\begin{aligned} \sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \Big| &\mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) \\ &- \mathbb{P}(\eta_t^{-1/2}(\theta_t - \theta_\star) \in B) \Big| \\ \leq \; & T_1 + T_2 + T_3, \end{aligned}$$

where we have defined

$$\begin{aligned} T_1 &= \mathsf{d}_{\mathsf{C}}^{\mathsf{b}}\big(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t), Y^{\mathsf{b}}\big), \\ T_2 &= \mathsf{d}_{\mathsf{C}}\big(\eta_t^{-1/2}(\theta_t - \theta_\star), Y\big), \\ T_3 &= \sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \left| \mathbb{P}(Y \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B) \right| . \end{aligned}$$

The term T_1 relates to the Gaussian approximation in the bootstrap world from Theorem 5. We estimate it using the decomposition (25), moment bounds for D_t^{b}, and the result of Shao and Zhang [2022, Theorem 2.1]. The term T_2 is controlled by the Gaussian approximation result for the original iterate (Theorem 3). The term T_3 is handled using matrix concentration bounds for $\|\Sigma_t^{\mathsf{b}} - \Sigma_t\|$ presented in Proposition 6. Combining the bounds implies the result.

Table 1: Coverage for nominal confidence level $\alpha = 0.95$ with $\gamma_H = 0$. Standard deviations are shown as subscripts.

T	SDB (ours)	EQ (ours)	PE
2000	$0.773_{\pm 0.013}$	$0.780_{\pm 0.013}$	$0.754_{\pm 0.013}$
6000	$0.936_{\pm 0.008}$	$0.941_{\pm 0.007}$	$0.931_{\pm 0.008}$
10000	$0.938_{\pm 0.008}$	$0.947_{\pm 0.007}$	$0.938_{\pm 0.008}$
14000	$0.941_{\pm 0.007}$	$0.946_{\pm 0.007}$	$0.941_{\pm 0.007}$

Table 2: Coverage for nominal confidence level $\alpha = 0.95$ with $\gamma_H = 0.2$. Standard deviations are shown as subscripts.

T	SDB (ours)	EQ (ours)	PE
2000	$0.398_{\pm 0.015}$	$0.404_{\pm 0.015}$	$0.376_{\pm 0.015}$
6000	$0.959_{\pm 0.006}$	$0.961_{\pm 0.006}$	$0.959_{\pm 0.006}$
10000	$0.946_{\pm 0.007}$	$0.949_{\pm 0.007}$	$0.946_{\pm 0.007}$
14000	$0.950_{\pm 0.007}$	$0.950_{\pm 0.007}$	$0.952_{\pm 0.007}$

Discussion. The counterpart of Theorem 4 has been conjectured without the proof for the Polyak-Ruppert averaged federated SGD estimates in Bonnerjee et al. [2026]. To our knowledge, we provide the first rigorous proof on the accuracy of Gaussian approximation with multiplier bootstrap in the federated setting. Importantly, approximation rates in Theorem 4 are of order up to $1/\sqrt{t}$ (up to logarithmic factors). As already emphasized, this rate is due to the fact that one can bypass approximating the asymptotic covariance matrix Σ_∞ as suggested earlier in Li et al. [2022]. Importantly, normal approximation with $\mathcal{N}(0, \Sigma_\infty)$ is bypassed even in the proof of Theorem 4, and is expected to yield slower approximation rates due to the lower bound (23).

7 NUMERICAL EXPERIMENTS

Numerical Illustration. We now numerically illustrate our normal approximation and bootstrap confidence intervals for FedLSA. To this end, we instantiate a Garnet environment Archibald et al. [1995], Geist et al. [2014] with $n = 30$ states, feature dimension $d = 5$, $a = 2$ actions, and branching factor $b = 2$. We generate $N = 5$ heterogeneous agents by perturbing this common Garnet MDP, following Mangold et al. [2024]. We run $R = 1024$ trajectories and construct confidence intervals for a random one-dimensional projection $u^\top \theta_t$. For each trajectory, we then check whether $\theta_\star$ lies within the interval, average this indicator over the R trajectories, and report the result as coverage, with the standard error shown as a subscript. For each run of the bootstrap methods, we compute 256 bootstrap trajectories.

Based on our results, we propose two methods (i) empirical bootstrap quantiles (EQ), and (ii) bootstrap standard deviation intervals (SDB). We compare these methods against a plug-in Gaussian approximation baseline, based on an estimated asymptotic covariance (PE). We provide full experimental details, additional confidence levels, and extended tables in Appendix D.

We report results for nominal confidence level $\alpha = 0.95$ using step size $\gamma_\eta = 0.6$ and two different regimes of local updates: in Table 1 with a constant number of local updates ($\gamma_H = 0$), and in Table 2 with an increasing number of local updates ($\gamma_H = 0.2$). In both cases, our methods (SDB and EQ) achieve better coverage during early iterations ($T = 2000$ and $T = 6000$), highlighting the superiority of these methods in regimes that are far from the asymptotic behavior. In contrast, as the number of total samples grow, all methods reach comparable coverage. This indicates that the plug-in covariance estimate improves as more samples are collected, and the algorithm reaches its asymptotic behavior. Finally, we note that using a constant number of local step sizes may allow for identifying confidence intervals with low precision earlier, while using an increasing number of local step sizes allows to reach good coverage with fewer communications.

8 CONCLUSION

We derived non-asymptotic Berry-Esseen–type bounds for the last iterate of FedLSA, providing finite-sample Gaussian approximation guarantees under heterogeneous data and communication constraints. Our results quantify the joint impact of the number of agents, step-size decay, and the schedule of local updates on the normal approximation rate. To derive these results, we established a new analysis of the convergence of FedLSA with decreasing step size and increasing number of local iterations, providing the first bounds on higher-order moments for this method. Based on this result, we established the validity of an online multiplier bootstrap procedure for FedLSA, which approximates the distribution of the normalized last iterate without requiring estimation of the asymptotic covariance matrix. Our results enable practical construction of confidence intervals and uncertainty quantification in federated settings.

Acknowledgements

The work of Ilya Levin, Maksim Shuklin, and Sergey Samsonov is an output of a research project HSE-BR-2025-019 implemented as part of the Basic Research Program at HSE University. This research was supported in part through computational resources of HPC facilities at HSE University Kostenetskiy et al. [2021]. The work of Paul Mangold is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005).

References

T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. On the generation of Markov decision processes. *The*

Journal of the Operational Research Society, 46(3):354–361, 1995.

S. Barsov and V. Ulyanov. Estimates for the closeness of Gaussian measures. *Dokl. Akad. Nauk SSSR*, 291(2): 273–277, 1986.

V. Bentkus. On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012.

Soham Bonnerjee, Sayar Karmakar, and Wei Biao Wu. Sharp gaussian approximations for decentralized federated learning. *Advances in Neural Information Processing Systems*, 38:157224–157270, 2026.

Bogdan Butyrin, Eric Moulines, Alexey Naumov, Sergey Samsonov, Qi-Man Shao, and Zhuo-Song Zhang. Improved Central Limit Theorem and Bootstrap Approximations for Linear Stochastic Approximation. *arXiv preprint arXiv:2510.12375*, 2025.

Xiangyu Chang, Xi Chen, Zehua Lai, He Li, Zhihong Liu, and Yichen Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *Journal of the American Statistical Association*, 0(0):1–14, 2026.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534): 708–719, 2021.

Louis H. Y. Chen and Qi-Man Shao. Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli*, 13(2):581–599, 2007.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251 – 273, 2020.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

Thinh T Doan. Local stochastic approximation: A unified view of federated learning and distributed multi-task reinforcement learning algorithms. *arXiv preprint arXiv:2006.13460*, 2020.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34:30063–30074, 2021.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19 (78):1–21, 2018.

M. V. Fedoryuk. *Metod perevala*. Izdat. "Nauka", Moscow, 1977.

Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.

Matthieu Geist, Bruno Scherrer, et al. Off-policy learning with eligibility traces: a survey. *J. Mach. Learn. Res.*, 15 (1):289–333, 2014.

Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

Eduard Gorbunov, Filip Hanzely, and Peter Richtarik. Local sgd: Unified theory and new efficient methods. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3556–3564. PMLR, 13–15 Apr 2021.

Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pages 1055–1092. PMLR, 2023.

Jia Gu and Song Xi Chen. Statistical inference for decentralized federated learning. *The Annals of Statistics*, 52 (6):2931–2955, 2024.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 4519–4529. PMLR, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International conference on machine learning*, pages 10997–11057. PMLR, 2022.

PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing, 2021.

Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local sgd. In *Conference on Learning Theory*, pages 1613–1661. PMLR, 2022.

Rui Liu and Alex Olshevsky. Distributed td (0) with almost no communication. *IEEE Control Systems Letters*, 7: 2892–2897, 2023.

Xingtu Liu. Central limit theorems for asynchronous averaged q-learning. *arXiv preprint arXiv:2509.18964*, 2025.

Ruichen Luo, Sebastian U Stich, Samuel Horváth, and Martin Takáč. Revisiting localsgd and scaffold: Improved rates and missing analysis. *arXiv preprint arXiv:2501.04443*, 2025.

Paul Mangold, Sergey Samsonov, Safwan Labbi, Ilya Levin, Reda Alami, Alexey Naumov, and Eric Moulines. Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning. *Advances in Neural Information Processing Systems*, 37:13927–13981, 2024.

Paul Mangold, Alain Durmus, Aymeric Dieuleveut, Sergey Samsonov, and Eric Moulines. Refined analysis of constant step size federated averaging and federated Richardson-Romberg extrapolation. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 5023–5031, 2025a.

Paul Mangold, Alain Oliviero Durmus, Aymeric Dieuleveut, and Eric Moulines. Scaffold with stochastic gradients: New analysis with linear speed-up. In *Forty-second International Conference on Machine Learning*, 2025b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr, 2017.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

F.W.J. Olver and W. Rheinbolt. *Asymptotics and Special Functions*. Academic Press, 2014. ISBN 9781483267449.

A. Osekowski. *Sharp Martingale and Semimartingale Inequalities*. Monografie Matematyczne 72. Birkhäuser Basel, 1 edition, 2012. ISBN 3034803699,9783034803694.

Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

A. S. Poznyak. *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques*. Elsevier, Oxford, 2008.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. *Advances in Neural Information Processing Systems*, 37: 12408–12460, 2024a.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4511–4547. PMLR, 2024b.

Sergey Samsonov, Marina Sheshukova, Eric Moulines, and Alexey Naumov. Statistical inference for linear stochastic approximation with markovian noise. *Advances in Neural Information Processing Systems*, 38:174565–174626, 2026.

Qi-Man Shao and Zhuo-Song Zhang. Berry–Esseen bounds for multivariate nonlinear statistics with applications to m-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

Marina Sheshukova, Sergey Samsonov, Denis Belomestny, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for stochastic gradient descent. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Galen R Shorack. *Probability for statisticians*. Springer, 2000.

Vladimir Spokoiny and Mayya Zhilova. Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, 43(6):2653 – 2675, 2015. doi: 10.1214/15-AOS1355.

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

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Hoi-To Wai. On the convergence of consensus algorithms with markovian noise and gradient bias. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4897–4902. IEEE, 2020.

Han Wang, Aritra Mitra, Hamed Hassani, George J. Pappas, and James Anderson. Federated TD learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*, 2024a.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *Transactions on Machine Learning Research*, 2024b.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Weichen Wu, Gen Li, Yuting Wei, and Alessandro Rinaldo. Statistical inference for temporal difference learning with linear function approximation. *arXiv preprint arXiv:2410.16106*, 2024.

Weichen Wu, Yuting Wei, and Alessandro Rinaldo. Uncertainty quantification for Markov chain induced martingales with application to temporal difference learning. *arXiv preprint arXiv:2502.13822*, 2025.

Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Online bootstrap inference with nonconvex stochastic gradient descent estimator. *arXiv preprint arXiv:2306.02205*, 2023.

Feng Zhu, Aritra Mitra, and Robert W. Heath. Achieving tighter finite-time rates for heterogeneous federated stochastic approximation under markovian sampling. *Transactions on Machine Learning Research*, 2026.

APPENDIX

TABLE OF CONTENTS

A PROOF FOR MOMENT BOUNDS

We are trying to solve the following linear equation system across N clients $c = 1 \dots N$:

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

where $\bar{\mathbf{A}} = \frac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}} = \frac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{b}}^c$.

Let $\theta_\star$ be the unique solution of $\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}$ and $\theta_\star^c$ the unique solution of $\bar{\mathbf{A}}^c\theta_\star^c = \bar{\mathbf{b}}^c$. Both $\bar{\mathbf{A}}^c$ and $\bar{\mathbf{b}}^c$ are not observed directly in our setting. Instead, we could observe $A^c(Z_t^c)$ and $b^c(Z_t^c)$ during t-th iteration at c-th client. For $z \in \mathsf{Z}$ we denote the stochastic part of the matrices and vectors $A^c(z)$ and $b^c(z)$

$$\widetilde{\mathbf{A}}^c(z) = \mathbf{A}^c(z) - \bar{\mathbf{A}}^c\,, \quad \widetilde{\mathbf{b}}^c(z) = \mathbf{b}^c(z) - \bar{\mathbf{b}}^c\,, \tag{27}$$

as well as the noise representations

$$\varepsilon^c(z) = \widetilde{\mathbf{A}}^c(z)\theta_\star^c - \widetilde{\mathbf{b}}^c(z)\,, \qquad \omega^c(z) = \widetilde{\mathbf{A}}^c(z)\theta_\star - \widetilde{\mathbf{b}}^c(z)\,.$$

The first corresponds to the intrinsic stochastic noise around the local solution, while the second measures the stochastic perturbation evaluated at the global solution. To control stochastic fluctuations, we define uniform bounds

$$\|\varepsilon\|_\infty = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\varepsilon^c(z)\|\,, \quad \mathrm{C}_{\mathbf{A}} = \max_{c\in[N]}\sup_{z\in\mathsf{Z}}\|\mathbf{A}^c(z)\|\,.$$

and the two following quantities, which measure variance

$$\bar{\sigma}^2_{\mathsf{het}} = \mathbb{E}_c\big[\|\Sigma^c_{\mathbf{A}}\|\;\|\theta^c_\star - \theta_\star\|^2\big]\;, \qquad \bar{\sigma}^2_\varepsilon = \mathbb{E}_c[\mathrm{Tr}(\Sigma^c_\varepsilon)]\;, \qquad \bar{\sigma}^2_{\mathbf{A}} = \frac{1}{N}\sum_{c=1}^{N}\mathbb{E}[\|\Sigma^c_{\mathbf{A}}\|^2] \tag{28}$$

Finally, we define the two following quantities, that define first and second order heterogeneity

$$\zeta_1^2 = \frac{1}{N}\sum_{c=1}^{N}\|\bar{\mathbf{A}}^c(\theta^c_\star - \theta_\star)\|^2\,, \qquad \zeta_2^2 = \frac{1}{N}\sum_{c=1}^{N}\|\bar{\mathbf{A}}^c - \bar{\mathbf{A}}\|^2\,. \tag{29}$$

A.1 EXPANSION OF THE ERROR OF FEDLSA

We first express a single local update of FEDLSA in terms of the local error $\theta^c_{t,h} - \theta^c_\star$. Using $\bar{\mathbf{A}}^c\theta^c_\star = \bar{\mathbf{b}}^c$, we obtain

$$\begin{aligned}\theta^c_{t,h} - \theta^c_\star &= (\mathrm{I} - \eta_{t,h}\mathbf{A}^c(Z^c_{t,h}))(\theta^c_{t,h-1} - \theta^c_\star) - \eta_{t,h}\left(\mathbf{A}^c(Z^c_{t,h})\theta^c_\star - \mathbf{b}^c(Z^c_{t,h})\right) \\ &= (\mathrm{I} - \eta_{t,h}\mathbf{A}^c(Z^c_{t,h}))(\theta^c_{t,h-1} - \theta^c_\star) - \eta_{t,h}\varepsilon^c(Z^c_{t,h})\,.\end{aligned} \tag{30}$$

Iterating the recursion over $h = 1, \dots, H_t$ yields

$$\theta^c_{t,H_t} - \theta^c_\star = \Gamma^c_{t,1:H_t}(\theta^c_{t,0} - \theta^c_\star) - \sum_{h=1}^{H_t}\eta_{t,h}\Gamma^c_{t,h+1:H_t}\varepsilon^c(Z^c_{t,h})\,, \tag{31}$$

where we defined the following matrix products

$$\Gamma^c_{t,l:r} = \prod_{h=l}^{r}(\mathrm{I} - \eta_{t,h}\mathbf{A}^c(Z^c_{t,h}))\,, \qquad \Gamma^c_t = \Gamma^c_{t,1:H_t}\,, \qquad \Gamma^{\mathsf{avg}}_t = \frac{1}{N}\sum_{c=1}^{N}\Gamma^c_t\,. \tag{32}$$

We now derive the global recursion for the error $\theta_t - \theta_\star$. Since $\theta_t = \frac{1}{N}\sum_{c=1}^{N} \theta^c_{t,H_t}$, we write

$$\theta_t - \theta_\star = \frac{1}{N}\sum_{c=1}^{N} \theta^c_{t,H_t} - \theta_\star = \frac{1}{N}\sum_{c=1}^{N}(\theta^c_{t,H_t} - \theta^c_\star) + \frac{1}{N}\sum_{c=1}^{N}(\theta^c_\star - \theta_\star).$$

Substituting (31) and using $\theta^c_{t,0} = \theta_{t-1}$ gives

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

Rearranging terms, we obtain the compact global error recursion

$$\begin{aligned}\theta_t - \theta_\star &= \Gamma^{\mathsf{avg}}_t(\theta_{t-1} - \theta_\star) + \frac{1}{N}\sum_{c=1}^{N}(\mathrm{I} - \Gamma^c_t)(\theta^c_\star - \theta_\star) \\ &\quad - \frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H_t}\eta_{t,h}\Gamma^c_{t,h+1:H_t}\varepsilon^c(Z^c_{t,h}).\end{aligned}$$

We may decompose $\theta_t - \theta_\star$ as

$$\theta_t - \theta_\star = \Gamma^{\mathsf{avg}}_t(\theta_{t-1} - \theta_\star) + \rho^{\mathsf{avg}}_t + \tau^{\mathsf{avg}}_t + \varphi^{\mathsf{avg}}_t\,, \tag{33}$$

where we introduced following notations for FedLSA's deterministic bias ρ^{avg}_t and its fluctuations τ^{avg}_t, as well as the fluctuations of the updates φ^{avg}_t,

$$\rho^{\mathsf{avg}}_t = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{I} - G^c_t)(\theta^c_\star - \theta_\star)\,, \quad \tau^{\mathsf{avg}}_t = \frac{1}{N}\sum_{c=1}^{N}(G^c_t - \Gamma^c_{t,1:H_t})(\theta^c_\star - \theta_\star)\,, \tag{34}$$

$$\varphi^{\mathsf{avg}}_t = -\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H_t}\eta_{t,h}\Gamma^c_{t,h+1:H_t}\varepsilon^c(Z^c_{t,h})\,, \tag{35}$$

as well as the following matrix notations, representing the expected value of the $\Gamma^c_{t,l:r}$, Γ^c_t, and Γ^{avg}_t matrices

$$G^c_{t,l:r} = \prod_{h=l}^{r}(\mathrm{I} - \eta_{t,h}\bar{\mathbf{A}}^c)\,, \qquad G^c_t = G_{t,1:H_t}\,, \qquad G^{\mathsf{avg}}_t = \frac{1}{N}\sum_{c=1}^{N} G^c_t\,. \tag{36}$$

Note that

$$\mathbb{E}[\Gamma^c_{t,l:r}] = G^c_{t,l:r}, \qquad \mathbb{E}[\Gamma^c_t] = G^c_t\,.$$

Unrolling (33), we obtain the following decomposition of the global iterate's error at step t,

$$\begin{aligned}\theta_t - \theta_\star &= \prod_{s=1}^{t}\Gamma^{\mathsf{avg}}_s(\theta_0 - \theta_\star) + \sum_{s=1}^{t}\prod_{j=s+1}^{t}\Gamma^{\mathsf{avg}}_j(\rho^{\mathsf{avg}}_s + \tau^{\mathsf{avg}}_s + \varphi^{\mathsf{avg}}_s) \\ &= \tilde{\theta}^{(\mathrm{tr})}_t + \tilde{\theta}^{(\text{bi, bi})}_t + \tilde{\theta}^{(\text{fl, bi})}_t + \tilde{\theta}^{(\text{fl})}_t\,.\end{aligned} \tag{37}$$

Where we introduced the notations

$$\tilde{\theta}^{(\mathrm{tr})}_t = \prod_{s=1}^{t}\Gamma^{\mathsf{avg}}_s\left\{\theta_0 - \theta_\star\right\} \tag{38}$$

$$\tilde{\theta}^{(\text{bi, bi})}_t = \sum_{s=1}^{t}\mathbb{E}\Big[\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_i\Big]\rho^{\mathsf{avg}}_s \tag{39}$$

$$\tilde{\theta}^{(\text{fl, bi})}_t = \sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_i\tau^{\mathsf{avg}}_s + \Delta_{s,t}\rho^{\mathsf{avg}}_s \tag{40}$$

$$\tilde{\theta}^{(\text{fl})}_t = \sum_{s=1}^{t}\Big\{\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_i \cdot \varphi^{\mathsf{avg}}_s\Big\}\,, \tag{41}$$

where we also defined $\Delta_{s,t} = \left\{\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}}\right\} - \mathbb{E}\left[\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}}\right]$, the fluctuations of the matrix product Γ_i^{avg}.

Finally, we define the past and future filtrations

$$\mathcal{F}_{s,h}^{-} = \sigma\big(Z_{t,k}^c : t < s \text{ or } (t = s,\ k \le h)\big)\,, \tag{42}$$
$$\mathcal{F}_{s,h}^{+} = \sigma\big(Z_{t,k}^c : t > s \text{ or } (t = s,\ k \ge h)\big)\,. \tag{43}$$

Here, $\mathcal{F}_{s,h}^{-}$ represents the noise up to iteration (s,h), while $\mathcal{F}_{s,h}^{+}$ represents the future noise starting from (s,h).

A.2 PROOFS FOR THE SECOND MOMENT BOUND $\mathbb{E}^{1/2}[\|\theta_t - \theta_\star\|^2]$

Transient term. First, we give a bound on the transient term $\tilde{\theta}_t^{(\mathsf{tr})}$, which quantifies the impact of the initialization.

Lemma 1. *Assume A1(2), A2, and let $\eta_s \le \eta_{\infty,2}$. Then, it holds*

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^2] \le \prod_{s=1}^{t}(1-\eta_s a)^{H_s}\|\theta_0 - \theta_\star\|\,.$$

Proof. Expanding $\tilde{\theta}_t^{(\mathrm{tr})}$ using (8) gives

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathrm{tr})}\|^2] = \mathbb{E}^{1/2}\left[\Big\|\prod_{s=1}^{t}\Gamma_s^{\mathsf{avg}}\cdot(\theta_0-\theta_\star)\Big\|^2\right] = \mathbb{E}^{1/2}\left[\Big\|\frac{1}{N}\sum_{c=1}^{N}\Gamma_{t,1:H_t}^c\prod_{s=1}^{t-1}\Gamma_s^{\mathsf{avg}}\cdot(\theta_0-\theta_\star)\Big\|^2\right].$$

Using (1) recursively then allows to bound

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathrm{tr})}\|^2] \le \frac{1}{N}\sum_{c=1}^{N}\mathbb{E}^{1/2}\left[\Big\|\Gamma_{t,1:H_t}\prod_{s=1}^{t-1}\Gamma_s^{\mathsf{avg}}\cdot(\theta_0-\theta_\star)\Big\|^2\right] \le \prod_{s=1}^{t}(1-\eta_s a)^{H_s}\|\theta_0-\theta_\star\|\,,$$

which is the result of the lemma. □

Bound on heterogeneity. To assess the impact of heterogeneity, we first give a bound on the accumulated heterogeneity between two communications.

Lemma 2. *Assume A2, A1, and let $\eta_s \le \eta_\infty$ for all $s \ge 0$, then*

$$\|\rho_t^{\mathsf{avg}}\| \le \frac{\eta_t^2 H_t^2}{2}\zeta_1\zeta_2\,.$$

Proof. We start by recalling the definition (34) of ρ_t^{avg}, and expanding the difference of matrices using Lemma 22

$$\rho_t^{\mathsf{avg}} = \frac{1}{N}\sum_{c=1}^{N}(\mathrm{I}-G_t^c)(\theta_\star^c-\theta_\star) = \frac{1}{N}\sum_{c=1}^{N}\sum_{h=0}^{H_t-1}\eta_t(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^h\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)\,.$$

where we used the fact that $G_t^c = (\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^{H_t}$ is defined in (36). Moreover, since $\sum_{c=1}^{N}\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)=0$, we can further write

$$\begin{aligned}\rho_t^{\mathsf{avg}} &= \frac{1}{N}\sum_{c=1}^{N}\sum_{h=0}^{H_t-1}\eta_t\Big((\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^h-(\mathrm{I}-\eta_t\bar{\mathbf{A}})^h\Big)\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)\\ &= \frac{\eta_t^2}{N}\sum_{c=1}^{N}\sum_{h=0}^{H_t-1}\sum_{\ell=1}^{h-1}(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^{\ell-1}(\bar{\mathbf{A}}^c-\bar{\mathbf{A}})(\mathrm{I}-\eta_t\bar{\mathbf{A}})^{h-\ell-1}\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)\,.\end{aligned}$$

Taking the norm of this equality, using the triangle inequality, bounding matrices using $\|\mathrm{I}-\eta_t\bar{\mathbf{A}}^c\| \leq 1$, and computing the sum, we obtain

$$\begin{aligned}\|\rho_t^{\mathsf{avg}}\| &\leq \frac{\eta_t^2 H_t(H_t-1)}{2N}\sum_{c=1}^N \|\bar{\mathbf{A}}^c-\bar{\mathbf{A}}\|\|\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)\| \\ &\leq \frac{\eta_t^2 H_t(H_t-1)}{2}\Big(\frac{1}{N}\sum_{c=1}^N\|\bar{\mathbf{A}}^c-\bar{\mathbf{A}}\|^2\Big)^{1/2}\Big(\frac{1}{N}\sum_{c=1}^N\|\bar{\mathbf{A}}^c(\theta_\star^c-\theta_\star)\|^2\Big)^{1/2},\end{aligned}$$

where we used the Cauchy-Schwarz inequality in the second inequality. The result follows from the definition of heterogeneity (29) □

And we can track the propagation of this bias through successive communications.

Bound on the bias. Next, we provide the bound on $\tilde{\theta}_t^{(\mathsf{bi},\mathsf{bi})}$ second moment bounds.

Lemma 3. *Under assumptions **A**1(2), **A**2 and **S**2 it holds that:*

$$\|\tilde{\theta}_t^{(\mathsf{bi},\mathsf{bi})}\| \leq \zeta_1\zeta_2\sum_{s=1}^t \eta_s^2 H_s^2 \exp\left(-a\sum_{i=s+1}^t \eta_i H_i\right).$$

Proof. By the definition of $\tilde{\theta}_t^{(\mathsf{bi},\,\mathsf{bi})}$ in (9)

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathsf{bi},\,\mathsf{bi})}\|^2] = \Big\|\sum_{s=1}^t \mathbb{E}\Big[\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\Big]\rho_s^{\mathsf{avg}}\Big\| = \Big\|\sum_{s=1}^t \mathbb{E}\Big[\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\,\rho_s^{\mathsf{avg}}\Big]\Big\|.$$

By applying **A**1 recursively along with Lemma 2, we have

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathsf{bi},\,\mathsf{bi})}\|^2] \leq \sum_{s=1}^t\prod_{i=s+1}^t(1-\eta_i a)^{H_i}\,\frac{\eta_s^2H_s^2}{2}\zeta_1\zeta_2 \leq \zeta_1\zeta_2\sum_{s=1}^t\eta_s^2H_s^2\exp\left(-a\sum_{i=s+1}^t\eta_iH_i\right),$$

which is the result of the lemma. □

Bound on fluctuations. We then bound the fluctuations of the algorithm, which decompose in a term $\tilde{\theta}_t^{(\mathsf{fl})}$, corresponding to the fluctuations of the updates, and a term $\tilde{\theta}_t^{(\mathsf{fl},\,\mathsf{bi})}$ which quantifies the fluctuations of the bias. We start with $\tilde{\theta}_t^{(\mathsf{fl})}$, showing that the algorithm's variance scales in $1/N$.

Lemma 4. *Assume **A**1(2), **A**2 and let $\eta_s \leq \eta_{\infty,2}$ for all $s \geq 0$. Then, it holds*

$$\mathbb{E}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^2] \leq \sum_{s=1}^t \exp(-2aS_{s+1:t})\cdot(H_s\eta_s^2\wedge\frac{\eta_s}{a})\cdot\frac{\bar{\sigma}_\varepsilon^2}{N}.$$

Proof. Defining $X_s = \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}$ for $s\geq 0$, we can expand (11) as

$$\mathbb{E}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^2] = \mathbb{E}\Big[\Big\|\sum_{s=1}^t\prod_{i=s+1}^t\Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\Big\|^2\Big] = \mathbb{E}\Big[\Big\|\sum_{s=1}^t X_s\Big\|^2\Big] = \mathbb{E}\Big[\sum_{s,r=1}^t X_s^\top X_r\Big]. \tag{44}$$

Assuming $r > s$ we get the following:

$$\begin{aligned}\mathbb{E}\Big[X_s^\top X_r\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[X_s^\top X_r\Big|\mathcal{F}_{s+1,1}^+\Big]\Big] = \mathbb{E}\left[\mathbb{E}\Big[(\varphi_r^{\mathsf{avg}})^\top\cdot\prod_{i=r+1}^t\Gamma_i^{\mathsf{avg}}\times\prod_{j=s+1}^t\Gamma_j^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\Big|\mathcal{F}_{s+1,1}^+\Big]\right] \\ &= \mathbb{E}\left[(\varphi_r^{\mathsf{avg}})^\top\cdot\prod_{i=r+1}^t\Gamma_i^{\mathsf{avg}}\times\prod_{j=s+1}^t\Gamma_j^{\mathsf{avg}}\cdot\underbrace{\mathbb{E}\Big[\varphi_s^{\mathsf{avg}}\Big|\mathcal{F}_{s+1,1}^+\Big]}_{0}\right] = 0.\end{aligned}$$

Plugging this in (44) gives

$$\mathbb{E}[\|\tilde{\theta}_t^{(\mathrm{fl})}\|^2] = \sum_{s=1}^{t} \mathbb{E}\Big[\Big\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \cdot \varphi_s^{\mathsf{avg}} \Big\|^2\Big] .$$

We now bound each term of this sum using **A**1 conditionally, which gives

$$\begin{aligned}
\mathbb{E}^{1/2}\Big[\Big\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \cdot \varphi_s^{\mathsf{avg}} \Big\|^2\Big] &\le \frac{1}{N}\sum_{c=1}^{N} \mathbb{E}^{1/2}\Big[\Big\| \Gamma_{t,1:H_t}^{c} \prod_{i=s+1}^{t-1} \Gamma_i^{\mathsf{avg}} \cdot \varphi_s^{\mathsf{avg}} \Big\|^2\Big] \\
&\le (1-\eta_t a)^{H_t} \mathbb{E}^{1/2}\Big[\Big\| \prod_{i=s+1}^{t-1} \bar{\Gamma}_i \cdot \varphi_s^{\mathsf{avg}} \Big\|^2\Big] \\
&\le \prod_{j=s+1}^{t} (1-\eta_j a)^{H_j} \mathbb{E}^{1/2}[\|\varphi_s^{\mathsf{avg}}\|^2] . \qquad (45)
\end{aligned}$$

Finally, we bound $\mathbb{E}^{1/2}[\|\varphi_s^{\mathsf{avg}}\|^2]$. As $\mathbb{E}[\Gamma_{s,h+1:H_s}^{c}\varepsilon^c(Z_{s,h}^c)|\mathcal{F}_{s,h+1}^{+}] = 0$,

$$\begin{aligned}
\mathbb{E}[\|\varphi_s^{\mathsf{avg}}\|^2] &= \frac{1}{N^2}\sum_{c=1}^{N}\sum_{h=1}^{H_s} \mathbb{E}[\|\eta_s \Gamma_{s,h+1:H_s}^{c}\varepsilon^c(Z_{s,h}^c)\|^2] \\
&\le \frac{1}{N^2}\sum_{c=1}^{N}\sum_{h=1}^{H_s} \eta_s^2 (1-\eta_s a)^{2(H_s-h)} \mathbb{E}[\|\varepsilon^c(Z_{s,h}^c)\|^2] \\
&= \frac{\eta_s^2}{N^2}\sum_{c=1}^{N} \mathbb{E}[\|\varepsilon^c(Z_{s,h}^c)\|^2] \cdot \sum_{h=1}^{H_s} (1-\eta_s)^{2(H_s-h)} .
\end{aligned}$$

Since $\sum_{h=0}^{H_s-1}(1-\eta_s a)^{2h} \le H_s \wedge \frac{1}{\eta_s a}$ and $\frac{1}{N}\sum_{c=1}^{N} \mathbb{E}[\|\varepsilon^c(Z_{s,h}^c)\|^2] = \mathbb{E}_c[\mathrm{Tr}(\Sigma_\varepsilon^c)] = \bar{\sigma}_\varepsilon^2$, we have

$$\mathbb{E}[\|\varphi_s^{\mathsf{avg}}\|^2] \le \frac{\bar{\sigma}_\varepsilon^2}{N} \cdot (H_s\eta_s^2 \wedge \frac{\eta_s}{a}) . \qquad (46)$$

Plugging (46) in (45) and using $\prod_{j=s+1}^{t}(1-\eta_j a)^{H_j} \le \exp(-aS_{s+1:t})$ gives the result. □

Fluctuations of the bias . We now bound $\tilde{\theta}_t^{(\mathrm{fl,\,bi})}$, the fluctuations of the bias. We show that it also scales in $1/N$.

Lemma 5. *Assume* **A**1(2), **A**2 *and let $\eta_s \le \eta_{\infty,2}$, for all $s \ge 0$, be a decreasing sequence. Then, it holds*

$$\mathbb{E}^{1/2}[\|\tilde{\theta}_t^{(\mathrm{fl,bi})}\|^2] \le \sqrt{\frac{4\bar{\sigma}_{het}^2}{N}} \sqrt{\sum_{s=1}^{t} \eta_s^2 H_s \exp(-2a\sum_{r=s}^{t}\eta_r H_r)} + \sqrt{\frac{4\bar{\sigma}_{\mathbf{A}}^2 \zeta_1^2 \zeta_2^2}{Nae}} \sqrt{\sum_{s=1}^{t} \eta_s^5 H_s^4 \exp\big(-a\sum_{r=s+1}^{t}\eta_r H_r\big)} .$$

Proof. Starting from (10), we have

$$\tilde{\theta}_t^{(\mathrm{fl,\,bi})} = \underbrace{\sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\tau_s^{\mathsf{avg}}}_{\mathbf{T_1}} + \underbrace{\sum_{s=1}^{t}\Big\{\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}} - \mathbb{E}\Big[\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\Big]\Big\}\rho_s^{\mathsf{avg}}}_{\mathbf{T_2}} . \qquad (47)$$

Bound on $\mathbf{T_1}$. We start by deriving a bound on $\tau_s^{\mathsf{avg}} = \frac{1}{N}\sum_{c=1}^{N}(G_s^c - \Gamma_s^c)(\theta_\star^c - \theta_\star)$,

$$\mathbb{E}[\|\tau_s^{\mathsf{avg}}\|^2] = \frac{1}{N^2}\sum_{c=1}^{N} \underbrace{\mathbb{E}[\|(G_s^c - \Gamma_s^c)(\theta_\star^c - \theta_\star)\|^2]}_{\mathbf{U_c}} . \qquad (48)$$

Using Lemma 22 to rewrite the matrix product difference, we have

$$\begin{aligned}\mathbf{U_c} &= \mathbb{E}\Big[\Big\|\Big(\prod_{h=1}^{H_t}(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)-\prod_{h=1}^{H_t}(\mathrm{I}-\eta_t\mathbf{A}^c(Z^c_{t,h}))\Big)(\theta^c_\star-\theta_\star)\Big\|^2\Big]\\ &= \eta_t^2\mathbb{E}\Big[\Big\|\sum_{h=1}^{H_t}(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^{h-1}(\bar{\mathbf{A}}^c-\mathbf{A}^c(Z^c_{t,h}))\Gamma^c_{t,h+1:H_t}(\theta^c_\star-\theta_\star)\Big\|^2\Big]\,.\end{aligned}$$

Noticing that $\mathbb{E}[(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^{h-1}(\mathbf{A}^c(Z^c_{t,h})-\bar{\mathbf{A}}^c)\Gamma^c_{t,h+1:H_t}|\mathcal{F}^+_{t,h+1}]=0$, we write

$$\begin{aligned}\mathbf{U_c} &= \eta_t^2\sum_{h=1}^{H_t}\mathbb{E}[\|(\mathrm{I}-\eta_t\bar{\mathbf{A}}^c)^{h-1}\widetilde{\mathbf{A}}^c(Z^c_{t,h})\Gamma^c_{t,h+1:H_t}(\theta^c_\star-\theta_\star)\|^2]\\ &\le \eta_t^2\sum_{h=1}^{H_t}(1-\eta_t a)^{2(h-1)}\mathbb{E}\Big[(\theta^c_\star-\theta_\star)^\top\Gamma^{c\top}_{t,h+1:H_t}\tilde{A}^c(Z^c_{t,h})^\top\tilde{A}^c(Z^c_{t,h})\Gamma^c_{t,h+1:H_t}(\theta^c_\star-\theta_\star)\Big]\,.\end{aligned}$$

As $v:=\Gamma^c_{t,h+1:H_t}(\theta^c_\star-\theta_\star)$ is $\mathcal{F}^+_{t,h+1}$-measurable, the law of total expectation gives

$$\mathbb{E}[v^\top\tilde{A}^c(Z^c_{t,h})^\top\tilde{A}^c(Z^c_{t,h})v]=\mathbb{E}\Big[v^\top\cdot\mathbb{E}[\tilde{A}^c(Z^c_{t,h})^\top\tilde{A}^c(Z^c_{t,h})|\mathcal{F}^+_{t,h+1}]\cdot v\Big]=\mathbb{E}\Big[v^\top\Sigma^c_{\mathbf{A}}v\Big]\,.$$

Then we can estimate $\mathbb{E}[v^\top\Sigma^c_{\mathbf{A}}v]$ by using the fact that $\Sigma^c_{\mathbf{A}}$ is symmetric

$$\mathbb{E}[v^\top\Sigma^c_{\mathbf{A}}v]=\mathbb{E}[\|(\Sigma^c_{\tilde{A}})^{1/2}v\|^2]\le\|\Sigma^c_{\mathbf{A}}\|\cdot\mathbb{E}[\|v\|^2]\le(1-\eta_t a)^{2(H_t-h)}\|\Sigma^c_{\mathbf{A}}\|\cdot\|\theta^c_\star-\theta_\star\|^2\,.$$

Hence, we obtain the bound

$$\mathbf{U_c}\le\eta_t^2\sum_{h=1}^{H_t}(1-\eta_t a)^{2(H_t-1)}\|\Sigma^c_{\mathbf{A}}\|\|\theta^c_\star-\theta_\star\|^2=H_t\eta_t^2(1-\eta_t a)^{2(H_t-1)}\|\Sigma^c_{\mathbf{A}}\|\|\theta^c_\star-\theta_\star\|^2\,. \tag{49}$$

Plugging (49) in (48), we obtain the following bound on τ^{avg}_s

$$\mathbb{E}[\|\tau^{\mathsf{avg}}_s\|^2]\le\frac{H_s\eta_s^2(1-\eta_s a)^{2(H_s-1)}}{N^2}\sum_{c=1}^N\|\Sigma^c_{\mathbf{A}}\|\|\theta^c_\star-\theta_\star\|^2=H_s\eta_s^2(1-\eta_s a)^{2(H_s-1)}\frac{\bar{\sigma}^2_{\mathsf{het}}}{N}\,. \tag{50}$$

Here, we recall that $\bar{\sigma}^2_{\mathsf{het}}=\frac{1}{N}\sum_{c=1}^N\|\Sigma^c_{\mathbf{A}}\|\|\theta^c_\star-\theta_\star\|^2$. Now, using the definition of $S_{s+1:t}$, we have the following bound

$$\mathbb{E}[\|\mathbf{T_1}\|^2]=\sum_{s=1}^t\mathbb{E}[\|\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\cdot\tau^{\mathsf{avg}}_s\|^2]\le\sum_{s=1}^t\exp(-2aS_{s+1:t})\mathbb{E}[\|\tau^{\mathsf{avg}}_s\|^2]\,.$$

Using (50) and $\eta_s a\le 1/2$, we get

$$\begin{aligned}\mathbb{E}[\|\mathbf{T_1}\|^2]&\le\sum_{s=1}^t\exp(-2aS_{s+1:t})(1-\eta_s a)^{2(H_s-1)}H_s\eta_s^2\frac{\bar{\sigma}^2_{\mathsf{het}}}{N}\\ &\le 4\sum_{s=1}^t\exp(-2aS_{s:t})H_s\eta_s^2\frac{\bar{\sigma}^2_{\mathsf{het}}}{N}\,.\end{aligned} \tag{51}$$

Bound on $\mathbf{T_2}$. Now we need to bound $\mathbf{T_2}=\sum_{s=1}^t\Delta_{s,t}\rho^{\mathsf{avg}}_s$, where we recall that $\Delta_{s,t}=\{\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\}-\mathbb{E}[\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i]$. Let's begin with a single summand $\Delta_{s,t}\rho^{\mathsf{avg}}_s$, which can be rewritten as

$$\begin{aligned}\Delta_{s,t}\rho^{\mathsf{avg}}_s&=\Big(\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i-\prod_{i=s+1}^t G^{\mathsf{avg}}_i\Big)\rho^{\mathsf{avg}}_s\\ &=\Big(\sum_{i=s+1}^t\Big\{\prod_{r=i+1}^t\Gamma^{\mathsf{avg}}_r\Big\}\cdot\Big\{\Gamma^{\mathsf{avg}}_i-G^{\mathsf{avg}}_i\Big\}\cdot\Big\{\prod_{r=s+1}^{i-1}G^{\mathsf{avg}}_r\Big\}\Big)\rho^{\mathsf{avg}}_s\,.\end{aligned}$$

We first bound the norm of $(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u$ for an arbitrary vector $u \in \mathbb{R}^d$ and $i \geq 0$. By clients' independence, we have

$$\begin{aligned}
\mathbb{E}\big[\|(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u\|^2\big] &= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}(\Gamma_{i,1:H_i}^c - G_{i,1:H_i}^c)u\Big\|^2\Big] \\
&= \frac{1}{N^2}\sum_{c=1}^{n}\mathbb{E}\big[\|(\Gamma_{i,1:H_i}^c - G_{i,1:H_i}^c)u\|^2\big] .
\end{aligned}$$

Then, Lemma 22 gives

$$\mathbb{E}[\|(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u\|^2] = \frac{\eta_i^2}{N^2}\sum_{c=1}^{N}\sum_{h=1}^{H_i}\mathbb{E}[\|(\mathrm{I} - \eta_i\bar{\mathbf{A}}^c)^{h-1}\widetilde{\mathbf{A}}^c(Z_{i,h}^c)\Gamma_{i,h+1:H_i}^c u\|^2] .$$

By applying conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{i+1,1}^+]$, we can rewrite it as:

$$\begin{aligned}
\mathbb{E}[\|(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u\|^2] &= \frac{\eta_i^2}{N^2}\sum_{c=1}^{N}\sum_{h=1}^{H_i}(1-\eta_i a)^{2(h-1)}\mathbb{E}[\|(\Sigma_{\mathbf{A}}^c)^{\frac{1}{2}}\Gamma_{i,h+1:H_i}^c u\|^2] \\
&\leq (1-\eta_i a)^{2(H_i-1)}\frac{\eta_i^2 H_i}{N}\bar{\sigma}_{\mathbf{A}}^2\|u\|^2 .
\end{aligned}$$

Using this knowledge, we can finally estimate $\Delta_{s,t}\rho_s^{\mathsf{avg}}$'s squared norm

$$\begin{aligned}
\mathbb{E}[\|\Delta_{s,t}\rho_s^{\mathsf{avg}}\|^2] &= \mathbb{E}\Big[\Big\|\sum_{i=s+1}^{t}\{\prod_{r=i+1}^{t}\Gamma_r^{\mathsf{avg}}\}\cdot(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})\cdot\{\prod_{r=s+1}^{i-1}G_r^{\mathsf{avg}}\}\rho_s^{\mathsf{avg}}\Big\|^2\Big] \\
&\leq \sum_{i=s+1}^{t}\frac{\eta_i^2 H_i}{N}\bar{\sigma}_{\mathbf{A}}^2\prod_{r=i+1}^{t}(1-\eta_r a)^{2H_r}(1-\eta_i a)^{2H_i-2}\prod_{r=s+1}^{i-1}(1-\eta_r a)^{2H_r}\|\rho_s^{\mathsf{avg}}\|^2 \\
&\leq 4\Big\{\prod_{r=s+1}^{t}(1-\eta_r a)^{2H_r}\Big\}\sum_{i=s+1}^{t}\frac{\eta_i^2 H_i}{N}\bar{\sigma}_{\mathbf{A}}^2\|\rho_s^{\mathsf{avg}}\|^2 .
\end{aligned}$$

where we used $1 - \eta_i a \geq 1/2$. Using the fact that the step size decreases, and the fact that for all $c > 0$ and $x \geq 0$, $x\exp(-2cx) \leq 1/(ec)\exp(-cx)$, we obtain the following bound

$$\begin{aligned}
\mathbb{E}[\|\Delta_{s,t}\rho_s^{\mathsf{avg}}\|^2] &\leq \frac{4\eta_s\bar{\sigma}_{\mathbf{A}}^2}{N}\|\rho_s^{\mathsf{avg}}\|^2\exp\big(-2a\sum_{r=s+1}^{t}\eta_r H_r\big)\sum_{r=s+1}^{t}\eta_r H_r \\
&\leq \frac{4\eta_s\bar{\sigma}_{\mathbf{A}}^2}{Nae}\|\rho_s^{\mathsf{avg}}\|^2\exp\big(-a\sum_{r=s+1}^{t}\eta_r H_r\big) . \qquad (52)
\end{aligned}$$

Using the reverse martingale structure, (52), and Lemma 2, we obtain a bound on $\mathbb{E}[\|T_2\|^2]$

$$\begin{aligned}
\mathbb{E}[\|\mathbf{T_2}\|^2] &\leq \frac{4\bar{\sigma}_{\mathbf{A}}^2}{Nae}\sum_{s=1}^{t}\eta_s\|\rho_s^{\mathsf{avg}}\|^2\exp\big(-a\sum_{r=s+1}^{t}\eta_r H_r\big) \\
&\leq \frac{4\bar{\sigma}_{\mathbf{A}}^2}{Nae}\zeta_1^2\zeta_2^2\sum_{s=1}^{t}\eta_s^5 H_s^4\exp\big(-a\sum_{r=s+1}^{t}\eta_r H_r\big) . \qquad (53)
\end{aligned}$$

The result follows from taking the root of the expected squared norm of (47), decomposing it two using Minkowski's inequality, and bounding each term using (51) and (53). □

Convergence results. Combining our results, we can now give an upper bound on the error and prove Theorem 1.

A.2.1 Proof of the Theorem 1

Proof. Using the decomposition (37) and applying Minkowski's inequality, we obtain

$$\mathbb{E}^{1/2}\big[\|\theta_t-\theta_\star\|^2\big] \le \mathbb{E}^{1/2}\big[\|\tilde{\theta}_t^{(\mathrm{tr})}\|^2\big] + \mathbb{E}^{1/2}\big[\|\tilde{\theta}_t^{(\mathrm{bi,bi})}\|^2\big] + \mathbb{E}^{1/2}\big[\|\tilde{\theta}_t^{(\mathrm{fl,bi})}\|^2\big] + \mathbb{E}^{1/2}\big[\|\tilde{\theta}_t^{(\mathrm{fl})}\|^2\big]$$

The four terms are controlled respectively by Lemma 1, Lemma 3, Lemma 5, and Lemma 4. Combining these bounds yields the claimed result. □

When the step size and number of local updates are constant, we recover a slightly improved variant of the Theorem A.6 of Mangold et al. [2024].

Corollary 3 (Constant step size)**.** *Assume* ***A**1(2),* ***A**2 and* ***S**1. Then, it holds*

$$\mathbb{E}^{1/2}[\|\theta_t-\theta_\star\|^2] \le (1-\eta a)^{Ht}\|\theta_0-\theta_\star\| + \frac{\eta H\zeta_1\zeta_2}{a} + \sqrt{\frac{\eta\bar{\sigma}_\varepsilon^2}{Na}} + \sqrt{\frac{\eta\bar{\sigma}_{het}^2}{Na}} + \sqrt{\frac{4\eta^4H^3\bar{\sigma}_{\mathbf{A}}^2\zeta_1^2\zeta_2^2}{Na^2e}}\,,$$

Proof. By Theorem 1 with fixed step size and number of local updates, we have

$$\begin{aligned}
&\mathbb{E}^{1/2}[\|\theta_t-\theta_\star\|^2]\\
&\le (1-\eta a)^{Ht}\|\theta_0-\theta_\star\| + \zeta_1\zeta_2\sum_{s=1}^t \eta^2H^2\exp\Big(-a\eta H(t-s)\Big) + \sqrt{\frac{\eta^2H\bar{\sigma}_\varepsilon^2}{N}}\sqrt{\sum_{s=1}^t\exp(-2a\eta Ht)}\\
&\quad + \sqrt{\frac{4\bar{\sigma}_{\text{het}}^2}{N}}\sqrt{\sum_{s=1}^t \eta^2H\exp(-2a\eta^2H(t-s))} + \sqrt{\frac{4\bar{\sigma}_{\mathbf{A}}^2\zeta_1^2\zeta_2^2}{Nae}}\sqrt{\sum_{s=1}^t\eta^5H^4\exp\big(-a\eta H(t-s)\big)}\,.
\end{aligned}$$

Bounding the sums, for $u>0$, of $\sum_{s=1}^t\exp(-u(t-s)) \le (1-\exp(-u))^{-1} \le 1/u$ gives the result. □

Finally, with decreasing step sizes and number of local updates, we obtain

Corollary 4. *Assume* ***A**1(2),* ***A**2 and* ***S**2, meaning that the step sizes and local updates satisfy, $\eta_t = \frac{\eta}{(t+1)^{\gamma_\eta}}$, $H_t = \lceil H(t+1)^{\gamma_H}\rceil$, for some $\eta, H>0$ and $0<\gamma_\eta<1$ and $0<\gamma_H<1$, such that $\gamma_\eta \ge \gamma_H$ and $\gamma=\gamma_\eta-\gamma_H \ne 1$. Then, for all $t\ge 1$, the following bound holds*

$$\begin{aligned}
\mathbb{E}^{1/2}\big[\|\theta_t-\theta_\star\|^2\big] &\le \exp\Big(-\frac{a\eta H}{1-\gamma}(t+1)^{1-\gamma}\Big)\|\theta_0-\theta_\star\| + \frac{8\zeta_1\zeta_2\eta H}{a}\frac{1}{(t+1)^\gamma} + \sqrt{\frac{8\eta\bar{\sigma}_\varepsilon^2}{Na}}\frac{1}{(t+1)^{\gamma_\eta/2}}\\
&\quad + \sqrt{\frac{32\eta\bar{\sigma}_{het}^2}{Na}}\frac{1}{(t+1)^{\gamma_\eta/2}} + \sqrt{\frac{32\eta^4H^3\bar{\sigma}_{\mathbf{A}}^2\zeta_1^2\zeta_2^2}{Nae}}\frac{1}{(t+1)^{(4\gamma_\eta-3\gamma_H)/2}}\,.
\end{aligned}$$

Proof. We apply Lemma 25 to the result of Theorem 1, thus obtain the bound. □

A.3 PROOFS FOR HIGH-ORDER MOMENT BOUNDS $\mathbb{E}^{1/p}[\|\theta_t-\theta_\star\|^p]$.

To prove Theorem 2, we establish moment bounds for each component in the decomposition $\tilde{\theta}_t^{(\mathrm{tr})}$, $\tilde{\theta}_t^{(\mathrm{fl})}$, $\tilde{\theta}_t^{(\mathrm{fl,bi})}$, and $\tilde{\theta}_t^{(\mathrm{bi,bi})}$.

Bound on the transient term. We begin with a bound on the p-th moment of the transient component.

Lemma 6. *Assume* ***A**1(p),* ***A**2, and let $\eta_s \le \eta_{\infty,p}$. Then, it holds*

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{tr})}\|^p] \le \exp\Big(-a\sum_{s=1}^t\eta_sH_s\Big)\|\theta_0-\theta_\star\|\,.$$

Proof. We start with bounding the transient term $\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^p]$. It holds that

$$\mathbb{E}^{1/p}\Big[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^p\Big] = \mathbb{E}^{1/p}\Big[\Big\|\prod_{s=1}^{t}\Gamma_s^{\mathsf{avg}}\{\theta_0-\theta_\star\}\Big\|^p\Big] = \mathbb{E}^{1/p}\Big[\Big\|\Big(\frac{1}{N}\sum_{c=1}^{N}\Gamma_t^c\Big)\prod_{s=1}^{t-1}\Gamma_s^{\mathsf{avg}}\{\theta_0-\theta_\star\}\Big\|^p\Big]\,.$$

Applying Minkowski's inequality, we have

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^p] \le \frac{1}{N}\sum_{c=1}^{N}\mathbb{E}^{1/p}\Big[\Big\|\Gamma_t^c\cdot\prod_{s=1}^{t-1}\Gamma_s^{\mathsf{avg}}\{\theta_0-\theta_\star\}\Big\|^p\Big]\,.$$

By using **A**1 conditionally, we get

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^p] \le (1-\eta_t a)^{H_s}\mathbb{E}^{1/p}\Big[\Big\|\prod_{s=1}^{t-1}\Gamma_s^{\mathsf{avg}}\{\theta_0-\theta_\star\Big\|^p\Big] \le \prod_{s=1}^{t}(1-\eta_s a)^{H_s}\|\theta_0-\theta_\star\|\,,$$

which gives the following bound on the transient term

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{tr})}\|^p] \le \exp\Big(-a\sum_{s=1}^{t}\eta_s H_s\Big)\|\theta_0-\theta_\star\|\,.$$

□

Bound on the fluctuation term.

Lemma 7. *Assume **A**1(p), **A**2, and let $\eta_s \le \eta_{\infty,p}$. Then, it holds*

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^p] \le p\left(\sum_{s=1}^{t}\exp\Big(-2a\sum_{i=s+1}^{t}\eta_i H_i\Big)N^{-1}p^2\|\varepsilon\|_\infty^2(\eta_s^2 H_s\wedge\frac{\eta_s}{a})\right)^{1/2}\,.$$

Proof. We now bound the fluctuation term $\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^p]$. By definition of $\tilde{\theta}_t^{(\mathsf{fl})}$, we have

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^p] = \mathbb{E}^{1/p}\Big[\Big\|\sum_{s=1}^{t}\Big\{\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\Big\}\Big\|^p\Big]\,.$$

Notice that the φ_t^{avg} are of zero expectation, and independent of each other. Thus, $\{\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\}$ forms a martingale difference w.r.t the filtration $\mathcal{F}_{i+1,1}^{+}$. Using Burkholder's inequality (Osekowski, 2012, Theorem 8.6), we then obtain the following bound

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^p] \le p\left(\mathbb{E}^{2/p}\Big[\Big(\sum_{s=1}^{t}\Big\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\Big\|^2\Big)^{p/2}\Big]\right)^{1/2}$$

Using Minkowski's inequality along with **A**1, we have

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathsf{fl})}\|^p] \le p\left(\sum_{s=1}^{t}\mathbb{E}^{2/p}\Big[\Big\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\cdot\varphi_s^{\mathsf{avg}}\Big\|^p\Big]\right)^{1/2} \le p\left(\sum_{s=1}^{t}\exp\Big(-2a\sum_{i=s+1}^{t}\eta_i H_i\Big)\mathbb{E}^{2/p}[\|\varphi_s^{\mathsf{avg}}\|^p]\right)^{1/2}\,.$$

Now we have to estimate $\mathbb{E}^{1/p}[\|\varphi_s^{\mathsf{avg}}\|^p]$. Recall that $\varphi_s^{\mathsf{avg}} = \frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H_s}\eta_s\Gamma_{s,h+1:H_s}^c\varepsilon^c(Z_{s,h}^c)$. Therefore, applying Burkholder's inequality and Minkowski's inequality again gives

$$\begin{aligned}\mathbb{E}^{1/p}[\|\varphi_s^{\mathsf{avg}}\|^p] &\le \frac{p}{N}\left(\mathbb{E}^{2/p}\Big[\Big(\sum_{c=1}^{N}\sum_{h=1}^{H_s}\eta_s^2\|\Gamma_{s,h+1:H_s}^c\varepsilon^c(Z_{s,h}^c)\|^2\Big)^{p/2}\Big]\right)^{1/2}\\ &\le \frac{p}{N}\left(\sum_{c=1}^{N}\sum_{h=1}^{H_s}\eta_s^2\mathbb{E}^{2/p}[\|\Gamma_{s,h+1:H_s}\varepsilon^c(Z_{s,h}^c)\|^p]\right)^{1/2}\,.\end{aligned}$$

Using $\mathbf{A}1$ and the definition of $\|\varepsilon\|_\infty$

$$\mathbb{E}^{1/p}[\|\varphi_s^{\mathsf{avg}}\|^p] \le \frac{p\eta_s}{N}\left(\sum_{c=1}^N\sum_{h=1}^{H_s}(1-\eta_s)^{2(H_s-h)}\|\varepsilon\|_\infty^2\right)^{1/2} \le pN^{-1/2}\eta_s\|\varepsilon\|_\infty\sqrt{(H_s \wedge \frac{1}{\eta_s a})}\,,$$

where $\sum_{h=1}^{H_s}(1-\eta_s)^{2(H_s-h)}$ is bounded by $H_s \wedge \frac{1}{\eta_s a}$ exactly as in Lemma 4. Finally, we bound

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{fl})}\|^p] \le p\left(\sum_{s=1}^t \exp\Big(-2a\sum_{i=s+1}^t \eta_i H_i\Big) N^{-1}p^2\|\varepsilon\|_\infty^2(\eta_s^2 H_s \wedge \frac{\eta_s}{a})\right)^{1/2}.$$

□

Bound on the bias. The bias term, $\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{bi,\,bi})}\|^p]$, is deterministic, it's p-th moment coincides with $\|\tilde{\theta}_t^{(\mathrm{bi,\,bi})}\|$, and it follows from Lemma 3, which states

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{bi,\,bi})}\|^p] = \|\tilde{\theta}_t^{(\mathrm{bi,\,bi})}\| \le \zeta_1\zeta_2\sum_{s=1}^t \eta_s^2 H_s^2 \exp\Big(-a\sum_{i=s+1}^t \eta_i H_i\Big)\,.$$

Bound on the fluctuations of the bias.

Lemma 8. *Assume $\mathbf{A}1(p)$, $\mathbf{A}2$, and let $\eta_s \le \eta_{\infty,p}$. Then, it holds*

$$\mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{fl,bi})}\|^p] \le 2\,\mathrm{C}_{\mathbf{A}}\,\zeta_3 p^2 N^{-1/2}\left(\sum_{s=1}^t \eta_s^2 H_s \exp\Big(-2a\sum_{i=s}^t \eta_i H_i\Big)\right)^{1/2}$$
$$+\frac{\mathrm{C}_{\mathbf{A}}\,p^3\zeta_1\zeta_2}{\sqrt{Na\mathrm{e}}}\sqrt{\sum_{s=1}^t \eta_s^5 H_s^4 \exp\Big(-a\sum_{i=s+1}^t \eta_i H_i\Big)}\,.$$

Proof. Recall the definition of $\tilde{\theta}_t^{(\mathrm{fl,\,bi})}$ from (10):

$$\tilde{\theta}_t^{(\mathrm{fl,\,bi})} = \underbrace{\sum_{s=1}^t\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\tau_s^{\mathsf{avg}}}_{\mathbf{T_1}} + \underbrace{\sum_{s=1}^t\Big\{\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \mathbb{E}\Big[\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}\Big]\Big\}\rho_s^{\mathsf{avg}}}_{\mathbf{T_2}}\,.$$

Bound on $\mathbf{T_1}$. We start by deriving a $p-$th moment of $\tau_s^{\mathsf{avg}} = \frac{1}{N}\sum_{c=1}^N(G_s^c - \Gamma_s^c)(\theta_\star^c - \theta_\star)$. Since the clients have independent noise, we can apply Burkholder's inequality to bound

$$\mathbb{E}^{1/p}[\|\tau_s^{\mathsf{avg}}\|^p] \le \frac{1}{N}\Big(\sum_{c=1}^N \underbrace{\mathbb{E}^{2/p}[\|(G_s^c-\Gamma_s^c)(\theta_\star^c-\theta_\star)\|^p]}_{\mathbf{U^2_{c,p}}}\Big)^{1/2}.$$

Now we use Lemma 22 to estimate $\mathbf{U_{c,p}}$:

$$\mathbf{U_{c,p}} = \mathbb{E}^{1/p}\Big[\Big\|\Big(\prod_{h=1}^{H_s}(\mathrm{I}-\eta_s\bar{\mathbf{A}}^c) - \prod_{h=1}^{H_s}(\mathrm{I}-\eta_s\mathbf{A}^c(Z_{s,h}^c))\Big)(\theta_\star^c-\theta_\star)\Big\|^p\Big]$$
$$= \eta_s\mathbb{E}^{1/p}\Big[\Big\|\sum_{h=1}^{H_s}(\mathrm{I}-\eta_s\bar{\mathbf{A}}^c)^{h-1}(\bar{\mathbf{A}}^c - \mathbf{A}^c(Z_{s,h}^c))\Gamma_{s,h+1:H_s}^c(\theta_\star^c-\theta_\star)\Big\|^p\Big]\,.$$

As $\mathbb{E}[(\mathrm{I}-\eta_s\bar{\mathbf{A}}^c)^{h-1}(\mathbf{A}^c(Z^c_{s,h})-\bar{\mathbf{A}}^c)\Gamma^c_{s,h+1:H_s}|\mathcal{F}^+_{s,h+1}]=0$, we can apply Burkholder's inequality here

$$
\begin{aligned}
\mathbf{U_{c,p}} &\leq p\eta_s\left(\sum_{h=1}^{H_s}\mathbb{E}^{2/p}\Big[\Big\|(\mathrm{I}-\eta_s\bar{\mathbf{A}}^c)^{h-1}(\bar{\mathbf{A}}^c-\mathbf{A}^c(Z^c_{s,h}))\Gamma^c_{s,h+1:H_s}(\theta^c_\star-\theta_\star)\Big\|^p\Big]\right)^{1/2} \\
&\leq p\eta_s\left(\sum_{h=1}^{H_s}(1-\eta_s a)^{2(h-1)}\,\mathrm{C}^2_{\mathbf{A}}(1-\eta_s a)^{2(H_s-h)}\|\theta^c_\star-\theta_\star\|^2\right)^{1/2} \\
&= p\eta_s\,\mathrm{C}_{\mathbf{A}}\left(\sum_{h=1}^{H_s}(1-\eta_s a)^{2(H_s-1)}\right)^{1/2}\|\theta^c_\star-\theta_\star\| = p\eta_s\,\mathrm{C}_{\mathbf{A}}(1-\eta_s a)^{H_s-1}H_s^{1/2}\|\theta^c_\star-\theta_\star\|
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}^{1/p}[\|\tau^{\mathsf{avg}}_s\|^p] &\leq \mathrm{C}_{\mathbf{A}}\,pN^{-1/2}\eta_s H_s^{1/2}(1-\eta_s a)^{H_s-1}\frac{\sum_{c=1}^N\|\theta^c_\star-\theta_\star\|^2}{N} \\
&= \mathrm{C}_{\mathbf{A}}\,\zeta_3 pN^{-1/2}\eta_s H_s^{1/2}(1-\eta_s a)^{H_s-1}\,.
\end{aligned}
$$

Now we are ready to estimate $\mathbf{T_1}$. We use here that $\{\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\tau_s\}_{s=1}^t$ forms a martingale difference, therefore we apply Burkholder's inequality

$$
\begin{aligned}
\mathbb{E}^{1/p}[\|\mathbf{T_1}\|^p] &= \mathbb{E}^{1/p}\Big[\Big\|\sum_{s=1}^t\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\,\tau^{\mathsf{avg}}_s\Big\|^p\Big] \leq p\left(\sum_{s=1}^t\mathbb{E}^{2/p}\Big[\Big\|\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\tau^{\mathsf{avg}}_s\Big\|^p\Big]\right)^{1/2} \\
&\leq p\left(\sum_{s=1}^t\prod_{i=s+1}^t(1-\eta_i a)^{2H_i}\,\mathbb{E}^{2/p}[\|\tau^{\mathsf{avg}}_s\|^p]\right)^{1/2} = 2\,\mathrm{C}_{\mathbf{A}}\,\zeta_3 p^2N^{-1/2}\left(\sum_{s=1}^t\prod_{i=s}^t(1-\eta_i a)^{2H_i}\,\eta_s^2H_s\right)^{1/2}
\end{aligned}
$$

We can rewrite this estimate as

$$
\mathbb{E}^{1/p}[\|\mathbf{T_1}\|^p] \leq 2\,\mathrm{C}_{\mathbf{A}}\,\zeta_3 p^2N^{-1/2}\left(\sum_{s=1}^t\eta_s^2H_s\exp\Big(-2a\sum_{i=s}^t\eta_iH_i\Big)\right)^{1/2}
$$

Bound on $\mathbf{T_2}$. Now we estimate $p-$th moment of $\mathbf{T_2}=\sum_{s=1}^t\Delta_{s,t}\rho^{\mathsf{avg}}_s$, where we recall that $\Delta_{s,t}=\{\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i\}-\mathbb{E}[\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i]$. We start from single summand $\Delta_{s,t}\rho^{\mathsf{avg}}_s$

$$
\Delta_{s,t}\rho^{\mathsf{avg}}_s = \Big(\prod_{i=s+1}^t\Gamma^{\mathsf{avg}}_i-\prod_{i=s+1}^tG^{\mathsf{avg}}_i\Big)\rho^{\mathsf{avg}}_s = \Big(\sum_{i=s+1}^t\Big\{\prod_{r=i+1}^t\Gamma^{\mathsf{avg}}_r\Big\}\cdot\Big\{\Gamma^{\mathsf{avg}}_i-G^{\mathsf{avg}}_i\Big\}\cdot\Big\{\prod_{r=s+1}^{i-1}G^{\mathsf{avg}}_r\Big\}\Big)\rho^{\mathsf{avg}}_s\,.
$$

We first bound moment of $(\Gamma^{\mathsf{avg}}_i-G^{\mathsf{avg}}_i)u$ for an arbitrary vector $u\in\mathbb{R}^d$ and $i\geq 0$. By applying Burkholder's inequality

$$
\begin{aligned}
\mathbb{E}\big[\|(\Gamma^{\mathsf{avg}}_i-G^{\mathsf{avg}}_i)u\|^p\big] &= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^N(\Gamma^c_{i,1:H_i}-G^c_{i,1:H_i})u\Big\|^p\Big] \\
&\leq \Big(\frac{p}{N}\Big)^p\left(\mathbb{E}^{2/p}\Big[\Big(\sum_{c=1}^N\|(\Gamma^c_{i,1:H_i}-G^c_{i,1:H_i})u\|^2\Big)^{p/2}\Big]\right)^{p/2} \\
&\leq \Big(\frac{p}{N}\Big)^p\left(\sum_{c=1}^N\mathbb{E}^{2/p}[\|(\Gamma^c_{i,1:H_i}-G^c_{i,1:H_i})u\|^p]\right)^{p/2}\,.
\end{aligned}
$$

Using Lemma 22, we get

$$\begin{aligned}
\mathbb{E}\big[\|(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u\|^p\big] &\le \Big(\frac{p\eta_i}{N}\Big)^p \Bigg(\sum_{c=1}^{N} \mathbb{E}^{2/p}\Big[\Big\|\sum_{h=1}^{H_i}(\mathrm{I}-\eta_i\bar{\mathbf{A}}^c)^{h-1}\widetilde{\mathbf{A}}^c(Z_{i,h}^c)\Gamma_{i,h+1:H_i}^c u\Big\|^p\Big]\Bigg)^{p/2} \\
&\le \Big(\frac{p\eta_i}{N}\Big)^p \Bigg(\sum_{c=1}^{N}\sum_{h=1}^{H_i} \mathbb{E}^{2/p}\Big[\Big\|(I-\eta_i\bar{\mathbf{A}}^c)^{h-1}\widetilde{\mathbf{A}}^c(Z_{i,h}^c)\Gamma_{i,h+1:H_i}^c u\Big\|^p\Big]\Bigg)^{p/2} \\
&\le \Big(\frac{p\eta_i}{N}\Big)^p \Big(NH_i(1-\eta_i a)^{2(h-1)}\,\mathrm{C}_{\mathbf{A}}^2(1-\eta_i a)^{2(H_i-h)}\Big)^{p/2} ,
\end{aligned}$$

or, in other words,

$$\mathbb{E}^{1/p}[\|(\Gamma_i^{\mathsf{avg}} - G_i^{\mathsf{avg}})u\|^p] \le \frac{2\,\mathrm{C}_{\mathbf{A}}\, p\eta_i\sqrt{H_i}(1-\eta_i a)^{H_i}\|u\|}{\sqrt{N}} .$$

Now we can bound $\Delta_{s,t}\rho_s^{\mathsf{avg}}$

$$\begin{aligned}
\mathbb{E}^{1/p}[\|\Delta_{s,t}\rho_s^{\mathsf{avg}}\|^p] &= \mathbb{E}^{1/p}\Big[\Big\|\sum_{i=s+1}^{t}\{\prod_{r=i+1}^{t}\Gamma_r^{\mathsf{avg}}\}\cdot(\Gamma_i^{\mathsf{avg}}-G_i^{\mathsf{avg}})\cdot\{\prod_{r=s+1}^{i-1}G_r^{\mathsf{avg}}\}\rho_s^{\mathsf{avg}}\Big\|^p\Big] \\
&\le p\Bigg(\sum_{i=s+1}^{t}\mathbb{E}^{2/p}\Big[\Big\|\{\prod_{r=i+1}^{t}\Gamma_r^{\mathsf{avg}}\}\cdot(\Gamma_i^{\mathsf{avg}}-G_i^{\mathsf{avg}})\cdot\{\prod_{r=s+1}^{i-1}G_r^{\mathsf{avg}}\}\rho_s^{\mathsf{avg}}\Big\|^p\Big]\Bigg)^{1/2} \\
&\le p\Bigg(\sum_{i=s+1}^{t}\prod_{r=s+1,r\neq i}^{t}(1-\eta_r a)^{2H_r}\frac{4\,\mathrm{C}_{\mathbf{A}}^2\,p^2\eta_i^2H_i(1-\eta_i a)^{2H_i}\|\rho_s^{\mathsf{avg}}\|^2}{N}\Bigg)^{1/2} \\
&\le \frac{2\,\mathrm{C}_{\mathbf{A}}\,p^2\|\rho_s^{\mathsf{avg}}\|}{\sqrt{N}}\Bigg(\prod_{r=s+1}^{t}(1-\eta_r a)^{2H_r}\Bigg)^{1/2}\Bigg(\sum_{i=s+1}^{t}\eta_i^2H_i\Bigg)^{1/2}
\end{aligned}$$

As $\eta_i \ge \eta_s$ for all $i > s$, we can rewrite the bound as

$$\mathbb{E}^{1/p}[\|\Delta_{s,t}\rho^{\mathsf{avg}}\|^p] \le \frac{2\,\mathrm{C}_{\mathbf{A}}\,p^2\|\rho_s^{\mathsf{avg}}\|}{\sqrt{N}}\eta_s^{1/2}\exp\Big(-a\sum_{i=s+1}^{t}\eta_iH_i\Big)\sqrt{\sum_{i=s+1}^{t}\eta_iH_i}$$

Now, finally, we can bound $\mathbf{T_2}$. Remark that $\mathbb{E}^{1/p}[\|\mathbf{T_2}\|^p] \le p\left(\sum_{s=1}^{t}\mathbb{E}^{2/p}[\|\Delta_{s,t}\rho_s^{\mathsf{avg}}\|^p]\right)^{1/2}$, and thus

$$\begin{aligned}
\mathbb{E}^{1/p}[\|\mathbf{T_2}\|^p] &\le p\Bigg(\sum_{s=1}^{t}\frac{4\,\mathrm{C}_{\mathbf{A}}^2\,p^4\|\rho_s^{\mathsf{avg}}\|^2}{N}\eta_s\exp\Big(-2a\sum_{i=s+1}^{t}\eta_iH_i\Big)\sum_{i=s+1}^{t}\eta_iH_i\Bigg)^{1/2} \\
&\le \frac{2\,\mathrm{C}_{\mathbf{A}}\,p^3}{\sqrt{N}}\sqrt{\sum_{s=1}^{t}\|\rho_s^{\mathsf{avg}}\|^2\eta_s\exp\Big(-2a\sum_{i=s+1}^{t}\eta_iH_i\Big)\sum_{i=s+1}^{t}\eta_iH_i} \\
&\le \frac{\mathrm{C}_{\mathbf{A}}\,p^3\zeta_1\zeta_2}{\sqrt{N}}\sqrt{\sum_{s=1}^{t}\eta_s^5H_s^4\exp\Big(-2a\sum_{i=s+1}^{t}\eta_iH_i\Big)\sum_{i=s+1}^{t}\eta_iH_i} .
\end{aligned}$$

Then we use the fact that $x\exp(-2cx) \le 1/(ec)\exp(-cx)$ for $c > 0$ and $x \ge 0$ and get:

$$\mathbb{E}^{1/p}[\|\mathbf{T_2}\|^p] \le \frac{\mathrm{C}_{\mathbf{A}}\,p^3\zeta_1\zeta_2}{\sqrt{Na\mathrm{e}}}\sqrt{\sum_{s=1}^{t}\eta_s^5H_s^4\exp\Big(-a\sum_{i=s+1}^{t}\eta_iH_i\Big)} .$$

Combining $\mathbf{T_1}$ and $\mathbf{T_2}$ finishes the proof. ∎

Convergence result. Now, we are ready to state the main result for the moment bounds $\mathbb{E}^{1/p}[\|\theta_t - \theta_\star\|^p]$.

A.3.1 Proof of the Theorem 2

Proof. We start from the decomposition of the last iterate (37). Then, by applying Minkowski's inequality, we have

$$\mathbb{E}^{1/p}[\|\theta_t - \theta_\star\|^p] \le \mathbb{E}^{1/p}[\|\tilde{\theta}_t^{\mathrm{tr}}\|^p] + \mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{bi,\,bi})}\|^p] + \mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{fl,\,bi})}\|^p] + \mathbb{E}^{1/p}[\|\tilde{\theta}_t^{(\mathrm{fl})}\|^p]$$

By controlling each of the terms by using Lemma 3, Lemma 6, Lemma 7 and Lemma 8 we get the desired result. □

Note that the bound in Theorem 2 can be rewritten in a slightly simpler form. Indeed, the last two terms in the bound can be controlled by $(1+t)^{-\gamma}$ up to a multiplicative constant. This simplified representation will be useful in the Gaussian approximation analysis for the multiplier bootstrap.

Corollary 5. *Under the assumptions of Theorem 2 and* **S**2*, for any* $t \ge 1$*, we have*

$$\mathbb{E}^{1/p}[\|\theta_t - \theta_\star\|^p] \le \exp\left(-a\sum_{s=1}^{t}\eta_s H_s\right)\|\theta_0 - \theta_\star\| + \mathrm{c}_{last,1}\, p^3 N^{-1/2}(1+t)^{-\gamma_\eta/2} + \mathrm{c}_{last,2}(1+t)^{-\gamma}\,, \tag{54}$$

where $\mathrm{c}_{last,1}$ *and* $\mathrm{c}_{last,2}$ *are defined in* (55)*.*

Proof. We introduce the constants

$$\begin{aligned}
\mathrm{c}_{\mathrm{last},1} &= (2\|\varepsilon\|_\infty + 2\,\mathrm{C}_{\mathbf{A}}\,\zeta_3 + 4(ae)^{-1/2}\,\mathrm{C}_{\mathbf{A}}\,\zeta_1\zeta_2)((\eta H)^{1/2} + a^{-1/2}L^{-1/2}(1+(a\eta H)^{\frac{\gamma+\gamma_\eta}{2(1-\gamma)}}))\,,\\
\mathrm{c}_{\mathrm{last},2} &= \zeta_1\zeta_2(\eta^{1/2}H + a^{-1/2}L^{1/2}(1+(a\eta H)^{\frac{\gamma}{1-\gamma}}))\,.
\end{aligned} \tag{55}$$

Applying Lemma 25, we obtain the bound (54). □

B NORMAL APPROXIMATION FOR FEDLSA

In this section, we provide the proof of Theorem 3, which establishes a Berry–Esseen type bound for the last iterate of FEDLSA.

For notational simplicity, we slightly abuse notation and write $Z_{i,j,c} := Z^c_{i,j}$ and $\theta_{i,j,c} := \theta^c_{i,j}$.

B.1 ERROR DECOMPOSITION AND LINEAR STATISTICS FOR GAUSSIAN APPROXIMATION

Firstly, for any $t \in \{1,\ldots,T\}$ and $h \in \{1,\ldots,H\}$, we unroll the recursion in (30), as follows

$$\theta^c_{t,h} - \theta^c_\star = (\mathrm{I} - \eta_t \mathbf{A}^c(Z_{t,h,c}))(\theta^c_{t,h-1} - \theta^c_\star) - \eta_t \varepsilon^c(Z_{t,h,c}) \tag{56}$$

$$= \Gamma^c_{t,h}(\theta^c_{t,0} - \theta^c_\star) - \eta_t \sum_{j=1}^{h}\Gamma^c_{t,j+1:h}\varepsilon^c(Z_{t,j,c})\,, \tag{57}$$

where for any $j \ge 1$ and $\theta \in \mathbb{R}^d$, $z \in \mathsf{Z}$, we recall the definition of the following values

$$\begin{aligned}
&\varepsilon^c(\theta, z) = \widetilde{\mathbf{A}}^c(z)\theta - \widetilde{\mathbf{b}}^c(z)\,,\\
&G^c_{t,\ell:r} = (\mathrm{I} - \eta_t\bar{\mathbf{A}}^c)^{r-\ell+1}, \quad \ell \le r\,,\\
&G^c_{t,\ell:r} = \mathrm{I}, \quad \ell > r\,,\\
&G^c_{t,j} = G^c_{t,1:j}\,.
\end{aligned}$$

Also, we set the following quantities

$$\Gamma^{\mathsf{avg}}_{t,h} = N^{-1}\sum_{c=1}^{N}\Gamma^c_{t,1:h}\,, \quad G^{\mathsf{avg}}_{t,h} = N^{-1}\sum_{c=1}^{N}G^c_{t,h}\,.$$

Now, we note that

$$
\begin{aligned}
N^{-1}\sum_{c=1}^{N}\{\theta^c_{t,h}-\theta_\star\} &= \Gamma^{\mathsf{avg}}_{t,h}(\theta_{t-1}-\theta_\star)+\Delta^{\mathsf{avg}}_{t,h}-\eta_t\mathcal{E}^{\mathsf{avg}}_{t,h} \\
&= \Gamma^{\mathsf{avg}}_{t,h}\prod_{i=1}^{t-1}\Gamma^{\mathsf{avg}}_{i}(\theta_0-\theta_\star)+\Gamma^{\mathsf{avg}}_{t,h}\sum_{s=1}^{t-1}\prod_{i=s+1}^{t-1}\Gamma^{\mathsf{avg}}_{i}(\Delta^{\mathsf{avg}}_{s,H_s}-\eta_s\mathcal{E}^{\mathsf{avg}}_{s,H_s})+\Delta^{\mathsf{avg}}_{t,h}-\eta_t\mathcal{E}^{\mathsf{avg}}_{t,h}\,,
\end{aligned}
$$

where we have set, for any $h\geq 0$,

$$
\Delta^{\mathsf{avg}}_{t,h}=N^{-1}\sum_{c=1}^{N}(\mathrm{I}-\Gamma^c_{t,h})(\theta^c_\star-\theta_\star)\,,\quad \text{and}\quad \mathcal{E}^{\mathsf{avg}}_{t,h}=N^{-1}\sum_{c=1}^{N}\sum_{s=1}^{h}\Gamma^c_{t,s+1:h}\varepsilon^c(Z_{t,s,c})\,.
$$

Particularly, for $h=H_t$, we obtain the last iterate decomposition, that is,

$$
\theta_t-\theta_\star=\prod_{i=1}^{t}\Gamma^{\mathsf{avg}}_{i}(\theta_0-\theta_\star)+\sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_{i}(\Delta^{\mathsf{avg}}_{s,H_s}-\eta_s\mathcal{E}^{\mathsf{avg}}_{s,H_s})\,.
$$

In this decomposition, the dominant term is the sum of the noise variables $\mathcal{E}^{\mathsf{avg}}_{s,H_s}$. We isolate it in the decomposition

$$
\begin{aligned}
\theta_t-\theta_\star &= \underbrace{-N^{-1}\sum_{s=1}^{t}\eta_s\prod_{i=s+1}^{t}G^{\mathsf{avg}}_{i}\sum_{c=1}^{N}\sum_{h=1}^{H_s}G^c_{s,h+1:H_s}\varepsilon^c(Z_{s,h,c})}_{\text{Linear statistics}}+\prod_{i=1}^{t}\Gamma^{\mathsf{avg}}_{i}(\theta_0-\theta_\star)+\sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_{i}\Delta^{\mathsf{avg}}_{s,H_s} \\
&\quad -N^{-1}\sum_{s=1}^{t}\eta_s\left(\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_{i}-\prod_{i=s+1}^{t}G^{\mathsf{avg}}_{i}\right)\sum_{c=1}^{N}\sum_{h=1}^{H_s}\Gamma_{s,h+1:H_s}\varepsilon^c(Z_{s,h,c}) \\
&\quad -N^{-1}\sum_{s=1}^{t}\eta_s\prod_{i=s+1}^{t}G^{\mathsf{avg}}_{i}\sum_{c=1}^{N}\sum_{h=1}^{H_s}(\Gamma_{s,h+1:H_s}-G_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})
\end{aligned}
\tag{58}
$$

However, the underbraced term is not the complete leading linear term. To get the correct leading term we have to extract the hidden from the first glance linear statistics from the summand $\sum_{s=1}^{t}\prod_{i=s+1}^{t}\Gamma^{\mathsf{avg}}_{i}\Delta^{\mathsf{avg}}_{s,H_s}$. Recalling the definitions from (34), we have

$$
\Delta^{\mathsf{avg}}_{s,H_s}=\rho^{\mathsf{avg}}_{s}+\tau^{\mathsf{avg}}_{s}\,.
$$

Here, $\rho^{\mathsf{avg}}_{t}=\frac{1}{N}\sum_{c=1}^{N}(\mathrm{I}-G^c_s)(\theta^c_\star-\theta_\star)$ represents the deterministic bias and does not contain any random noise, while $\tau^{\mathsf{avg}}_{s}=\frac{1}{N}\sum_{c=1}^{N}(G^c_s-\Gamma^c_{t,1:H_s})(\theta^c_\star-\theta_\star)$ contains linear noise from the $(G^c_s-\Gamma^c_{t,1:H_s})$. Therefore, to extract full linear statistics from $\theta_t-\theta_\star$ we need to linearize τ^{avg}_{s}.

Linearization of τ^{avg}_{s}. We need the following linearization to establish the correct leading term. Recalling the definitions from (34), we have

$$
\Delta^{\mathsf{avg}}_{s,H_s}=\rho^{\mathsf{avg}}_{s}+\tau^{\mathsf{avg}}_{s}\,.
$$

We aim to linearize the τ^{avg}_{s} term in order to extract the leading linear statistic. To this end, we linearize terms of the form

$$
\begin{aligned}
G^c_s-\Gamma^c_s &= \eta_s\sum_{i=1}^{H_s}G^c_{s,1:i-1}\widetilde{\mathbf{A}}^c(Z_{s,i,c})\Gamma^c_{s,i+1:H_s} \\
&= \underbrace{\eta_s\sum_{i=1}^{H_s}G^c_{s,1:i-1}\widetilde{\mathbf{A}}^c(Z_{s,i,c})}_{U^c_s}+\underbrace{\eta_s\sum_{i=1}^{H_s}G^c_{s,1:i-1}\widetilde{\mathbf{A}}^c(Z_{s,i,c})(\Gamma^c_{s,i+1:H_s}-\mathrm{I})}_{R^c_s}\,.
\end{aligned}
\tag{59}
$$

First, we bound the term U_s^c as

$$\mathbb{E}[\|U_s^c\|^2] \leq \eta_s^2 \sum_{i=1}^{H_s} (1-\eta_s a)^{2(i-1)} \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \leq \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \eta_s^2 H_s \,.$$

Then, we write

$$\mathbb{E}[\|R_s^c\|^2] \leq \eta_s^2 \sum_{i=1}^{H_s} (1-\eta_s a)^{2(i-1)} \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \mathbb{E}[\|\mathrm{I} - \Gamma_{s,i+1:H_s}^c\|^2] \,.$$

Next, we bound $\mathbb{E}^{1/2}[\|\mathrm{I} - \Gamma_{s,i+1:H_s}^c\|^2] \leq \sum_{j=i+1}^{H_s} \eta_s \, \mathrm{C}_{\mathbf{A}} (1-\eta_s a)^{j-i-1} \leq \eta_s H_s \, \mathrm{C}_{\mathbf{A}}$.Thus, we obtain

$$\mathbb{E}[\|R_s^c\|^2] \leq \mathrm{C}_{\mathbf{A}}^2 \, \eta_s^4 H_s^3 \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \,. \tag{60}$$

Leading linear statistics. The correct leading linear statistics defined in (58) consists of two terms

$$M_t = -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} G_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c}) \tag{61}$$

$$+ N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} G_{s,1:h-1}^c \widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star^c - \theta_\star) \,. \tag{62}$$

We can decompose M_t in four terms,

$$M_t = M_{t,1} + M_{t,2} + M_{t,3} + M_{t,4} \,,$$

where

$$M_{t,1} = -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} \varepsilon^c(Z_{s,h,c}) \,,$$

$$M_{t,2} = N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} \widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star^c - \theta_\star) \,,$$

$$M_{t,3} = -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (G_{s,h+1:H_s}^c - \mathrm{I}) \varepsilon^c(Z_{s,h,c}) \,,$$

$$M_{t,4} = N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (G_{s,1:h-1}^c - \mathrm{I}) \widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star^c - \theta_\star) \,.$$

From the definition (3) of ε^c, we see that the first and second terms can be combined into

$$\hat{M}_t = M_{t,1} + M_{t,2} = -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} \varepsilon^c(\theta_\star, Z_{s,h,c}) \,.$$

Let us denote

$$\Sigma_*^c = \mathbb{E}[\varepsilon^c(\theta_\star, Z)(\varepsilon^c(\theta_\star, Z))^T] \,, \quad \Sigma_*^{\mathsf{avg}} = N^{-1} \sum_{c=1}^{N} \Sigma_*^c \,. \tag{63}$$

Then, we define the covariance matrix, corresponding to the term $M_{t,1} + M_{t,2}$, as

$$\hat{\Sigma}_t = \mathbb{E}[\hat{M}_t \hat{M}_t^T] = N^{-1} \eta_t^{-1} \sum_{s=1}^{t} \eta_s^2 H_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \Sigma_* \left(\prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \right)^T \,, \tag{64}$$

which is the leading component of the FedLSA covariance. According to Lemma 13, the terms $M_{t,3}$ and $M_{t,4}$ are residual.

We may notice that in the case when $H = \mathrm{const}$, the leading component of FedLSA covariance $\hat{\Sigma}_t$ coincides with that of Bonnerjee et al. [2026, Theorem 2.2].

Final error decomposition for Gaussian approximation. We now define

$$\Sigma_t = \mathbb{E}[M_t M_t^T] \,, \tag{65}$$

which allows us to write the central error decomposition for the normalized statistic $\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star)$

$$\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star) = W_t + D_t \,, \tag{66}$$

where we define

$$W_t = \Sigma_t^{-1/2} M_t = \sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N} \upsilon_{s,h,c} \,, \tag{67}$$

$$D_t = D_{t,1} + D_{t,2} + D_{t,3} + D_{t,4} + D_{t,5} + D_{t,6} \,, \tag{68}$$

with

$$\upsilon_{s,h,c} = N^{-1}\eta_t^{-1/2}\eta_s\Sigma_t^{-1/2}\prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\left\{G_{s,1:h-1}^c \widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star^c - \theta_\star) - G_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\right\} \,,$$

$$D_{t,1} = \eta_t^{-1/2}\Sigma_t^{-1/2}\prod_{i=1}^{t}\Gamma_i^{\mathsf{avg}}(\theta_0 - \theta_\star) \,,$$

$$D_{t,2} = N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s=1}^{t}\prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\sum_{c=1}^{N} R_s^c(\theta_\star^c - \theta_\star) \,,$$

$$D_{t,3} = \eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s=1}^{t}\Big(\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\Big)\Delta_s^{\mathsf{avg}} \,,$$

$$D_{t,4} = \eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s=1}^{t}\prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\rho_s^{\mathsf{avg}} \,,$$

$$D_{t,5} = N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s=1}^{t}\eta_s\Big(\prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\Big)\sum_{c=1}^{N}\sum_{h=1}^{H_s}\Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c}) \,,$$

$$D_{t,6} = N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s=1}^{t}\eta_s\prod_{i=s+1}^{t} G_i^{\mathsf{avg}}\sum_{c=1}^{N}\sum_{h=1}^{H_s}(G_{s,h+1:H_s}^c - \Gamma_{s,h+1:H_s}^c)\varepsilon^c(Z_{s,h,c}) \,.$$

Theorem 3 Now we are ready to proof Theorem 3 – the main result of this section.

B.2 PROOF OF THE THEOREM 3

Applying Shao and Zhang [2022, Theorem 2.1] to the decomposition (66), we have

$$\mathsf{d}_\mathsf{C}(\eta_t^{-1/2}\Sigma_t^{-1/2}(\theta_t - \theta_\star), Y) \le \underbrace{259 d^{1/2}\Upsilon_t}_{R_1} + \underbrace{2\mathbb{E}[\|W_t\|\|D_t\|]}_{R_2} + \underbrace{2\sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\mathbb{E}[\|\upsilon_{s,h,c}\|\|D_t - D_t^{(s,h,c)}\|]}_{R_3} \,, \tag{69}$$

where $\Upsilon_t = \sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\mathbb{E}[\|\upsilon_{s,h,c}\|^3]$ and $D_t^{(s,h,c)}$ is a counterpart of D_t, in which we have replaced $Z_{s,h,c}$ by its independent copy $Z'_{s,h,c}$. In the same way, we define $D_{t,i}^{(s,h,c)}$ for $i \in \{1,2,3,4\}$.

Thus, to obtain the desired result we have to bound R_1, R_2 and R_3.

Bound on R_1

At first, we note that

$$\begin{aligned}
\|\upsilon_{s,h,c}\| &\leq \mathrm{C}_{\mathbf{\Sigma}}\, N^{-1/2}\eta_t^{-1/2}\eta_s \prod_{i=s+1}^{t}(1-\eta_i a)^{H_i}((1-\eta_s a)^{H_s-h}\|\varepsilon\|_\infty + (1-\eta_s a)^{h-1}\,\mathrm{C}_{\mathbf{A}}\,\|\theta_\star^c-\theta_\star\|) \\
&\leq \mathrm{C}_{\mathbf{\Sigma}}\, N^{-1/2}\eta_t^{-1/2}\eta_s \exp\left(-a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)((1-\eta_s a)^{H_s-h}\|\varepsilon\|_\infty + (1-\eta_s a)^{h-1}\,\mathrm{C}_{\mathbf{A}}\,\|\theta_\star^c-\theta_\star\|) \\
&\leq \mathrm{C}_{\mathbf{\Sigma}}\, N^{-1/2}\eta_t^{-1/2}\eta_s \exp\left(-a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\|\theta_\star^c-\theta_\star\|)\,.
\end{aligned}$$

Thus, applying Lemma 25, we get

$$\begin{aligned}
\Upsilon_t = \sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\mathbb{E}[\|\upsilon_{s,h,c}\|^3] &\leq \mathrm{C}_{\mathbf{\Sigma}}^3\, N^{-1/2}\eta_t^{-3/2}\sum_{s=1}^{t}\eta_s^3 H_s \exp\left(-3a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)^3 \\
&\leq 2(\eta^2 H + La^{-1}(1+(3a\eta H)^{\frac{2\gamma_\eta+\gamma}{1-\gamma}}))\,\mathrm{C}_{\mathbf{\Sigma}}^3\, N^{-1/2}\eta^{1/2}(1+t)^{-\gamma_\eta/2}(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)^3\,, \qquad (70)
\end{aligned}$$

where we have set

$$\zeta_3^2 = N^{-1}\sum_{c=1}^{N}\|\theta_\star^c-\theta_\star\|^2\,.$$

Bound on R_2

Firstly, we establish the bound on Δ_j^{avg} for any $1 \leq j \leq t$. By the definition of Δ_j^{avg}, we have

$$\Delta_j^{\mathsf{avg}} = N^{-1}\sum_{c=1}^{N}(\mathrm{I}-\Gamma_{j,H_j}^c)(\theta_\star^c-\theta_\star) = \tau_j^{\mathsf{avg}} + \rho_j^{\mathsf{avg}}\,.$$

Then, using Lemma 2 and (50), we obtain

$$\mathbb{E}[\|\Delta_j^{\mathsf{avg}}\|^2] \leq 2\mathbb{E}[\|\tau_j^{\mathsf{avg}}\|^2] + 2\mathbb{E}[\|\rho_j^{\mathsf{avg}}\|^2] \leq 2\bar{\sigma}_{\mathsf{het}}^2 N^{-1}\eta_j^2 H_j + \zeta_1^2\zeta_2^2\eta_j^4 H_j^4 \quad . \qquad (71)$$

Now, Applying Hölders inequality, we obtain

$$\mathbb{E}[\|W_t\|\|D_t\|] \leq \mathbb{E}^{1/2}[\|W_t\|^2]\mathbb{E}^{1/2}[\|D_t\|^2]\,. \qquad (72)$$

Firstly, we note that $\mathbb{E}[\|W_t\|^2] = d$. Therefore, we can proceed with the D_t. For $D_{t,1}$, we have

$$\begin{aligned}
\mathbb{E}^{1/2}[\|D_{t,1}\|^2] &\leq \mathrm{e}^{(1-\gamma)^{-1}}\,\mathrm{C}_{\mathbf{\Sigma}}\,\eta_t^{-1/2}\exp\left(-a(1-\gamma)^{-1}\eta H(1+t)^{1-\gamma}\right)\|\theta_0-\theta_\star\| \\
&\leq \mathrm{c}_{D,1}\,\mathrm{C}_{\mathbf{\Sigma}}\exp\left(-\frac{a}{2(1-\gamma)}\eta H(1+t)^{1-\gamma}\right)\|\theta_0-\theta_\star\|\,, \qquad (73)
\end{aligned}$$

where we have used that $x^{1/2}e^{-cx^{1-\gamma}} \leq (ce(1-\gamma))^{-\frac{1}{2(1-\gamma)}}e^{-cx^{1-\gamma}/2}$ for any $x, c > 0$ and have set $\mathrm{c}_{D,1} = 2\mathrm{e}^{(1-\gamma)^{-1}}(ae\eta H)^{-\frac{1}{2(1-\gamma)}}\eta^{1/2}$. Using the martingale-difference structure of the sequence $\{R_s^c, 1 \leq s \leq t\}$ for any c, (60) and applying Lemma 25, we obtain

$$\begin{aligned}
\mathbb{E}[\|D_{t,2}\|^2] &\leq \mathrm{C}_{\mathbf{A}}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^4 H_s^3\exp\left(-2a\eta H\sum_{\ell=s+1}^{t}(1+\ell)^{-\gamma}\right)\sum_{c=1}^{N}\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\|\theta_\star^c-\theta_\star\|^2 \\
&\leq \mathrm{c}_{D,2}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2(1+t)^{-2\gamma}\,, \qquad (74)
\end{aligned}$$

where

$$\mathrm{c}_{D,2}^2 = 8a^{-1}\,\mathrm{C}_{\mathbf{A}}^2\,\zeta_4^2((\eta H)^3 + La^{-1}(\eta H)^2(1+(2a\eta H)^{\frac{3\gamma+\gamma_\eta}{1-\gamma}}))$$

and we have set

$$\zeta_4^2 = \frac{1}{N} \sum_{c=1}^{N} \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \|\theta_\star^c - \theta_\star\|^2 .$$

Then, again using the martingale-difference structure of the sequence $\{\left(\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^t G_i^{\mathsf{avg}}\right)\tau_s^{\mathsf{avg}}, 1 \leq s \leq t\}$, Lemma 16 and (71), we obtain

$$\begin{aligned}
\mathbb{E}[\|D_{t,3}\|^2] &\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N \eta_t^{-1} \sum_{s=1}^t \mathbb{E}[\| \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^t G_i^{\mathsf{avg}} \|^2] \mathbb{E}[\|\Delta_s^{\mathsf{avg}}\|^2] \\
&\leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, \eta_t^{-1} \sum_{s=1}^t \eta_s \mathbb{E}[\|\Delta_s^{\mathsf{avg}}\|^2] \exp\left(-2a \sum_{\ell=s+1}^t \eta_\ell H_\ell\right) \\
&\leq 4a^{-1} \bar{\sigma}_{\mathbf{A}}^2 \zeta_3^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} \eta_t^{-1} \sum_{s=1}^t \eta_s^3 H_s \exp\left(-2a \sum_{\ell=s+1}^t \eta_\ell H_\ell\right) \\
&+ 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2 \zeta_1^2 \zeta_2^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, \eta_t^{-1} \sum_{s=1}^t \eta_s^5 H_s^4 \exp\left(-2a \sum_{\ell=s+1}^t \eta_\ell H_\ell\right) \\
&\leq \mathrm{c}_{D,3,1}^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} (1+t)^{-\gamma_\eta} + \mathrm{c}_{D,3,2}^2\, \mathrm{C}_{\mathbf{\Sigma}}^2 (1+t)^{-3\gamma} .
\end{aligned} \tag{75}$$

where

$$\begin{aligned}
\mathrm{c}_{D,3,1}^2 &= 64 a^{-2} \bar{\sigma}_{\mathbf{A}}^2 \zeta_3^2 (\eta^2 H + L a^{-1} \eta (1 + (2a\eta H)^{\frac{\gamma+2\gamma_\eta}{1-\gamma}})) , \\
\mathrm{c}_{D,3,2}^2 &= 256 a^{-2} \bar{\sigma}_{\mathbf{A}}^2 \zeta_1^2 \zeta_2^2 ((\eta H)^4 + L a^{-1} (\eta H)^3 (1 + (2a\eta H)^{\frac{4\gamma+\gamma_\eta}{1-\gamma}})) .
\end{aligned}$$

For $D_{t,4}$, we have

$$\mathbb{E}^{1/2}[\|D_{t,4}\|^2] \leq \mathrm{c}_{D,4}\, \mathrm{C}_{\mathbf{\Sigma}}\, N^{1/2} (1+t)^{-\gamma} , \tag{76}$$

where $\mathrm{c}_{D,4} = 8a^{-1}\eta H$. For the term $D_{t,5}$, we note that the sequence $\{(\prod_{i=s+1}^t G_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}) \sum_{h=1}^{H_s} \Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c}), 1 \leq s \leq t\}$ is martingale-difference w.r.t the filtration $\mathcal{F}_s^-$. Then, applying Lemma 16 and using (52), we get

$$\begin{aligned}
\mathbb{E}[\|D_{t,5}\|^2] &\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} \eta_t^{-1} \sum_{s=1}^t \sum_{c=1}^N \eta_s^2 \mathbb{E}[\|(\prod_{i=s+1}^t G_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}}) \sum_{h=1}^{H_s} \Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\|^2] \\
&\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} \eta_t^{-1} \sum_{s=1}^t \sum_{c=1}^N \eta_s^2 \mathbb{E}[\| \prod_{i=s+1}^t G_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} \|^2] \mathbb{E}[\| \sum_{h=1}^{H_s} \Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\|^2] \\
&\leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-2} \eta_t^{-1} \sum_{s=1}^t \sum_{c=1}^N \eta_s^3 \mathbb{E}[\| \sum_{h=1}^{H_s} \Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\|^2] \exp\left(-2a \sum_{\ell=s+1}^t \eta_\ell H_\ell\right) .
\end{aligned} \tag{77}$$

Now, using the fact that for any $s \in \{1, \ldots, t\}$ the sequence $\{\Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c}), 1 \leq h \leq H_s\}$ is martingale-difference w.r.t filtration $\mathcal{F}_{s,h}^+$, we get

$$\mathbb{E}[\| \sum_{h=1}^{H_s} \Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\|^2] = \sum_{h=1}^{H_s} \mathbb{E}[\|\Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\|^2] \leq \mathrm{Tr}(\Sigma_\varepsilon^c) H_s . \tag{78}$$

Combining (77), (78), and using Lemma 25, we obtain

$$\begin{aligned}
\mathbb{E}[\|D_{t,5}\|^2] &\leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2 \bar{\sigma}_\varepsilon^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} \eta_t^{-1} \sum_{s=1}^t \eta_s^3 H_s \exp\left(-2a \sum_{\ell=s+1}^t \eta_\ell H_\ell\right) \\
&\leq \mathrm{c}_{D,5}^2\, \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1} (1+t)^{-\gamma_\eta} ,
\end{aligned} \tag{79}$$

where

$$\mathrm{c}_{D,5}^2 = 4a^{-2}\bar{\sigma}_{\mathbf{A}}^2\bar{\sigma}_{\varepsilon}^2(\eta^2 H + La^{-1}\eta(2a\eta H)^{\frac{\gamma+2\gamma_\eta}{1-\gamma}})\,.$$

For the last term $D_{t,6}$, we note that again for any $s \in \{1,\ldots,t\}$ the sequence $\{(G^c_{s,h+1:H_s} - \Gamma^c_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c}), 1 \leq h \leq H_s\}$ is martingale-difference w.r.t filtration $\mathcal{F}^+_{s,h}$. Thus, applying Lemma 25, we get

$$\begin{aligned}
\mathbb{E}[\|D_{t,6}\|^2] &= \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\sum_{s=1}^{t}\sum_{c=1}^{N}\eta_s^2\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\sum_{h=1}^{H_s}\mathbb{E}[\|(G^c_{s,h+1:H_s} - \Gamma^c_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})\|^2] \\
&\leq \bar{\sigma}_{\text{noise}}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2\,\eta_t^{-1}\sum_{s=1}^{t}\eta_s^4 H_s^2\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right) \\
&\leq \mathrm{c}_{D,6}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2(1+t)^{-(\gamma+\gamma_\eta)}\,,
\end{aligned} \tag{80}$$

where we have set

$$\begin{aligned}
\bar{\sigma}_{\text{noise}}^2 &= N^{-1}\sum_{c=1}^{N}\mathrm{Tr}(\Sigma_\varepsilon^c)\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\,, \\
\mathrm{c}_{D,6}^2 &= 4a^{-1}\bar{\sigma}_{\text{noise}}^2(\eta^3 H^2 + La^{-1}\eta^2 H(2a\eta H)^{\frac{2(\gamma+\gamma_\eta)}{1-\gamma}})\,.
\end{aligned}$$

and used that

$$\begin{aligned}
\mathbb{E}[\|(G^c_{s,h+1:H_s} - \Gamma^c_{s,h+1:H_s})u\|^2] &= \eta_s^2\sum_{\ell=h+1}^{H_s}\mathbb{E}[\|(\mathrm{I} - \eta_s\bar{\mathbf{A}}^c)^{\ell-1}\widetilde{\mathbf{A}}^c(Z_{s,\ell,c})\Gamma^c_{s,\ell+1:H_s}u\|^2] \\
&\leq \eta_s^2\sum_{\ell=h+1}^{H_s}(1-\eta_s a)^{2(\ell-1)}\mathrm{Tr}(\Sigma^c_{\mathbf{A}})(1-\eta_s a)^{2(H_s-\ell)}\|u\|^2 \\
&\leq \eta_s^2 H_s(1-\eta_s a)^{2(H_s-1)}\mathrm{Tr}(\Sigma^c_{\mathbf{A}})\|u\|^2\,.
\end{aligned}$$

Bound on R_3

To bound each term in R_3, we can again use Hölders inequality, that is,

$$\mathbb{E}[\|\upsilon_{s,h,c}\|\|D_t - D_t^{(s,h,c)}\|] \leq \mathbb{E}^{1/2}[\|\upsilon_{s,h,c}\|^2]\mathbb{E}^{1/2}[\|D_t - D_t^{(s,h,c)}\|^2]\,.$$

Applying Minkowski's inequality, we get

$$\begin{aligned}
\mathbb{E}^{1/2}[\|D_t - D_t^{(s,h,c)}\|^2] &\leq \mathbb{E}^{1/2}[\|D_{t,1} - D_{t,1}^{(s,h,c)}\|^2] + \mathbb{E}^{1/2}[\|D_{t,2} - D_{t,2}^{(s,h,c)}\|^2] \\
&\quad + \mathbb{E}^{1/2}[\|D_{t,3} - D_{t,3}^{(s,h,c)}\|^2] + \mathbb{E}^{1/2}[\|D_{t,4} - D_{t,4}^{(s,h,c)}\|^2] \\
&\quad + \mathbb{E}^{1/2}[\|D_{t,5} - D_{t,5}^{(s,h,c)}\|^2] + \mathbb{E}^{1/2}[\|D_{t,6} - D_{t,6}^{(s,h,c)}\|^2]\,.
\end{aligned}$$

Also, we set

$$R_{3,i} = \sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\mathbb{E}^{1/2}[\|\upsilon_{s,h,c}\|^2]\,\mathbb{E}^{1/2}[\|D_{t,i} - D_{t,i}^{(s,h,c)}\|^2]$$

Perturbations of $D_{t,1}$.

Now, we should bound each perturbation term separately. For the first term, applying Lemma 22, Minkowski's inequality, we obtain

$$
\begin{aligned}
\mathbb{E}^{1/2}[\|D_{t,1} - D_{t,1}^{(s,h,c)}\|^2] &= \mathbb{E}^{1/2}[\|\eta_t^{-1/2}\Sigma_t^{-1/2}\left(\prod_{i=1}^{t}\Gamma_i^{\mathsf{avg}} - \prod_{i=1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\right)(\theta_0 - \theta_\star)\|^2] \\
&\leq \mathrm{C}_{\boldsymbol{\Sigma}}\, \eta_t^{-1/2}\mathbb{E}^{1/2}[\|\left(\prod_{i=1}^{t}\Gamma_i^{\mathsf{avg}} - \prod_{i=1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\right)(\theta_0 - \theta_\star)\|^2] \\
&\leq \mathrm{C}_{\boldsymbol{\Sigma}}\, \eta_t^{-1/2}\mathbb{E}^{1/2}[\|\sum_{\ell=1}^{t}\prod_{i=1}^{\ell-1}\Gamma_i^{\mathsf{avg}}\cdot(\Gamma_\ell^{\mathsf{avg}} - \Gamma_\ell^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})})\prod_{i=\ell+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}(\theta_0 - \theta_\star)\|^2] \\
&\leq \mathrm{e}\,\mathrm{C}_{\boldsymbol{\Sigma}}\, \eta_t^{-1/2}\exp\left(-a\sum_{i=1}^{t}\eta_i H_i\right)\sum_{\ell=1}^{t}\mathbb{E}^{1/2}[\|\left(\Gamma_\ell^{\mathsf{avg}} - \Gamma_\ell^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\right)(\theta_0 - \theta_\star)\|^2] \\
&= \mathrm{e}\,\mathrm{C}_{\boldsymbol{\Sigma}}\, N^{-2}\eta_t^{-1/2}\exp\left(-a\sum_{i=1}^{t}\eta_i H_i\right)\mathbb{E}^{1/2}[\|\left(\Gamma_{s,H_s}^{c} - \Gamma_{s,H_s}^{c,(s,h,c)}\right)(\theta_0 - \theta_\star)\|^2]\,.
\end{aligned}
$$

According to Lemma 15, we have

$$
\begin{aligned}
\mathbb{E}^{1/2}[\|D_{t,1} - D_{t,1}^{(s,h,c)}\|^2] &\leq 4\mathrm{e}\,\mathrm{C}_{\boldsymbol{\Sigma}}\, N^{-1/2}\eta_t^{-1/2}\eta_s\exp\left(-a\sum_{\ell=1}^{t}\eta_\ell H_\ell\right)\sqrt{\mathrm{Tr}(\Sigma_{\mathbf{A}}^{c})}\|\theta_0 - \theta_\star\| \\
&\leq 4\mathrm{e}\,\mathrm{C}_{\boldsymbol{\Sigma}}\, N^{-1/2}\eta_t^{-1/2}\eta_s\exp\left(-\frac{a\eta H}{1-\gamma}(1+t)^{1-\gamma}\right)\sqrt{\mathrm{Tr}(\Sigma_{\mathbf{A}}^{c})}\|\theta_0 - \theta_\star\|\,.
\end{aligned}
$$

Thus, for $R_{3,1}$, we obtain

$$
\begin{aligned}
R_{3,1} &\leq 4\mathrm{e}\,\mathrm{C}_{\boldsymbol{\Sigma}}^2\, \eta_t^{-1}\exp\left(-\frac{a\eta H}{1-\gamma}(1+t)^{1-\gamma}\right)\sum_{s=1}^{t}\eta_s^2 H_s\exp\left(-a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)(\bar{\sigma}_{\mathbf{A}}\|\varepsilon\|_\infty + \zeta_4\,\mathrm{C}_{\mathbf{A}})\|\theta_0 - \theta_\star\| \\
&\leq 4\mathrm{e}(2\eta H + La^{-1}(1 + (a\eta H)^{\frac{3\gamma_\eta}{2(1-\gamma)}}))\,\mathrm{C}_{\boldsymbol{\Sigma}}^2\exp\left(-\frac{a\eta H}{1-\gamma}(1+t)^{1-\gamma}\right)(\bar{\sigma}_{\mathbf{A}}\|\varepsilon\|_\infty + \zeta_4\,\mathrm{C}_{\mathbf{A}})\|\theta_0 - \theta_\star\| \\
&= \mathrm{c}_1\,\mathrm{C}_{\boldsymbol{\Sigma}}^2\exp\left(-\frac{a\eta H}{1-\gamma}(1+t)^{1-\gamma}\right)\,, \qquad (81)
\end{aligned}
$$

where we have set

$$
\mathrm{c}_1 = 8\mathrm{e}(\eta H + La^{-1}(1 + (a\eta H)^{\frac{3\gamma_\eta}{2(1-\gamma)}}))(\bar{\sigma}_{\mathbf{A}}\|\varepsilon\|_\infty + \zeta_4\,\mathrm{C}_{\mathbf{A}})\|\theta_0 - \theta_\star\|\,.
$$

Perturbations of $D_{t,2}$. Now, we can proceed with the perturbations. For the perturbations of $D_{t,2}$, we have

$$
\mathbb{E}[\|D_{t,2} - D_{t,2}^{(s,h,c)}\|^2] \leq 2\,\mathrm{C}_{\mathbf{A}}^2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^4 H_s^3\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma_{\mathbf{A}}^{c})\|\theta_\star^{c} - \theta_\star\|^2\,,
$$

since $R_{s'}^{c'} - R_{s'}^{c',(s,h,c)} = 0$ for $c' \neq c$ and $s' \neq s$. This leads to the following estimation in the final bound

$$
\begin{aligned}
R_{3,2} &\leq 2\,\mathrm{C}_{\mathbf{A}}\,\mathrm{C}_{\boldsymbol{\Sigma}}^2\, \eta_t^{-1}\sum_{s=1}^{t}\eta_s^3 H_s^{5/2}\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\zeta_4(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3) \\
&\leq \mathrm{c}_2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2(1+t)^{-(3\gamma-\gamma_\eta)/2}\,, \qquad (82)
\end{aligned}
$$

where we also used the Cauchy-Schwartz inequality to bound $N^{-1}\sum_{c=1}^{N}\sqrt{\mathrm{Tr}(\Sigma_{\mathbf{A}}^{c})}\|\theta_\star^{c} - \theta_\star\|(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\|\theta_\star^{c} - \theta_\star\|)$ and defined

$$
\mathrm{c}_2 = 8\,\mathrm{C}_{\mathbf{A}}(\eta^2 H^{5/2} + La^{-1}(1 + (2a\eta H)^{\frac{5}{2}+\frac{\gamma_\eta}{2\gamma}}))\zeta_4(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)\,.
$$

Perturbations of $D_{t,3}$.

For the perturbations of $D_{t,3}$, we have the decomposition $D_{t,3} - D_{t,3}^{(s,h,c)} = D_{t,3}^1 + D_{t,3}^2$, where we define

$$D_{t,3}^1 = \eta_t^{-1/2} \Sigma_t^{-1/2} \sum_{s'=1}^{t} (\prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} G_i^{\mathsf{avg}})(\Delta_{s'}^{\mathsf{avg}} - \Delta_{s'}^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})})$$
$$D_{t,3}^2 = \eta_t^{-1/2} \Sigma_t^{-1/2} \sum_{s'=1}^{t} (\prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}) \Delta_{s'}^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \,.$$

We start with estimating $\mathbb{E}^{1/2}[\|D_{t,3}^1\|^2]$. As $\Delta_{s'}^{\mathsf{avg}} = \Delta_{s'}^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}$ for $s' \neq s$, we have

$$\begin{aligned}
\mathbb{E}[\|D_{t,3}^1\|^2] &\leq \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, \eta_t^{-1} \mathbb{E}[\| \left(\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \right) (\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}) \|^2] \\
&\leq \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, \eta_t^{-1} \mathbb{E}[\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \|^2] \mathbb{E}[\|\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] \,,
\end{aligned}$$

where we have noticed that $\prod_{i=s+1}^{t} \Gamma_\ell^{\mathsf{avg}} - \prod_{i=s+1}^{t} G_\ell^{\mathsf{avg}}$ and $\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}$ are independent. Now, applying Lemma 16 and Lemma 17, we obtain

$$\begin{aligned}
\mathbb{E}[\|D_{t,3}^1\|^2] &\leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, \eta_t^{-1} \eta_s \exp\left(-a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \mathbb{E}[\|\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] \\
&\leq \mathrm{c}_{3,1}^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-2} \eta_t^{-1} \eta_s^3 \exp\left(-a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \|\theta_\star^c - \theta_\star\|^2 \,.
\end{aligned}$$

where we have set

$$\mathrm{c}_{3,1}^2 = 32 a^{-1} \bar{\sigma}_{\mathbf{A}}^2 \,.$$

For the second term, applying Lemma 18, using the martingale-difference structure of the sequence $\{\Big(\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} -$

$\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\Big) \Delta_{s'}^{\mathsf{avg}}, 1 \leq s' \leq t\}$, (71) and Lemma 25 together with (71), we get

$$
\begin{aligned}
\mathbb{E}[\|D_{t,3}^2\|^2] &\leq \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, \eta_t^{-1} \sum_{s'=1}^{s-1} \mathbb{E}[\| \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \|^2] \mathbb{E}[\|\Delta_{s'}^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] \\
&\leq 4\mathrm{e} \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1} \eta_t^{-1} \eta_s^2 \sum_{s'=1}^{s-1} \mathbb{E}[\|\Delta_{s'}^{\mathsf{avg}}\|^2] \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \\
&\leq 8\mathrm{e}\bar{\sigma}_{\mathsf{het}}^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-2} \eta_t^{-1} \eta_s^2 \sum_{s'=1}^{s-1} \eta_{s'}^2 H_{s'} \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \\
&+ 4\mathrm{e} \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, \zeta_1^2 \zeta_2^2 N^{-1} \eta_t^{-1} \eta_s^2 \sum_{s'=1}^{s-1} \eta_{s'}^4 H_{s'}^4 \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \\
&\leq 8\mathrm{e} \, \mathrm{C}_{\mathbf{A}}^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-2} \eta_t^{-1} \eta_s^2 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \sum_{s'=1}^{s-1} \eta_{s'}^2 H_{s'} \exp\left(-2a \sum_{\ell=s'+1}^{s} \eta_\ell H_\ell\right) \\
&+ 4\mathrm{e}\zeta_1^2 \zeta_2^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1} \eta_t^{-1} \eta_s^2 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \sum_{s'=1}^{s-1} \eta_{s'}^4 H_{s'}^4 \exp\left(-2a \sum_{\ell=s'+1}^{s} \eta_\ell H_\ell\right) \\
&\leq 32\mathrm{e}a^{-1} \, \mathrm{C}_{\mathbf{A}}^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-2} \eta_t^{-1} \eta_s^3 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \\
&+ 256\mathrm{e}a^{-1} \zeta_1^2 \zeta_2^2 \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1} (\eta H)^3 \eta_t^{-1} \eta_s^2 (1+s)^{-3\gamma} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \,,
\end{aligned}
$$

where we used that $\gamma > 1/2$ for the second term. Therefore, we have

$$
\begin{aligned}
\mathbb{E}^{1/2}[\|D_{t,3} - D_{t,3}^{(s,h,c)}\|^2] &\leq (\mathrm{c}_{3,2} + \mathrm{c}_{3,1} \|\theta_\star^c - \theta_\star\|) \, \mathrm{C}_{\boldsymbol{\Sigma}} \, N^{-1} \eta_t^{-1/2} \eta_s^{3/2} \exp\left(-a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sqrt{\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)} \\
&+ \mathrm{c}_{3,3} \, \mathrm{C}_{\boldsymbol{\Sigma}} \, N^{-1/2} (\eta H)^{3/2} \eta_t^{-1/2} \eta_s (1+s)^{-3\gamma/2} \exp\left(-a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sqrt{\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)} \,,
\end{aligned}
$$

where $\mathrm{c}_{3,2}^2 = 32\mathrm{e}a^{-1} \, \mathrm{C}_{\mathbf{A}}^2$ and $\mathrm{c}_{3,3}^2 = 256\mathrm{e}a^{-1}\zeta_1^2\zeta_2^2$. Now, using this bound and applying Lemma 25, we get

$$
\begin{aligned}
R_{3,3} &\leq 4(\mathrm{c}_{3,2}^{1/2} \, \bar{\sigma}_\varepsilon + \mathrm{c}_{3,1}^{1/2} \, \zeta_4) \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1/2} \eta_t^{-1} \sum_{s=1}^{t} \eta_s^{5/2} H_s \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) (\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}} \, \zeta_3) \\
&+ \mathrm{c}_{3,3} \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 (\eta H)^{3/2} \eta_t^{-1} \sum_{s=1}^{t} \eta_s^2 H_s (1+s)^{-3\gamma/2} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) (\bar{\sigma}_\varepsilon \|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}} \, \zeta_4) \\
&\leq \mathrm{c}_{3,4} \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1/2} (1+t)^{-\gamma_\eta/2} + \mathrm{c}_{3,5} \, \mathrm{C}_{\boldsymbol{\Sigma}}^2 \, N^{-1} (1+t)^{-3\gamma/2} \,,
\end{aligned}
\tag{83}
$$

where we have set

$$
\begin{aligned}
\mathrm{c}_{3,4} &= 2(\mathrm{c}_{3,2}^{1/2} \, \bar{\sigma}_\varepsilon + \mathrm{c}_{3,1}^{1/2} \, \zeta_4)(\eta^{3/2} H + L a^{-1} \eta^{1/2} (1 + (2a\eta H)^{1+\frac{3\gamma_\eta}{2\gamma}})(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}} \, \zeta_3) \,, \\
\mathrm{c}_{3,5} &= 2 \, \mathrm{c}_{3,3}((\eta H)^{5/2} + L a^{-1} (\eta H)^{3/2} (1 + (2a\eta H)^{5/2+\gamma_\eta/\gamma})(\bar{\sigma}_\varepsilon \|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}} \, \zeta_4) \,.
\end{aligned}
$$

Perturbations of $D_{t,5}$.

Note that for $D_{t,5}$, we have the following decomposition

$$
\begin{aligned}
D_{t,5} &= N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s',h',c'}\eta_{s'}\Big(\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}-\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg}}\Big)\Gamma^{c'}_{s',h'+1:H_{s'}}\varepsilon^{c'}(Z_{s',h',c'}) \\
&+ N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s',h',c'}\eta_{s'}\Big(\prod_{i=s'+1}^{t}G_i^{\mathsf{avg}}-\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\Big)(\Gamma^{c'}_{s',h'+1:H_{s'}}-\Gamma^{c',(s,h,c)}_{s',h'+1:H_{s'}})\varepsilon^{c'}(Z_{s',h',c'}) \\
&+ N^{-1}\eta_t^{-1/2}\eta_s\Sigma_t^{-1/2}(\prod_{i=s+1}^{t}G_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})})\Gamma^{c,(s,h,c)}_{s,h+1:H_s}(\varepsilon^{c}(Z_{s,h,c})-\varepsilon^{c}(Z'_{s,h,c})) \\
&+ \sum_{s',h',c'}\eta_{s'}(\prod_{i=s'+1}^{t}G_i^{\mathsf{avg}}-\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})})\Gamma^{c',(s,h,c)}_{s',h'+1:H_{s'}}\varepsilon^{c'}(Z^{s,h,c}_{s',h',c'}) = T_{5,1}+T_{5,2}+T_{5,3}+T_{5,4}\,. \qquad (84)
\end{aligned}
$$

We can see that in the perturbation $D_{t,5}-D_{t,5}^{(s,h,c)}$ there are only three terms T_1, T_2 and T_3. Thus, we have

$$
\mathbb{E}[\|D_{t,5}-D_{t,5}^{(s,h,c)}\|^2] \le 4\mathbb{E}[\|T_{5,1}\|^2]+4\mathbb{E}[\|T_{5,2}\|^2]+4\mathbb{E}[\|T_{5,3}\|^2]\,.
$$

Further, we bound these terms separately. For $T_{5,1}$, using the martingale-difference structure of the sequences $\{\Big(\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\Big)\Gamma^{c}_{s,h+1:H_s}\varepsilon^{c}(Z_{s,h,c}), s\in[t]\}$ and $\{\Gamma^{c}_{s,h+1:H_s}\varepsilon^{c}(Z_{s,h,c}), h\in[H_s]\}$, applying Lemma 18 and Lemma 25, we get

$$
\begin{aligned}
&\mathbb{E}[\|T_{5,1}\|^2] \\
&\le \mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-1}\eta_t^{-1}\sum_{s'=1}^{t}\sum_{c'=1}^{N}\eta_{s'}^2\mathbb{E}[\|\sum_{h'=1}^{H_{s'}}\Big(\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}-\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg}}\Big)\Gamma^{c'}_{s',h'+1:H_{s'}}\varepsilon^{c'}(Z_{s',h',c'})\|^2] \\
&\le \mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-1}\eta_t^{-1}\sum_{s'=1}^{t}\sum_{c'=1}^{N}\eta_{s'}^2\mathbb{E}[\|\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}-\prod_{i=s'+1}^{t}\Gamma_i^{\mathsf{avg}}\|^2]\sum_{h'=1}^{H_{s'}}\mathbb{E}[\|\Gamma^{c'}_{s',h'+1:H_{s'}}\varepsilon^{c'}(Z_{s',h',c'})\|^2] \\
&\le 4\mathrm{e}\bar{\sigma}_\varepsilon^2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-2}\eta_t^{-1}\eta_s^2\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma^c_{\mathbf{A}})\sum_{s'=1}^{s-1}\eta_{s'}^2H_{s'}\exp\left(-2a\sum_{\ell=s'+1}^{s}\eta_\ell H_\ell\right) \\
&\le \mathrm{c}_{5,1}^2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-2}\eta_t^{-1}\eta_s^3\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma^c_{\mathbf{A}})\,,
\end{aligned}
$$

where

$$
\mathrm{c}_{5,1}^2 = 32\mathrm{e}a^{-1}L\bar{\sigma}_\varepsilon^2(1+(2a\eta H)^{\frac{\gamma_\eta+\gamma}{1-\gamma}})\,.
$$

For the term $T_{5,2}$, we note that $\Gamma^{c'}_{s',h'+1:H_{s'}}-\Gamma^{c',(s,h,c)}_{s',h'+1:H_{s'}}=0$ for $s'\neq s$. Hence, applying Lemma 16 and Lemma 15, we get

$$
\begin{aligned}
\mathbb{E}[\|T_{5,2}\|^2] &\le \mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-1}\eta_t^{-1}\eta_s^2\mathbb{E}[\|\prod_{i=s+1}^{t}G_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\|^2]\sum_{h'=1}^{h-1}\mathbb{E}[\|(\Gamma^{c}_{s,h'+1:H_{s'}}-\Gamma^{c,(s,h,c)}_{s,h'+1:H_s})\varepsilon^{c}(Z_{s,h',c})\|^2] \\
&\le 2a^{-1}\bar{\sigma}_{\mathbf{A}}^2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-2}\eta_t^{-1}\eta_s^3\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\sum_{h'=1}^{h-1}\mathbb{E}[\|\Gamma^{c}_{s,h'+1:H_{s'}}-\Gamma^{c,(s,h,c)}_{s,h'+1:H_s}\|^2]\mathrm{Tr}(\Sigma^c_\varepsilon) \\
&\le \mathrm{c}_{5,2}^2\,\mathrm{C}_{\boldsymbol{\Sigma}}^2 N^{-2}\eta_t^{-1}\eta_s^5 H_s\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma^c_{\mathbf{A}})\mathrm{Tr}(\Sigma^c_\varepsilon)\,,
\end{aligned}
$$

where we have set $\mathrm{c}_{5,2}^2 = 32a^{-1}\bar{\sigma}_{\mathbf{A}}^2$. For the term $T_{5,3}$, applying Lemma 16, we obtain

$$
\begin{aligned}
\mathbb{E}[\|T_{5,3}\|^2] &\leq 2\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^2\mathbb{E}[\|\prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}}\|^2]\mathrm{Tr}(\Sigma_\varepsilon^c) \\
&\leq \mathrm{c}_{5,3}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-2}\eta_t^{-1}\eta_s^3 \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma_\varepsilon^c)\,,
\end{aligned}
$$

where we define $\mathrm{c}_{5,3}^2 = 4a^{-1}\bar{\sigma}_{\mathbf{A}}^2$. Thus, combining the above bounds, we get

$$
\mathbb{E}[\|D_{t,5} - D_{t,5}^{(s,h,c)}\|^2] \leq \mathrm{c}_{5,4}^2\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-2}\eta_t^{-1}\eta_s^3 \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)(1+\mathrm{Tr}(\Sigma_{\mathbf{A}}^c))\mathrm{Tr}(\Sigma_\varepsilon^c)\,,
$$

where we have set

$$
\mathrm{c}_{5,4}^2 = 4\,\mathrm{c}_{5,1}^2 + 4\,\mathrm{c}_{5,2}^2 + 4\,\mathrm{c}_{5,3}^2\ .
$$

Thus, again applying Lemma 25, we get

$$
\begin{aligned}
R_{3,5} &\leq 2\,\mathrm{c}_{5,4}\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1/2}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^{5/2}H_s \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)(\bar{\sigma}_\varepsilon + \bar{\sigma}_{\text{noise}})(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3) \\
&\leq \mathrm{c}_{5,5}\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1/2}(1+t)^{-\gamma_\eta/2}\,,
\end{aligned} \tag{85}
$$

where

$$
\mathrm{c}_{5,5} = 4\,\mathrm{c}_{5,4}(\eta^{3/2}H + L\eta^{1/2}(1+(2a\eta H)^{\frac{2\gamma+3\gamma_\eta}{1-\gamma}}))(\bar{\sigma}_\varepsilon + \bar{\sigma}_{\text{noise}})(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)\,.
$$

Perturbations of $D_{t,6}$.

Firstly, we note that

$$
\begin{aligned}
D_{t,6} - D_{t,6}^{(s,h,c)} &= N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\sum_{s',h',c'}\eta_{s'}\prod_{i=s'+1}^{t} G_i^{\mathsf{avg}}(\Gamma_{s',h'+1:H_{s'}}^{c',(s,h,c)} - \Gamma_{s',h'+1:H_{s'}}^{c'})\varepsilon^c(Z_{s',h',c'}) \\
&\quad + N^{-1}\eta_t^{-1/2}\Sigma_t^{-1/2}\eta_s\prod_{i=s+1}^{t} G_i^{\mathsf{avg}}(G_{s,h+1:H_s}^c - \Gamma_{s,h+1:H_s}^{c,(s,h,c)})(\varepsilon^c(Z_{s,h,c}) - \varepsilon^c(Z'_{s,h,c})) \\
&= T_{6,1} + T_{6,2}\,.
\end{aligned} \tag{86}
$$

We can see that the first term is zero for $s' \neq s$, $c' \neq c$ and $h' \geq h$. Thus, applying Lemma 15, we get

$$
\begin{aligned}
\mathbb{E}[\|T_{6,1}\|^2] &\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^2 \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\sum_{h'=1}^{h-1}\mathbb{E}[\|\Gamma_{s,h'+1:H_s}^{c,(s,h,c)} - \Gamma_{s,h'+1:H_s}^c\|^2]\mathrm{Tr}(\Sigma_\varepsilon^c) \\
&\leq 16\,\mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^4 H_s \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\mathrm{Tr}(\Sigma_\varepsilon^c)\,.
\end{aligned}
$$

The second term $T_{6,2}$ ca be bounded, as

$$
\begin{aligned}
\mathbb{E}[\|T_{6,2}\|^2] &\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^2 \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathbb{E}[\|G_{s,h+1:H_s}^c - \Gamma_{s,h+1:H_s}^c\|^2]\mathrm{Tr}(\Sigma_\varepsilon^c) \\
&\leq \mathrm{C}_{\mathbf{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^4 H_s \exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\mathrm{Tr}(\Sigma_\varepsilon^c)\,.
\end{aligned}
$$

Combining together the last two inequalities, we obtain

$$\mathbb{E}[\|D_{t,6} - D_{t,6}^{(s,h,c)}\|^2] \leq 34\, \mathrm{C}_{\boldsymbol{\Sigma}}^2\, N^{-1}\eta_t^{-1}\eta_s^4 H_s \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\mathrm{Tr}(\Sigma_\varepsilon^c)\,.$$

Thus, we have

$$\begin{aligned} R_{3,6} &\leq 17\bar{\sigma}_{\text{noise}}\, \mathrm{C}_{\boldsymbol{\Sigma}}^2\, \eta_t^{-1} \sum_{s=1}^{t} \eta_s^3 H_s^{3/2} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \\ &\leq \mathrm{c}_6 (1+t)^{-(\gamma+\gamma_\eta)/2}\,, \end{aligned} \tag{87}$$

where we have set

$$\mathrm{c}_6 = 68\bar{\sigma}_{\text{noise}}(\eta^2 H^{3/2} + La^{-1}\eta H^{1/2}(1 + (2a\eta H)^{\frac{3(\gamma+\gamma_\eta)}{2(1-\gamma)}}))\,.$$

Finalized Berry-Essen bound We return to the decomposition (69). We get

$$R_1 \leq \mathrm{c}_{R,1}\, \mathrm{C}_{\boldsymbol{\Sigma}}^3\, N^{-1/2}\eta^{1/2}(1+t)^{-\gamma_\eta/2}\,.$$

where $\mathrm{c}_{R,1} = 259 d^{1/2} 8 a^{-1}\eta^{1/2}$. For R_2, we apply (72) and combine together (73), (74), (75), (76), (79), (80), thus we obtain

$$\begin{aligned} R_2 \leq \mathrm{c}_{D,1}\, d^{1/2} N^{1/2} \exp\left(-a\eta H t/2\right)\|\theta_0 - \theta_\star\| &+ (\mathrm{c}_{D,2} + \mathrm{c}_{D,3,2} + \mathrm{c}_{D,4}) d^{1/2} N^{1/2} (1+t)^{-\gamma} \\ &+ (\mathrm{c}_{D,3,1} + \mathrm{c}_{D,5} + \mathrm{c}_{D,6}) d^{1/2} N^{-1/2} (1+t)^{-\gamma_\eta/2}\,. \end{aligned}$$

For R_3, we use the bounds (81), (82), (83), (85), (87), and get

$$\begin{aligned} R_3 \leq \mathrm{c}_1 \exp\left(-\frac{a\eta H}{1-\gamma}(1+t)^{1-\gamma}\right)&\|\theta_0 - \theta_\star\| + \mathrm{c}_2 (1+t)^{-(3\gamma-\gamma_\eta)/2} \\ + \mathrm{c}_{3,5}(1+t)^{-3\gamma/2} &+ (\mathrm{c}_{3,4} + \mathrm{c}_{5,5} + \mathrm{c}_6) N^{-1/2}(1+t)^{-\gamma_\eta/2}\,, \end{aligned}$$

where $\gamma = \gamma_\eta - \gamma_H$.

B.3 LOWER BOUND FOR GAUSSIAN APPROXIMATION

To illustrate the inherent limitations of Gaussian approximation, we consider a simple 1-dimensional linear stochastic approximation with a single agent and constant local updates ($H_t = 1$ for all $t \geq 1$):

$$\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}, \qquad \bar{\mathbf{A}} = 1, \quad \bar{\mathbf{b}} = 0.$$

Comparing with the general setting, here we have $\gamma_H = 0$ and $\gamma_\eta = \gamma \in (0,1)$. The unbiased stochastic estimates are given by

$$\mathbf{A}(Z_j) = 1 + \xi_j, \qquad \mathbf{b}(Z_j) = \bar{\mathbf{b}},$$

where $\xi_j \sim \mathcal{N}(0,1)$. Then, FEDLSA reduces to

$$\theta_t = \theta_{t-1} - \eta_t(\mathbf{A}(Z_t)\theta_{t-1} - \mathbf{b}(Z_t)), \quad \theta_0 = 0. \tag{88}$$

Unrolling the recursion gives

$$\theta_t = -\sum_{j=1}^{t} \eta_j \prod_{\ell=j+1}^{t} (1-\eta_\ell)\xi_j,$$

and its variance under step-size scaling satisfies

$$\eta_t^{-1/2}\theta_t \sim \mathcal{N}(0, v_t), \quad v_t := \mathrm{Var}[\eta_t^{-1/2}\theta_t] = \eta_t^{-1} \sum_{j=1}^{t} \eta_j^2 \prod_{\ell=j+1}^{t} (1-\eta_\ell)^2. \tag{89}$$

Under this setting, we can state the following results

Proposition 3. *Assume that $\eta_1 = 0$ and $\eta_\ell = (1+\ell)^{-\gamma}$ with $0 < \gamma < 1$. For 1-dimensional single agent* FedLSA *with $\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}, \bar{\mathbf{A}} = 1, \bar{\mathbf{b}} = 0$ and $\mathbf{A}(Z_j) = 1 + \xi_j, \mathbf{b}(Z_j) = \bar{\mathbf{b}}$ with $\xi_j \sim \mathcal{N}(0,1)$ it holds that*

$$\lim_{t\to\infty} v_t := \lim_{t\to\infty} \mathrm{Var}[\eta_t^{-1/2}\theta_t] = \frac{1}{2}\,,$$

where v_t is defined in (89). Moreover, for any $t \geq 4$, it can be shown that

$$\left|v_t - \frac{1}{2}\right| \geq \mathrm{C}_{\mathbf{v}}(t^{\gamma-1} + t^{-\gamma})\,,$$

for some positive constant $\mathrm{C}_{\mathbf{v}} > 0$ that depends only on γ.

From Proposition 3 it follows that the limiting variance of v_t is $\sigma_\infty^2 = \frac{1}{2}$. Moreover, Proposition 3 gives us the lower bound for $|v_t - \sigma_\infty^2| = |v_t - \frac{1}{2}| \geq \mathrm{C}_{\mathbf{v}}(t^{\gamma-1} + t^{-\gamma})$. Consequently, by Devroye et al. [2018, Theorem 1.1], the convex distance between $\eta_t^{-1/2}(\theta_t - \theta_\star)$ and limiting distribution $\mathcal{N}(0, \sigma_\infty^2)$ satisfies

$$\mathsf{d}_{\mathsf{C}}\big(\eta_t^{-1/2}(\theta_t - \theta_\star), \mathcal{N}(0,\sigma_\infty^2)\big) \gtrsim t^{\gamma-1} + t^{-\gamma},$$

and the statement follows.

Proof of Proposition 3. We begin by noting that

$$\eta_j^2 = \frac{1}{2}\eta_j(1 - (1-\eta_j)^2) + \frac{1}{2}\eta_j^3\,.$$

Denoting $P_{j,t} := \prod_{\ell=j+1}^{t}(1-\eta_\ell)^2$, we can write

$$\begin{aligned}
\eta_t^{-1} v_t &= \frac{1}{2}\sum_{j=1}^{t}\eta_j(P_{j,t} - P_{j-1,t}) + \frac{1}{2}\sum_{j=1}^{t}\eta_j^3 P_{j,t} \\
&= \frac{1}{2}S_{t,1} + \frac{1}{2}S_{t,2}\,.
\end{aligned}$$

We first consider the term $S_{t,1}$. By applying Abel's summation by parts, we obtain

$$\begin{aligned}
\sum_{j=1}^{t}\eta_j(P_{j,t} - P_{j-1,t}) &= \eta_t P_{t,t} - \eta_1 P_{0,t} + \sum_{j=1}^{t-1}(\eta_j - \eta_{j+1})P_{j,t} \\
&= \eta_t + \sum_{j=1}^{t-1}(\eta_j - \eta_{j+1})P_{j,t} - \eta_1 P_{0,t}\,.
\end{aligned}$$

Using the inequalities $(j+1)^{-\gamma} - (j+2)^{-\gamma} \geq \frac{\gamma}{2}(j+1)^{-\gamma-1}$ and $\log(1-x) \geq -4x$ for $x \in [0, 1/2]$, we have

$$\begin{aligned}
\sum_{j=1}^{t-1}(\eta_j - \eta_{j+1})P_{j,t} &\geq \frac{\gamma}{2}\sum_{j=1}^{t-1}(j+1)^{-\gamma-1}\exp\Big(-4\sum_{\ell=j+1}^{t}\eta_\ell\Big) \\
&\geq \frac{\gamma}{2}\sum_{j=1}^{t-1}(j+1)^{-\gamma-1}\exp\Big(-\frac{4}{1-\gamma}\big(t^{1-\gamma} - (j+1)^{1-\gamma}\big)\Big) \\
&= \frac{\gamma}{2}\exp\Big(-\frac{4}{1-\gamma}t^{1-\gamma}\Big)\sum_{j=1}^{t-1}(j+1)^{-\gamma-1}\exp\Big(\frac{4}{1-\gamma}(j+1)^{1-\gamma}\Big)\,.
\end{aligned}$$

Since the function $f(x) = (x+1)^{-\gamma-1}\exp\big(\frac{4}{1-\gamma}(x+1)^{1-\gamma}\big)$ is strictly increasing for $x \geq 1$, we can bound the sum by the corresponding integral:

$$
\begin{aligned}
\sum_{j=1}^{t-1}(\eta_j-\eta_{j+1})P_{j,t} &\geq \frac{\gamma}{2}\exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)\int_1^{t-1}(x+1)^{-\gamma-1}\exp\left(\frac{4}{1-\gamma}(x+1)^{1-\gamma}\right)\mathrm{d}x \\
&= \frac{\gamma}{2}t^{-\gamma}\exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)\int_{2/t}^{1}u^{-(1+\gamma)}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}t^{1-\gamma}\right)\mathrm{d}u \\
&\geq \frac{\gamma}{2}t^{-\gamma}\exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)\int_{1/2}^{1}u^{-(1+\gamma)}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}t^{1-\gamma}\right)\mathrm{d}u\,.
\end{aligned}
$$

For $t \geq 4$ we may estimate asymptotic behaviour of the integral using Laplace approximation introduced in, for instance, Fedoryuk [1977, Theorem 1.2] and Olver and Rheinbolt [2014]:

$$
\frac{\gamma}{2}t^{-\gamma}\int_{1/2}^{1}u^{-(1+\gamma)}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}t^{1-\gamma}\right)\mathrm{d}u = \frac{\gamma}{8}t^{-1}\left(1+\mathcal{O}(t^{\gamma-1})\right)\,.
$$

Therefore, as $\eta_1 P_{0,t} \leq \prod_{\ell=1}^{t}(1-\eta_\ell)^2$ is exponentially small, we may write

$$
\begin{cases}
S_{t,1} = \eta_t + R_{t,1}\,, \\
R_{t,1} \geq \mathrm{C}_{\mathbf{v},\mathbf{1}}\, t^{-1}
\end{cases}
$$

for some positive constant $\mathrm{C}_{\mathbf{v},\mathbf{1}}$ depending only on γ.

Next, we consider $S_{t,2}$. Following a similar approach as for $S_{t,1}$ and approximating the sum with an integral, we have

$$
\begin{aligned}
S_{t,2} = \sum_{j=1}^{t}\eta_j^3 P_{j,t} &\geq \sum_{j=1}^{t}(1+j)^{-3\gamma}\exp\left(-4\sum_{\ell=j+1}^{t}\eta_\ell\right) \\
&\geq \exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)\sum_{j=1}^{t}(j+1)^{-3\gamma}\exp\left(\frac{4}{1-\gamma}(j+1)^{1-\gamma}\right) \\
&\geq \exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)\int_1^t(x+1)^{-3\gamma}\exp\left(\frac{4}{1-\gamma}(x+1)^{1-\gamma}\right) \\
&= \exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)(t+1)^{1-3\gamma}\int_{2/(t+1)}^{1}u^{-3\gamma}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}(t+1)^{1-\gamma}\right)\mathrm{d}u \\
&\geq \exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)(t+1)^{1-3\gamma}\int_{1/2}^{1}u^{-3\gamma}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}(t+1)^{1-\gamma}\right)\mathrm{d}u\,.
\end{aligned}
$$

Here we also use the fact that function $f(x) = (x+1)^{-3\gamma}\exp\left(\frac{4}{1-\gamma}\right)$ is strictly increasing on $[1,\infty)$. Now we determine the asymptotic behaviour of the following integral using Laplace approximation

$$
\exp\left(-\frac{4}{1-\gamma}t^{1-\gamma}\right)(t+1)^{1-3\gamma}\int_{1/2}^{1}u^{-3\gamma}\exp\left(\frac{4}{1-\gamma}u^{1-\gamma}(t+1)^{1-\gamma}\right)\mathrm{d}u = \frac{1}{4}t^{-2\gamma}(1+\mathcal{O}(t^{\gamma-1}))\,.
$$

Finally, taking $\mathrm{C}_{\mathbf{v}} := \min(\mathrm{C}_{\mathbf{v},\mathbf{1}}, \mathrm{C}_{\mathbf{v},\mathbf{2}})$ yields the desired lower bound

$$
\left|v_t - \frac{1}{2}\right| \geq \mathrm{C}_{\mathbf{v}}\left(t^{\gamma-1}+t^{-\gamma}\right)\,.
$$

□

C MULTIPLIER BOOTSTRAP VALIDITY

We start with the expression for t_0 in **A**4.

A' 4. *We assume that*

$$t \geq t_0 = \max\left(N, \mathrm{e}^3, \left(\frac{4C_{\infty,1}}{2\lambda_{\min}(\bar{\mathbf{A}})\lambda_{\min}(\Sigma_\infty)}\right)^{1/(1-\gamma)}, \left(\frac{4C_{\infty,2}}{2\lambda_{\min}(\bar{\mathbf{A}})\lambda_{\min}(\Sigma_\infty)}\right)^{1/\gamma}, n_0\right)$$

where n_0 satisfies

$$\begin{cases} \left(\frac{2\,\mathrm{c}_U\,\eta}{N^3}\right)^{1/2}(1+n_0)^{-\gamma_\eta/2}\sqrt{\log(2dn_0)} + \frac{U_{\max}\eta}{3N^2}(1+n_0)^{-\gamma_\eta}\log(2dn_0) & \leq N^{-1}\lambda_{\min}(\Sigma_*^{\mathsf{avg}})/2\,, \\ \bar{\mathrm{c}}^{\mathsf{b}}_{M,2}N^{-2}\log^6(n_0)(1+n_0)^{-\gamma_\eta} + \bar{\mathrm{c}}^{\mathsf{b}}_{M,3}N^{-1}\log^3(n_0)(1+n_0)^{-2\gamma} & \leq N^{-1}\lambda_{\min}(\Sigma_*^{\mathsf{avg}})/2\,, \end{cases}$$

for constants $\bar{\mathrm{c}}^{\mathsf{b}}_{M,2}, \bar{\mathrm{c}}^{\mathsf{b}}_{M,3}$ defined in Lemma 14 and $\mathrm{c}_U, U_{\max}$ from Lemma 12.

We set the events under which we will consider the normal approximation in bootstrap world

$$\begin{aligned} &\Omega_1 = \{Z \in \mathcal{Z} : \forall c \in [N]\,, \|(\mathrm{I} - \eta\mathbf{A}^c(Z))u\| \leq \mathrm{e}(1-\eta a)\|u\|\}\,, \\ &\Omega_2 = \{(\mathbb{E}^{\mathsf{b}}[\|D_t^{\mathsf{b}}\|^p])^{1/p} \leq \bar{\mathrm{c}}_1^{\mathsf{b}} p^6\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, N^{-1}(1+t)^{1/p-\gamma_\eta/2} + \bar{\mathrm{c}}_2^{\mathsf{b}} p^4\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, N^{-1/2}(1+t)^{1/p-\gamma}\}\,, \\ &\Omega_3 = \{\|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \lesssim_{\log_t} N^{-3/2}t^{-\gamma_\eta/2} + N^{-1}t^{-2\gamma}\} \\ &\Omega_0 = \Omega_1 \cap \Omega_2 \cap \Omega_3\,. \end{aligned} \tag{90}$$

In addition, throughout this section we use the notation

$$\mathcal{F}_{s,h}^{\mathsf{b},+} = \sigma(w_{t,k,c}, Z_{t,k,c} \mid t > s \text{ or } t = s \text{ and } k \geq h)\,, \quad \mathcal{F}_{s,h}^{\mathsf{b},-} = \sigma(w_{t,k,c}, Z_{t,k,c} \mid t < s \text{ or } t = s \text{ and } k \leq h)\,.$$

We start with the definition of the bootstrap procedure. We consider a set of random variables $\mathcal{W}^t = \{w_{s,h,c}, s \in [t], h \in [H_s], c \in [N]\}$, where $w_{s,h,c}$ are i.i.d and independent of $\mathcal{Z}^t = \{Z_{s,h,c}, s \in [t], h \in [H_s], c \in [N]\}$, with $\mathbb{E}[w_{1,1,1}] = 1$, $\mathrm{Var}[w_{1,1,1}] = 1$, and for any $p > 2$, we set $\mathbb{E}^{1/p}[|w_{s,h,c} - 1|^p] = m_p < \infty$. Then, we denote $\mathbb{P}^{\mathsf{b}} = \mathbb{P}(\cdot|\mathcal{Z}^t)$ and $\mathbb{E}^{\mathsf{b}} = \mathbb{E}(\cdot|\mathcal{Z}^t)$. According to multiplier bootstrap procedure, we slightly modify the iterations (6) introducing multiplicative noise, that is,

$$\theta_{t,h}^{\mathsf{b},c} = \theta_{t,h-1}^{\mathsf{b},c} - \eta_t w_{t,h,c}(\mathbf{A}^c(Z_{t,h,c})\theta_{t,h-1}^{\mathsf{b},c} - \mathbf{b}^c(Z_{t,h,c}))\,. \tag{91}$$

C.1 PROOF OF THE THEOREM 4

Proof. The goal of the Theorem 4 is to estimate the quantity

$$\sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \left|\mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) - \mathbb{P}(\eta_t^{-1/2}(\theta_t - \theta_\star) \in B)\right|\,.$$

From now, we restrict ourselves to the event Ω_0. Restricting to this event, we obtain that, with triangle inequality

$$\begin{aligned} &\sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \left|\mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) - \mathbb{P}(\eta_t^{-1/2}(\theta_t - \theta_\star) \in B)\right| \\ &\quad \leq \mathsf{d}_{\mathsf{C}}^{\mathsf{b}}\big(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t), Y^{\mathsf{b}}\big) + \mathsf{d}_{\mathsf{C}}\big(\eta_t^{-1/2}(\theta_t - \theta_\star), Y\big) + \sup_{B \in \mathrm{Conv}(\mathbb{R}^d)} \left|\mathbb{P}(Y \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B)\right|\,. \end{aligned}$$

Now we control the first term using Theorem 5, second one using Theorem 3 and the third one using Lemma 21 together with

$$\|\Sigma_t^{-1/2}\Sigma_t^{\mathsf{b}}\Sigma_t^{-1/2} - \mathrm{I}\| \leq \|\Sigma_t^{-1}\|\|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \lesssim_{\log_t} \mathrm{C}_{\mathbf{\Sigma}}^2(N^{-1/2}t^{-\gamma_\eta/2} + t^{-2\gamma})\,.$$

We show that $\mathbb{P}(\Omega_0) \geq 1 - \frac{4}{t}$ in Lemma 9. □

Lemma 9. *Let $\Omega_0 = \Omega_1 \cap \Omega_2 \cap \Omega_3$ be the event defined in (90). Then*

$$\mathbb{P}(\Omega_0) \geq 1 - \frac{4}{t},$$

where $t \geq 1$ denotes the number of global iterations of `FedLSA`.

Proof. We first bound the probability of Ω_2^c. By Markov's inequality,

$$\begin{aligned}\mathbb{P}(\Omega_2^c) &\leq \frac{\mathbb{E}[\|D_t^{\mathsf{b}}\|^p]}{\left(\bar{c}_1^{\mathsf{b}} p^6 \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}} \, N^{-1/2}(1+t)^{1/p-\gamma_\eta/2} + \bar{c}_2^{\mathsf{b}} p^4 \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}(1+t)^{1/p-\gamma}\right)^p} \\ &\leq \frac{\mathbb{E}[\|D_t^{\mathsf{b}}\|^p]}{t\left(\bar{c}_1^{\mathsf{b}} p^6 \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}} \, N^{-1/2}(1+t)^{-\gamma_\eta/2} + \bar{c}_2^{\mathsf{b}} p^4 \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}(1+t)^{-\gamma}\right)^p} \leq \frac{1}{t}.\end{aligned}$$

Thus, $\mathbb{P}(\Omega_2) \geq 1 - \frac{1}{t}$.

Next, applying Lemma 10 with $p = \log(Nt)$ and using a union bound over $c \in [N]$, we obtain that on the event Ω_1,

$$\|(\mathrm{I} - \eta \mathbf{A}^c(z))u\| \leq \mathrm{e}(1 - \eta a)\|u\|,$$

and $\mathbb{P}(\Omega_1) \geq 1 - \frac{1}{t}$.

Further, by Proposition 6, we have

$$\mathbb{P}(\Omega_3) \geq 1 - \frac{2}{t}.$$

Finally, applying the union bound to the events Ω_1, Ω_2, and Ω_3 yields

$$\mathbb{P}(\Omega_0) = \mathbb{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \geq 1 - \frac{4}{t},$$

which concludes the proof. □

C.2 RATE OF GAUSSIAN APPROXIMATION IN THE BOOTSTRAP WORLD

Now, our aim is to quantify the convergence rate of the convex distance in the bootstrap world

$$\mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t), Y^{\mathsf{b}}) = \sup_{B \in \mathsf{C}(\mathbb{R}^d)} \left| \mathbb{P}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t) \in B) - \mathbb{P}^{\mathsf{b}}(Y^{\mathsf{b}} \in B) \right| ,$$

where $Y^{\mathsf{b}} \sim \mathrm{N}(0, \Sigma_t^{\mathsf{b}})$ and Σ_t^{b} is defined in (92). We start with the decomposition

$$\begin{aligned}\theta_{t,h}^{\mathsf{b},c} - \theta_{t,h}^c &= (\mathrm{I} - \eta_t \mathbf{A}^c(Z_{t,h,c}))(\theta_{t,h-1}^{\mathsf{b},c} - \theta_{t,h-1}^c) \\ &\quad + \eta_t(\varepsilon^c(\theta_{t,h-1}^{\mathsf{b},c}, Z_{t,h,c}) + \bar{\mathbf{A}}^c \theta_{t,h-1}^{\mathsf{b},c} - \bar{\mathbf{b}}^c) - \eta_t w_{t,h,c}(\varepsilon^c(\theta_{t,h-1}^{\mathsf{b},c}, Z_{t,h,c}) + \bar{\mathbf{A}}^c \theta_{t,h-1}^{\mathsf{b},c} - \bar{\mathbf{b}}^c) \\ &= (\mathrm{I} - \eta_t \mathbf{A}^c(Z_{t,h,c}))(\theta_{t,h-1}^{\mathsf{b},c} - \theta_{t,h-1}^c) - \eta_t(w_{t,h,c} - 1)(\varepsilon^c(Z_{t,h,c}) + \mathbf{A}^c(Z_{t,h,c})(\theta_{t,h-1}^{\mathsf{b},c} - \theta_\star^c)) \\ &= (\mathrm{I} - \eta_t w_{t,h,c} \mathbf{A}^c(Z_{t,h,c}))(\theta_{t,h-1}^{\mathsf{b},c} - \theta_{t,h-1}^c) - \eta_t(w_{t,h,c} - 1)\tilde{\varepsilon}_{t,h}^c \, .\end{aligned}$$

Here, we define

$$\tilde{\varepsilon}_{t,h}^c = \varepsilon^c(Z_{t,h,c}) + \mathbf{A}^c(Z_{t,h,c})(\theta_{t,h-1}^c - \theta_\star^c) \, .$$

We also define the bootstrap counterparts of random matrices product

$$\Gamma_{t,\ell:r}^{\mathsf{b},c} = \prod_{h=l}^{r}(\mathrm{I} - \eta_t w_{t,h,c} \mathbf{A}^c(Z_{t,h,c})) \, , \quad \Gamma_t^{\mathsf{b},c} = \Gamma_{t,1:H_t}^{\mathsf{b},c} \, , \quad \Gamma_t^{\mathsf{b},\mathsf{avg}} = N^{-1} \sum_{c=1}^{N} \Gamma_t^{\mathsf{b},c} \, .$$

Setting $h = H_t$ and unrolling the recursion, we get

$$\theta_{t,H_t}^{\mathsf{b},c} - \theta_{t,H_t}^c = \Gamma_{t,H_t}^{\mathsf{b},c}(\theta_{t,0}^{\mathsf{b},c} - \theta_{t,0}^c) - \eta_t \sum_{h=1}^{H_t}(w_{t,h,c} - 1)\Gamma_{t,h+1:H_t}^{\mathsf{b},c} \tilde{\varepsilon}_{t,h}^c \, ,$$

Averaging over clients, we obtain

$$
\begin{aligned}
\theta_t^{\mathsf{b}} - \theta_t &= N^{-1} \sum_{c=1}^{N} \{\theta_{t,H_t}^{\mathsf{b},c} - \theta_{t,H_t}^{c}\} \\
&= \Gamma_t^{\mathsf{b},\mathsf{avg}}(\theta_{t-1}^{\mathsf{b}} - \theta_{t-1}) - \eta_t \underbrace{N^{-1} \sum_{c=1}^{N} \sum_{h=1}^{H_t} (w_{t,h,c} - 1) \Gamma_{t,h+1:H_t}^{\mathsf{b},c} \tilde{\varepsilon}_{t,h}^{c}}_{\mathcal{E}_t^{\mathsf{b},\mathsf{avg}}} \\
&= \prod_{i=1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}}(\theta_0^{\mathsf{b}} - \theta_0) - \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \mathcal{E}_s^{\mathsf{b},\mathsf{avg}} \,.
\end{aligned}
$$

Again, in this decomposition we isolate the leading term with $\varepsilon^c(Z_{t,h,c})$, that is,

$$
\begin{aligned}
\theta_t^{\mathsf{b}} - \theta_t = & \underbrace{-N^{-1} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^c(Z_{s,h,c})}_{\text{Leading term}} + \prod_{i=1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}}(\theta_0^{\mathsf{b}} - \theta_0) \\
& + N^{-1} \sum_{s=1}^{t} \eta_s \Big(\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \Big) \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^c(Z_{s,h,c}) \\
& + N^{-1} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) (\Gamma_{s,h+1:H_s}^{c} - \Gamma_{s,h+1:H_s}^{\mathsf{b},c}) \varepsilon^c(Z_{s,h,c}) \\
& - N^{-1} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{\mathsf{b},c} \mathbf{A}^c(Z_{s,h,c}) (\theta_{s,h-1}^{c} - \theta_\star^{c}) \,.
\end{aligned}
$$

The second term here vanish since $\theta_0^{\mathsf{b}} = \theta_0$. Further, we can extract the heterogeneous bias from the last term, and get the completed leading term

$$
\begin{aligned}
M_t^{\mathsf{b}} &= -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^c(Z_{s,h,c}) \\
&\quad - N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \mathbf{A}^c(Z_{s,h,c}) (\theta_\star - \theta_\star^{c}) \\
&\quad = -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^c(\theta_\star, Z_{s,h,c}) \,, \\
\Sigma_t^{\mathsf{b}} &= \mathbb{E}^{\mathsf{b}}[M_t^{\mathsf{b}} (M_t^{\mathsf{b}})^T] \,.
\end{aligned}
\tag{92}
$$

Analogously to the case of M_t, we can further specify the leading term of M_t^{b}, as,

$$
\begin{aligned}
M_t^{\mathsf{b}} &= -N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \varepsilon^c(\theta_\star, Z_{s,h,c}) \\
&\quad + N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \left(\prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \right) \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \varepsilon^c(\theta_\star, Z_{s,h,c}) \\
&\quad + N^{-1} \eta_t^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) (\mathrm{I} - \Gamma_{s,h+1:H_s}^{c}) \varepsilon^c(\theta_\star, Z_{s,h,c}) \\
&\quad = M_{t,1}^{\mathsf{b}} + M_{t,2}^{\mathsf{b}} + M_{t,3}^{\mathsf{b}} \,.
\end{aligned}
$$

Therefore, we can state our final decomposition of self-normalized statistic in the boostrap wolrd $\eta_t^{-1/2} (\Sigma_t^{\mathsf{b}})^{-1/2} (\theta_t^{\mathsf{b}} - \theta_t)$ in the following form

$$
\eta_t^{-1/2} (\Sigma_t^{\mathsf{b}})^{-1/2} (\theta_t^{\mathsf{b}} - \theta_t) = W_t^{\mathsf{b}} + D_t^{\mathsf{b}} \,, \tag{93}
$$

where we have set

$$W_t^{\mathsf{b}} = (\Sigma_t^{\mathsf{b}})^{-1/2} M_t^{\mathsf{b}} = \sum_{s=1}^{t} \sum_{h=1}^{H_s} \sum_{c=1}^{N} \upsilon_{s,h,c}^{\mathsf{b}} \,,$$

$$\upsilon_{s,h,c}^{\mathsf{b}} = -N^{-1} \eta_t^{-1/2} \eta_s (\Sigma_t^{\mathsf{b}})^{-1/2} \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^{c}(\theta_\star, Z_{s,h,c}) \,,$$

and $D_t^{\mathsf{b}} = D_{t,1}^{\mathsf{b}} + D_{t,2}^{\mathsf{b}} + D_{t,3}^{\mathsf{b}} + D_{t,4}^{\mathsf{b}} + D_{t,5}^{\mathsf{b}}$, with

$$D_{t,1}^{\mathsf{b}} = \eta_t^{-1/2} N^{-1} (\Sigma_t^{\mathsf{b}})^{-1/2} \sum_{s=1}^{t} \eta_s \Big(\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b,avg}} \Big) \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{c} \varepsilon^{c}(Z_{s,h,c}) \,,$$

$$D_{t,2}^{\mathsf{b}} = \eta_t^{-1/2} N^{-1} (\Sigma_t^{\mathsf{b}})^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b,avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) (\Gamma_{s,h+1:H_s}^{c} - \Gamma_{s,h+1:H_s}^{\mathsf{b},c}) \varepsilon^{c}(Z_{s,h,c}) \,,$$

$$D_{t,3}^{\mathsf{b}} = -\eta_t^{-1/2} N^{-1} (\Sigma_t^{\mathsf{b}})^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b,avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{\mathsf{b},c} \mathbf{A}^{c}(Z_{s,h,c}) (\theta_{s,h-1}^{c} - \theta_\star) \,,$$

$$D_{t,4}^{\mathsf{b}} = \eta_t^{-1/2} N^{-1} (\Sigma_t^{\mathsf{b}})^{-1/2} \sum_{s=1}^{t} \eta_s \Big(\prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b,avg}} \Big) \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) \Gamma_{s,h+1:H_s}^{\mathsf{b},c} \mathbf{A}^{c}(Z_{s,h,c}) (\theta_\star - \theta_\star^{c}) \,,$$

$$D_{t,5}^{\mathsf{b}} = \eta_t^{-1/2} N^{-1} (\Sigma_t^{\mathsf{b}})^{-1/2} \sum_{s=1}^{t} \eta_s \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c} - 1) (\Gamma_{s,h+1:H_s}^{c} - \Gamma_{s,h+1:H_s}^{\mathsf{b},c}) \mathbf{A}^{c}(Z_{s,h,c}) (\theta_\star - \theta_\star^{c}) \,.$$

Using the decomposition (93) together with Shao and Zhang [2022, Theorem 2.1], we are now ready to state the main result of this section.

Theorem 5. *(Gaussian approximation in the "boostrap world") Let $Y^{\mathsf{b}} \sim \mathrm{N}(0, \Sigma_t^{\mathsf{b}})$ and $\eta H < \beta_\infty$. Under the assumptions **A**1($\log(Nt)$), **A**2, **S**2, **A**3 and **A**4, on the event Ω_0, we have*

$$\mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(\eta_t^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t), Y^{\mathsf{b}}) \le M_1^{\mathsf{b}} N^{-1/2} \log^6(t)(1+t)^{-\gamma_\eta/2} + M_2^{\mathsf{b}} \log^4(t)(1+t)^{-\gamma} \,, \tag{94}$$

where the event Ω_0 i defined in (90) and $\mathbb{P}(\Omega_0) \ge 1 - \frac{4}{t}$.

Proof. We start by applying Proposition 5 to the statistics $X = W_t^{\mathsf{b}}$ and $Y = D_t^{\mathsf{b}}$, that is,

$$\mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(\eta_t^{-1/2}(\Sigma_t^{\mathsf{b}})^{-1/2}(\theta_t^{\mathsf{b}} - \theta_t), Z^{\mathsf{b}}) \le \mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(W_t^{\mathsf{b}}, Z^{\mathsf{b}}) + 2c_d^{p/(p+1)} (\mathbb{E}^{\mathsf{b}}[\|D_t^{\mathsf{b}}\|^p])^{1/(1+p)} \,. \tag{95}$$

where $Z^{\mathsf{b}} \sim \mathrm{N}(0, \mathrm{I})$. Therefore, applying Shao and Zhang [2022, Theorem 2.1], we estimate

$$\mathsf{d}_{\mathsf{C}}^{\mathsf{b}}(W_t^{\mathsf{b}}, Y^{\mathsf{b}}) \le 259 d^{1/2} \Upsilon_t^{\mathsf{b}} \,,$$

where $\Upsilon_t^{\mathsf{b}} = \sum_{s=1}^{t} \sum_{h=1}^{H_s} \sum_{c=1}^{N} \mathbb{E}^{\mathsf{b}}[\|\upsilon_{s,h,c}\|^3]$. We note that for any $s \in [t], h \in [H_s]$ and $c \in [N]$, using Proposition 6, we have

$$\|\upsilon_{s,h,c}\| \le \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}} \, N^{-1/2} \eta_t^{-1/2} \eta_s |w_{s,h,c} - 1| \|\varepsilon\|_\infty \,.$$

Thus, applying Lemma 25, on the event Ω_1, we get

$$\begin{aligned}
\Upsilon_t^{\mathsf{b}} &\le \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}^3 \, N^{-3/2} \eta_t^{-3/2} \sum_{s=1}^{t} \eta_s^3 \sum_{h=1}^{H_s} \sum_{c=1}^{N} \mathbb{E}[|w_{s,h,c} - 1|^3] \| \prod_{i=1}^{t} \Gamma_i^{\mathsf{avg}} \|^3 \| \Gamma_{s,h+1:H_s}^{c} \|^3 \|\varepsilon\|_\infty^3 \\
&\le W_{\max}^3 \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}^3 \, N^{-1/2} \eta_t^{-3/2} \sum_{s=1}^{t} \eta_s^3 H_s \exp\left(-3a \sum_{i=s+1}^{t} \eta_i H_i \right) \|\varepsilon\|_\infty^3 \\
&\le 2 W_{\max}^3 (\eta^2 H + L a^{-1} \eta (1 + (3a\eta H)^{\frac{2\gamma_\eta + \gamma}{1-\gamma}})) \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}^3 \, N^{-1/2} (1+t)^{-\gamma_\eta/2} \|\varepsilon\|_\infty^3 \\
&= \mathrm{c}_{\Upsilon}^{\mathsf{b}} \, \mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}^3 \, N^{-1/2} (1+t)^{-\gamma_\eta/2} \|\varepsilon\|_\infty^3 \,.
\end{aligned}$$

Then, on the event Ω_2 with $p = \log(t) - 1$, applying Proposition 4, we bound the second term of (95), as

$$
\begin{aligned}
(\mathbb{E}^{\mathsf{b}}[\|D_t^{\mathsf{b}}\|^p])^{1/(p+1)} &\le (\bar{\mathrm{c}}_1^{\mathsf{b}} p^6 \,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, N^{-1/2}(1+t)^{1/p-\gamma_\eta/2} + \bar{\mathrm{c}}_2^{\mathsf{b}} p^4 \,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}(1+t)^{1/p-\gamma})^{p/(p+1)} \\
&\le (1+\bar{\mathrm{c}}_1^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}})^{p/(p+1)} p^6 N^{-1/2}(1+t)^{1/(p+1)-\gamma_\eta/2}(1+t)^{\frac{\gamma_\eta}{2(p+1)}} \\
&+ (1+\bar{\mathrm{c}}_2^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}})^{p/(p+1)} p^4 (1+t)^{1/(p+1)-\gamma+\gamma/(p+1)} \\
&\le \mathrm{e}^{2+\gamma_\eta/2}(1+\bar{\mathrm{c}}_1^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}) N^{-1/2}\log^6(t)(1+t)^{-\gamma_\eta/2} \\
&+ \mathrm{e}^{3/2+\gamma}(1+\bar{\mathrm{c}}_2^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}})\log^4(t)(1+t)^{-\gamma}\,.
\end{aligned}
$$

Setting the constants

$$
\begin{aligned}
M_1^{\mathsf{b}} &= \mathrm{c}_{\Upsilon}^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}^3\,\|\varepsilon\|_\infty^3 + \mathrm{e}^{2+\gamma_\eta/2}(1+\bar{\mathrm{c}}_1^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}})\,, \\
M_2^{\mathsf{b}} &= \mathrm{e}^{3/2+\gamma}(1+\bar{\mathrm{c}}_2^{\mathsf{b}}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}})\,,
\end{aligned}
$$

we obtain the result (94). □

In the following proposition, we assume that the expectation is taken over the probability space generated by $\mathcal{W}^t$ and $\mathcal{Z}^t$.

Proposition 4. *Under Assumptions **A**1, **A**2(p), **S**2, **A**3, and **A**4, we get*

$$
\mathbb{E}^{1/p}[\|D_t^{\mathsf{b}}\|^p] \le \bar{\mathrm{c}}_1^{\mathsf{b}} p^6 N^{-1/2}(1+t)^{-\gamma_\eta/2} + \bar{\mathrm{c}}_2^{\mathsf{b}} p^4 (1+t)^{-\gamma}\,,
$$

where the constants $\bar{\mathrm{c}}_1$, $\bar{\mathrm{c}}_2$ are defined in (96).

Proof. Firstly, we apply the Minkowski's inequality, that is,

$$
\mathbb{E}^{1/p}[\|D_t^{\mathsf{b}}\|^p] \le \mathbb{E}^{1/p}[\|D_{t,1}^{\mathsf{b}}\|^p] + \mathbb{E}^{1/p}[\|D_{t,2}^{\mathsf{b}}\|^p] + \mathbb{E}^{1/p}[\|D_{t,3}^{\mathsf{b}}\|^p] + \mathbb{E}^{1/p}[\|D_{t,4}^{\mathsf{b}}\|^p] + \mathbb{E}^{1/p}[\|D_{t,5}^{\mathsf{b}}\|^p]\,.
$$

To bound $D_{t,1}^{\mathsf{b}}$, using that the sequence $\{\left(\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{b,avg}}\right)(w_{s,h,c}-1)\Gamma_{s,h+1:H_s}^c \varepsilon^c(Z_{s,h,c})\}$ is martingale-difference w.r.t filtration $\mathcal{F}_{s,h}^{\mathsf{b},+}$ and applying Burkholder's inequality [Osekowski, 2012, Theorem 8.6], together with Lemma 25 and Lemma 20, we get

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|D_{t,1}^{\mathsf{b}}\|^p] \\
&\le p\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, N^{-1/2}\eta_t^{-1/2}\left(\sum_{s=1}^t \eta_s^2 \mathbb{E}^{2/p}[\|\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{b,avg}}\|^p]\mathbb{E}^{2/p}[\|\sum_{c=1}^N\sum_{h=1}^{H_s}(w_{s,h,c}-1)\Gamma_{s,h+1:H}^c\varepsilon^c(Z_{s,h,c})\|^p]\right)^{1/2} \\
&\le W_{\max} p^3\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,\eta_t^{-1/2}\|\varepsilon\|_\infty\left(\sum_{s=1}^t \eta_s^2 H_s \mathbb{E}^{2/p}[\|\prod_{i=s+1}^t \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^t \Gamma_i^{\mathsf{b,avg}}\|^p]\right)^{1/2} \\
&\le \mathrm{e}^{1/2}a^{-1/2}\,\mathrm{C}_{\mathbf{A}}\,W_{\max}p^6\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,N^{-1/2}\eta_t^{-1/2}\|\varepsilon\|_\infty\left(\sum_{s=1}^t \eta_s^2 H_s \exp\left(-a\sum_{\ell=s+1}^t \eta_\ell H_\ell\right)\right)^{1/2} \\
&\le \mathrm{c}_1^{\mathsf{b}} p^6 N^{-1/2}(1+t)^{-\gamma_\eta/2}\,,
\end{aligned}
$$

where, again applying Burkholder's inqequality, we used the bound

$$
\begin{aligned}
\mathbb{E}^{1/p}[\|\sum_{c=1}^N\sum_{h=1}^{H_s}(w_{s,h,c}-1)\Gamma_{s,h+1:H}^c\varepsilon^c(Z_{s,h,c})\|^p] &\le p\left(\sum_{c=1}^N \mathbb{E}^{2/p}[\|\sum_{h=1}^{H_s}(w_{s,h,c}-1)\Gamma_{s,h+1:H_s}^c\varepsilon^c(Z_{s,h,c})\|^p]\right)^{1/2} \\
&\le W_{\max}p^2 N^{1/2} H_s^{1/2}\|\varepsilon\|_\infty\,,
\end{aligned}
$$

and have set

$$
\mathrm{c}_1^{\mathsf{b}} = 2\mathrm{e}^{1/2}a^{-1/2}\,\mathrm{C}_{\mathbf{A}}\,W_{\max}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}(\eta^{1/2}H^{1/2} + a^{-1/2}L^{1/2}(1+(a\eta H)^{\frac{\gamma+\gamma_\eta}{2(1-\gamma)}}))\|\varepsilon\|_\infty\,.
$$

Further, using the martingale-difference structure of the sequence $\{\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{b},\mathsf{avg}}(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c}), 1\leq s\leq t\}$ w.r.t filtration $\mathcal{F}^{\mathsf{b},+}_{s,h}$, applying Burkholder's inequality and Lemma 25, we bound $D_{t,2}$, as

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,2}\|^p] \\
&\quad\leq p\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,N^{-1/2}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p]\mathbb{E}^{2/p}[\|\sum_{c=1}^{N}\sum_{h=1}^{H_s}(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})\|^p]\right)^{1/2} \\
&\quad\leq \mathrm{C}_{\mathbf{A}}\,W_{\max}p^4\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^4H_s^2\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\right)^{1/2}\|\varepsilon\|_\infty \\
&\quad\leq \mathrm{c}^{\mathsf{b}}_2\,p^4(1+t)^{-(\gamma_\eta+\gamma)/2}\,,
\end{aligned}
$$

where, applying Burkholder's inequality and the fact that the sequence $\{(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c}), 1\leq h\leq H_s\}$is martingale-difference w.r.t filtration $\mathcal{F}^{\mathsf{b},+}_{s,h}$, we used the bound

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|\sum_{c=1}^{N}\sum_{h=1}^{H_s}(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})\|^p] \\
&\qquad\qquad\leq p\left(\sum_{c=1}^{N}\mathbb{E}^{2/p}[\|\sum_{h=1}^{H_s}(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})\|^p]\right)^{1/2} \\
&\qquad\qquad\leq \mathrm{C}_{\mathbf{A}}\,W_{\max}p^3N^{1/2}\eta_sH_s\|\varepsilon\|_\infty\,,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|\sum_{h=1}^{H_s}(w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c})\|^p] \\
&\qquad\qquad\leq p\|\varepsilon\|_\infty\left(\sum_{h=1}^{H_s}\mathbb{E}^{2/p}[|w_{s,h,c}-1|^p]\mathbb{E}^{2/p}[\|\Gamma^c_{s,h+1:H_s}-\Gamma^{\mathsf{b},c}_{s,h+1:H_s}\|^p]\right)^{1/2} \\
&\qquad\qquad\leq \mathrm{C}_{\mathbf{A}}\,W_{\max}p^2\eta_sH_s\|\varepsilon\|_\infty\,.
\end{aligned}
$$

Here, we also have set

$$
\mathrm{c}^{\mathsf{b}}_2 = 4\,\mathrm{C}_{\mathbf{A}}\,W_{\max}\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}(\eta H + a^{-1/2}L^{1/2}(\eta H)^{1/2}(1+(2a\eta H)^{\frac{\gamma+\gamma_\eta}{1-\gamma}}))\|\varepsilon\|_\infty\,.
$$

Now, in order to bound $D^{\mathsf{b}}_{t,4}$, we note that the sequence $\{\left(\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{b},\mathsf{avg}}\right)(w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H_s}\varepsilon^c(Z_{s,h,c}), 1\leq h\leq H_s\}$ is martingale-difference w.r.t the filtration $\mathcal{F}^{\mathsf{b},+}_{s,h}$. Then, applying Burkholder's inequality together with Lemma 25 and Lemma 20, defining

$$
S^{\mathsf{b}}_{s,4} = \mathbb{E}^{2/p}[\|\sum_{c=1}^{N}\sum_{h=1}^{H_s}(w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H}\mathbf{A}^c(Z_{s,h,c})(\theta_\star-\theta^c_\star)\|^p]\,,
$$

we obtain

$$
\begin{aligned}
\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,4}\|^p] &\leq p\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,N^{-1/2}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p]S^{\mathsf{b}}_{s,4}\right)^{1/2} \\
&\leq \mathrm{C}_{\mathbf{A}}\,\zeta_3W_{\max}p^3\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2H_s\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p]\right)^{1/2} \\
&\leq e^{1/2}a^{-1/2}\,\mathrm{C}^2_{\mathbf{A}}\,\zeta_3W_{\max}p^6\,\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\,N^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2H_s\exp\left(-a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\right)^{1/2} \\
&\leq \mathrm{c}^{\mathsf{b}}_4\,p^6N^{-1/2}(1+t)^{-\gamma_\eta/2}\,,
\end{aligned}
$$

where we have used that

$$\begin{aligned}
&\mathbb{E}^{2/p}[\| \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H} \mathbf{A}^c(Z_{s,h,c})(\theta_\star - \theta^c_\star)\|^p] \\
&\qquad\qquad \leq p \left(\sum_{c=1}^{N} \mathbb{E}^{2/p}[\| \sum_{h=1}^{H_s} (w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H_s} \mathbf{A}^c(Z_{s,h,c})(\theta_\star - \theta^c_\star)\|^p] \right)^{1/2} \\
&\qquad\qquad \leq \mathrm{C}_{\mathbf{A}}\, \zeta_3 W_{\max} p^2 N^{1/2} H_s^{1/2}\,,
\end{aligned}$$

and have set

$$\mathrm{c}^{\mathsf{b}}_4 = 2\mathrm{e}^{1/2} a^{-1/2} \zeta_3\, \mathrm{C}^2_{\mathbf{A}}\, W_{\max}\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} (\eta^{1/2} H^{1/2} + a^{-1/2} L^{1/2} (1 + (a\eta H)^{\frac{\gamma+\gamma_\eta}{2(1-\gamma)}}))\,.$$

Finally, to bound $D_{t,5}$, we again use the martingale-difference structure of the underlying sequence $\{\prod_{i=s+1}^{t} \Gamma^{\mathsf{b},\mathsf{avg}}_i (w_{s,h,c} - 1)(\Gamma^c_{s,h+1:H_s} - \Gamma^{\mathsf{b},c}_{s,h+1:H_s})\varepsilon^c(Z_{s,h,c}), 1 \leq s \leq t\}$ w.r.t the filtration $\mathcal{F}^{\mathsf{b},+}_{s,h}$ and apply Burkholder's inequality together with Lemma 25, thus, denoting

$$S^{\mathsf{b}}_{s,5} = \mathbb{E}^{2/p}[\| \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s} - \Gamma^{\mathsf{b},c}_{s,h+1:H_s}) \mathbf{A}^c(Z_{s,h,c})(\theta_\star - \theta^c_\star)\|^p]\,,$$

we obtain

$$\begin{aligned}
\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,5}\|^p] &\leq p\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, N^{-1/2} \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^2 \mathbb{E}^{2/p}[\| \prod_{i=s+1}^{t} \Gamma^{\mathsf{avg}}_i \|^p] S^{\mathsf{b}}_{s,5} \right)^{1/2} \\
&\leq \mathrm{C}^2_{\mathbf{A}}\, \zeta_3 W_{\max} p^4\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^4 H_s^2 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \\
&\leq \mathrm{c}^{\mathsf{b}}_5\, p^4 (1+t)^{-(\gamma_\eta+\gamma)/2}\,,
\end{aligned}$$

where we have used the bound

$$\begin{aligned}
&\mathbb{E}^{2/p}[\| \sum_{c=1}^{N} \sum_{h=1}^{H_s} (w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s} - \Gamma^{\mathsf{b},c}_{s,h+1:H_s}) \mathbf{A}^c(Z_{s,h,c})(\theta_\star - \theta^c_\star)\|^p] \\
&\qquad\qquad \leq p \left(\sum_{c=1}^{N} \mathbb{E}^{2/p}[\| \sum_{h=1}^{H_s} (w_{s,h,c}-1)(\Gamma^c_{s,h+1:H_s} - \Gamma^{\mathsf{b},c}_{s,h+1:H_s}) \mathbf{A}^c(Z_{s,h,c})(\theta_\star - \theta^c_\star)\|^p] \right)^{1/2} \\
&\qquad\qquad \leq W_{\max} \zeta_3\, \mathrm{C}^2_{\mathbf{A}}\, p^3 N^{1/2} \eta_s H_s\,,
\end{aligned}$$

and have set

$$\mathrm{c}^{\mathsf{b}}_5 = 4 W_{\max}\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}}\, \zeta_3\, \mathrm{C}^2_{\mathbf{A}} (\eta H + a^{-1/2} L^{1/2} (\eta H)^{1/2} (1 + (2a\eta H)^{\frac{\gamma+\gamma_\eta}{1-\gamma}}))\,.$$

For the term $D^{\mathsf{b}}_{t,3}$, we first note that

$$\theta^c_{s,h-1} - \theta_\star = \Gamma^c_{s,h-1}(\theta_{s-1} - \theta_\star) - \eta_s \sum_{j=1}^{h-1} \Gamma^c_{s,j+1:h-1} \varepsilon^c(Z_{s,h,c}) + (\mathrm{I} - \Gamma^c_{s,h-1})(\theta^c_\star - \theta_\star)\,.$$

This decomposition separates $D^{\mathsf{b}}_{t,3}$ into three terms, that is, $D^{\mathsf{b}}_{t,3} = D^{\mathsf{b}}_{t,3,1} + D^{\mathsf{b}}_{t,3,2} + D^{\mathsf{b}}_{t,3,3}$. For the term $D_{t,3,1}$, using the martingale-difference structure of the sequence $\{\prod_{i=s+1}^{t} \Gamma^{\mathsf{b},\mathsf{avg}}_i (w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H_s} \mathbf{A}^c(Z_{s,h,c})\Gamma^c_{s,h-1}(\theta_{s-1} - \theta_\star)\}$ w.r.t the filtration $\mathcal{F}^{\mathsf{b},+}_{s,h}$ and applying Burkholder's inequality together with Corollary 5, setting

$$S^{\mathsf{b}}_{s,3,1} = \sum_{c=1}^{N} \sum_{h=1}^{H_s} \mathbb{E}^{2/p}[\|(w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H_s} \mathbf{A}^c(Z_{s,h,c})\Gamma^c_{s,h-1}(\theta_{s-1} - \theta_\star)\|^p]\,,$$

we get

$$
\begin{aligned}
&\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,3,1}\|^p] \\
&\quad \leq p^3 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, N^{-1/2} \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^2 S^{\mathsf{b}}_{s,3,1} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \\
&\quad \leq \mathrm{C}_{\mathbf{A}} \, W_{\max} p^3 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^2 H_s \mathbb{E}^{2/p}[\|\theta_{s-1} - \theta_\star\|^p] \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \\
&\quad \leq 4\mathrm{e} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} p^3 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, \eta_t^{-1/2} \left(\exp\left(-2a \sum_{\ell=1}^{t} \eta_\ell H_\ell \right) \sum_{s=1}^{t} \eta_s^2 H_s \right)^{1/2} \\
&\quad + 4 \, \mathrm{c}_{\text{last},1} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} p^6 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, N^{-1/2} \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^2 H_s (1+s)^{-\gamma_\eta} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \\
&\quad + 4 \, \mathrm{c}_{\text{last},2} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} p^3 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^2 H_s (1+s)^{-2\gamma} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \\
&\quad \leq \mathrm{c}^{\mathsf{b}}_{3,1,2} \, p^6 N^{-1/2} (1+t)^{-\gamma_\eta/2} + (\mathrm{c}^{\mathsf{b}}_{3,1,1} + \mathrm{c}^{\mathsf{b}}_{t,3,1,3}) p^3 (1+t)^{-\gamma} \, .
\end{aligned}
$$

where we have used that $x^{1/2} e^{-cx} \leq (c\mathrm{e})^{-1/2} e^{-cx/2}$ together with the bound

$$
\eta_t^{-1/2} \exp\left(-\frac{a}{2} \sum_{\ell=1}^{t} \eta_\ell H_\ell \right) \leq e^{a/(1-\gamma)} \left(\frac{2(\gamma_\eta + \gamma)}{a \mathrm{e} \eta H} \right)^{\frac{\gamma_\eta + \gamma}{1-\gamma}} \eta^{-1/2} (1+t)^{-\gamma} \, ,
$$

and have set

$$
\begin{aligned}
\mathrm{c}^{\mathsf{b}}_{3,1,1} &= 4 a^{-1/2} e^{1/2 + a/(1-\gamma)} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} \left(\frac{2(\gamma_\eta + \gamma)}{a \mathrm{e}} \right)^{\frac{\gamma_\eta + \gamma}{1-\gamma}} \eta^{-1/2} \|\theta_0 - \theta_\star\| \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, , \\
\mathrm{c}^{\mathsf{b}}_{3,1,2} &= 8 \, \mathrm{c}_{\text{last},1} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} ((\eta H)^{1/2} + a^{-1/2} L^{1/2} (1 + (2a\eta H)^{\frac{\gamma + 2\gamma_\eta}{2(1-\gamma)}})) \, , \\
\mathrm{c}^{\mathsf{b}}_{3,1,3} &= 8 \, \mathrm{c}_{\text{last},2} \, \mathrm{C}_{\mathbf{A}} \, W_{\max} \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} ((\eta H)^{1/2} + a^{-1/2} L^{1/2} (1 + (2a\eta H)^{\frac{3\gamma + \gamma_\eta}{2(1-\gamma)}})) \, .
\end{aligned}
$$

Similarly, for the term $D^{\mathsf{b}}_{t,3,2}$, applying Burkholder's inequality together with Lemma 25, setting

$$
S^{\mathsf{b}}_{s,3,2} = \sum_{c=1}^{N} \sum_{h=1}^{H_s} \mathbb{E}^{2/p}[\|(w_{s,h,c} - 1) \Gamma^{\mathsf{b},c}_{s,h+1:H_s} \mathbf{A}^c(Z_{s,h,c}) \sum_{j=1}^{h-1} \Gamma^c_{s,j+1:h-1} \varepsilon^c(Z_{s,h,c})\|^p] \, ,
$$

we obtain

$$
\begin{aligned}
\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,3,2}\|^p] &\leq p^3 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, N^{-1/2} \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^4 \mathbb{E}^{2/p}[\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \|^p] S^{\mathsf{b}}_{s,3,2} \right)^{1/2} \\
&\leq \mathrm{C}_{\mathbf{A}} \, W_{\max} p^4 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \, \eta_t^{-1/2} \left(\sum_{s=1}^{t} \eta_s^4 H_s^2 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell \right) \right)^{1/2} \|\varepsilon\|_\infty \\
&\leq \mathrm{c}^{\mathsf{b}}_{3,2} \, p^4 \, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} (1+t)^{-(\gamma_\eta + \gamma)/2} \, ,
\end{aligned}
$$

where we have used that

$$
\mathbb{E}^{1/p}[\| \sum_{j=1}^{h-1} \Gamma^c_{s,j+1:h-1} \varepsilon^c(Z_{s,h,c}) \|^p] \leq p \left(\sum_{j=1}^{h-1} \mathbb{E}^{2/p}[\| \Gamma^c_{s,j+1:h-1} \varepsilon^c(Z_{s,h,c}) \|^p] \right)^{1/2} \leq p H_s^{1/2} \|\varepsilon\|_\infty \, ,
$$

and have set

$$\mathrm{c}^{\mathsf{b}}_{3,2} = 4W_{\max}\,\mathrm{C}_{\mathbf{A}}\,\mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}(\eta H)^{1/2}(1+(2a\eta H)^{\frac{\gamma+\gamma_\eta}{1-\gamma}}))\|\varepsilon\|_{\infty}\;.$$

Finally, to bound $D^{\mathsf{b}}_{t,3,3}$, we again apply Burkholder's inequality and Lemma 25, thus, setting

$$S^{\mathsf{b}}_{s,3,3} = \sum_{c=1}^{N}\sum_{h=1}^{H_s}\mathbb{E}^{2/p}[\|(w_{s,h,c}-1)\Gamma^{\mathsf{b},c}_{s,h+1:H_s}\mathbf{A}^c(Z_{s,h,c})(\mathrm{I}-\Gamma^c_{s,h-1})(\theta^c_\star-\theta_\star)\|^p]\,,$$

we get

$$\begin{aligned}
\mathbb{E}^{1/p}[\|D^{\mathsf{b}}_{t,3,3}\|^p] &\leq p^3\,\mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}\,N^{-1/2}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}\Gamma^{\mathsf{b},\mathsf{avg}}_i\|^p]S^{\mathsf{b}}_{s,3,3}\right)^{1/2} \\
&\leq W_{\max}\zeta_3\,\mathrm{C}^2_{\mathbf{A}}\,p^3\,\mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}\left(\sum_{s=1}^{t}\eta_s^4H_s^3\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\right)^{1/2} \\
&\leq \mathrm{c}^{\mathsf{b}}_{3,3}\,p^3\,\mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}(1+t)^{-\gamma}\,,
\end{aligned}$$

where we defined

$$\mathrm{c}^{\mathsf{b}}_{3,3} = 8W_{\max}\zeta_3\,\mathrm{C}^2_{\mathbf{A}}\,\mathrm{C}_{\boldsymbol{\Sigma}^{\mathsf{b}}}((\eta H)^{3/2}+a^{-1/2}L^{1/2}\eta H(1+(2a\eta H)^{\frac{3\gamma+\gamma_\eta}{2(1-\gamma)}}))\;.$$

Finally, setting

$$\begin{aligned}
\bar{\mathrm{c}}^{\mathsf{b}}_1 &= \mathrm{c}^{\mathsf{b}}_1+\mathrm{c}^{\mathsf{b}}_{3,1,3}+\mathrm{c}^{\mathsf{b}}_4 \\
\bar{\mathrm{c}}^{\mathsf{b}}_2 &= \mathrm{c}^{\mathsf{b}}_{3,1,1}+\mathrm{c}^{\mathsf{b}}_2+\mathrm{c}^{\mathsf{b}}_{3,1,3}+\mathrm{c}^{\mathsf{b}}_{3,2}+\mathrm{c}^{\mathsf{b}}_{3,3}+\mathrm{c}^{\mathsf{b}}_5\,,
\end{aligned} \tag{96}$$

we complete the proof. □

D EXPERIMENTAL RESULTS

In this section, we provide the experimental results for the normal approximation of FedLSA. Code is available in `https://github.com/levensons/FedLSA_NormApprox`. We consider the Garnet environment Archibald et al. [1995], Geist et al. [2014] with $n = 30$ states, dimension $d = 5$, number of actions $a = 2$, and branching number $b = 2$. We then generate $N = 5$ heterogeneous Garnet environments with small perturbations. The process for generating heterogeneous environments follows that described in Mangold et al. [2024]. In our experiment, we first sample a unit vector $u \in \mathbb{S}_{d-1}$ onto which we project the obtained confidence sets. Then, we sample $R = 1024$ trajectories $\{Z^r_{s,h,c}, s \in [t], h \in [H_s], c \in [N]\}_{r=1}^{1024}$ and, for each of them, additionally sample $N_{\mathsf{b}} = 256$ bootstrap trajectories. The bootstrap weights $\{w_{s,h,c}, s \in [t], h \in [H_s], c \in [N]\}$ are i.i.d. and defined as

$$w_{s,h,c} = \frac{\tilde{w}_{s,h,c}-\mathbb{E}[\tilde{w}_{s,h,c}]}{\sqrt{\mathrm{Var}[\tilde{w}_{s,h,c}]}}\,,\quad \tilde{w}_{s,h,c}\sim \mathrm{Beta}(0.5,2)\,.$$

We therefore consider several algorithms for constructing confidence sets for the last iterate θ_t. Then, based on R samples, we estimate the probability that $\theta_\star$ belongs to the confidence sets and report it as coverage.

Plug-in Estimate (PE). As a baseline, we consider the plug-in estimate in a way similar to Li et al. [2022]. First, we approximate the noise covariance and design matrix at step t as

$$\widehat{\Sigma}^{\mathsf{avg}}_{*,t} = N^{-1}\bar{H}_t^{-1}\sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\varepsilon^c(\theta_s,Z_{s,h,c})(\varepsilon^c(\theta_s,Z_{s,h,c}))^T\,,\quad \bar{\mathbf{A}}_t = N^{-1}\bar{H}_t^{-1}\sum_{s=1}^{t}\sum_{h=1}^{H_s}\sum_{c=1}^{N}\mathbf{A}^c(Z_{s,h,c})\,,$$

where $\bar{H}_t = \sum_{s=1}^{H_t}H_s$. Afterwards, we solve the Poisson equation with Σ^{avg}_* replaced by $\widehat{\Sigma}^{\mathsf{avg}}_{*,t}$; that is, we find $\widehat{\Sigma}_{\infty,t}$ such that

$$\bar{\mathbf{A}}_t\widehat{\Sigma}_\infty+\widehat{\Sigma}_\infty\bar{\mathbf{A}}_t^T = N^{-1}\widehat{\Sigma}^{\mathsf{avg}}_{*,t}\,.$$

Finally, we consider the confidence interval $u^T\theta_t \pm z_{\alpha/2}\eta_t^{1/2}\sqrt{u^T\hat{\Sigma}_{\infty,t}u}$, where $z_{\alpha/2}$ is the $\alpha/2$-quantile of $\mathcal{N}(0,1)$.

Table 3: Coverage comparison for different methods with confidence levels $\alpha \in \{0.95, 0.9, 0.8\}$ and $\gamma_H = 0$. Standard deviations are shown as subscripts.

	$\alpha = 0.95$			$\alpha = 0.9$			$\alpha = 0.8$		
T	SDB	EQ	PE	SDB	EQ	PE	SDB	EQ	PE
2000	$0.773_{\pm 0.013}$	$0.780_{\pm 0.013}$	$0.754_{\pm 0.013}$	$0.676_{\pm 0.015}$	$0.683_{\pm 0.015}$	$0.652_{\pm 0.015}$	$0.547_{\pm 0.016}$	$0.544_{\pm 0.016}$	$0.517_{\pm 0.016}$
6000	$0.936_{\pm 0.008}$	$0.941_{\pm 0.007}$	$0.931_{\pm 0.008}$	$0.870_{\pm 0.011}$	$0.880_{\pm 0.010}$	$0.866_{\pm 0.011}$	$0.778_{\pm 0.013}$	$0.778_{\pm 0.013}$	$0.768_{\pm 0.013}$
10000	$0.938_{\pm 0.008}$	$0.947_{\pm 0.007}$	$0.938_{\pm 0.008}$	$0.888_{\pm 0.010}$	$0.890_{\pm 0.010}$	$0.883_{\pm 0.010}$	$0.786_{\pm 0.013}$	$0.792_{\pm 0.013}$	$0.775_{\pm 0.013}$
14000	$0.941_{\pm 0.007}$	$0.946_{\pm 0.007}$	$0.941_{\pm 0.007}$	$0.895_{\pm 0.010}$	$0.900_{\pm 0.009}$	$0.888_{\pm 0.010}$	$0.798_{\pm 0.013}$	$0.804_{\pm 0.012}$	$0.789_{\pm 0.013}$

Table 4: Coverage comparison for different methods with confidence levels $\alpha \in \{0.95, 0.9, 0.8\}$ and $\gamma_H = 0.2$. Standard deviations are shown as subscripts.

	$\alpha = 0.95$			$\alpha = 0.9$			$\alpha = 0.8$		
T	SDB	EQ	PE	SDB	EQ	PE	SDB	EQ	PE
2000	$0.398_{\pm 0.015}$	$0.404_{\pm 0.015}$	$0.376_{\pm 0.015}$	$0.291_{\pm 0.014}$	$0.291_{\pm 0.014}$	$0.268_{\pm 0.014}$	$0.173_{\pm 0.012}$	$0.177_{\pm 0.012}$	$0.161_{\pm 0.011}$
6000	$0.959_{\pm 0.006}$	$0.961_{\pm 0.006}$	$0.959_{\pm 0.006}$	$0.914_{\pm 0.009}$	$0.926_{\pm 0.008}$	$0.917_{\pm 0.009}$	$0.817_{\pm 0.012}$	$0.820_{\pm 0.012}$	$0.814_{\pm 0.012}$
10000	$0.946_{\pm 0.007}$	$0.949_{\pm 0.007}$	$0.946_{\pm 0.007}$	$0.903_{\pm 0.009}$	$0.907_{\pm 0.009}$	$0.907_{\pm 0.009}$	$0.815_{\pm 0.012}$	$0.825_{\pm 0.012}$	$0.818_{\pm 0.012}$
14000	$0.950_{\pm 0.007}$	$0.950_{\pm 0.007}$	$0.952_{\pm 0.007}$	$0.892_{\pm 0.010}$	$0.894_{\pm 0.010}$	$0.892_{\pm 0.010}$	$0.788_{\pm 0.013}$	$0.789_{\pm 0.013}$	$0.788_{\pm 0.013}$

Empirical Quantiles (EQ). As described in Section 6, we construct the confidence interval using the empirical quantiles of the bootstrap samples.

Standard Deviation-Based Confidence Intervals (SDB). We estimate the standard deviation $\hat{\sigma}_t^{\mathsf{b}}(u)$ of the bootstrap samples $\{\theta_t^b, b \in [N_{\mathsf{b}}]\}$ and consider the confidence interval $u^T\theta_t \pm z_{\alpha/2}\eta_t^{1/2}\hat{\sigma}_t^{\mathsf{b}}(u)$.

Discussion. In Table 3 and Table 4, we compare coverage across different confidence levels and iteration lengths. In our experiments, we consider a decreasing step size with $\gamma_\eta = 0.6$. We also consider two regimes: one with a constant number of local updates, $H = 20$, and another in which the number of local updates increases at the rate $\gamma_H = 0.2$ with $H = 5$. In the case $\gamma_H = 0$, we can see that the bootstrap-based methods outperform the baseline even for a small number of iterations and, for $T = 6000$, produce decent coverage. On the other hand, when we consider an increasing number of local updates, for large T it becomes comparable to, or slightly better than, the bootstrap-based methods. This can be explained by the fact that, as the number of samples increases, the PE estimate becomes sufficiently accurate to provide good coverage.

E TECHNICAL RESULTS

In this section we collect several auxiliary results that will be used in the proofs of the main theorems.

Proposition 5. *Let ν be a standard Gaussian measure in $\mathbb{R}^d$. Then for any random vectors X, Y taking values in $\mathbb{R}^d$, and any $p \geq 1$,*

$$\sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}(X + Y \in B) - \nu(B)| \leq \sup_{B \in \mathsf{C}(\mathbb{R}^d)} |\mathbb{P}(X \in B) - \nu(B)| + 2c_d^{p/(p+1)}\mathbb{E}^{1/(p+1)}[\|Y\|^p], \tag{97}$$

where c_d is the isoperimetric constant of class $\mathsf{C}(\mathbb{R}^d)$.

Proof. Proof and definition for the isoperimetric constant c_d can be found in Sheshukova et al. [2026, Proposition 1]. □

Lemma 10. *Fix $\delta \in (0, 1/\mathrm{e}^2)$ and let Y be a positive random variable, such that $\mathbb{E}^{1/p}[Y^p] \leq C_1 + C_2 p$ for any $2 \leq p \leq \log(1/\delta)$. Then it holds with probability at least $1 - \delta$, that*

$$Y \leq \mathrm{e}C_1 + \mathrm{e}C_2 \log(1/\delta) \,.$$

Proof. For the complete proof, see Samsonov et al. [2024a, Lemma 1]. □

Lemma 11. *Under Assumptions A1, A2(2), S2 and A3, choosing $\eta H \leq \beta_\infty$ for some $\beta_\infty > 0$, we have that*

$$\|\Sigma_t - \Sigma_\infty\| \leq C_{\infty,1}(1+t)^{\gamma-1} + C_{\infty,2}(1+t)^{-\gamma}\,,$$

where the matrix Σ_∞ is the solution of the Lyapunov's equation

$$\bar{\mathbf{A}}\Sigma_\infty + \Sigma_\infty \bar{\mathbf{A}}^T = N^{-1}\Sigma_*^{\mathsf{avg}}\,, \tag{98}$$

where Σ_*^{avg} is defined in (63).

Proof. Due to Poznyak [2008, Lemma 9.1], the solution of Lyapunov's equation (98) exists since the matrix $\bar{\mathbf{A}}$ is Hurwitz. Let us consider only the leading part of the term M_t, that is, $\hat{M}_t = M_{t,1} + M_{t,2}$. Using the notation $P_s^t = \prod_{i=s+1}^t G_i^{\mathsf{avg}}$, we define

$$\hat{\Sigma}_t = \mathbb{E}[\hat{M}_t\hat{M}_t^T] = N^{-1}\eta_t^{-1}\sum_{s=1}^t \eta_s^2 H_s P_s^t \Sigma_*^{\mathsf{avg}}(P_s^t)^T$$

To estimate the difference between Σ_t and Σ_∞, we write the recursion

$$\begin{aligned}
\hat{\Sigma}_{t+1} &= \eta_t\eta_{t+1}^{-1}G_{t+1}^{\mathsf{avg}}\hat{\Sigma}_t(G_{t+1}^{\mathsf{avg}})^T + N^{-1}\eta_{t+1}H_{t+1}\Sigma_*^{\mathsf{avg}} \\
&= \eta_t\eta_{t+1}^{-1}\{\hat{\Sigma}_t + \hat{\Sigma}_t(G_{t+1}^{\mathsf{avg}} - \mathrm{I})^T + (G_{t+1}^{\mathsf{avg}} - \mathrm{I})\hat{\Sigma}_t\} \\
&+ \eta_t\eta_{t+1}^{-1}(G_{t+1}^{\mathsf{avg}} - \mathrm{I})\hat{\Sigma}_t(G_{t+1}^{\mathsf{avg}} - \mathrm{I})^T + N^{-1}\eta_{t+1}H_{t+1}\Sigma_*^{\mathsf{avg}} \\
&= \hat{\Sigma}_t - \eta_{t+1}H_{t+1}\{\bar{\mathbf{A}}\hat{\Sigma}_t + \hat{\Sigma}_t\bar{\mathbf{A}}^T - N^{-1}\Sigma_*^{\mathsf{avg}}\} \\
&+ (\eta_t\eta_{t+1}^{-1} - 1)(\hat{\Sigma}_t - \hat{\Sigma}_t\bar{R}_{t+1}^T - \bar{R}_{t+1}\hat{\Sigma}_t) \\
&+ \eta_t\eta_{t+1}^{-1}(G_{t+1}^{\mathsf{avg}} - \mathrm{I})\hat{\Sigma}_t(G_{t+1}^{\mathsf{avg}} - \mathrm{I})^T\,,
\end{aligned}$$

where we have used the identity

$$\begin{aligned}
(\mathrm{I} - \eta_{t+1}\bar{\mathbf{A}}^c)^{H_{t+1}} - \mathrm{I} &= -\eta_{t+1}\bar{\mathbf{A}}^c\sum_{h=1}^{H_{t+1}}(\mathrm{I} - \eta_{t+1}\bar{\mathbf{A}}^c)^{h-1} \\
&= -\eta_{t+1}H_{t+1}\bar{\mathbf{A}}^c - \underbrace{\eta_{t+1}\bar{\mathbf{A}}^c\sum_{h=1}^{H_{t+1}}\{(\mathrm{I} - \eta_{t+1}\bar{\mathbf{A}}^c)^{h-1} - \mathrm{I}\}}_{R_{t+1}^c}\,.
\end{aligned}$$

Now, the error $E_t = \hat{\Sigma}_t - \Sigma_\infty$ evolves as

$$\begin{aligned}
E_{t+1} &= E_t - \eta_{t+1}H_{t+1}\{\bar{\mathbf{A}}E_t + E_t\bar{\mathbf{A}}^T\} \\
&+ (\eta_t\eta_{t+1}^{-1} - 1)(\hat{\Sigma}_t - \hat{\Sigma}_t\bar{R}_{t+1}^T - \bar{R}_{t+1}\hat{\Sigma}_t) + \eta_t\eta_{t+1}^{-1}(G_{t+1}^{\mathsf{avg}} - \mathrm{I})\hat{\Sigma}_t(G_{t+1}^{\mathsf{avg}} - \mathrm{I})^T \\
&= E_t - \eta_{t+1}H_{t+1}\{\bar{\mathbf{A}}E_t + E_t\bar{\mathbf{A}}^T\} + r_{t+1}\,.
\end{aligned}$$

To quantify the convergence, we transition the recursion into the vectorized form. Therefore, we get

$$\mathrm{vec}(E_{t+1}) = (\mathrm{I} - \eta_{t+1}H_{t+1}((\mathrm{I}\otimes\bar{\mathbf{A}}) + (\bar{\mathbf{A}}\otimes\mathrm{I})))\mathrm{vec}(E_t) + \mathrm{vec}(r_{t+1}) = F_{t+1}\mathrm{vec}(E_t) + \mathrm{vec}(r_{t+1})\,.$$

Clearly, the matrix $(\mathrm{I}\otimes\bar{\mathbf{A}}) + (\bar{\mathbf{A}}\otimes\mathrm{I})$ is Hurwitz. Therefore, due to the Durmus et al. [2021, Proposition 1], we have

$$\|F_t\| \leq \kappa_Q^{1/2}(1-\eta H b)^{1/2}\,, \quad \eta_t H_t \leq \beta_\infty\,.$$

Now, to bound R_t, we apply Lemma 22, thus

$$\|R_t\| \leq \mathrm{C}_{\mathbf{A}}\,\eta_t^2 H_t^2\,.$$

Therefore, using Lemma 13 and the bound $\|G_t^{\mathsf{avg}} - \mathrm{I}\| \leq \mathrm{C}_{\mathbf{A}}\,\eta_t H_t$, the remainder term r_t can be bounded, as

$$\|r_{t+1}\| \leq 8\,\mathrm{c}_{\hat{\Sigma}}\,\gamma_\eta(1 + \mathrm{C}_{\mathbf{A}}(\eta H)^2)(1+t)^{-1} + 2\,\mathrm{C}_{\mathbf{A}}^2\,\mathrm{c}_{\hat{\Sigma}}(\eta_{t+1}H_{t+1})^2\,.$$

Using that $\|\mathrm{vec}(r_{t+1})\| \leq d^{1/2}\|r_{t+1}\|$, we get

$$
\begin{aligned}
\|\mathrm{vec}(E_{t+1})\| &\leq \kappa_Q^{1/2} \prod_{i=1}^{t+1} (1-\eta_i H_i b)^{1/2} \|\mathrm{vec}(\hat{\Sigma}_1 - \Sigma_\infty)\| + \kappa_Q^{1/2} d^{1/2} \sum_{j=1}^{t+1} \prod_{i=j+1}^{t+1} (1-\eta_i H_i b)^{1/2} \|r_j\| \\
&\leq \kappa_Q^{1/2} \exp\left(-\frac{b}{2}\sum_{\ell=1}^{t+1} \eta_\ell H_\ell\right) \|\mathrm{vec}(\hat{\Sigma}_1 - \Sigma_\infty)\| \\
&+ 16\, \mathrm{c}_{\hat{\Sigma}}\, \kappa_Q^{1/2} \gamma_\eta (1 + \mathrm{C}_{\mathbf{A}} (\eta H)^2) d^{1/2} \sum_{j=1}^{t+1} (1+j)^{-1} \exp\left(-\frac{b}{2}\sum_{\ell=j+1}^{t+1} \eta_\ell H_\ell\right) \\
&+ 2\, \mathrm{C}_{\mathbf{A}}^2\, \mathrm{c}_{\hat{\Sigma}}\, \kappa_Q^{1/2} d^{1/2} \sum_{j=1}^{t+1} \eta_j^2 H_j^2 \exp\left(-\frac{b}{2}\sum_{\ell=j+1}^{t+1} \eta_\ell H_\ell\right) \\
&\leq \kappa_Q^{1/2} \exp\left(-\frac{b}{2}\sum_{\ell=1}^{t+1} \eta_\ell H_\ell\right) \|\mathrm{vec}(\hat{\Sigma}_1 - \Sigma_\infty)\| + M_{\Sigma,1}(2+t)^{\gamma-1} + M_{\Sigma,2}(2+t)^{-\gamma} \,,
\end{aligned}
$$

where

$$
\begin{aligned}
M_{\Sigma,1} &= 16\, \mathrm{c}_{\hat{\Sigma}}\, \kappa_Q^{1/2} \gamma_\eta (1 + \mathrm{C}_{\mathbf{A}}(\eta H)^2)(1 + 2b^{-1} L (1 + ((b/2)\eta H)^{\frac{1+\gamma}{1-\gamma}})) d^{1/2} \,, \\
M_{\Sigma,2} &= 8\, \mathrm{C}_{\mathbf{A}}^2\, \mathrm{c}_{\hat{\Sigma}}\, \kappa_Q^{1/2} ((\eta H)^2 + b^{-1} L \eta H (1 + ((b/2)\eta H)^{\frac{3\gamma}{1-\gamma}})) d^{1/2} \,.
\end{aligned}
$$

Finally, using (105), (106), the bound $xe^{-cx} \leq (ce)^{-1}$ and that $\|E_t\| \leq \|E_t\|_F$, we get

$$
\|\Sigma_t - \Sigma_\infty\| \leq \|\Sigma_t - \hat{\Sigma}_t\| + \|\hat{\Sigma}_t - \Sigma_\infty\| \leq C_{\infty,1}(1+t)^{\gamma-1} + C_{\infty,2}(1+t)^{-\gamma} \,,
$$

where we have used that $\|\mathrm{vec}(\Sigma_\infty)\| \leq (2\lambda_{\min}(\bar{\mathbf{A}}))^{-1} N^{-1} \|\mathrm{vec}(\Sigma_*^{\mathsf{avg}})\|$, and have set

$$
C_{\infty,1} = \kappa_Q^{1/2} \left(\frac{be}{2(1-\gamma)}\eta H\right)^{-1} e^{\frac{b}{1-\gamma}\eta H} d^{1/2} (2\eta H + (2\lambda_{\min}(\bar{\mathbf{A}}))^{-1}) N^{-1} \|\Sigma_*^{\mathsf{avg}}\| + M_{\Sigma,1} \,, \tag{99}
$$

$$
C_{\infty,2} = \mathrm{c}_{\Sigma,2} + 2(\mathrm{c}_{M,3}^2 + \mathrm{c}_{M,4}^2) \,. \tag{100}
$$

□

Lemma 12. *Under Assumptions **A**1, **A**2(2), **S**2, **A**3, and **A**4, we have*

$$
\|\Sigma_t^{-1/2}\| \leq N^{1/2}\, \mathrm{C}_{\mathbf{\Sigma}} \,, \tag{101}
$$

where Σ_t is defined in (65) and

$$
\mathrm{C}_{\mathbf{\Sigma}} = \sqrt{2}(\lambda_{\min}(\Sigma_*^{\mathsf{avg}}))^{-1/2} \,.
$$

Proof. Applying Lidski's inequality, we obtain

$$
\lambda_{\min}(\Sigma_t) \geq \lambda_{\min}(\Sigma_\infty) - \|\Sigma_t - \Sigma_\infty\|
$$

Thus, under **A**4, we get that

$$
\|\Sigma_t - \Sigma_\infty\| \leq 2\lambda_{\min}(\bar{\mathbf{A}}) \frac{\lambda_{\min}(\Sigma_\infty)}{2} \,.
$$

Finally, using the relation $\lambda_{\min}(\Sigma_\infty) \leq \|\Sigma_\infty\| \leq (2\lambda_{\min}(\bar{\mathbf{A}}))^{-1} N^{-1} \lambda_{\min}(\Sigma_*^{\mathsf{avg}})$, the result (101) follows. □

Proposition 6. *Under assumptions **A**1, **S**2 and **A**4, with probability at least $1 - 2/t$, we have*

$$
\|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \lesssim_{\log_t} N^{-3/2} t^{-\gamma_\eta/2} + N^{-1} t^{-2\gamma} \,, \tag{102}
$$

where Σ_t^{b} is defined in (65). Moreover, on the event (102), under (assumption on t), we have

$$
\|(\Sigma_t^{\mathsf{b}})^{-1/2}\| \leq N^{1/2}\, \mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} \,,
$$

and

$$
\mathrm{C}_{\mathbf{\Sigma}^{\mathsf{b}}} = 2(\lambda_{\min}(\Sigma_*^{\mathsf{avg}}))^{-1/2} \,.
$$

Proof. Firstly, we note that

$$\|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \leq \|\Sigma_t - \hat{\Sigma}_t\| + \|\hat{\Sigma}_t - \hat{\Sigma}_t^{\mathsf{b}}\| + \|\hat{\Sigma}_t^{\mathsf{b}} - \Sigma_t^{\mathsf{b}}\| \,,$$

where $\hat{\Sigma}_t = \mathbb{E}[\hat{M}_t \hat{M}_t^T]$ and $\hat{\Sigma}_t^{\mathsf{b}} = \mathbb{E}^{\mathsf{b}}[M_{t,1}^{\mathsf{b}} M_{t,1}^{\mathsf{b}})^T]$. Then, we note

$$\hat{\Sigma}_t - \hat{\Sigma}_t^{\mathsf{b}} = N^{-2}\eta_t^{-1} \sum_{s=1}^{t} \sum_{c=1}^{N} \sum_{h=1}^{H_s} \underbrace{\eta_s^2 \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} \{\varepsilon^c(\theta_\star, Z_{s,h,c})(\varepsilon^c(\theta_\star, Z_{s,h,c}))^T - \Sigma_*^c\}(\prod_{i=s+1}^{t} G_i^{\mathsf{avg}})^T}_{U_{s,h,c}} \,.$$

We note that $\mathbb{E}[U_{s,h,c}] = 0$ and it can be bounded, as

$$\|U_{s,h,c}\| \leq \eta_s^2 \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right)(\|\varepsilon\|_\infty^2 + \|\Sigma_*^c\|) \leq U_{\max}\eta_t^2 \,.$$

where we have set

$$U_{\max} = \left(9 + \left(\frac{2\gamma_\eta}{2a\eta H \mathrm{e}}\right)^{\frac{2\gamma_\eta}{1-\gamma}} \exp\left(\frac{6a\eta H}{1-\gamma}\right)\right)(\|\varepsilon\|_\infty^2 + \max_{c\in[N]} \|\Sigma_*^c\|)$$

Hence, applying Lemma 25, we obtain

$$\|\sum_{s,h,c} \mathbb{E}[U_{s,h,c}U_{s,h,c}^T]\| \leq 2N \sum_{s=1}^{t} \eta_s^4 H_s \exp\left(-4a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right)(\|\varepsilon\|_\infty^4 + \sigma_*^2) \leq \mathrm{c}_U \, N\eta_t^3 \,,$$

where

$$\mathrm{c}_U = 4(\eta H + a^{-1}L(1 + (4a\eta H)^{\frac{\gamma+3\gamma_\eta}{1-\gamma}}))(\|\varepsilon\|_\infty^4 + \sigma_*^2)$$

Hence, we can apply the matrix Bernstein inequality Tropp [2015, Theorem 6.1.1], which implies

$$\mathbb{P}(\|\hat{\Sigma}_t - \hat{\Sigma}_t^{\mathsf{b}}\| \geq x) = \mathbb{P}(\|\sum_{s,h,c} U_{s,h,c}\| \geq N^2\eta_t x) \leq 2d \exp\left(-\frac{N^3 x^2}{\eta_t(2\,\mathrm{c}_U + 2U_{\max}Nx/3)}\right) .$$

Equivalently, with probability at least $1-\delta$, it holds that

$$\|\hat{\Sigma}_t - \hat{\Sigma}_t^{\mathsf{b}}\| \leq \left(\frac{2\,\mathrm{c}_U\,\eta}{N^3}\right)^{1/2} (1+t)^{-\gamma_\eta/2}\sqrt{\log(2d/\delta)} + \frac{U_{\max}\eta}{3N^2}(1+t)^{-\gamma_\eta}\log(2d/\delta) \,. \tag{103}$$

Combining the results (103), (104) and (107) with $\delta = 1/t$, we obtain (102).

In order to prove the second part of the lemma, we apply Lidski's inequality, and get

$$\lambda_{\min}(\Sigma_t^{\mathsf{b}}) \geq \lambda_{\min}(\Sigma_t) - \|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \geq \lambda_{\min}(\Sigma_\infty)/4 \,.$$

Then, by **A**4, we choose n_0 such that for $t \geq n_0$, we have $\|\Sigma_t - \Sigma_t^{\mathsf{b}}\| \leq \lambda_{\min}(\Sigma_\infty)/4$. Also, under this assumption, we have $\lambda_{\min}(\Sigma_t) \geq \lambda_{\min}(\Sigma_\infty)/2$ as stated in Lemma 12. □

Lemma 13. *Under Assumptions* ***A****1, **A**2*$(\log(t))$*, **S**2, **A**3, and **A**4, we have that*

$$\mathbb{E}[\|M_t\|^2] \leq \mathrm{c}_M \, N^{-1} \,.$$

Particularly, we get that

$$\|\hat{\Sigma}_t\| = \|\mathbb{E}[\hat{M}_t \hat{M}_t^T]\| \leq \mathrm{c}_{\hat{\Sigma}} \, N^{-1} \,,$$

where $\hat{M}_t = M_{t,1} + M_{t,2}$*, and*

$$\|\Sigma_t - \hat{\Sigma}_t\| \leq (\mathrm{c}_{M,3}^2 + \mathrm{c}_{M,4}^2)N^{-1}(1+t)^{-2\gamma} \,. \tag{104}$$

Proof. Applying again Lemma 22 and Lemma 25, we get

$$\begin{aligned}\mathbb{E}[\|M_{t,3}\|^2] &\leq N^{-2}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^2\sum_{c=1}^{N}\sum_{h=1}^{H_s}\mathbb{E}[\|(G^c_{s,h+1:H_s}-\mathrm{I})\varepsilon^c(Z_{s,h,c})\|^2]\exp\Big(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\Big)\\ &\leq \mathrm{C}_{\mathbf{A}}^2\,\bar{\sigma}_\varepsilon^2 N^{-1}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^4 H_s^3\exp\Big(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\Big)\\ &\leq 4\,\mathrm{C}_{\mathbf{A}}^2\,\bar{\sigma}_\varepsilon^2((\eta H)^3+a^{-1}L(\eta H)^2(1+(2a\eta H)^{-\frac{3\gamma+\gamma_\eta}{1-\gamma}}))N^{-1}(1+t)^{-2\gamma}\\ &= \mathrm{c}_{M,3}^2\,N^{-1}(1+t)^{-2\gamma}\,,\end{aligned}\tag{105}$$

and

$$\begin{aligned}\mathbb{E}[\|M_{t,4}\|^2] &\leq N^{-2}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^2\sum_{c=1}^{N}\sum_{h=1}^{H_s}\mathbb{E}[\|(G^c_{s,1:h-1}-\mathrm{I})\widetilde{\mathbf{A}}^c(Z_{s,h,c})(\theta_\star-\theta^c_\star)\|^2]\exp\Big(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\Big)\\ &\leq 4\bar{\sigma}_{\mathsf{het}}^2((\eta H)^3+a^{-1}L(\eta H)^2(1+(2a\eta H)^{-\frac{3\gamma+\gamma_\eta}{1-\gamma}}))N^{-1}(1+t)^{-2\gamma}\\ &= \mathrm{c}_{M,4}^2\,N^{-1}(1+t)^{-2\gamma}\,.\end{aligned}\tag{106}$$

Similarly, we get that

$$\begin{aligned}\mathbb{E}[\|M_{t,1}\|^2] &\leq \bar{\sigma}_\varepsilon^2 N^{-1}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^2 H_s\exp\Big(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\Big)\\ &\leq 2\bar{\sigma}_\varepsilon^2(\eta H+a^{-1}L(1+(2a\eta H)^{\frac{\gamma+\gamma_\eta}{1-\gamma}}))N^{-1}=\mathrm{c}_{M,1}\,N^{-1}\,,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\|M_{t,2}\|^2] &\leq \bar{\sigma}_{\mathsf{het}}^2 N^{-1}\eta_t^{-1}\sum_{s=1}^{t}\eta_s^2 H_s\exp\Big(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\Big)\\ &\leq 2\bar{\sigma}_{\mathsf{het}}^2(\eta H+a^{-1}L(1+(2a\eta H)^{\frac{\gamma+\gamma_\eta}{1-\gamma}}))N^{-1}=\mathrm{c}_{M,2}\,N^{-1}\,.\end{aligned}$$

Finally, setting $\mathrm{c}_M=\mathrm{c}_{M,1}+\mathrm{c}_{M,2}+\mathrm{c}_{M,3}+\mathrm{c}_{M,4}$ and $\mathrm{c}_{\hat{\Sigma}}=\mathrm{c}_{M,1}+\mathrm{c}_{M,2}$, we get the result. □

Lemma 14. *Under Assumptions **A**1, **A**2($\log(t)$), **S**2, **A**3, and **A**4, we have that*

$$\mathbb{E}^{1/p}[\|\Sigma_t^{\mathsf{b}}-\hat{\Sigma}_t^{\mathsf{b}}\|^p]\leq 2^{13}\,\mathrm{c}_{M,2}^{\mathsf{b}}\,p^6N^{-2}(1+t)^{-\gamma_\eta}+2^8\,\mathrm{c}_{M,3}^{\mathsf{b}}\,p^3N^{-1}(1+t)^{-2\gamma}\,,$$

where $\hat{\Sigma}_t=\mathbb{E}[M_{t,1}^{\mathsf{b}}(M_{t,1}^{\mathsf{b}})^T]$. Moreover, with probability at least $1-\delta$, we have

$$\|\Sigma_t^{\mathsf{b}}-\hat{\Sigma}_t^{\mathsf{b}}\|\lesssim\bar{\mathrm{c}}_{M,2}^{\mathsf{b}}N^{-2}\log^6(1/\delta)(1+t)^{-\gamma_\eta}+\bar{\mathrm{c}}_{M,3}^{\mathsf{b}}N^{-1}\log^3(1/\delta)(1+t)^{-2\gamma}\,.\tag{107}$$

Proof. We note that

$$\mathbb{E}^{1/p}[\|\Sigma_t^{\mathsf{b}}-\hat{\Sigma}_t^{\mathsf{b}}\|^p]\leq\mathbb{E}^{1/p}[\|(M_{t,2}^{\mathsf{b}}+M_{t,3}^{\mathsf{b}})\|^{2p}]\leq 2\mathbb{E}^{1/p}[\|M_{t,2}^{\mathsf{b}}\|^{2p}]+2\mathbb{E}^{1/p}[\|M_{t,3}^{\mathsf{b}}\|^{2p}]\,.$$

Now, applying Burkholder's inequality to the sequence $\{(\prod_{i=s+1}^t G_i^{\mathsf{avg}}-\prod_{i=s+1}^t\Gamma_i^{\mathsf{avg}})(w_{s,h,c}-1)\varepsilon^c(\theta_\star,Z_{s,h,c})\}$, we get

$$\begin{aligned}&\mathbb{E}^{1/p}[\|M_{t,2}^{\mathsf{b}}\|^p]\\ &\leq p^3N^{-1}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2\sum_{c=1}^{N}\sum_{h=1}^{H_s}\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}G_i^{\mathsf{avg}}-\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\|^p]\mathbb{E}^{2/p}[\|(w_{s,h,c}-1)\varepsilon^c(\theta_\star,Z_{s,h,c})\|^p]\right)^{1/2}\\ &\leq 4a^{-1}\,\mathrm{C}_{\mathbf{A}}\,W_{\max}p^6N^{-1}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^3H_s\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\right)^{1/2}(\|\varepsilon\|_\infty+\mathrm{C}_{\mathbf{A}}\,\zeta_3)\\ &\leq \mathrm{c}_{M,2}^{\mathsf{b}}\,p^6N^{-1}(1+t)^{-\gamma_\eta/2}\,,\end{aligned}$$

where

$$\mathrm{c}^{\mathsf{b}}_{M,2} = 8a^{-1}W_{\max}\,\mathrm{C}_{\mathbf{A}}(\eta H^{1/2} + a^{-1/2}L^{1/2}\eta^{1/2}(1 + (2a\eta H)^{\frac{\gamma+2\gamma}{2(1-\gamma)}}))(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)\,,$$

and we used that

$$\mathbb{E}^{1/p}[\| \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}}\|^p] \le 2a^{-1/2}\bar{\sigma}_{\mathbf{A}} N^{-1/2}\eta_s^{1/2}\exp\left(-a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right).$$

Again, applying Burkholder's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}^{1/p}[\|M^{\mathsf{b}}_{t,3}\|^p] \\
&\le p^3 N^{-1}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^2\sum_{c=1}^{N}\sum_{h=1}^{H_s}\mathbb{E}^{2/p}[\|\prod_{i=s+1}^{t}\Gamma_i^{\mathsf{avg}}\|^p]\mathbb{E}^{2/p}[\|(w_{s,h,c}-1)(\mathrm{I}-\Gamma^c_{s,h+1:H_s})\varepsilon^c(\theta_\star, Z_{s,h,c})\|^p]\right)^{1/2} \\
&\le 2\,\mathrm{C}_{\mathbf{A}}\,W_{\max}p^3N^{-1/2}\eta_t^{-1/2}\left(\sum_{s=1}^{t}\eta_s^4H_s^3\exp\left(-2a\sum_{\ell=s+1}^{t}\eta_\ell H_\ell\right)\right)^{1/2}(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3) \\
&\le \mathrm{c}^{\mathsf{b}}_{M,3}\,p^3N^{-1/2}(1+t)^{-\gamma}\,,
\end{aligned}$$

where

$$\mathrm{c}^{\mathsf{b}}_{M,3} = 2\,\mathrm{C}_{\mathbf{A}}\,W_{\max}((\eta H)^{3/2} + a^{-1/2}L^{1/2}\eta H(1 + (2a\eta H)^{\frac{3\gamma+\gamma_\eta}{2(1-\gamma)}}))(\|\varepsilon\|_\infty + \mathrm{C}_{\mathbf{A}}\,\zeta_3)\,,$$

and we also used that

$$\mathbb{E}^{1/p}[\|\mathrm{I}-\Gamma^c_{s,h-1}\|^p] \le \mathrm{C}_{\mathbf{A}}\,\eta_s H_s\,.$$

To prove (107), we apply Markov's inequality with $p = \log(t)$. ☐

Lemma 15. *Assume* ***A****1(2),* ***A****2 and* ***S****2. For any $k \neq c$ and $n \neq t$, we have $\Gamma^k_{n,\alpha:\beta} - \Gamma^{k,(t,j,c)}_{n,\alpha:\beta} = 0$. For the case $k = c$ and $n = t$, we obtain the bound*

$$\mathbb{E}^{1/2}[\|\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta}\|^2] \le 4\eta_t \mathrm{e}^{-a\eta_t(\beta-\alpha+1)}\sqrt{\mathsf{Tr}(\Sigma^{\mathsf{c}}_{\mathbf{A}})},$$

where $1 \le \alpha \le \beta \le H_n$.

Proof. Decompose $\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta}$ using Lemma 22

$$\begin{aligned}
&\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta} = \\
&= \sum_{i=\alpha}^{\beta}\left\{\prod_{\ell=\alpha}^{i-1}(\mathrm{I}-\eta_t\mathbf{A}^c(Z^c_{t,\ell}))\right\}\eta_t\left(\mathbf{A}^c(Z^{(t,j,c)}_{t,i,c}) - \mathbf{A}^c(Z_{t,i,c})\right)\left\{\prod_{\ell=i+1}^{\beta}(\mathrm{I}-\eta_t\mathbf{A}^c(Z^{(t,j,c)}_{t,\ell,c}))\right\}
\end{aligned}$$

For any $i \neq j$ the corresponding summand is zero. Therefore

$$\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta} = \left\{\prod_{\ell=\alpha}^{j-1}(\mathrm{I}-\eta_t\mathbf{A}^c(Z^c_{t,\ell}))\right\}\eta_t\left(\mathbf{A}^c(Z^{(t,j,c)}_{t,j,c}) - \mathbf{A}^c(Z_{t,j,c})\right)\left\{\prod_{\ell=j+1}^{\beta}(\mathrm{I}-\eta_t\mathbf{A}^c(Z^{(t,j,c)}_{t,\ell,c}))\right\}$$

Using assumption **A**1, we get

$$\mathbb{E}^{1/2}[\|\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta}\|^2] \le \eta_t(1-\eta_t a)^{\beta-\alpha}\mathbb{E}^{1/2}[\|\mathbf{A}^c(Z^{(t,j,c)}_{t,j,c}) - \mathbf{A}^c(Z_{t,j,c})\|^2]$$

As $\mathbf{A}^c(Z^{(t,j,c)}_{t,j,c}) - \mathbf{A}^c(Z_{t,j,c}) = \widetilde{\mathbf{A}}^c(Z^{(t,j,c)}_{t,j,c}) - \widetilde{\mathbf{A}}^c(Z_{t,j,c})$ then, using Minkowski's inequality, we get

$$\begin{aligned}
\mathbb{E}^{1/2}[\|\Gamma^c_{t,\alpha:\beta} - \Gamma^{c,(t,j,c)}_{t,\alpha:\beta}\|^2] &\le \frac{2}{1-\eta_t a}\eta_t(1-\eta_t a)^{\beta-\alpha+1}\mathbb{E}^{1/2}[\|\widetilde{\mathbf{A}}^c(Z_{t,j,c})\|^2] \\
&\le 4\eta_t\mathrm{e}^{-a\eta_t(\beta-\alpha+1)}\sqrt{\mathrm{Tr}(\Sigma^c_{\mathbf{A}})}
\end{aligned}$$

☐

Lemma 16. *Under assumptions **A**1(2), **A**2 and **S**2 with $\gamma_\eta \geq \gamma_H$, for any $s \in \{0, \ldots, t-1\}$, we have*

$$\mathbb{E}[\| \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \|^2] \leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2 N^{-1} \eta_s \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right).$$

Proof. Using Lemma 22, we also note that the sequence $\{ \prod_{j=s+1}^{i-1} G_j^{\mathsf{avg}} (G_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{avg}}) \prod_{j=i}^{t} \Gamma_j^{\mathsf{avg}}, s+1 \leq i \leq t\}$ is martingale-difference, thus

$$\begin{aligned}
\mathbb{E}[\| \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \|^2] &= \sum_{i=s+1}^{t} \mathbb{E}[\| \prod_{j=s+1}^{i-1} G_j^{\mathsf{avg}} (G_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{avg}}) \prod_{j=i}^{t} \Gamma_j^{\mathsf{avg}} \|^2] \\
&\leq \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sum_{i=s+1}^{t} \exp(2a\eta_i H_i) \mathbb{E}[\|G_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{avg}}\|^2].
\end{aligned}$$

Again, applying Lemma 22 to the last term, we obtain

$$\begin{aligned}
\mathbb{E}[\|G_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{avg}}\|^2] &\leq N^{-2} \eta_i^2 \sum_{c=1}^{N} \sum_{h=1}^{H_i} \exp(-2a\eta_i(H_i - h - 1)) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \\
&\leq \bar{\sigma}_{\mathbf{A}}^2 N^{-1} \eta_i^2 H_i .
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\mathbb{E}[\| \prod_{i=s+1}^{t} G_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} \|^2] &\leq \mathrm{e} \bar{\sigma}_{\mathbf{A}}^2 N^{-1} \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sum_{i=s+1}^{t} \eta_i^2 H_i \\
&\leq 2a^{-1} \bar{\sigma}_{\mathbf{A}}^2 N^{-1} \eta_s \exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right),
\end{aligned}$$

where we used the inequality $k \exp(-2ck) \leq (c\mathrm{e})^{-1} \exp(-ck)$. □

Lemma 17. *Under assumptions **A**1(2), **A**2 and **S**2 with $\gamma_\eta \geq \gamma_H$, for any $s \in \{1, \ldots, t\}, h \in \{1, \ldots, H_s\}, c \in \{1, \ldots, N\}$, we have*

$$\mathbb{E}[\|\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] \leq 16 N^{-2} \eta_s^2 \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \|\theta_\star^c - \theta_\star\|^2 .$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}[\|\Delta_s^{\mathsf{avg}} - \Delta_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] &= N^{-2} \mathbb{E}[\|(\Gamma_s^{c,(s,h,c)} - \Gamma_s^c)(\theta_\star^c - \theta_\star)\|^2] \\
&\leq 16 N^{-2} \eta_s^2 \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) \|\theta_\star^c - \theta_\star\|^2 .
\end{aligned}$$

where we have applied Lemma 15. □

Lemma 18. *Under assumptions **A**1(2), **A**2 and **S**2 with $\gamma_\eta \geq \gamma_H$, for any $s \in \{1, \ldots, t\}, h \in \{1, \ldots, H_s\}, c \in \{1, \ldots, N\}$ and $s' < s$, we have*

$$\mathbb{E}[\| \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \|^2] \leq 4\mathrm{e} N^{-2} \eta_s^2 \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c) .$$

Proof. Using Minkowski's inequality and Lemma 22, we note that

$$\mathbb{E}[\| \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \|^2] \leq \mathbb{E}[\| \prod_{i=s'+1}^{s-1} \Gamma_i^{\mathsf{avg}} (\Gamma_s^{\mathsf{avg}} - \Gamma_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}) \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \|^2],$$

since $\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} = 0$ for $i \neq s$. Thus, we get

$$\mathbb{E}[\| \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s'+1}^{t} \Gamma_i^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})} \|^2] \leq \mathrm{e}\mathbb{E}[\|\Gamma_s^{\mathsf{avg}} - \Gamma_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right)$$
$$\leq 4\mathrm{e} N^{-2} \eta_s^2 \exp\left(-2a \sum_{\ell=s'+1}^{t} \eta_\ell H_\ell\right) \mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\,.$$

where we have used that

$$\mathbb{E}[\|\Gamma_s^{\mathsf{avg}} - \Gamma_s^{\mathsf{avg},(\mathsf{s},\mathsf{h},\mathsf{c})}\|^2] = N^{-2}\mathbb{E}[\|\Gamma_{s,H_s}^c - \Gamma_{s,H_s}^{c,(s,h,c)}\|^2] \leq 4N^{-2}\eta_s^2 \mathrm{Tr}(\Sigma_{\mathbf{A}}^c)\,.$$

□

Lemma 19. *Under the assumptions **A1**(2), **A2** and **S2**, for any $1 \leq i \leq t$, we have*

$$\mathbb{E}^{1/p}[\|\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p] \leq \mathrm{C}_{\mathbf{A}}\, m_p p^2 N^{-1/2} \eta_i H_i^{1/2}\,.$$

Proof. Applying Burkholder's inequality, we get

$$\mathbb{E}^{1/p}[\|\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p] \leq pN^{-1}\left(\sum_{c=1}^{N} \mathbb{E}^{2/p}[\|\Gamma_i^c - \Gamma_i^{\mathsf{b},c}\|^p]\right)^{1/2}$$
$$\leq \mathrm{C}_{\mathbf{A}}\, m_p p^2 N^{-1/2} \eta_i H_i^{1/2}\,.$$

where, using the martingale-difference structure of the sequence $\{(w_{s,h,c}-1)\Gamma_{i,1:h-1}^c \mathbf{A}^z(Z_{s,h,c})\Gamma_{i,h+1:H_i}^c, 1 \leq h \leq H_i\}$ w.r.t filtration $\mathcal{F}_{i,h}^{\mathsf{b}}$, we have that

$$\mathbb{E}^{1/p}[\|\Gamma_i^c - \Gamma_i^{\mathsf{b},c}\|^p] \leq p\eta_i \left(\sum_{h=1}^{H_i} \mathbb{E}^{2/p}[\|(w_{s,h,c}-1)\Gamma_{i,1:h-1}^c \mathbf{A}^c(Z_{s,h,c})\Gamma_{i,h+1:H_i}^{\mathsf{b},c}\|^p]\right)^{1/2}$$
$$\leq \mathrm{C}_{\mathbf{A}}\, m_p p \eta_i H_i^{1/2}\,.$$

□

Lemma 20. *Under the assumptions **A1**(2), **A2** and **S2**, we have*

$$\mathbb{E}^{1/p}[\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \|^p] \leq \mathrm{e}^{1/2} a^{-1/2}\, \mathrm{C}_{\mathbf{A}}\, m_p p^3 N^{-1/2} \exp\left(-\frac{a}{2} \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right).$$

Proof. Using Lemma 22, the fact that the sequence $\{\prod_{j=s+1}^{i-1} \Gamma_i^{\mathsf{avg}}(\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{b},\mathsf{avg}}) \prod_{j=i}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}}, s+1 \leq i \leq t\}$ it martingale-difference w.r.t filtration $\mathcal{F}_{s,h}^{\mathsf{b},+}$ and applying Burkholder's inequality, together with Lemma 19, we obtain

$$\mathbb{E}^{1/p}[\| \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{avg}} - \prod_{i=s+1}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \|^p] \leq p\left(\sum_{i=s+1}^{t} \mathbb{E}^{2/p}[\| \prod_{j=s+1}^{i-1} \Gamma_i^{\mathsf{avg}}(\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{b},\mathsf{avg}}) \prod_{j=i}^{t} \Gamma_i^{\mathsf{b},\mathsf{avg}} \|^p]\right)^{1/2}$$
$$\leq \mathrm{e}p\left(\exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sum_{i=s+1}^{t} \mathbb{E}^{2/p}[\|\Gamma_i^{\mathsf{avg}} - \Gamma_i^{\mathsf{b},\mathsf{avg}}\|^p]\right)^{1/2}$$
$$\leq \mathrm{e}\, \mathrm{C}_{\mathbf{A}}\, m_p p^3 N^{-1/2} \left(\exp\left(-2a \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right) \sum_{i=s+1}^{t} \eta_i^2 H_i\right)^{1/2}$$
$$\leq \mathrm{e}^{1/2} a^{-1/2}\, \mathrm{C}_{\mathbf{A}}\, m_p p^3 N^{-1/2} \exp\left(-\frac{a}{2} \sum_{\ell=s+1}^{t} \eta_\ell H_\ell\right).$$

where we have used that $k\exp(-2ck) \leq (c\mathrm{e})^{-1}\exp(-ck)$. □

Below we state the Gaussian comparison inequality due to Devroye et al. [2018], see also Barsov and Ulyanov [1986].

Lemma 21. *Let Σ_1 and Σ_2 be positive definite covariance matrices in $\mathbb{R}^{d\times d}$. Let $X \sim \mathrm{N}(0,\Sigma_1)$ and $Y \sim \mathrm{N}(0,\Sigma_2)$. Then*

$$\mathsf{d}_{\mathsf{TV}}(X,Y) \le \frac{3}{2}\|\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2} - \mathrm{I}\|_{\mathsf{F}}\,.$$

Here we present some useful technical lemmas we use during the proofs of our main results.

Lemma 22. *For any matrix-valued sequences $(A_i)_{i=1}^N$ and $(B_i)_{i=1}^N$ it holds that:*

$$\prod_{i=1}^N A_i - \prod_{i=1}^N B_i = \sum_{i=1}^N \Big\{\prod_{j=1}^{i-1} A_j\Big\}(A_i - B_i)\Big\{\prod_{j=i+1}^N B_j\Big\}$$

Lemma 23. *Assume $\eta_t = c_0(t_0+t)^{-\gamma}$ for some $\gamma \in [1/2,1)$. Then:*

$$\sum_{i=s+1}^t \eta_i \le \frac{\eta}{1-\gamma}\left((1+t)^{1-\gamma} - (1+s)^{1-\gamma}\right)$$

Proof. By definition of η_t:

$$\sum_{i=s+1}^t \eta_i = \sum_{i=s+1}^t \eta(1+i)^{-\gamma} = \eta \sum_{k=s+2}^{1+t} k^{-\gamma}$$

As $f(x) = x^{-\gamma}$ is a decreasing function for $x \ge 0$, then:

$$c_0 \sum_{k=s+2}^{1+t} k^{-\gamma} \le \eta \int_{1+s}^{1+t} x^{-\gamma}\mathrm{d}x = \frac{\eta}{1-\gamma}((1+t)^{1-\gamma} - (1+s)^{1-\gamma})\,.$$

□

Lemma 24. *Assume $\gamma \in (0,1)$ and $1 \le \ell \le r \in \mathbb{N}$. Then,*

$$\frac{(r+2)^{1-\gamma} - (\ell+1)^{1-\gamma}}{1-\gamma} \le \sum_{k=\ell}^r (1+k)^{-\gamma} \le \frac{(r+1)^{1-\gamma} - \ell^{1-\gamma}}{1-\gamma}\,.$$

Proof. As $f(x) = (1+x)^{-\gamma}$ is a decreasing function, then

$$\sum_{k=\ell}^r (1+k)^{-\gamma} \le \int_{\ell-1}^r (1+x)^{-\gamma}\mathrm{d}x = \int_\ell^{r+1} x^{-\gamma}\mathrm{d}x = \frac{(r+1)^{1-\gamma} - \ell^{1-\gamma}}{1-\gamma}\,.$$

Similarly, we can give a lower bound of our function as,

$$\sum_{k=\ell}^r (1+k)^{-\gamma} \ge \int_\ell^{r+1} (1+x)^{-\gamma}\mathrm{d}x = \int_{\ell+1}^{r+2} (1+x)^{-\gamma}\mathrm{d}x = \frac{(r+2)^{1-\gamma} - (\ell+1)^{1-\gamma}}{1-\gamma}$$

□

Lemma 25. *For any $t \ge 1$ and $\beta < 1$, it holds that*

$$\sum_{s=1}^{t-1} s^{-\alpha} \exp\{-u \sum_{\ell=s+1}^t (1+\ell)^{-\beta}\} \le Lu^{-1}(1+u^{-\frac{\alpha}{1-\beta}})t^{\beta-\alpha}\,,$$

where L is defined in (108).

Proof. We set $m = \lceil \frac{t}{2} \rceil$. To estimate

$$S_t := \sum_{s=1}^{t-1} s^{-\alpha} \exp\left(-u \sum_{\ell=s+1}^{t} (1+\ell)^{-\beta}\right),$$

we split this sum into two parts:

$$S_t = \underbrace{\sum_{s=1}^{m} s^{-\alpha} \exp\left(-u \sum_{\ell=s+1}^{t} (1+\ell)^{-\beta}\right)}_{S_{t,1}} + \underbrace{\sum_{s=m+1}^{t} s^{-\alpha} \exp\left(-u \sum_{\ell=s+1}^{t} (1+\ell)^{-\beta}\right)}_{S_{t,2}}.$$

Estimating $S_{t,1}$. Here, we note that for $s \leq m$:

$$\sum_{\ell=s+1}^{t} (1+\ell)^{-\beta} \geq \sum_{\ell=m+1}^{t} (1+\ell)^{-\beta} \geq \frac{t}{2} t^{-\beta} = \frac{1}{2} t^{1-\beta}.$$

Then

$$S_{t,1} \leq \exp\left(-\frac{u}{2} t^{1-\beta}\right) \sum_{s=1}^{m} s^{-\alpha} \leq t \exp\left(-\frac{u}{2} t^{1-\beta}\right).$$

We can see that the function $g(x) = x^{1+\alpha-\beta} \exp\left(-v x^{1-\beta}\right)$ for $x \geq 1$ can be uniformly bounded by the value

$$\begin{aligned} L_0(v) = \left(\frac{1+\alpha-\beta}{v\mathrm{e}(1-\beta)}\right)^{\frac{1+\alpha-\beta}{1-\beta}} &\leq \mathrm{e}^{-1}\left(1+\frac{\alpha}{1-\beta}\right)^{1+\frac{\alpha}{1-\beta}} v^{-\left(1+\frac{\alpha}{1-\beta}\right)} \\ &\leq L_1 v^{-\left(1+\frac{\alpha}{1-\beta}\right)}, \end{aligned}$$

where we have set $L_1 = \mathrm{e}^{-1}\left(1+\frac{\alpha}{1-\beta}\right)^{1+\frac{\alpha}{1-\beta}}$. Therefore, we obtain the bound

$$S_{t,1} \leq L_0(u/2) t^{\beta-\alpha}.$$

Estimating $S_{t,2}$. Here, we know that $s^{-\alpha} \leq (t/2)^{-\alpha}$. That is, we get

$$\begin{aligned} S_{t,2} \leq 2^{\alpha} t^{-\alpha} \sum_{s=m+1}^{t-1} \exp\left(-u(t-s)(1+t)^{-\beta}\right) &\leq 2^{\alpha} t^{-\alpha} e^{-u(1+t)^{-\beta}} \left(1 - e^{-u(1+t)^{-\beta}}\right)^{-1} \\ &\leq 2^{\alpha+\beta} u^{-1} t^{\beta-\alpha}, \end{aligned}$$

where we have used that $\frac{1}{1-e^{-x}} \leq \frac{a}{1-e^{-a}} x^{-1}$ for any $x \leq a$, that is, in our case $a = 2^{-\beta} u$. Finally, setting

$$L = 2^{\alpha+\beta} + 2^{1+2\alpha} L_1, \tag{108}$$

we obtain the result. □