Convergence Guarantees for Federated SARSA with Local Training and Heterogeneous Agents

Paul Mangold [1] **Eloïse Berthier** [2] **Eric Moulines** [1 3]

Abstract

We present a novel theoretical analysis of Federated SARSA (FedSARSA) with linear function approximation and local training. We establish convergence guarantees for FedSARSA in the presence of heterogeneity, both in local transitions and rewards, providing the first sample and communication complexity bounds in this setting. At the core of our analysis is a new, exact multi-step error expansion for single-agent SARSA, which is of independent interest. Our analysis precisely quantifies the impact of heterogeneity, demonstrating the convergence of FedSARSA with multiple local updates. Crucially, we show that FedSARSA achieves linear speed-up with respect to the number of agents, up to higher-order terms due to Markovian sampling. Numerical experiments support our theoretical findings.

1. Introduction

Federated reinforcement learning (FRL) (Zhuo et al., 2019) allows multiple agents to collaboratively learn a policy without exchanging raw data. By sharing information, agents can accelerate training by leveraging one another's experience. This paradigm is particularly valuable when communication, storage, or privacy constraints preclude direct data sharing. Yet, effective learning in such settings remains difficult. Communication is costly, and while federated methods try to mitigate this by relying on local updates, these updates can cause client drift when environment dynamics are different from one agent to another.

Despite rapid progress in FRL (Qi et al., 2021; Jin et al., 2022; Khodadadian et al., 2022), little attention has been given to the federated counterpart of the classical on-policy method SARSA. This method, which simultaneously updates and evaluates a policy, is fundamental in RL. Most existing works have studied either federated policy evaluation with TD learning (Wang et al., 2024; Mitra et al., 2024; Mangold et al., 2024; Beikmohammadi et al., 2025), methods like Q-learning (Khodadadian et al., 2022; Woo et al., 2025; Zheng et al., 2024), or policy gradient (Yang et al., 2024; Lan et al., 2025; Labbi et al., 2025a). A notable exception is Zhang et al. (2024), who analyzed the FedSARSA algorithm and showed that federated training can reduce variance. However, when agents are heterogeneous, their convergence rate is affected by a persistent bias, which remains even with a single local update.

In this paper, we present a novel analysis of the FedSARSA algorithm *with multiple local training steps*, establishing explicit finite-time convergence bounds. We also introduce a federated variant of fitted SARSA (Zou et al., 2019), where the policy is updated only at communication rounds, guaranteeing that all agents are using the same policy at all times. Our analysis begins with a new framework for single-agent SARSA, based on a novel expansion of the error along the trajectory that separates transient and fluctuation components, and tightly characterizes the impact of Markovian noise. We then extend this novel analysis to the federated setting, establishing sharp convergence guarantees where we quantify explicitly the impact of heterogeneity in local updates. Unlike prior work (Zhang et al., 2024), our results do not rely on an averaged environment assumption, and directly characterize the local drift arising from multiple heterogeneous local steps, allowing for a tighter characterization of the impact of agent heterogeneity on FedSARSA.

An important feature of our analysis is the characterization of the limiting point, to which FedSARSA converges. This allows for a rigorous analysis of the algorithm when both transition kernels and reward functions differ across agents. The main contributions of our work are the following:

1. We develop a novel analysis of federated SARSA, precisely characterizing its convergence point and deriving explicit finite-time bounds that quantify the effect of environmental heterogeneity. We provide the first sample complexity bounds for FedSARSA, showing that it achieves *linear speed-up* in the number of agents. This

[1]CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France [2]Unité d'Informatique et d'Ingénierie des Systèmes, ENSTA, Institut Polytechnique de Paris, 91120 Palaiseau, France [3]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE. Correspondence to: Paul Mangold <paul.mangold@polytechnique.edu>.

Preprint. January 29, 2026.

follows from a careful analysis of the impact of Markovian noise, showing that higher-order variance terms scale proportionally to the chain's mixing time τ_{mix}.

2. At the core of our analysis lies a new analysis of single-agent SARSA based on a refined error decomposition and a sharp characterization of the impact of Markovian noise, which is of independent interest.
3. We provide numerical illustrations validating our theory and empirically demonstrating the linear speed-up of FedSARSA. We also provide a JAX implementation of a deep FedSARSA variant, demonstrating the applicability of our method to federated deep RL.

The paper is organized as follows: we introduce FRL in Section 3, and discuss related work in Section 2. We present our analysis of single-agent SARSA in Section 4, and our results on FedSARSA in Section 5. We then illustrate our theory numerically in Section 6.

Notations. We denote by $\langle \cdot\,, \cdot \rangle$ the Euclidean inner product and by $\| \cdot \|$ its associated norm. All vectors are column vectors. We let I be the $d \times d$ identity matrix and e_j the j-th vector of the canonical basis of $\mathbb{R}^d$. For a matrix $A \in \mathbb{R}^{d,d}$, we denote $A_{i,j}$ its i,j-th coordinate, and for a vector $b \in \mathbb{R}^d$, we denote b_i its i-th coordinate. For two sequences $(u_n)_{n\geq 0}$ and $(v_n)_{n\geq 0}$, we write $u_n \lesssim v_n$ if there exists $c > 0$ such that $u_n \leq c v_n$ for all $n \geq 0$, and $u_n \approx v_n$ if both $u_n \lesssim v_n$ and $v_n \lesssim u_n$. For a closed convex set $\mathcal{W} \subseteq \mathbb{R}^d$, $\Pi_{\mathcal{W}}$ denotes the projection onto $\mathcal{W}$. Finally, for a set X, $\mathcal{P}(X)$ denotes the set of probability measures on the measurable space $(X, \mathcal{B}(X))$, where $\mathcal{B}(X)$ is the Borel σ-field of X.

2. Related Work

SARSA and RL. *(Tabular.)* The SARSA algorithm (short for State–Action–Reward–State–Action), introduced by Rummery & Niranjan (1994), is a classical *on-policy* reinforcement learning method. Singh et al. (2000) proved its convergence in the tabular case, showing that if the policy becomes greedy while maintaining exploration (the greedy-in-the-limit with infinite exploration condition), the state-action values converge to the unique optimal solution. This result extended earlier convergence proofs for *off-policy* Q-learning (Watkins & Dayan, 1992).

(Linear function approximation.) In large or continuous state spaces, SARSA is combined with linear function approximation (LFA) for the state-action function using an embedding in $\mathbb{R}^d$. Tsitsiklis & Van Roy (1997) established the convergence of TD(0) to a fixed point of a linear equation; subsequent works analyzed it in both asymptotic and finite-time regimes (Tsitsiklis & Van Roy, 1997; Bhandari et al., 2018; Samsonov et al., 2024). SARSA generalizes TD learning (Vamvoudakis et al., 2021; Meyn, 2022), estimating the state-action value of the current policy via TD updates while improving the policy. De Farias & Van Roy (2000) proved the existence of a solution for SARSA using fixed-point arguments. Related convergence results have also been established for Q-learning (Melo et al., 2008) and actor-critic (Wu et al., 2020; Barakat et al., 2022).

(Convergence Rates for SARSA.) Extending TD to SARSA requires a policy improvement operator Imp_β (formally defined in (2)), which updates the policy from the parameters learned via TD. In analyses of single-agent SARSA, it is commonly assumed that Imp_β satisfies a Lipschitz condition, with $\mathrm{C}_{\mathrm{lip}} \geq 0$,

$$|\,\mathrm{Imp}_\beta(\theta_1)(a|s) - \mathrm{Imp}_\beta(\theta_2)(a|s)| \leq \mathrm{C}_{\mathrm{lip}} \|\theta_1 - \theta_2\|,$$

for all $s, a \in \mathcal{S} \times \mathcal{A}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$. Setting $\mathrm{C}_{\mathrm{lip}} = 0$ recovers TD(0), without policy improvement. Perkins & Precup (2002) proved convergence of a SARSA variant with LFA, using multiple TD updates between improvements, small $\mathrm{C}_{\mathrm{lip}}$, and ε-greedy policies. Melo et al. (2008) later established the first asymptotic convergence proof for SARSA with LFA. Under a similar Lipschitz assumption on Imp_β, Zou et al. (2019) gave a non-asymptotic convergence analysis of SARSA with LFA, using projection onto a bounded ball at each step, akin to TD(0) with Markovian noise (Bhandari et al., 2018). More recently, Zhang et al. (2023) studied SARSA with larger $\mathrm{C}_{\mathrm{lip}}$, identifying a *chattering* phenomenon (Gordon, 1996; 2000).

Federated Reinforcement Learning. Federated reinforcement learning (FRL) (Zhuo et al., 2019; Qi et al., 2021) generalizes federated learning (McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020; Ogier du Terrail et al., 2022) to sequential decision making. Early theoretical work concentrated on policy evaluation, particularly federated TD–type methods under generative and online-interaction setting, establishing finite-time convergence guarantees and often linear-in-agents speedups (Khodadadian et al., 2022; Dal Fabbro et al., 2023; Tian et al., 2024). A parallel line of research addresses environment heterogeneity (agents facing distinct MDPs), characterizing how aggregation must adapt to model mismatch and under what conditions collaboration remains beneficial (Jin et al., 2022; Woo et al., 2025). Beyond evaluation, tabular setting has received complexity analyses. For Q-learning, recent results establish linear speedups together with sharp—often optimal or near-optimal—sample and communication bounds (Khodadadian et al., 2022; Woo et al., 2025; Zheng et al., 2024; Salgia & Chi, 2024). For value-iteration–style methods, federated regret bounds with linear speedups and explicit heterogeneity terms are now available (Labbi et al., 2025b). On the policy-gradient side (generative/simulator-oracle setting), theory now encompasses both natural policy gradient and actor–critic methods (Lan et al., 2023; Yang et al., 2024; Jor-

dan et al., 2024). In contrast, the online-interaction setting in FRL remains comparatively underexplored. A notable exception is Zhang et al. (2024), who analyze federated SARSA under agent heterogeneity and prove convergence to a heterogeneity-dependent neighborhood of the optimum: convergence cannot, in general, be made arbitrarily precise.

3. Background on Federated SARSA

Federated Reinforcement Learning. In FRL, $N > 0$ agents collaborate to learn a single shared policy. Formally, each agent's environment is modeled as a Markov Decision Process (MDP), yielding N MDPs $\{(\mathcal{S}, \mathcal{A}, P^{(c)}, r^{(c)}, \gamma)\}_{c\in\{1,\dots,N\}}$, with a common state space $\mathcal{S}$, action space $\mathcal{A}$, and discount factor $\gamma \in (0,1)$. Each agent $c \in \{1,\dots,N\}$ has its own transition kernel $P^{(c)}$, where $P^{(c)}(\cdot|s,a)$ denotes the probability of transitioning from state $s \in \mathcal{S}$ after action $a \in \mathcal{A}$, and a deterministic reward function $r^{(c)} : \mathcal{S}\times\mathcal{A} \to [0,1]$. State-action pairs are embedded in $\mathbb{R}^d$ via a feature map $\phi : (s,a) \mapsto \phi(s,a)$. For a policy π_θ parameterized by $\theta \in \mathbb{R}^d$, we denote by $P_\theta^{(c)}$ the induced state transition kernel and by $\mu_\theta^{(c)}$ the stationary distribution satisfying $\mu_\theta^{(c)} P_\theta^{(c)} = \mu_\theta^{(c)}$. In this context, heterogeneity lies both in the transition kernel and the rewards. We measure it using ϵ_p and ϵ_r, defined as

$$\begin{aligned}\epsilon_p &\triangleq \sup_{\varrho\in\mathcal{P}(\mathcal{S}\times\mathcal{A})}\sup_{c,c'\in\{1,\dots,N\}} \|\varrho P^{(c)} - \varrho P^{(c')}\|_{\mathrm{TV}} \ , \\ \epsilon_r &\triangleq \sup_{s,a\in\mathcal{S}\times\mathcal{A}}\sup_{c,c'\in\{1,\dots,N\}} |r^{(c)}(s,a) - r^{(c')}(s,a)| \ ,\end{aligned} \tag{1}$$

where ϵ_p measures heterogeneity in the transition dynamics and ϵ_r heterogeneity in the rewards. This measure of heterogeneity is classical in federated RL, and has been used in many prior works (Wang et al., 2024; Zhang et al., 2024).

Federated SARSA. SARSA combines TD learning (Sutton, 1988) with policy improvement: TD updates estimate the state-action value function, which is then used to update the policy. In its federated version (FedSARSA), agents collaboratively learn a shared policy. Each agent performs several local TD updates, after which a central server aggregates the local estimators to update the global policy.

We approximate the state-action value function by a linear model $Q_\theta : (s,a) \mapsto \phi(s,a)^\top\theta$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$, where ϕ is a fixed embedding of state-action pairs. At global iteration $t \geq 0$ and local iteration $h \geq 0$, the parameter $\theta_{t,h}^{(c)} \in \mathbb{R}^d$ of agent $c \in \{1,\dots,N\}$ is updated via the local TD rule

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

where, for $h \geq 0$, $Z_{t,h}^{(c)} = (S_{t,h}^{(c)}, A_{t,h}^{(c)}, S_{t,h+1}^{(c)}, A_{t,h+1}^{(c)})$ takes values in $\mathsf{Z} \triangleq (\mathcal{S}\times\mathcal{A})^{\times 2}$, and represents the current and the next observed state-action pairs. The TD error is defined at iteration t,h by $\mathbf{A}^{(c)}(Z_{t,h+1}^{(c)})\theta_{t,h}^{(c)} + \mathbf{b}^{(c)}(Z_{t,h+1}^{(c)})$ where for $z = (s,a,s',a') \in \mathsf{Z}$, define $\mathbf{A}^{(c)}$ and $\mathbf{b}^{(c)}$ as

$$\begin{aligned}\mathbf{A}^{(c)}(z) &= \phi(s,a)\left(\gamma\phi(s',a')^\top - \phi(s,a)^\top\right) \\ \mathbf{b}^{(c)}(z) &= \phi(s,a)r^{(c)}(s,a) \ .\end{aligned}$$

After $H > 0$ local TD updates are performed, the local parameters are sent to the server, which averages them and projects the result on a compact convex set $\mathcal{W} \subseteq \mathbb{R}^d$,

$$\theta_{t+1} = \mathrm{proj}_{\mathcal{W}}(\bar\theta_{t+1}), \text{ where } \bar\theta_{t+1} = \tfrac{1}{N}\textstyle\sum_{c=1}^N \theta_{t,H}^{(c)} \ .$$

The global, shared policy is then updated using the softmax of the approximated state-action value

$$\mathrm{Imp}_\beta \colon \theta \mapsto \pi_\theta, \text{where } \pi_\theta(a|s) \propto \exp(\beta\phi(s,a)^\top\theta) \tag{2}$$

for all $s,a \in \mathcal{S}\times\mathcal{A}$, and $\beta > 0$ is referred to as the *sharpness* of the policy. Note that this policy improvement operator is $\mathrm{C_{lip}}$-Lipschitz, with $\mathrm{C_{lip}} = \beta$, that is that for θ,θ', we have, for $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$|\pi_\theta(a|s) - \pi_{\theta'}(a|s)| \leq \mathrm{C_{lip}}\|\theta - \theta'\| \ .$$

We give the pseudo-code for FedSARSA in Algorithm 1.

Assumptions. In the following, we assume the feature map to be bounded, which gives uniform bounds on $\mathbf{A}^{(c)}(\cdot)$ and $\mathbf{b}^{(c)}(\cdot)$ and restricts the diameter of $\mathcal{W}$.

A 1. *The state-action feature map ϕ is such that $\sup_{s,a\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\| \leq 1$. This gives the almost sure bounds, for any $c \in \{1,\dots,N\}$,*

$$\|\mathbf{A}^{(c)}\| \leq \mathrm{C}_A \triangleq (1+\gamma) \ , \quad \|\mathbf{b}^{(c)}\| \leq \mathrm{C}_b \triangleq 1 \ ,$$

where for $z \in \mathsf{Z}$, $\mathbf{A}^{(c)}(z) \in \mathbb{R}^{d\times d}$ and $\mathbf{b}^{(c)}(z) \in \mathbb{R}^d$, we defined the norms $\|\mathbf{A}^{(c)}\|^2 = \sum_{1\leq i,j\leq d}\sup_{z\in\mathsf{Z}}\mathbf{A}_{i,j}^{(c)}(z)^2$ and $\|\mathbf{b}^{(c)}\|^2 = \sum_{1\leq i\leq d}\sup_{z\in\mathsf{Z}}\mathbf{b}_i^{(c)}(z)^2$.

Given a policy π_θ with stationary distribution $\mu_\theta^{(c)}$ over the state space, we define for each agent $c \in \{1,\dots,N\}$ a stationary distribution $\nu_\theta^{(c)}$ on Z. A tuple (s,a,s',a') is drawn by sampling $s \sim \mu_\theta^{(c)}$, $a \sim \pi_\theta(\cdot|s)$, $s' \sim P^{(c)}(\cdot|s,a)$, and $a' \sim \pi_\theta(\cdot|s')$. We then define the expectations of $\mathbf{A}^{(c)}(\cdot)$ and $\mathbf{b}^{(c)}(\cdot)$ with respect to this distribution.

$$\begin{aligned}\bar{\mathbf{A}}^{(c)}(\theta) &= \mathbb{E}_{z\sim\nu_\theta}\big[\phi(s,a)\big(\gamma\phi(s',a')^\top - \phi(s,a)^\top\big)\big] \\ \bar{\mathbf{b}}^{(c)}(\theta) &= \mathbb{E}_{z\sim\nu_\theta}\big[\phi(s,a)r^{(c)}(s,a)\big] \ ,\end{aligned} \tag{3}$$

where the key difference with TD learning (Samsonov et al., 2024; Mangold et al., 2024) is that $z \sim \nu_\theta$ depends on the parameter currently being optimized θ. From these definitions, we see that a limit point $\theta_\star \in \mathbb{R}^d$ of the FedSARSA algorithm must satisfy the equation

$$\tfrac{1}{N}\textstyle\sum_{c=1}^N \bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star + \tfrac{1}{N}\sum_{c=1}^N \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \, . \tag{4}$$

Algorithm 1 FedSARSA: Federated State-Action-Reward-State-Action

Input: step sizes $\eta_t > 0$, initial parameters θ_0, projection set $\mathcal{W}$, number of local steps $H > 0$, number of communications $T > 0$, initial distribution ϱ over states
Initialize first state $s_{-1,H}^{(c)} \sim \varrho$ for $c \in \{1, \dots, N\}$ and initial policy $\pi_{\theta_0} = \mathrm{Imp}_\beta(Q_{\theta_0}$
for step $t = 0$ to $T - 1$ **do**
  for agent $c = 1$ to N **do**
    Initialize $\theta_{t,0}^{(c)} = \theta_t$, take first action $a_{t,0}^{(c)} \sim \pi_{\theta_t}(\cdot|s_{t-1,H}^{(c)})$
    for step $h = 0$ to $H - 1$ **do**
      Take action $a_{t,h+1}^{(c)} \sim \pi_{\theta_t}(\cdot|s_{t,h}^{(c)})$, observe reward $r^{(c)}(s_{t,h}^{(c)}, a_{t,h}^{(c)})$, next state $s_{t,h+1}^{(c)}$
      Compute $\delta_{t,h}^{(c)} = r^{(c)}(s_{t,h+1}^{(c)}, a_{t,h+1}^{(c)}) + \gamma\phi(s_{t,h+1}^{(c)}, a_{t,h+1}^{(c)})^\top\theta_{t,h}^{(c)} - \phi(s_{t,h}^{(c)}, a_{t,h}^{(c)})^\top\theta_{t,h}^{(c)}$
      Update $\theta_{t,h+1}^{(c)} = \theta_{t,h}^{(c)} + \eta_t\delta_{t,h}^{(c)}\phi(s_{t,h}^{(c)}, a_{t,h}^{(c)})$
    end for
  end for
  Compute $\bar{\theta}_{t+1} = \frac{1}{N}\sum_{c=1}^N \theta_{t,H}^{(c)}$ and update of global parameter $\theta_{t+1} = \Pi\left(\bar{\theta}_{t+1}\right)$
  Policy improvement $\pi_{\theta_{t+1}} = \mathrm{Imp}_\beta(Q_{\theta_{t+1}})$
end for
Return: θ_T

We discuss the existence of such points in Section 5. We assume that the solutions of (4) and $\mathcal{W}$ satisfy the following.

A 2. *There exists $a > 0$ such that, for any $\theta_\star$ satisfying (4) and $c \in \{1, \dots, N\}$, $\bar{\mathbf{A}}^{(c)}(\theta_\star)$ is negative definite and the largest eigenvalue of $1/2(\bar{\mathbf{A}}^{(c)}(\theta_\star) + \bar{\mathbf{A}}^{(c)}(\theta_\star)^\top)$ is $-a$.*

A 3. *The set $\mathcal{W} \subseteq \mathbb{R}^d$ is large enough so that there exists a $\theta_\star \in \mathcal{W}$ satisfying (4). Moreover, there exists $\mathrm{C}_{\text{proj}} > 0$ such that $\sup_{\theta\in\mathcal{W}} \|\theta\| \leq \mathrm{C}_{\text{proj}}$.*

Under A 3, we also define the following two quantities, which bound intermediate FedSARSA's iterates and updates,

$$\widetilde{\mathrm{C}}_{\text{proj}} = 4\mathrm{C}_{\text{proj}} + 1 \text{ and } \mathrm{G} \triangleq \mathrm{C}_A\widetilde{\mathrm{C}}_{\text{proj}} + \mathrm{C}_b \ . \tag{5}$$

The next two assumptions define the Markovian property of the noise and the variance of the Markov chains at stationarity, and are classical in the analysis of RL methods.

A 4. *The kernels $P_\theta^{(c)}$ have invariant distributions $\mu_\theta^{(c)}$ and are uniformly geometrically ergodic. There exists $\tau_{\text{mix}} \geq 0$ such that, for all distributions ϱ, ϱ' over Z, and $h \geq 0$,*

$$\|\varrho(P_\theta^{(c)})^h - \varrho'(P_\theta^{(c)})^h\|_{\text{TV}} \leq (1/4)^{\lfloor h/\tau_{\text{mix}}\rfloor} \ .$$

Following previous work on SARSA (Zou et al., 2019; Zhang et al., 2024), we assume the policy improvement operator's Lipschitz constant C_{lip} is not too large.

A 5. *The constant C_{lip} is such that $80\mathrm{G}\mathrm{C}_{\text{lip}}|\mathcal{A}|\tau_{\text{mix}} \leq a$.*

This assumption is classical in finite-sample analysis of SARSA and FedSARSA (Zhang et al., 2024). To our knowledge, the only theoretical work that relaxes this assumption is Zhang et al. (2023). However, without this assumption, SARSA does not necessarily converge, and one can only show that SARSA's iterates remain in a bounded domain. In this work, we aim to study the *convergence* of FedSARSA, requiring this assumption: extending our results to larger Lipschitz constants is a promising direction for future research on federated SARSA.

4. Single-Agent SARSA

First, we present our novel analytical framework for the single-agent SARSA algorithm, that is Algorithm 1 with $N = 1$, or Algorithm 2 in Appendix D. In the single-agent case, we have $\bar{\mathbf{A}}^{(1)}(\theta) = \bar{\mathbf{A}}(\theta)$, and we can define the solution $\theta_\star$ as the vector that satisfies

$$\bar{\mathbf{A}}^{(1)}(\theta_\star)\theta_\star + \bar{\mathbf{b}}^{(1)}(\theta_\star) = 0 \ . \tag{6}$$

Existing analysis (Zou et al., 2019) guarantees that such a $\theta_\star$ exists, and that it is unique under assumptions A1-5. We propose a novel decomposition of the error of SARSA as

$$\begin{aligned}\theta_{t,h+1}^{(1)} - \theta_\star = \left(\mathrm{I} + \eta_t\bar{\mathbf{A}}^{(1)}(\theta_\star)\right)(\theta_{t,h}^{(1)} - \theta_\star) \\ + \eta_t\varphi_{t,h}^{(1)} + \eta_t\varepsilon_{t,h}^{(1)}(Z_{t,h+1}^{(1)}) \ ,\end{aligned} \tag{7}$$

where we introduced the notations

$$\begin{aligned}\varphi_{t,h}^{(1)} &= (\bar{\mathbf{A}}^{(1)}(\theta_t) - \bar{\mathbf{A}}^{(1)}(\theta_\star))\theta_{t,h}^{(1)} + \bar{\mathbf{b}}^{(1)}(\theta_t) - \bar{\mathbf{b}}^{(1)}(\theta_\star) \\ \varepsilon_{t,h}^{(1)}(Z_{t,h+1}^{(1)}) &= (\mathbf{A}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{A}}^{(1)}(\theta_t))\theta_{t,h}^{(1)} \\ &\quad + \mathbf{b}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{b}}^{(1)}(\theta_t) \ .\end{aligned} \tag{8}$$

The term $\varphi_{t,h}^{(1)}$ accounts for the discrepancy between the current policy parameterized by θ_t and the optimal one, parameterized by $\theta_\star$. The second term $\varepsilon_{t,h}^{(1)}(Z_{t,h+1}^{(1)})$ is a

fluctuation term. Now, we define the following matrices, which will appear in all subsequent error decompositions,

$$\Gamma^{(1)}_{t,k:h} = \left(\mathrm{I} + \eta_t \bar{\mathbf{A}}^{(1)}(\theta_\star)\right)^{k-h+1} \text{ , for } k, h \geq 0 \text{ , } \quad (9)$$

with the convention that $\Gamma^{(1)}_{t,k:h} = \mathrm{I}$ when $h < k$. Unrolling (7) gives the following decomposition.

Claim 1. *Let $t \geq 0$. The updates of SARSA at block t can be written as*

$$\theta^{(1)}_{t,H} - \theta_\star = \Gamma^{(1)}_{t,1:H}(\theta_t - \theta_\star) + \sum_{h=1}^{H} \eta_t \Gamma^{(1)}_{t,h+1:H} \varphi^{(1)}_{t,h-1} + \sum_{h=1}^{H} \eta_t \Gamma^{(1)}_{t,h+1:H} \varepsilon^{(1)}_{t,h-1}(Z^{(1)}_{t,h}) \text{ . } \quad (10)$$

We prove this claim in Appendix C.1. We are now ready to prove the following lemma, which shows that SARSA reduces the error between consecutive blocks of updates.

Lemma 4.1. *Assume A1–5. Let $t \geq 0$, assume that the step size satisfies $\eta_t H \mathrm{C}_A \leq 1/6$. Then, it holds that*

$$\mathbb{E}[\|\theta^{(1)}_{t,H} - \theta_\star\|^2] \leq (1 - \tfrac{\eta_t a H}{4})\|\theta_t - \theta_\star\|^2 + c_1 \eta_t^2 H \tau_{\mathrm{mix}} \mathrm{G}^2 + \delta \tfrac{c_1 \eta_t \tau^2_{\mathrm{mix}} \mathrm{G}^2}{Ha} + \tfrac{c_1 \eta_t^3 H \tau^2_{\mathrm{mix}} \mathrm{G}^2 \mathrm{C}_A^2}{a} \text{ , }$$

where $c_1 > 0$ is an absolute constant and $\delta = 0$ if episodes start in the stationary distribution and $\delta = 1$ otherwise.

We prove this lemma in Appendix D. The proof is based on the error expansion from Claim 1. The first term is a transient term, which decreases linearly towards zero, the second term is a fluctuation term and third term is an error term due to sampling from the "wrong" policy. The first term can be bounded using A2, and the third term's bound follows from A5. Due to the Markovian nature of the noise, the second term requires a very careful examination, in order to handle all the correlation between pairs of iterates. In Appendix A, we provide an analysis of this Markovian error, with tight bounds depending on H and τ_{mix}.

Note that, when $\delta = 1$, one of the error term scales in η_t / H: controlling it requires setting $H \geq \tau_{\mathrm{mix}}$. This is unavoidable, since when $\delta = 1$, it is necessary to do τ_{mix} updates to get close to the stationary distribution of the Markov chain. One can eliminate this term by skipping about τ_{mix} samples before updating θ_t, ensuring that the updates start in the stationary distribution with high probability. We can now state our main theorem in the single-agent setting, which gives a convergence rate for SARSA.

Theorem 4.2. *Assume A 1–5. Assume that the step size $\eta_t = \eta$ is constant and satisfies $\eta H \mathrm{C}_A \leq 1/5$ and that $H \geq \tau_{\mathrm{mix}}$. Then it holds that*

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim (1 - \tfrac{\eta a H}{4})^T \|\theta_0 - \theta_\star\|^2 + \tfrac{c_1 \eta \tau_{\mathrm{mix}} \mathrm{G}^2}{a} + \delta \tfrac{c_1 \tau^2_{\mathrm{mix}} \mathrm{G}^2}{H^2 a^2} + \tfrac{c_1 \eta^2 \tau^2_{\mathrm{mix}} \mathrm{G}^2 \mathrm{C}_A^2}{a^2} \text{ , }$$

where δ is defined in Lemma 4.1.

We state this theorem with explicit constants and prove it in Appendix D. It shows that SARSA converges linearly to a neighborhood of $\theta_\star$, and that the size of this neighborhood is determined by the step size, the variance of the updates, and the mixing time τ_{mix}.

Corollary 4.3. *Assume A 1–5. Let $\epsilon > 0$, set $\eta \approx \min\left(\frac{1}{\mathrm{C}_A}, \frac{a\epsilon^2}{\mathrm{G}^2 \tau_{\mathrm{mix}}}, \frac{a\epsilon}{\mathrm{G}\mathrm{C}_A \tau_{\mathrm{mix}}}\right)$ and $H \approx \max\left(1, \frac{\mathrm{G}\tau_{\mathrm{mix}}}{a\epsilon}\right)$, then SARSA reaches $\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \epsilon^2$ with*

$$TH \approx \max\left(\tfrac{\mathrm{C}_A}{a}, \tfrac{\mathrm{G}^2 \tau_{\mathrm{mix}}}{a^2 \epsilon^2}, \tfrac{\mathrm{C}_A \mathrm{G} \tau_{\mathrm{mix}}}{a^2 \epsilon}\right) \log\left(\tfrac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$$

samples and $T \gtrsim \frac{\mathrm{C}_A}{a} \log\left(\frac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$ policy updates.

The proof is in Appendix D. This shows that, with proper hyperparameter settings, single-agent SARSA reaches a solution with mean squared error ϵ^2 using $O(\log(1/\epsilon))$ policy improvements and $O(1/\epsilon^2 \log(1/\epsilon))$ samples. It highlights the relevance of keeping the policy constant during blocks, reducing the need for policy improvement steps, while keeping the same overall sample complexity.

Remark 4.4. Our analysis can be extended to the setting where the policy is updated after each sample, by bounding the difference between the samples obtained with the fixed policy π_{θ_t} and the updated policy $\pi_{\theta^{(1)}_{t,h}}$, as proposed by Zou et al. (2019). We refrain from extending our analysis to this setting, since, in federated settings, one may desire the policy to remain identical for all agents at all times.

5. Convergence of FedSARSA

We now present our main result, establishing the global convergence of the FedSARSA algorithm to a point $\theta_\star$. To this end, we first establish existence and uniqueness of $\theta_\star$, in Section 5. We then extend the methodology that we introduced in Section 4 to FedSARSA in Section 5, establishing the first convergence result for FedSARSA and the corresponding sample and communication complexity.

Limit Point of FedSARSA. To identify the limit of the FedSARSA algorithm, we consider the idealized, deterministic FedSARSA algorithm, where the local updates are replaced by their expected value, and only a single local step is performed. This gives the global parameter update

$$\theta_{t+1} = \mathrm{proj}_{\mathcal{W}}\left(\theta_t + \eta_t \kappa_t\right) \text{ , } \quad (11)$$
$$\text{where } \kappa_t = \tfrac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{A}}^{(c)}(\theta_t)\theta_t + \tfrac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{b}}^{(c)}(\theta_t) \text{ . }$$

This algorithm must converge to a point $\theta_\star$ that is a fixed point of (11). However, the existence of such a point $\theta_\star$ is not straightforward. The main difficulty lies in the fact that the matrices $\bar{\mathbf{A}}^{(c)}(\cdot)$ and $\bar{\mathbf{b}}^{(c)}(\cdot)$ depend on the current policy. Indeed, for a fixed policy parameter ω, finding $\omega_\star$ such that $\frac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{A}}^{(c)}(\omega)\omega_\star + \frac{1}{N}\sum_{c=1}^{N} \bar{\mathbf{b}}^{(c)}(\omega) = 0$ boils down to

the federated TD learning algorithm with linear approximation, which is known to converge (Mangold et al., 2024). In the next proposition, we extend this result to the fixed points of the FedSARSA update, establishing the existence and unicity of a solution of (4).

Proposition 5.1. *Assume A 1–5. There exists a unique parameter $\theta_\star \in \mathcal{W}$ such that*

$$\tfrac{1}{N}\sum_{c=1}^N \bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star + \tfrac{1}{N}\sum_{c=1}^N \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \ .$$

We postpone the proof to Appendix E.1. The definition of $\theta_\star$ as a solution of this equation is crucial. In other FRL works (Wang et al., 2024; Zhang et al., 2024), heterogeneity is handled by introducing a *virtual environment* using averaged transitions and rewards from all environments. Unfortunately, studying convergence to the optimal parameters that correspond to this average environment leads to non-vanishing bias. In contrast, our approach will allow to show convergence to arbitrary precision towards to the unique fixed point $\theta_\star$ defined in Proposition 5.1. The global fixed point $\theta_\star$ can be related to the local optimum.

Proposition 5.2. *Assume A 1–5. For any $c \in \{1,\dots,N\}$, assume that $\theta_\star^{(c)} \in \mathcal{W}$, then the local optimum $\theta_\star^{(c)}$ (defined analogously to (6)) satisfies*

$$\|\theta_\star^{(c)} - \theta_\star\| \lesssim (1+\tau_{\mathrm{mix}})(\epsilon_p\|\theta_\star\| + \epsilon_r) \ ,$$

where ϵ_p and ϵ_r are defined in (1).

This shows that when agents are increasingly homogeneous (i.e., ϵ_p and ϵ_r get closer to zero), the shared parameter $\theta_\star$ and the local optimums $\theta_\star^{(c)}$ become closer. The proof is postponed to Appendix E.1.

Convergence Rate of FedSARSA. Now that we identified the limit point of FedSARSA, we can decompose its error similarly to SARSA in Claim 1. To this end, we follow (9) and define the matrices $\Gamma_{t,k:h}^{(c)} \triangleq (\mathrm{I} + \eta_t \bar{\mathbf{A}}^{(c)}(\theta_\star))^{k-h+1}$ for $k \geq h \geq 0$ and $c \in \{1,\dots,N\}$, with the convention $\Gamma_{t,k:h}^{(c)} = \mathrm{I}$ for $k < h$, and $\theta_\star$ as defined in Proposition 5.1. We define, for $c \in \{1,\dots,N\}$,

$$\begin{aligned}\varphi_{t,h}^{(c)} &= (\bar{\mathbf{A}}^{(c)}(\theta_t) - \bar{\mathbf{A}}^{(c)}(\theta_\star))\theta_{t,h}^{(c)} + \bar{\mathbf{b}}^{(c)}(\theta_t) - \bar{\mathbf{b}}^{(c)}(\theta_\star) \\ \varepsilon_{t,h}^{(c)}(Z_{t,h+1}^{(c)}) &= (\mathbf{A}^{(c)}(Z_{t,h+1}^{(c)}) - \bar{\mathbf{A}}^{(c)}(\theta_t))\theta_{t,h}^{(c)} \\ &\qquad + \mathbf{b}^{(c)}(Z_{t,h+1}^{(c)}) - \bar{\mathbf{b}}^{(c)}(\theta_t) \ ,\end{aligned}$$

which are the federated counterparts of (8). We also define the local limit parameter $\vartheta_\star^{(c)}$ as the solution of the local equation, when samples are collected from the global optimal policy $\theta_\star$,

$$\bar{\mathbf{A}}^{(c)}(\theta_\star)\vartheta_\star^{(c)} + \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \ .$$

The point $\vartheta_\star^{(c)}$ is used solely for analysis purposes, and allows to measure heterogeneity through the two following quantities, which we relate to ϵ_p and ϵ_r from (1).

Proposition 5.3. *Assume A 1, A 2, A 4, and A 5. For $c \in \{1,\dots,N\}$, there exists $\zeta_{\bar{\mathbf{A}}}, \zeta_{\theta_\star} \geq 0$ such that*

$$\|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|^2 \leq \zeta_{\bar{\mathbf{A}}}^2 \ , \quad \|\bar{\mathbf{A}}^{(c)}(\theta_\star)(\vartheta_\star^{(c)} - \theta_\star)\|^2 \leq \zeta_{\theta_\star}^2 \ ,$$

where we introduced the constants $\zeta_{\bar{\mathbf{A}}} \triangleq 4\mathrm{C}_A(1+\tau_{\mathrm{mix}})\epsilon_p$ and $\zeta_{\theta_\star} \triangleq 6(1+\tau_{\mathrm{mix}})(\epsilon_p\|\theta_\star\| + \epsilon_r)$.

The proof is postponed to Appendix E.2. We measure two types of heterogeneity: $\zeta_{\bar{\mathbf{A}}}$ measures heterogeneity of the matrices $\bar{\mathbf{A}}^{(c)}(\theta_\star)$ themselves, while $\zeta_{\theta_\star}$ relates the discrepancy between the global and local solutions of TD learning when following the global optimal policy. Similarly to Claim 1, we obtain a decomposition of the federated update.

Claim 2. *For $t \geq 0$, the global updates of Algorithm 1 satisfy, before projection,*

$$\begin{aligned}\bar{\theta}_{t+1} - \theta_\star &= \tfrac{1}{N}\sum_{c=1}^N \Gamma_{1:H}^{(c)}(\theta_t - \theta_\star) + \Delta_{1:H} \\ &\quad + \tfrac{\eta_t}{N}\sum_{c=1}^N \sum_{h=1}^H \Gamma_{h+1:H}^{(c)}\Big(\varepsilon_{t,h-1}^{(c)}(Z_{t,h}^{(c)}) + \varphi_{t,h-1}^{(c)}\Big),\end{aligned}$$

where $\Delta_{1:H} \triangleq \frac{1}{N}\sum_{c=1}^N (\mathrm{I} - \Gamma_{1:H}^{(c)})(\vartheta_\star^{(c)} - \theta_\star)$ accounts for bias due to heterogeneity.

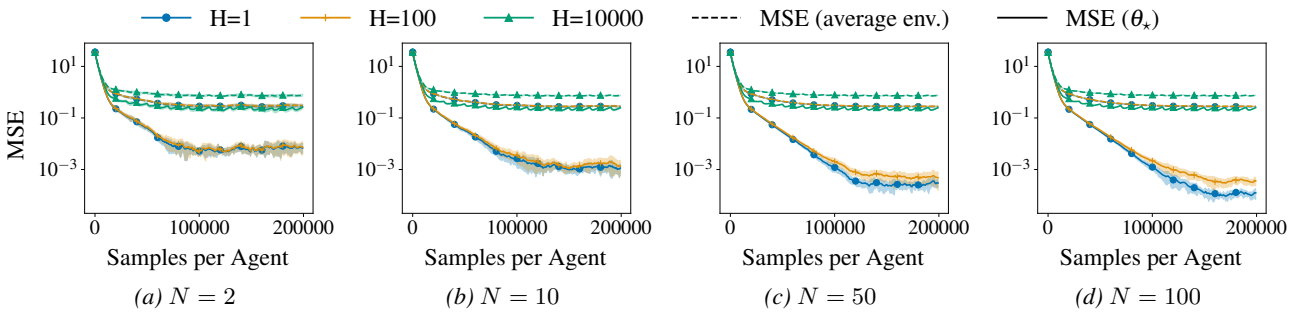
Analogously to Lemma 4.1 in the single-agent case, we bound the progress in-between policy updates.

Lemma 5.4. *Assume A 1, A 2, A 4, and A 5. Let $t \geq 0$, assume that the step size satisfies $\eta_t H \mathrm{C}_A \leq 1/6$. Then, it holds that, for some universal constant $c_2 > 0$,*

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1} - \theta_\star\|^2] &\leq (1 - \tfrac{\eta_t a H}{8})\|\theta_t - \theta_\star\|^2 + \tfrac{c_2\eta_t^3 H(H-1)^2}{a}\zeta_{\bar{\mathbf{A}}}^2\zeta_{\theta_\star}^2 \\ &\quad + \tfrac{c_2\eta_t^2 H\tau_{\mathrm{mix}}\mathrm{G}^2}{N} + \tfrac{c_2\delta\eta_t\mathrm{G}^2\tau_{\mathrm{mix}}^2}{Ha} + \tfrac{c_2\eta_t^3\mathrm{G}^2\mathrm{C}_A^2 H\tau_{\mathrm{mix}}^2}{a} \ ,\end{aligned}$$

where $\delta = 0$ if the $Z_{t,0}^{(c)}$ are sampled from the stationary distribution $\nu_{\theta_t}^{(c)}$ and $\delta = 1$ otherwise.

We prove this lemma in Appendix E.3. The proof essentially follows the same structure as the proof of Lemma 4.1, using the error decomposition from Claim 2. First, we note that the transient terms and error due to sampling from the "wrong" policy are handled in the same way. The analysis differs with the single-agent case in two crucial ways. First, environment heterogeneity induces an additional error term, which increases with ϵ_p and ϵ_r, scaling with the constants defined in Proposition 5.3. Second, the leading variance terms *decrease with the number of agents*: this allows FedSARSA to achieve reduced sample complexity per agent, which is essential in federated RL. Moreover, our sharp analysis technique allows us to show that higher-order terms (due to the Markovian nature of the noise), only increase with τ_{mix}, and not with H. This is in stark contrast with existing analyses (e.g., Zhang et al. (2024)), and allows to derive improved sample complexity. We now state our main theorem, assessing the convergence of FedSARSA.



(a) $N = 2$ *(b)* $N = 10$ *(c)* $N = 50$ *(d)* $N = 100$

Figure 1. MSE as a function of the number of communications. For each run, we report two errors: (i) in solid lines, the error $\mathbb{E}[\|\theta_t - \theta_\star\|^2]$ in estimating $\theta_\star$ as defined in Proposition 5.1, and (ii) in dashed lines, the error $\mathbb{E}[\|\theta_t - \chi_\star\|^2]$ in estimating $\chi_\star$, the limit of SARSA on the averaged environment. For each plot, we report the average over 10 runs and the corresponding standard deviation.

Theorem 5.5. *Assume A1, A2, A4, and A5, that the step size $\eta_t = \eta$ is constant and satisfies $\eta H \mathrm{C}_A \leq 1/5$ and that $H \geq \tau_{\mathrm{mix}}$. Then it holds that*

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim (1 - \tfrac{\eta a H}{8})^T \|\theta_0 - \theta_\star\|^2 + \tfrac{c_2 \eta^2 (H-1)^2}{a^2} \zeta_{\bar{\mathbf{A}}}^2 \zeta_{\theta_\star}^2 + \tfrac{c_2 \eta \tau_{\mathrm{mix}} \mathrm{G}^2}{Na} + \tfrac{c_2 \delta \mathrm{G}^2 \tau_{\mathrm{mix}}^2}{H^2 a^2} + \tfrac{c_2 \eta^2 \mathrm{G}^2 \mathrm{C}_A^2 \tau_{\mathrm{mix}}^2}{a^2} \quad ,$$

where $\delta = 0$ if the $Z_{t,0}^{(c)}$ are sampled from the stationary distribution $\nu_{\theta_t}^{(c)}$ and $\delta = 1$ otherwise.

The two main differences with the single-agent cases are that (i) leading variance terms decrease in $1/N$, and (ii) heterogeneity induces additional error, smaller than

$$\tfrac{\eta^2 (H-1)^2}{a^2} \zeta_{\bar{\mathbf{A}}}^2 \zeta_{\theta_\star}^2 \lesssim \tfrac{\eta^2 (H-1)^2}{a^2} \mathrm{C}_A^2 (1 + \tau_{\mathrm{mix}})^4 \epsilon_p^2 (\epsilon_p \|\theta_\star\| + \epsilon_r)^2.$$

This term decreases to zero with the step size η and disappears when $H = 1$, which is in stark contrast with existing work (Zhang et al., 2024), which exhibit a non-vanishing bias when $\epsilon_p \neq 0$ or $\epsilon_r \neq 0$. Interestingly, this term decreases to zero as kernel heterogeneity $\epsilon_p \to 0$, *even if the rewards are heterogeneous*, which is in line with previous observations in the literature (Zhu et al., 2024; Yang et al., 2024; Labbi et al., 2025b). We now state the sample and communication complexity of FedSARSA.

Corollary 5.6. *Assume A 1–5. Let $\epsilon > 0$. Set $\eta \approx \min\left(\frac{1}{\mathrm{C}_A}, \frac{Na\epsilon^2}{\mathrm{G}^2 \tau_{\mathrm{mix}}}, \frac{a\epsilon}{\mathrm{G}\mathrm{C}_A \tau_{\mathrm{mix}}}\right)$, and H such that $H \lesssim \frac{\tau_{\mathrm{mix}} \mathrm{G}}{\zeta_{\bar{\mathbf{A}}} \zeta_{\theta_\star}} \max\left(\frac{\mathrm{G}}{N\epsilon}, \mathrm{C}_A\right)$ and $H \gtrsim \frac{\delta \tau_{\mathrm{mix}} \mathrm{G}}{a\epsilon}$, then FedSARSA reaches $\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \epsilon^2$ with $T \gtrsim \max\left(\frac{\mathrm{C}_A}{a}, \frac{\zeta_{\bar{\mathbf{A}}} \zeta_{\theta_\star}}{a^2 \epsilon}\right) \log\left(\frac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$ communications, and*

$$TH \approx \max\left(\tfrac{\mathrm{C}_A}{a}, \tfrac{\mathrm{G}^2 \tau_{\mathrm{mix}}}{Na^2\epsilon^2}, \tfrac{\mathrm{G}\mathrm{C}_A \tau_{\mathrm{mix}}}{a^2 \epsilon}\right) \log\left(\tfrac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right) \quad (12)$$

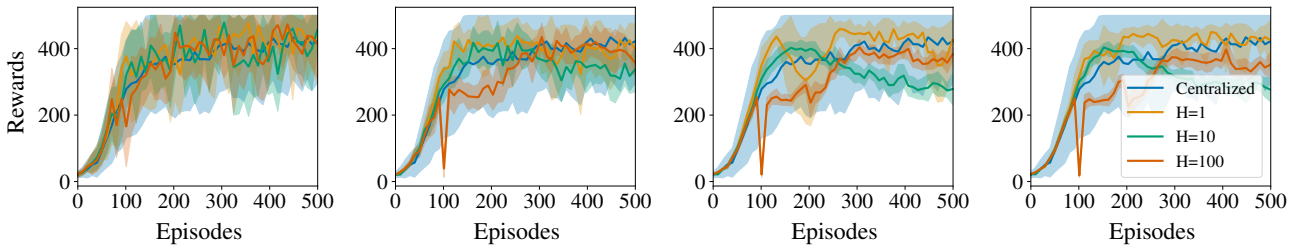
samples per agent.

We prove this corollary in Appendix E.3. This corollary shows that FedSARSA can exploit the experience of multiple agents to accelerate training, a property in FRL, typically known as *linear speed-up*. This effect is most pronounced when high precision is required and heterogeneity is moderate, in which case each agent takes $TH \approx \frac{\mathrm{G}^2 \tau_{\mathrm{mix}}}{Na^2\epsilon^2} \log\left(\frac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$ samples. In other regimes, other terms in the maximum may dominate. This occurs when (i) low-precision results suffice (i.e., large ϵ), or (ii) heterogeneity is high (i.e., large $\zeta_{\bar{\mathbf{A}}} \zeta_{\theta_\star}$).

Note that the linear speed-up is not unconditional and is limited by two phenomena that cannot be avoided. First, the algorithm cannot be faster than its deterministic counterpart, which gives the first term of the max in (12). Second, due to the Markovian nature of the noise, higher-order terms scaling in $\eta^2 \tau_{\mathrm{mix}}^2$ remain in the rate of Theorem 5.5), which gives the third term of the max in (12). Nonetheless, we stress that previous analyses exhibited a $\eta^2 \tau_{\mathrm{mix}} H$ terms, which we reduce to $\eta^2 \tau_{\mathrm{mix}}^2$: this is a direct consequence of our tighter analysis of Markovian noise. Finally, we note that in heterogeneous environments, the number of communications scales polynomially in $O(1/\epsilon)$. Reducing this to $O(\log(1/\epsilon))$ is a promising open question, which could be tackled by communicating in between policy updates, or using heterogeneity-correction methods such as Scaffold or Scafflsa (Karimireddy et al., 2020; Mangold et al., 2024).

6. Numerical Experiments

We now study tabular FedSARSA algorithm on synthetic problems, and propose an extension to deep RL. First, we generate 2 different instances of the Garnet environment (Archibald et al., 1995; Geist et al., 2014) with $|\mathcal{S}| = 10$ states and $|\mathcal{A}| = 3$ actions, where we connect each state with 2 other states with random transitions. For tabular experiments, we choose $N \in \{2, 10, 50, 100\}$ and equip half of the agents with the first environment, and the other half with the second one. For deep experiments, we use $N \in \{2, 10, 50, 100\}$ copies of CartPole, and use a deep network with two hidden layers to approximate the state-action value function. All experiments are run on a single laptop with an RTX 2000 Ada Generation Laptop GPU, and the code is available online at `https://github.com/pmangold/fed-sarsa`.



(a) $N = 2$ *(b)* $N = 10$ *(c)* $N = 50$ *(d)* $N = 100$

Figure 2. Rewards over the course of the learning for SARSA and FedSARSA on the CartPole environment for different numbers of agents N and numbers of local updates H. For each plot, we report the average over 10 runs and the corresponding standard deviation.

FedSARSA has linear speed-up. In Figure 1, we report the convergence of FedSARSA for different number of agents, with the same shared problem. In all experiments, we keep the same step size and the same number of local updates. As predicted by our theory, increasing the number of agents reduces the variance, allowing to reach solutions with higher precision. However, as the number of local steps gets larger, the bias of FedSARSA due to heterogeneity increases, and eventually prevents the algorithm from converging to satisfying precision. Remarkably, this phenomenon appears only when the number of local updates is very large, highlighting the relevance of FedSARSA.

FedSARSA converges even with multiple local steps. In Figure 1, we run the algorithm for $H \in \{1, 100, 10000\}$. For small values of H, FedSARSA has small heterogeneity bias, while for $H = 10000$, FedSARSA suffers from a large heterogeneity bias. This is in line with our theory, which shows that FedSARSA's bias increases with the number of local iterations, but remains small as long as H is small. We stress that this is the case until $H = 10000$ local updates, allowing for significant communication reduction.

FedSARSA converges to $\theta_\star$. We study the norm of two errors for the FedSARSA: (i) the error $\bar{\theta}_{t,h} - \theta_\star$, where $\bar{\theta}_{t,h}$ is the average of the local parameters $\theta_{t,h}^{(c)}$ for $c \in \{1, \dots, N\}$, in estimating $\theta_\star$, as defined in Proposition 5.1, and (ii) the error $\theta_t - \chi_\star$, where $\chi_\star$ is the point to which local SARSA converges when run on the *averaged environment*. In Figure 1, we plot the error as a function of the number of updates, for $H \in \{1, 100, 10000\}$. The error relative to $\theta_\star$ is reported in solid lines, and quickly becomes small, while the error relative to $\chi_\star$ remains large, demonstrating that FedSARSA does not converge to this point. This underlines the soundness of our analysis in showing convergence to the solution of the fixed-point equation defined in Proposition 5.1 rather than to the solution of a virtual environment.

Deep FedSARSA. To evaluate the soundness of the learned policy in real environments, we introduce a deep variance of FedSARSA for episodic environments. To this end, we propose a full-featured JAX (Google Research, 2018) implementation of SARSA and FedSARSA, supporting parallel environment executions using gymnax (Lange, 2022). Following common deep RL practice, we stabilize the learning with a small replay buffer, as well as a target Q network, which remains fixed during episodes, and is updated to the current value of the Q network at the end of episodes. We run experiments on the CartPole environment, which we report in Figure 2. Our results show that FedSARSA reaches large rewards faster and in a more stable way than SARSA, provided the number of local steps is not too large. When the number of local steps increases, the algorithm becomes biased due to noisy updates and heterogeneity, making rewards drop just after aggregation.

7. Conclusion and Discussion

We provide the first sample and communication complexity result for the FedSARSA algorithm in heterogeneous environments, assuming that the policy improvement operator is Lipschitz. Our results highlight that FedSARSA converges with arbitrary precision even with multiple local steps, and that it has linear speed-up. To conduct this analysis, we develop a novel analytical framework for single-agent SARSA, based on a novel, exact expansion of the algorithm's error over multiple updates, paired with a careful analysis of the impact of Markovian noise. We then characterize the point to which FedSARSA converges, highlighting that, contrary to common belief, the algorithm does not converge to the optimal parameter of a virtual averaged environment, but rather to the solution of an equation that depends on all environments. Together with numerical validation and extensions to deep RL, our results demonstrate that federated SARSA is both theoretically sound and practically applicable. Expanding our theoretical analysis to more general policy improvement operators constitutes a promising next step towards bridging the gap between the theoretical foundations of FedSARSA and its practical performance. Finally, when the number of local iterations increases, FedSARSA's iterations become increasingly biased due to noise and heterogeneity. This opens novel perspectives for debiasing FedSARSA in communication-constrained settings.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Archibald, T. W., McKinnon, K., and Thomas, L. C. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

Barakat, A., Bianchi, P., and Lehmann, J. Analysis of a target-based actor-critic algorithm with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 991–1040, 2022.

Beikmohammadi, A., Khirirat, S., Richtárik, P., and Magnússon, S. Collaborative value function estimation under model mismatch: A federated temporal difference analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 41–58. Springer, 2025.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pp. 1691–1692, 2018.

Dal Fabbro, N., Mitra, A., and Pappas, G. J. Federated td learning over finite-rate erasure channels: Linear speedup under markovian sampling. *IEEE Control Systems Letters*, 7:2461–2466, 2023.

De Farias, D. P. and Van Roy, B. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 105:589–608, 2000.

Douc, R., Moulines, E., Priouret, P., Soulier, P., Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov Chains: Basic Definitions*. Springer, 2018.

Geist, M., Scherrer, B., et al. Off-policy learning with eligibility traces: a survey. *Journal of Machine Learning Research*, 15(1):289–333, 2014.

Google Research, J. T. Jax: Composable transformations of python+numpy programs. https://github.com/google/jax, 2018.

Gordon, G. J. Chattering in sarsa (λ). *CMU Learning Lab Technical Report*, 1996.

Gordon, G. J. Reinforcement learning with function approximation converges to a region. *Advances in Neural Information Processing Systems*, 13, 2000.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37, 2022.

Jordan, P., Grötschla, F., Fan, F. X., and Wattenhofer, R. Decentralized federated policy gradient with byzantine fault-tolerance and provably fast convergence. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 964–972, 2024.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143, 2020.

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pp. 10997–11057, 2022.

Labbi, S., Mangold, P., Tiapkin, D., and Moulines, E. On global convergence rates for federated policy gradient under heterogeneous environment. *arXiv preprint arXiv:2505.23459*, 2025a.

Labbi, S., Tiapkin, D., Mancini, L., Mangold, P., and Moulines, E. Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025b.

Lan, G., Wang, H., Anderson, J., Brinton, C., and Aggarwal, V. Improved communication efficiency in federated natural policy gradient via ADMM-based gradient updates. *Advances in Neural Information Processing Systems*, 36: 59873–59885, 2023.

Lan, G., Han, D.-J., Hashemi, A., Aggarwal, V., and Brinton, C. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lange, R. T. gymnax: A JAX-based reinforcement learning environment library, 2022. URL http://github.com/RobertTLange/gymnax.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Mangold, P., Samsonov, S., Labbi, S., Levin, I., Alami, R., Naumov, A., and Moulines, E. Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning. *Advances in Neural Information Processing Systems*, 37:13927–13981, 2024.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Melo, F., Meyn, S., and Ribeiro, M. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 664–671, 2008.

Meyn, S. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

Mitra, A., Pappas, G. J., and Hassani, H. Temporal difference learning with compressed updates: Error-feedback meets reinforcement learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Ogier du Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., 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.

Perkins, T. and Precup, D. A convergent form of approximate policy iteration. *Advances in Neural Information Processing Systems*, 15, 2002.

Qi, J., Zhou, Q., Lei, L., and Zheng, K. Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence & Robotics*, 1(1):18–57, 2021.

Rio, E. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80. Springer, 2017.

Rummery, G. A. and Niranjan, M. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Salgia, S. and Chi, Y. The sample-communication complexity trade-off in federated Q-learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Samsonov, S., Tiapkin, D., Naumov, A., and Moulines, E. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4511–4547, 2024.

Singh, S., Jaakkola, T., Littman, M. L., and Szepesvári, C. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287–308, 2000.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Tian, H., Paschalidis, I. C., and Olshevsky, A. One-shot averaging for distributed TD(λ) under markov sampling. *IEEE Control Systems Letters*, 2024.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transations on Automatic Control*, 42(5), 1997.

Vamvoudakis, K. G., Wan, Y., Lewis, F. L., and Cansever, D. *Handbook of Reinforcement Learning and Control*. Springer, 2021.

Wang, H., Mitra, A., Hassani, H., Pappas, G. J., and Anderson, J. Federated TD learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Watkins, C. J. and Dayan, P. Q-learning. *Machine Learning*, 8:279–292, 1992.

Woo, J., Joshi, G., and Chi, Y. The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26(26): 1–85, 2025.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Yang, T., Cen, S., Wei, Y., Chen, Y., and Chi, Y. Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. *Advances in Neural Information Processing Systems*, 37:121304–121375, 2024.

Zhang, C., Wang, H., Mitra, A., and Anderson, J. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhang, S., Des Combes, R. T., and Laroche, R. On the convergence of sarsa with linear function approximation. In *International Conference on Machine Learning*, pp. 41613–41646, 2023.

Zheng, Z., Gao, F., Xue, L., and Yang, J. Federated Q-learning: Linear regret speedup with low communication cost. In *The Twelfth International Conference on Learning Representations*, pp. 57399–57443, 2024.

Zhu, F., Heath, R., and Mitra, A. Towards fast rates for federated and multi-task reinforcement learning. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 2658–2663. IEEE, 2024.

Zhuo, H. H., Feng, W., Lin, Y., Xu, Q., and Yang, Q. Federated deep reinforcement learning. *arXiv preprint arXiv:1901.08277*, 2019.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for SARSA with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Technical Results on Markov Chains

A.1. Markov kernels and canonical Markov chains

Throughout this section, P denotes a Markov transition kernel on a Polish space $(\mathsf{Z}, \mathcal{Z})$. We denote by $(\mathsf{Z}^{\mathbb{N}}, \mathcal{Z}^{\otimes\mathbb{N}})$ the set of Z-valued sequence endowed with the product σ-field and by $(Z_h)_{h\in\mathbb{N}}$ the canonical (or coordinate) process (see Chapter 3 (Douc et al., 2018)). For any probability distribution ϱ on $(\mathsf{Z}, \mathcal{Z})$, there exists a unique distribution $\mathbb{P}_\varrho$ on the canonical space such that the coordinate process $(Z_h)_{h\in\mathbb{N}}$ is a Markov chain with initial distribution ϱ and Markov kernel P. We denote by $\mathbb{E}_\varrho$ the corresponding expectation. For any bounded measurable function f on $(\mathsf{Z}, \mathcal{Z})$,

$$\mathbb{E}_\varrho[f(Z_0)] = \varrho(f), \quad \text{and} \quad \mathbb{E}_\varrho[f(Z_{h+1}) \,|\mathcal{F}_h] = Pf(Z_h), \quad \mathbb{P}_\varrho\text{-a.s.},$$

where $(\mathcal{F}_h = \sigma(Z_0, \ldots, Z_h))_{h\in\mathbb{N}}$ is the *canonical filtration* and for any bounded measurable function, $Pf(z) = \int_{\mathsf{Z}} P(z, \mathrm{d}z')f(z')$. For any two markov kernels P and Q, we denote by PQ the Markov kernel defined by $PQf(z) = \int_{\mathsf{Z}} P(z, \mathrm{d}z')Qf(z')$. Finally, we denote $k \in \mathbb{N}$, P^k the k-th power of P.

A.2. Exact coupling

Let $(Z_h)_{h\in\mathbb{N}}$ and $(Z'_h)_{h\in\mathbb{N}}$ be (discrete time) processes. By a *coupling* of (Z_h) and (Z'_h) we mean a simultaneous realizations of these processes on the same probability space $(\Omega, \bar{\mathcal{F}}, \bar{\mathbb{P}})$. We say that a *coupling is successful* if the two processes agree $\bar{\mathbb{P}}(Z_h = Z'_h, \text{for all } h \text{ large enough}) = 1$. Let ϱ and ϱ' be two probability distributions on $(\mathsf{Z}, \mathcal{Z})$. We say that $(\Omega, \bar{\mathcal{F}}, \bar{\mathbb{P}}, (\bar{Z}_h)_{h\in\mathbb{N}}, (\bar{Z}'_h)_{h\in\mathbb{N}}, \tau_{\text{couple}})$ is an *exact coupling* of $(\mathbb{P}_\varrho, \mathbb{P}_{\varrho'})$ if:

1. For all $A \in \mathcal{Z}^{\otimes\mathbb{N}}$, $\quad \bar{\mathbb{P}}((\bar{Z}_h)_{h\in\mathbb{N}} \in A) = \mathbb{P}_\varrho(A) \quad$ and $\quad \bar{\mathbb{P}}((\bar{Z'}_h)_{h\in\mathbb{N}} \in A) = \mathbb{P}_{\varrho'}(A)$.
2. $\mathbb{P}(\tau_{\text{couple}} < \infty) = 1$.
3. For all $h \in \mathbb{N}$, $\quad \bar{Z}_{h+\tau_{\text{couple}}} = \bar{Z}'_{h+\tau_{\text{couple}}}$.

Note that we allow ourselves a slight abuse of notation here, as the distribution $\bar{\mathbb{P}}$ depends on both ϱ and ϱ'. To avoid cumbersome notation, this dependence is kept implicit.

In particular, at all times h subsequent to the coupling time τ_{couple}, the two processes coincide $Z_h = Z'_h$. Theorem 19.3.9 of Douc et al. (2018) shows that:

Lemma A.1. *There exists a maximal exact coupling, i.e. an exact coupling $(\Omega, \bar{\mathcal{F}}, \mathbb{P}, (\bar{Z}_h)_{h\in\mathbb{N}}, (\bar{Z}'_h)_{h\in\mathbb{N}}, \tau_{\text{couple}})$, such that*

$$\mathbb{P}\left(\tau_{\text{couple}} > h\right) = (1/2)\|\varrho P^h - \varrho' P^h\|_{\text{TV}} \ .$$

Corollary A.2. *Assume that P is uniformly geometrically ergodic with mixing time τ_{mix}, i.e.*

$$(1/2)\|\varrho P^h - \varrho' P^h\|_{\text{TV}} \le (1/4)^{\lfloor h/\tau_{\text{mix}} \rfloor}.$$

Let $(\Omega, \bar{\mathcal{F}}, \mathbb{P}, (\bar{Z}_h)_{h\in\mathbb{N}}, (\bar{Z}'_h)_{h\in\mathbb{N}}, \tau_{\text{couple}})$ be an exact coupling of $\mathbb{P}_\varrho$ and $\mathbb{P}_{\varrho'}$. Then,

$$\mathbb{E}[\tau_{\text{couple}}] \le \frac{4\tau_{\text{mix}}}{3} \ , \qquad \textit{and} \quad \mathbb{E}[\tau^2_{\text{couple}}] \le \frac{20\tau^2_{\text{mix}}}{9} \ .$$

Proof. Note first that

$$\mathbb{E}[\tau_{\text{couple}}] = \sum_{h=0}^{\infty} \mathbb{P}\left(\tau_{\text{couple}} > h\right) = (1/2)\sum_{h=0}^{\infty} \|\varrho P^h - \nu\|_{\text{TV}} \le \sum_{h=0}^{\infty} (1/4)^{\lfloor h/\tau_{\text{mix}} \rfloor} = (4/3)\tau_{\text{mix}}.$$

The second inequality follows from

$$\mathbb{E}[\tau^2_{\text{couple}}] = \sum_{h=0}^{\infty} h^2 \mathbb{P}\left(\tau_{\text{couple}} = h\right) = \sum_{h=0}^{\infty} h^2 \big(\mathbb{P}\left(\tau_{\text{couple}} > h-1\right) - \mathbb{P}\left(\tau_{\text{couple}} > h\right)\big) \ ,$$

which gives, after reorganizing the terms,

$$\mathbb{E}[\tau^2_{\text{couple}}] = \sum_{h=0}^{\infty} \big((h+1)^2 - h^2\big)\mathbb{P}\left(\tau_{\text{couple}} > h\right) = \sum_{h=0}^{\infty} (2h+1)\mathbb{P}\left(\tau_{\text{couple}} > h\right) \ .$$

Using $\mathbb{P}(\tau_{\text{couple}} > h) = \|\varrho P^h - \nu\|_{\text{TV}} \leq (1/4)^{\lfloor h/\tau_{\text{mix}} \rfloor}$ and grouping the terms by blocks gives

$$\mathbb{E}[\tau_{\text{couple}}^2] \leq \sum_{h=0}^{\infty} (2h+1)(1/4)^{\lfloor h/\tau_{\text{mix}} \rfloor} = \tau_{\text{mix}}^2 \sum_{k=0}^{\infty} (2k+1)(1/4)^k = \tau_{\text{mix}}^2 \Big(\frac{8}{9} + \frac{4}{3}\Big) \ ,$$

and the second inequality of the corollary follows. □

Lemma A.3. *Assume that P is uniformly geometrically ergodic with unique invariant distribution ν, i.e. $\nu P = \nu$ and mixing time τ_{mix}. Let $f_h : \mathsf{Z} \to \mathbb{R}^d$, for $h \geq 0$, be functions such that $\|f\|_{2,\infty} = \sup_{h \geq 0} \sup_{z \in \mathsf{Z}} |f_h(z)| < +\infty$. Then, it holds that*

$$\Big\| \mathbb{E}_\varrho \Big[\textstyle\sum_{h=0}^{H-1} f_h(Z_h) \Big] \Big\| \leq \Big\| \mathbb{E}_\nu \Big[\textstyle\sum_{h=0}^{H-1} f_h(Z_h) \Big] \Big\| + \frac{8\|f\|_\infty \tau_{\text{mix}}}{3} \ ,$$
$$\mathbb{E}_\varrho \Big[\Big\| \textstyle\sum_{h=0}^{H-1} f_h(Z_h) \Big\|^2 \Big] \leq 2\mathbb{E}_\nu \Big[\Big\| \textstyle\sum_{h=0}^{H-1} f_h(Z_h) \Big\|^2 \Big] + \frac{160\|f\|_\infty^2 \tau_{\text{mix}}^2}{9} \ .$$

Proof. Let $(\Omega, \bar{\mathcal{F}}, (\bar{Z}_h)_{h \in \mathbb{N}}, (\bar{Y}_h)_{h \in \mathbb{N}}, \tau_{\text{couple}})$ an exact maximal coupling of $\mathbb{P}_\varrho$ and $\mathbb{P}_\nu$. Denote $S_H^Z = \sum_{h=0}^{H-1} f_h(\bar{Z}_h)$ and $S_H^Y = \sum_{h=0}^{H-1} f_h(\bar{Y}_h)$, and using the definition of τ_{couple}, we have

$$\begin{aligned} \mathbb{E}_\varrho \Big[\textstyle\sum_{h=0}^{H-1} f(Z_h) \Big] &= \bar{\mathbb{E}}\big[S_H^Y\big] + \bar{\mathbb{E}}\big[S_H^Z\big] - \bar{\mathbb{E}}\big[S_H^Y\big] \\ &= \mathbb{E}_\nu \Big[\sum_{h=0}^{H-1} f(Z_h) \Big] + \bar{\mathbb{E}} \Big[\sum_{h=0}^{\tau_{\text{couple}}} \big(f(\bar{Z}_h) - f(\bar{Y}_h) \big) \Big], \end{aligned} \tag{13}$$

since for $h \geq \tau_{\text{couple}}$, $f(\bar{Z}_h) - f(\bar{Y}_h) = 0$. The first inequality follows by taking the norm of (13), using the triangle inequality, the definition of $\|f\|_{2,\infty}$, and Corollary A.2. The second inequality follows from Young's inequality, which gives

$$\mathbb{E}_\varrho \Big[\Big\| \textstyle\sum_{h=0}^{H-1} f(Z_h) \Big\|^2 \Big] \leq 2\mathbb{E}_\nu \Big[\Big\| \textstyle\sum_{h=0}^{H-1} f(Z_h) \Big\|^2 \Big] + 2\bar{\mathbb{E}} \Big[\Big\| \textstyle\sum_{h=0}^{\tau_{\text{couple}}} f(\bar{Z}_h) - f(\bar{Y}_h) \Big\|^2 \Big] \ .$$

We then use Jensen's inequality, and the definition of $\|f\|_{2,\infty}$, to bound the second term as

$$\begin{aligned} 2\mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{\tau_{\text{couple}}} f(Z_h) - f(Y_h) \Big\|^2 \Big] &\leq 2\mathbb{E}_\varrho \Big[\tau_{\text{couple}} \sum_{h=0}^{\tau_{\text{couple}}} \|f(Z_h) - f(Y_h)\|^2 \Big] \\ &\leq 2\mathbb{E}_\varrho \Big[4\tau_{\text{couple}}^2 \|f\|_\infty^2 \Big] \ , \end{aligned}$$

and the result follows from Corollary A.2. □

A.3. Berbee's Lamma

We conclude this section by giving a simplified statement of the Berbee lemma. Consider the extended measurable space $\tilde{\mathsf{Z}}_{\mathbb{N}} = \mathsf{Z}^{\mathbb{N}} \times [0,1]$, equipped with the σ-field $\tilde{\mathsf{Z}}_{\mathbb{N}} = \mathsf{Z}^{\otimes \mathbb{N}} \otimes \mathcal{B}([0,1])$. For each probability measure ϱ on $(\mathsf{Z}, \mathcal{Z})$, we consider the probability measure $\tilde{\mathbb{P}}_\varrho = \mathbb{P}_\varrho \otimes \text{Unif}([0,1])$ and denote by $\tilde{\mathbb{E}}_\varrho$ the corresponding expected value. Finally, we denote by $(\tilde{Z}_k)_{k \in \mathbb{N}}$ the canonical process defined, for each $k \in \mathbb{N}$, by $\tilde{Z}_k : ((z_i)_{i \in \mathbb{N}}, u) \in \mathsf{Z}_{\mathbb{N}} \mapsto z_k$ and $U : ((z_i)_{i \in \mathbb{N}}, u) \in \mathsf{Z}_{\mathbb{N}} \mapsto u$. Under $\tilde{\mathbb{P}}_\varrho$, the process $\{\tilde{Z}_k\}_{k \in \mathbb{N}}$ is by construction a Markov chain with initial distribution ϱ and Markov kernel P, and is independent of U. The distribution of U under $\tilde{\mathbb{P}}_\varrho$ is uniform over $[0,1]$.

We first recall (Rio, 2017), Chapter 5. Let $\mathcal{A}$ and $\mathcal{B}$ two σ-fields of $(\Omega, \mathcal{T}, \mathbb{P})$. The β-mixing of $(\mathcal{A}, \mathcal{B})$ is defined by:

$$\beta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup \left\{ \sum_{i \in I} \sum_{j \in J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \right\}$$

the maximum being taken over all finite partitions $(A_i)_{i \in I}$ and $(B_j)_{j \in J}$ of Ω with the sets A_i in $\mathcal{A}$ and the sets B_j in $\mathcal{B}$.

Lemma A.4. *Let $\mathcal{A}$ be a σ-field in $(\Omega, \mathcal{T}, \mathbb{P})$ and X be a random variable with values in some Polish space. Let δ be a random variable with uniform distribution over $[0,1]$, independent of the σ-field generated by X and $\mathcal{A}$. Then there exists a random variable X^*, with the same law as X, independent of X, such that $\mathbb{P}(X \neq X^*) = \beta(\mathcal{A}, \sigma(X))$. Furthermore X^* is measurable with respect to the σ-field generated by $\mathcal{A}$ and (X, δ).*

Lemma A.5. *Assume that P is uniformly geometrically ergodic with mixing time τ_{mix}. Set $m \in \mathbb{N}$. Then, there exists a random process $(\tilde{Z}_k^\star)_{k\in\mathbb{N}}$ defined on $(\tilde{\mathsf{Z}}_{\mathbb{N}}, \tilde{\mathcal{Z}}_{\mathbb{N}}, \tilde{\mathbb{P}}_\varrho)$ such that for any $k \geq m$,*

1. *$\tilde{Z}_k^\star$ is independent of $\tilde{\mathcal{F}}_{k-m} = \sigma\{\tilde{Z}_\ell : \ell \leq k-m\}$;*
2. *$\tilde{\mathbb{P}}_\varrho(\tilde{Z}_k^\star \neq \tilde{Z}_k) \leq (1/4)^{\lfloor m/\tau_{\mathrm{mix}} \rfloor}$;*
3. *the random variables $\tilde{Z}_k^\star$ and $\tilde{Z}_k$ have the same distribution under $\tilde{\mathbb{P}}_\varrho$.*

Proof. We apply for each $k \in \mathbb{N}$ Lemma A.4 with $\Omega = \tilde{\mathsf{Z}}_{\mathbb{N}}$, $\mathbb{P} = \tilde{\mathbb{P}}_\varrho$, $\mathcal{A} = \sigma\{\tilde{Z}_\ell : \ell \geq k-m\}$ and $X = \tilde{Z}_k$. We conclude by using the bound for β-mixing coefficient given in (Douc et al., 2018), Theorem 3.3. □

Lemma A.6. *Assume that P is uniformly geometrically ergodic with mixing time τ_{mix}. Let $0 \leq m \leq k \in \mathbb{N}$ and $f : \mathsf{Z} \to \mathbb{R}$ and $g : \mathsf{Z}^{k-m} \to \mathbb{R}$ be two bounded measurable functions. Then, for any initial distribution ϱ,*

$$|\mathbb{E}_\varrho[f(Z_k)g(Z_{0:k-m})]| \leq \|f\|_\infty |\mathbb{E}_\varrho[g(Z_{0:k-m})]| + 2\|f\|_\infty \|g\|_\infty (1/4)^{\lfloor m/\tau_{\mathrm{mix}} \rfloor},$$

where, for any sequence $(u_\ell)_{\ell\in\mathbb{N}}$ and $0 \leq k \leq \ell$, we set $u_{k:\ell} = [u_k, \dots, u_\ell]$.

Proof. Using Lemma A.5, we get that

$$\begin{aligned} A := \mathbb{E}_\varrho[f(Z_k)g(Z_{0:k-m})] &= \tilde{\mathbb{E}}_\varrho[f(\tilde{Z}_k)g(\tilde{Z}_{0:k-m})] \\ &= \tilde{\mathbb{E}}_\varrho[f(\tilde{Z}_k^\star)g(\tilde{Z}_{0:k-m})] + \tilde{\mathbb{E}}_\varrho[\{f(\tilde{Z}_k) - f(\tilde{Z}_k^\star)\}g(\tilde{Z}_{0:k-m})] \end{aligned}$$

where by construction $\tilde{Z}_k^\star$ is independent of $\sigma\{\tilde{Z}_\ell : \ell \geq k-m\}$ under $\tilde{\mathbb{P}}_\varrho$. Hence:

$$\tilde{\mathbb{E}}_\varrho[f(\tilde{Z}_k^\star)g(\tilde{Z}_{0:k-m})] = \mathbb{E}_\varrho[f(Z_k)]\mathbb{E}_\varrho[g(Z_{0:k-m})],$$

where we have used that, under $\tilde{\mathbb{P}}_\varrho$, the law of $\tilde{Z}_k^\star$ and $\tilde{Z}_k$ coincide and $\tilde{\mathbb{E}}_\varrho[g(\tilde{Z}_{0:k-m})] = \mathbb{E}_\varrho[g(Z_{0:k-m})]$. Finally,

$$\begin{aligned} |\tilde{\mathbb{E}}_\varrho[\{f(\tilde{Z}_k) - f(\tilde{Z}_k^\star)\}g(\tilde{Z}_{0:k-m})]| &= |\tilde{\mathbb{E}}_\varrho[\{f(\tilde{Z}_k) - f(\tilde{Z}_k^\star)\}g(\tilde{Z}_{0:k-m})\mathbb{1}_{\{\tilde{Z}_k \neq \tilde{Z}_k^*\}}]| \\ &\leq 2\|f\|_\infty \|g\|_\infty \tilde{\mathbb{P}}_\varrho(\tilde{Z}_k \neq \tilde{Z}_k^\star) \ . \end{aligned}$$

The result follows. □

Lemma A.7. *Assume that P is uniformly geometrically ergodic with mixing time τ_{mix}. Let $0 \leq h \leq k \leq \ell \leq m$ and let $f_h, f_k, f_\ell, f_m : \mathsf{Z} \to \mathbb{R}$ be bounded measurable functions. Then, for any initial distribution ϱ, it holds that*

$$\begin{aligned} &\left|\mathbb{E}_\varrho\left[(f_h(Z_h) - \varrho P^h f_h)(f_k(Z_k) - \varrho P^k f_k)(f_\ell(Z_\ell) - \varrho P^\ell f_\ell)\right]\right| \\ &\qquad \leq 8\|f_h\|_\infty \|f_k\|_\infty \|f_\ell\|_\infty (1/4)^{\lfloor(\ell-k)/\tau_{\mathrm{mix}}\rfloor} (1/4)^{\lfloor(k-h)/\tau_{\mathrm{mix}}\rfloor} \ . \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\left|\mathbb{E}_\varrho\left[(f_h(Z_h) - \varrho P^h f_h)(f_k(Z_k) - \varrho P^k f_k)(f_\ell(Z_\ell) - \varrho P^\ell f_\ell)(f_m(Z_m) - \varrho P^m f_m)\right]\right| \\ &\leq 16\|f_\ell\|_\infty \|f_m\|_\infty \|f_h\|_\infty \|f_k\|_\infty (1/4)^{\lfloor(m-\ell)/\tau_{\mathrm{mix}}\rfloor} ((1/4)^{\lfloor(\ell-k)/\tau_{\mathrm{mix}}\rfloor} + (1/4)^{\lfloor(k-h)/\tau_{\mathrm{mix}}\rfloor}) \ . \end{aligned} \tag{15}$$

Proof. In this proof, we denote $\bar{f}_i = f_i - \varrho P^i f_i$, $i \in \{h, k, \ell, m\}$. We first prove the first part (14) of the lemma. Note that

$$\mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)\bar{f}_\ell(Z_\ell)] = \mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)P^{\ell-k}\bar{f}_\ell(Z_k)] \ .$$

Using the Chapman-Kolmogorov equation for Markov chain, we have that, for all $z \in \mathsf{Z}$,

$$P^{\ell-k}\bar{f}_\ell(z) = \delta_z P^{\ell-k} f_\ell - \varrho P^k P^{\ell-k} f_\ell \ . \tag{16}$$

This implies that

$$\begin{aligned}\|P^{\ell-k}\bar{f}_\ell\|_\infty &\le 2(1/4)^{\lfloor(\ell-k)/\tau_{\rm mix}\rfloor}\|f_\ell\|_\infty \ , \\ \|\bar{f}_k P^{\ell-k}\bar{f}_\ell\|_\infty &\le 4(1/4)^{\lfloor(\ell-k)/\tau_{\rm mix}\rfloor}\|f_k\|_\infty\|f_\ell\|_\infty \ .\end{aligned} \tag{17}$$

Noticing that $\mathbb{E}_\rho[\bar{f}_h(Z_h)] = 0$, we use Lemma A.6 to bound

$$\begin{aligned}|\mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)\bar{f}_\ell(Z_\ell)]| &\le \|\bar{f}_h\|_\infty\|\bar{f}_k P^{\ell-k}\bar{f}_\ell\|_\infty(1/4)^{\lfloor k-h/\tau_{\rm mix}\rfloor} \\ &\le 4\|\bar{f}_h\|_\infty\|f_k\|_\infty\|f_\ell\|_\infty(1/4)^{\lfloor(\ell-k)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(k-h)/\tau_{\rm mix}\rfloor} \ ,\end{aligned}$$

and (14) follows. To prove the second inequality (15), we first use the Markov property

$$\mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)\bar{f}_\ell(Z_\ell)\bar{f}_m(Z_m)] = \mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)\bar{f}_\ell(Z_\ell)P^{m-\ell}\bar{f}_m(Z_\ell)] \ .$$

By Lemma A.6, we obtain

$$\begin{aligned}&|\mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)\bar{f}_\ell(Z_\ell)\bar{f}_m(Z_m)]| \\ &\quad\le \|\bar{f}_\ell(Z_\ell)P^{m-\ell}\bar{f}_m(Z_\ell)\|_\infty\Big\{|\mathbb{E}_\varrho[\bar{f}_h(Z_h)\bar{f}_k(Z_k)]| + \|\bar{f}_h\|_\infty\|\bar{f}_k\|_\infty(1/4)^{\lfloor(\ell-k)/\tau_{\rm mix}\rfloor}\Big\} \\ &\quad\le 4\|f_\ell\|_\infty\|f_m\|_\infty\|\bar{f}_h\|_\infty\|\bar{f}_k\|_\infty(1/4)^{\lfloor(m-\ell)/\tau_{\rm mix}\rfloor}\Big\{(1/4)^{\lfloor(k-h)/\tau_{\rm mix}\rfloor} + (1/4)^{\lfloor(\ell-k)/\tau_{\rm mix}\rfloor}\Big\} \ ,\end{aligned}$$

where we also used the Markov property and proceeded as in the derivation of (17). □

A.4. Bounds on covariances

In this part, we will make extensive use of the following norms, for a sequence of matrices $F_h : \mathsf{Z} \to \mathbb{R}^{d\times d}$ and vectors $g_h : \mathsf{Z}^h \to \mathbb{R}^d$

$$\begin{aligned}|||g|||_{2,\infty} &\triangleq \sup_{h\ge 0}\|g_h\|_{2,\infty} \ , \quad \text{with} \quad \|g_h\|_{2,\infty} \triangleq \Big(\sum_{i=1}^d \sup_{z_{1:h}\in\mathsf{Z}^h}|g_{h,i}(z_{1:h})|^2\Big)^{1/2} , \\ |||F|||_{2,\infty} &\triangleq \sup_{h\ge 0}|||F_h|||_{2,\infty} \ , \quad \text{with} \quad |||F_h|||_{2,\infty} \triangleq \Big(\sum_{i=1}^d \|F_{h,i,:}\|_{2,\infty}^2\Big)^{1/2} \ .\end{aligned}$$

Lemma A.8. *Assume that P is uniformly geometrically ergodic with mixing time $\tau_{\rm mix}$. Let $f_h : \mathsf{Z} \to \mathbb{R}^d$ be uniformly bounded for $h \ge 0$. Then, for any initial distribution ϱ, we get*

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{h=0}^{H-1}\{f_h(Z_h) - \varrho P^h f_h\}\Big\|^2\Big] \le 15H\tau_{\rm mix}\|f\|_{2,\infty}^2 \ . \tag{18}$$

Proof. Expanding the square, we have

$$\begin{aligned}&\mathbb{E}_\varrho\Big[\Big\|\sum_{h=0}^{H-1}\{f_h(Z_h) - \varrho P^h f_h\}\Big\|^2\Big] \\ &= \underbrace{\sum_{h=0}^{H-1}\mathbb{E}_\varrho[\|f_h(Z_h) - \varrho P^h f_h\|^2]}_{\mathbf{A_1}} + \underbrace{2\sum_{0\le h<h'\le H}\mathbb{E}_\varrho[\langle f_h(Z_h) - \varrho P^h f_h \, , \, f_{h'}(Z_{h'}) - \varrho P^{h'} f_{h'}\rangle]}_{\mathbf{A_2}} \ .\end{aligned} \tag{19}$$

We bound the first term by $\mathbf{A_1} \le 4H\|f\|_\infty^2$. For the second term $\mathbf{A_2}$, we proceed coordinate by coordinate, using the triangle inequality and the Markov property, which yields

$$\Big|\mathbb{E}_\varrho\Big[\Big\langle f_h(Z_h) - \varrho P^h f_h \, , \, f_{h'}(Z_{h'}) - \varrho P^{h'} f_{h'}\Big\rangle\Big]\Big|$$

$$\leq \sum_{i=1}^{d} \left| \mathbb{E}_\varrho \Big[e_i^\top \big(f_h(Z_h) - \varrho P^h f_h \big) \cdot e_i^\top \big(f_{h'}(Z_{h'}) - \varrho P^{h'} f_{h'} \big) \Big] \right|$$
$$= \sum_{i=1}^{d} \left| \mathbb{E}_\varrho \Big[\big(f_{h,i}(Z_h) - \varrho P^h f_{h,i} \big) \cdot \big(\delta_{Z_h} P^{h'-h} e_i^\top f_{h'} - \varrho P^h P^{h'-h} f_{h',i} \big) \Big] \right| \ .$$

By ergodicity of the Markov chain, we obtain

$$\left| \mathbb{E}_\varrho \Big[\Big\langle f_h(Z_h) - \varrho P^h f_h \, , \, f_{h'}(Z_{h'}) - \varrho P^{h'} f_{h'} \Big\rangle \Big] \right| \leq \sum_{i=1}^{d} 2 \|f_{h,i}\|_\infty \cdot 2 \|f_{h',i}\|_\infty (1/4)^{\lfloor (h'-h)/\tau_{\mathrm{mix}} \rfloor} \ ,$$

where, $\|f_{k,i}\|_\infty = \sup_{z \in \mathsf{Z}} |f_{k,i}(z)|$ for $k \in \{h, h'\}$. Using the definition of $\|f\|_{2,\infty}$ then gives

$$|\mathbf{A_2}| \leq 8 \|f\|_{2,\infty}^2 \sum_{0 \leq h < h' \leq H} (1/4)^{\lfloor (h'-h)/\tau_{\mathrm{mix}} \rfloor} \leq \frac{32 \|f\|_{2,\infty}^2 \tau_{\mathrm{mix}} H}{3} \ ,$$

and the result follows by summing the bounds on $\mathbf{A_1}$ and $\mathbf{A_2}$. □

Corollary A.9. *Under the assumptions of Lemma A.8, it holds that,*

$$\mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{H-1} \{ f_h(Z_h) - \nu P^h f_h \} \Big\|^2 \Big] \leq 34 H \tau_{\mathrm{mix}} \|f\|_{2,\infty}^2 \ . \tag{20}$$

Proof. We can decompose the error as

$$\mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{H-1} \{ f_h(Z_h) - \nu P^h f_h \} \Big\|^2 \Big] \leq 2 \mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{H-1} \{ f_h(Z_h) - \nu P^h f_h \} \Big\|^2 \Big] + 2 \mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{H-1} \{ \nu P^h f_h - \varrho P^h f_h \} \Big\|^2 \Big] \ .$$

The first term can be bounded using Lemma A.8, while the second term can be bounded as

$$2 \mathbb{E}_\varrho \Big[\Big\| \sum_{h=0}^{H-1} \{ \nu P^h f_h - \varrho P^h f_h \} \Big\|^2 \Big] \leq 2H \sum_{h=0}^{H-1} \mathbb{E}_\varrho \Big[\Big\| \{ \nu P^h f_h - \varrho P^h f_h \} \Big\|^2 \Big] \leq \frac{8}{3} H \tau_{\mathrm{mix}} \|f\|_{2,\infty}^2 \ ,$$

where the second bound follows from the mixing property of the Markov chain. □

Lemma A.10. *Assume that P is uniformly geometrically ergodic with unique invariant distribution ν, i.e. $\nu P = \nu$ and mixing time τ_{mix}. Let $F_h : \mathsf{Z} \to \mathbb{R}^{d \times d}$ and $g_h : \mathsf{Z}^h \to \mathbb{R}^d$, for $h \geq 0$, be uniformly bounded. Then, it holds that*

$$\mathbb{E}_\varrho \Big[\Big\| \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \big\{ F_h(Z_h) - \varrho P^h F_h \big\} g_\ell(Z_{1:\ell}) \Big\|^2 \Big] \leq \frac{70}{3} H^3 \tau_{\mathrm{mix}} |\!|\!| F |\!|\!|_{2,\infty}^2 \sup_{h \in \{1, \ldots, H\}} \mathbb{E}[\|g_h(Z_{1:h})\|^2] \ ,$$
$$\mathbb{E}_\varrho \Big[\Big\| \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \big\{ F_h(Z_h) - \varrho P^h F_h \big\} g_\ell(Z_{1:\ell}) \Big\|^2 \Big] \leq \frac{70}{3} H^3 \tau_{\mathrm{mix}} |\!|\!| F |\!|\!|_{2,\infty}^2 |\!|\!| g |\!|\!|_{2,\infty}^2 \ .$$

Proof. We define $\bar{F}_h(z) = F_h(z) - \varrho P^h F_h$. Expanding the square, we have

$$\mathbb{E}_\varrho \Big[\Big\| \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \bar{F}_h(Z_h) g_\ell(Z_{1:\ell}) \Big\|^2 \Big]$$
$$= \sum_{h=1}^{H} \mathbb{E}_\varrho \Big[\Big\| \sum_{\ell=1}^{h-1} \bar{F}_h(Z_h) g_\ell(Z_{1:\ell}) \Big\|^2 \Big] + 2 \sum_{h'>h=1}^{H} \sum_{\ell=1}^{h-1} \sum_{\ell'=1}^{h'-1} \mathbb{E}_\varrho \Big[\Big\langle \bar{F}_h(Z_h) g_\ell(Z_{1:\ell}) \, , \, \bar{F}_{h'}(Z_{h'}) g_{\ell'}(Z_{1:\ell'}) \Big\rangle \Big]$$
$$= \sum_{h=1}^{H} \mathbb{E}_\varrho \Big[\Big\| \sum_{\ell=1}^{h-1} \bar{F}_h(Z_h) g_\ell(Z_{1:\ell}) \Big\|^2 \Big]$$

$$+ 2\sum_{h'>h=1}^{H}\sum_{\ell=1}^{h-1}\sum_{\ell'=1}^{h'-1}\underbrace{\mathbb{E}_\varrho\Big[\Big\langle \bar{F}_h(Z_h)g_\ell(Z_{1:\ell})\,, P^{h'-\max(h,\ell')}\bar{F}_{h'}(Z_{\max(h,\ell')})g_{\ell'}(Z_{1:\ell'})\Big\rangle\Big]}_{\mathbf{A}^{h,h'}_{\ell,\ell'}} \ .$$

To bound the first sum, we remark that, for $h \in \{1,\dots,H\}$,

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\bar{F}_h(Z_h)g_\ell(Z_{1:\ell})\Big\|^2\Big] \le h\sum_{\ell=1}^{h-1}\mathbb{E}_\varrho\Big[\Big\|\bar{F}_h(Z_h)g_\ell(Z_{1:\ell})\Big\|^2\Big] \le 4h\sum_{\ell=1}^{h-1}|||F|||^2_{2,\infty}\mathbb{E}_\varrho\big[\|g_\ell(Z_{1:\ell})\|^2\big] \ ,$$

where the second inequality follows by bounding F_h's operator norm by $|||F|||_{2,\infty}^2$. We thus obtain

$$\sum_{h=1}^{H}\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\bar{F}_h(Z_h)g_\ell(Z_{1:\ell})\Big\|^2\Big] \le 2H^3|||F|||^2_{2,\infty}\sup_{1\le\ell\le H}\mathbb{E}_\varrho\big[\|g_\ell(Z_{1:\ell})\|^2\big] \ . \tag{21}$$

To bound the second term, we remark that

$$\begin{aligned}
|\mathbf{A}^{h,h'}_{\ell,\ell'}| &\le \sum_{i=1}^{d}\Big|\mathbb{E}_\varrho\Big[e_i^\top\cdot(\bar{F}_h(Z_h)g_\ell(Z_{1:\ell}))\times e_i^\top\cdot(P^{h'-\max(h,\ell')}\bar{F}_{h'}(Z_{\max(h,\ell')})g_{\ell'}(Z_{1:\ell'}))\Big]\Big| \\
&= \sum_{i=1}^{d}\Big|\mathbb{E}_\varrho\Big[\sum_{j=1}^{d}\bar{F}_{h,i,j}(Z_h)g_{\ell,j}(Z_{1:\ell})\sum_{j=1}^{d}P^{h'-\max(h,\ell')}\bar{F}_{h',i,j}(Z_{\max(h,\ell')})g_{\ell',j}(Z_{1:\ell'}))\Big]\Big| \ .
\end{aligned}$$

By Jensen's then Hölder inequality, we have

$$\begin{aligned}
|\mathbf{A}^{h,h'}_{\ell,\ell'}| &\le \sum_{i=1}^{d}\mathbb{E}_\varrho^{1/2}\Big[\Big|\sum_{j=1}^{d}\bar{F}_{h,i,j}(Z_h)g_{\ell,j}(Z_{1:\ell})\Big|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\Big|\sum_{j=1}^{d}P^{h'-\max(h,\ell')}\bar{F}_{h',i,j}(Z_{\max(h,\ell')})g_{\ell',j}(Z_{1:\ell'}))\Big|^2\Big] \\
&\le \sum_{i=1}^{d}\mathbb{E}_\varrho^{1/2}\Big[\|\bar{F}_{h,i,:}(Z_h)\|^2\|g_{\ell,:}(Z_{1:\ell})\|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\|P^{h'-\max(h,\ell')}\bar{F}_{h',i,:}(Z_{\max(h,\ell')})\|^2\|g_{\ell',:}(Z_{1:\ell'})\|^2\Big] \\
&\le \sum_{i=1}^{d}\|\bar{F}_{h,i,:}\|_{2,\infty}\|P^{h'-\max(h,\ell')}\bar{F}_{h',i,:}\|_{2,\infty}\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell,:}(Z_{1:\ell})\|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell',:}(Z_{1:\ell'})\|^2\Big] \ ,
\end{aligned}$$

where we also used the Cauchy-Schwarz inequality in the second inequality. By the mixing property of the Markov chain, and using $\|\bar{F}_{h,i,:}\|_{2,\infty} \le 2\|F_{h,i,:}\|_{2,\infty}$ we obtain

$$\begin{aligned}
|\mathbf{A}^{h,h'}_{\ell,\ell'}| &\le 4\sum_{i=1}^{d}(1/4)^{\lfloor(h'-\max(h,\ell'))/\tau_{\mathrm{mix}}\rfloor}\|F_{h,i,:}\|_{2,\infty}\|F_{h',i,:}\|_{2,\infty}\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell,:}(Z_{1:\ell})\|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell',:}(Z_{1:\ell'})\|^2\Big] \\
&\le 4(1/4)^{\lfloor(h'-\max(h,\ell'))/\tau_{\mathrm{mix}}\rfloor}|||F_h|||_\infty|||F_{h'}|||_\infty\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell,:}(Z_{1:\ell})\|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell',:}(Z_{1:\ell'})\|^2\Big] \ .
\end{aligned}$$

Finally, summing over all ℓ, ℓ', we obtain

$$\sum_{\ell=1}^{h}\sum_{\ell'=1}^{h'}|\mathbf{A}^{h,h'}_{\ell,\ell'}| \le 4\sum_{\ell=0}^{h-1}\sum_{\ell'=0}^{h'-1}(1/4)^{\lfloor(h'-\max(h,\ell'))/\tau_{\mathrm{mix}}\rfloor}|||F|||^2_{2,\infty}\mathbb{E}_\varrho^{1/2}\Big[\|g_\ell(Z_{1:\ell})\|^2\Big]\mathbb{E}_\varrho^{1/2}\Big[\|g_{\ell'}(Z_{1:\ell'})\|^2\Big] \ .$$

We then bound each $\mathbb{E}^{1/2}[\|g_{\ell+1}(Z_{1:\ell})\|^2]$ by their supremum over $\ell \in \{1,\dots,H\}$, sum over all $h \ne h'$, and use the triangle inequality to obtain

$$\Big|2\sum_{h'>h}\sum_{\ell=1}^{h}\sum_{\ell'=1}^{h'}\mathbf{A}^{h,h'}_{\ell,\ell'}\Big| \le 8\sum_{h'>h}\sum_{\ell=0}^{h-1}\sum_{\ell'=0}^{h'-1}(1/4)^{\lfloor(h'-\max(h,\ell'))/\tau_{\mathrm{mix}}\rfloor}|||F|||^2_{2,\infty}\sup_{1\le m\le d}\mathbb{E}_\varrho\Big[\|g_m(Z_{1:m})\|^2\Big] \ .$$

We then remark that

$$\sum_{\ell=0}^{h-1}\sum_{\ell'=0}^{h'-1}(1/4)^{\lfloor(h'-\max(h,\ell'))/\tau_{\rm mix}\rfloor}\leq\sum_{\ell=0}^{h-1}\sum_{\ell'=0}^{h'-1}(1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}+(1/4)^{\lfloor(h'-\ell')/\tau_{\rm mix}\rfloor}\leq\frac{8h'\tau_{\rm mix}}{3}\ ,$$

and the bound on the first term follows by summing this inequality over $h'>h$, and combining the resulting inequality with (21). □

Corollary A.11. *Under the assumptions of Lemma A.10, it holds that*

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\big\{F_h(Z_h)-\nu F_h\big\}g_\ell(Z_{1:\ell})\Big\|^2\Big]\leq 48H^3\tau_{\rm mix}|||F|||_{2,\infty}^2\sup_{h\in\{1,\dots,H\}}\mathbb{E}[\|g_h(Z_{1:h})\|^2]\ ,$$
$$\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\big\{F_h(Z_h)-\nu F_h\big\}g_\ell(Z_{1:\ell})\Big\|^2\Big]\leq 48|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2H^3\tau_{\rm mix}\ .$$

Proof. Recentering the $F_h(Z_h)$'s on their expectation, and using the fact that $\nu=P\nu$, we get

$$\begin{aligned}&\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Big\{F_h(Z_h)-\nu F_h\Big\}g_\ell(Z_{1:\ell})\Big\|^2\Big]\\&\leq 2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\bar F_h(Z_h)g_\ell(Z_{1:\ell})\Big\|^2\Big]+2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Big\{\varrho P^hF_h-\nu P^hF_h\Big\}g_\ell(Z_{1:\ell})\Big\|^2\Big]\ .\end{aligned}\tag{22}$$

The first term is directly bounded by Lemma A.10. The second term can be bounded using Jensen's inequality together with the mixing property

$$\begin{aligned}2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Big\{\varrho P^hF_h-\nu P^hF_h\Big\}g_\ell(Z_{1:\ell})\Big\|^2\Big]&\leq 2H\sum_{h=1}^{H}h\sum_{\ell=1}^{h-1}|||F|||_{2,\infty}^2(1/4)^{\lfloor h/\tau_{\rm mix}\rfloor}\mathbb{E}[\|g_\ell(Z_{1:\ell})\|^2]\\&\leq\frac{4H^3\tau_{\rm mix}}{3}|||F|||_{2,\infty}^2\sup_{1\leq h\leq H}\mathbb{E}[\|g_h(Z_{1:h})\|^2]\ ,\end{aligned}$$

and the result follows. □

When g is centered and only depends on one of the Markov iterates, we have the refined bound.

Lemma A.12. *Assume that P is uniformly geometrically ergodic with mixing time $\tau_{\rm mix}$. Let $F_h:\mathsf{Z}\to\mathbb{R}^{d\times d}$ and $g_h:\mathsf{Z}\to\mathbb{R}^d$ be uniformly bounded for $h\geq 0$. Then, for any initial distribution ϱ, it holds that*

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{F_h(Z_h)-\varrho P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\leq 76|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2H^2\tau_{\rm mix}^2\ .$$

Proof. We set $\bar F_h=F_h-\varrho P^hF_h$ and $\bar g_\ell=g_\ell-\varrho P^\ell g_\ell$. Expanding the square, we have

$$\begin{aligned}&\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\bar F_h(Z_h)\bar g_\ell(Z_\ell)\Big\|^2\Big]\\&=\sum_{h=1}^{H}\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\bar F_h(Z_h)\bar g_\ell(Z_\ell)\Big\|^2\Big]+2\sum_{h<h'=1}^{H}\sum_{\ell=1}^{h-1}\sum_{\ell'=1}^{h'-1}\mathbb{E}_\varrho[\langle\bar F_h(Z_h)\bar g_\ell(Z_\ell)\,,\,\bar F_{h'}(Z_{h'})\bar g_{\ell'}(Z_{\ell'})\rangle]\ .\end{aligned}$$

Bound on the variance term. To bound the first term, we expand it as

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\bar F_h(Z_h)\bar g_\ell(Z_\ell)\Big\|^2\Big]$$

$$= \sum_{\ell=1}^{h-1} \mathbb{E}_\varrho[\|\bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\|^2] + 2\sum_{\ell<\ell'=1}^{h-1} \mathbb{E}_\varrho[\langle \bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\,,\, \bar{F}_h(Z_h)\bar{g}_{\ell'}(Z_{\ell'})\rangle] \ .$$

It is easily seen that $\mathbb{E}_\varrho[\|\bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\|^2] \le 4|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2$. On the other hand, we write

$$\mathbb{E}_\varrho[\langle \bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\,,\, \bar{F}_h(Z_h)\bar{g}_{\ell'}(Z_{\ell'})\rangle] = \sum_{i=1}^{d}\sum_{j,j'=1}^{d} \bar{F}_{h,i,j}(Z_h)\bar{g}_{\ell,j}(Z_\ell)\bar{F}_{h,i,j'}(Z_h)\bar{g}_{\ell',j'}(Z_{\ell'})$$

$$= \sum_{i=1}^{d}\sum_{j,j'=1}^{d} \overline{FF}_{h,i,j,j'}(Z_h)\bar{g}_{\ell,j}(Z_\ell)\bar{g}_{\ell',j'}(Z_{\ell'}) + \varrho P^h(\bar{F}_{h,i,j}\bar{F}_{h,i,j'})\bar{g}_{\ell,j}(Z_\ell)\bar{g}_{\ell',j'}(Z_{\ell'}) \ ,$$

where we denote $\overline{FF}_{h,i,j,j'}(Z_h) = \bar{F}_{h,i,j}(Z_h)\bar{F}_{h,i,j'}(Z_h) - \varrho P^h(\bar{F}_{h,i,j}\bar{F}_{h,i,j'})$. By the triangle inequality, Lemma A.7 and the mixing property, we obtain

$$\begin{aligned}
&|\mathbb{E}_\varrho[\langle \bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\,,\, \bar{F}_h(Z_h)\bar{g}_{\ell'}(Z_{\ell'})\rangle]| \\
&\le \sum_{i=1}^{d}\sum_{j,j'=1}^{d} 16\|F_{h,i,j}\|_\infty\|F_{h,i,j'}\|_\infty\|g_{\ell,j}\|_\infty\|g_{\ell',j'}\|_\infty(1/4)^{\lfloor(h-\ell')/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell'-\ell)/\tau_{\rm mix}\rfloor} \\
&\qquad + 16\|F_{h,i,j}\|_\infty\|F_{h,i,j'}\|_\infty\|g_{\ell,j}\|_\infty\|g_{\ell',j'}\|_\infty(1/4)^{\lfloor(\ell'-\ell)/\tau_{\rm mix}\rfloor} \ .
\end{aligned}$$

Using the definition of $|||\cdot|||_{2,\infty}$, we obtain

$$|\mathbb{E}_\varrho[\langle \bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\,,\, \bar{F}_h(Z_h)\bar{g}_{\ell'}(Z_{\ell'})\rangle]| \le 32|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2(1/4)^{\lfloor(\ell'-\ell)/\tau_{\rm mix}\rfloor} \ .$$

Summing over h, ℓ, and ℓ', this gives a bound on the first term

$$\begin{aligned}
\sum_{h=1}^{H}\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\bar{F}_h(Z_h)\bar{g}_\ell(Z_\ell)\Big\|^2\Big] &\le 4H^2|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2 + 2\sum_{h=1}^{H}\frac{4\cdot 32}{3}h|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2\tau_{\rm mix}^2 \\
&\le \frac{140}{3}H^2\tau_{\rm mix}^2|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2 \ . \qquad (23)
\end{aligned}$$

Bound on the covariance term. To study the covariance terms, we consider all different orderings of h, h', ℓ, and ℓ'. Note that $h' > h$, $h' > \ell'$, and $h > \ell$ by construction, there are thus three possible orderings: $h' > h > \ell' \ge \ell$, $h' > h > \ell > \ell'$, and $h' > \ell' \ge h > \ell$. We now treat each case separately. We set for $i, i', j, j' \in \{1, \dots, d\}$,

$$\mathbf{B}^{i,i',j,j'}_{h,\ell,h',\ell'} = \mathbb{E}_\nu[\bar{F}_{h,i,j}(Z_h)\bar{g}_{\ell,j}(Z_\ell)\bar{F}_{h',i',j'}(Z_{h'})\bar{g}_{\ell',j'}(Z_{\ell'})] \ .$$

Denoting $\mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'} = 16\|F_{h,i,j}\|_\infty\|g_{\ell,j}\|_\infty\|F_{h',i',j'}\|_\infty\|g_{\ell',j'}\|_\infty$.

(Case $h' > h > \ell' > \ell$.) Applying Lemma A.7, we get, since $\ell' > \ell$,

$$|\mathbf{B}^{i,i',j,j'}_{h,\ell,h',\ell'}| \le \mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\Big((1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell'-\ell)/\tau_{\rm mix}\rfloor} + (1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell')/\tau_{\rm mix}\rfloor}\Big) \ .$$

(Case $h' > h > \ell \ge \ell'$.) We simply need to switch the ℓ' and ℓ.

$$|\mathbf{B}^{i,i',j,j'}_{h,\ell,h',\ell'}| \le \mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\Big((1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell-\ell')/\tau_{\rm mix}\rfloor} + (1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell)/\tau_{\rm mix}\rfloor}\Big) \ .$$

(Case $h' > \ell' \ge h > \ell$.) Similarly to the first two cases, applying Lemma A.7

$$|\mathbf{B}^{i,i',j,j'}_{h,\ell,h',\ell'}| \le \mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\Big((1/4)^{\lfloor(h'-\ell')/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell)/\tau_{\rm mix}\rfloor} + (1/4)^{\lfloor(h'-\ell')/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell'-h)/\tau_{\rm mix}\rfloor}\Big) \ .$$

(Final bound.) Splitting the sum and combining the three above results, we obtain

$$\begin{aligned}\sum_{\ell=1}^{h-1}\sum_{\ell'=1}^{h'-1}\left|\mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\right| &\le \mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\sum_{h'>h>\ell'>\ell}(1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell'-\ell)/\tau_{\rm mix}\rfloor}+(1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell')/\tau_{\rm mix}\rfloor}\\ &+\mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\sum_{h'>h>\ell\ge\ell'}(1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell-\ell')/\tau_{\rm mix}\rfloor}+(1/4)^{\lfloor(h'-h)/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell)/\tau_{\rm mix}\rfloor}\\ &+\mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}\sum_{h'>\ell'\ge h>\ell}(1/4)^{\lfloor(h'-\ell')/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(h-\ell)/\tau_{\rm mix}\rfloor}+(1/4)^{\lfloor(h'-\ell')/\tau_{\rm mix}\rfloor}(1/4)^{\lfloor(\ell'-h)/\tau_{\rm mix}\rfloor}\\ &\le \mathbf{C}^{i,i',j,j'}_{h,\ell,h',\ell'}(4/3)^2H^2\tau_{\rm mix}^2\ .\end{aligned}$$

Summing over all i,i',j,j', and using the definitions of the norms, we obtain

$$2\sum_{h<h'=1}^{H}\sum_{\ell=1}^{h-1}\sum_{\ell'=1}^{h'-1}\mathbb{E}_\varrho[\langle\bar F_h(Z_h)\bar g_\ell(Z_\ell)\,,\,\bar F_{h'}(Z_{h'})\bar g_{\ell'}(Z_{\ell'})\rangle]\le\frac{256}{9}H^2\tau_{\rm mix}^2|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2\ ,$$

and the result follows by combining this inequality with (23). □

Corollary A.13. *Under the assumptions of Lemma A.12, with ν such that $\nu=P\nu$, we have*

$$\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{F_h(Z_h)-\nu P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\le 312H^2\tau_{\rm mix}^2|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2\ .$$

Proof. Using Young's inequality, we have

$$\begin{aligned}&\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{F_h(Z_h)-\nu P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\\ &\le 2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{F_h(Z_h)-\varrho P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\\ &\quad+2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{\varrho P^hF_h-\nu P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\ .\end{aligned}$$

The second term can be bounded using Lemma A.8, which gives

$$\begin{aligned}2\mathbb{E}_\varrho\Big[\Big\|\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\{\varrho P^hF_h-\nu P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]&\le 2H\sum_{h=1}^{H}\mathbb{E}_\varrho\Big[\Big\|\sum_{\ell=1}^{h-1}\{\varrho P^hF_h-\nu P^hF_h\}\{g_\ell(Z_\ell)-\varrho P^\ell g_\ell\}\Big\|^2\Big]\\ &\le 2H\sum_{h=1}^{H}2\cdot15\cdot4h\tau_{\rm mix}(1/4)^{\lfloor h/\tau_{\rm mix}\rfloor}|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2\\ &\le 160H^2\tau_{\rm mix}^2|||F|||_{2,\infty}^2|||g|||_{2,\infty}^2\ ,\end{aligned}$$

where we used $|||\varrho P^hF_h-\nu P^hF_h|||_{2,\infty}\le 2(1/4)^{\lfloor h/\tau_{\rm mix}\rfloor}|||F|||_{2,\infty}$ in the second inequality. The first term can be bounded using Lemma A.12, and the result follows. □

B. Technical Lemmas

B.1. Bounds on policy improvement

In subsequent proofs, we will require the following lemma, restated from Zou et al. (2019)'s Lemma 3. It allows to bound the difference between the two measures μ_{θ_1} and μ_{θ_2}, parameterized by two parameters θ_1,θ_2, as a function of the distance between these two parameters.

Lemma B.1 (from (Zou et al., 2019)). *Assume A1 and A4. Let μ_θ the invariant joint distribution of state and action after following policy π_θ. Then, for $\theta_1, \theta_2 \in \mathbb{R}^d$ we have:*

$$\|\mu_{\theta_1} - \mu_{\theta_2}\|_{\mathrm{TV}} \leq \mathrm{C}_\mu \|\theta_1 - \theta_2\| \ , \tag{24}$$

where $\mathrm{C}_\mu \triangleq \mathrm{C}_{\mathrm{lip}} |\mathcal{A}| \, (1 + 4\tau_{\mathrm{mix}})$.

Proof. Note that under our assumption A 4, the assumptions of Zou et al. (2019) are satisfied with $m = 4$ and $\rho = (1/4)^{1/\tau_{\mathrm{mix}}}$. Lemma 3 from Zou et al. (2019) then gives

$$\mathrm{C}_\mu \leq \mathrm{C}_{\mathrm{lip}} |\mathcal{A}| \left(1 + \lceil \log(1/m) \rceil + \frac{1}{1-\rho}\right) \leq \mathrm{C}_{\mathrm{lip}} |\mathcal{A}| \left(1 + \frac{1}{1 - \exp(-\log(4)/\tau_{\mathrm{mix}})}\right) \ ,$$

and the result follows from $\exp(-x) \leq 1 - x/2$ and $1/\log(4) \leq 2$. □

A crucial corollary of this lemma is that the matrix $\mathbf{A}^{(1)}(\cdot)$ and the vector $\mathbf{b}^{(1)}(\cdot)$ are Lipschitz.

Corollary B.2. *Assume A1 and A4. Then, for any $\theta \in \mathbb{R}^d$, the following property holds*

$$\|\mathbf{A}^{(1)}(\theta) - \mathbf{A}^{(1)}(\theta_\star)\| \leq \mathrm{C}_\mu \mathrm{C}_A \|\theta - \theta_\star\| \ , \quad \|\mathbf{b}^{(1)}(\theta) - \mathbf{b}^{(1)}(\theta_\star)\| \leq \mathrm{C}_\mu \mathrm{C}_b \|\theta - \theta_\star\| \ ,$$

where $\mathrm{C}_\mu = \mathrm{C}_{\mathrm{lip}} |\mathcal{A}| \, (1 + 4\tau_{\mathrm{mix}})$ *is defined in Lemma B.1.*

Proof. From (3), we have, with $\mathbf{A}^{(c)}(s, a, s', a') = \phi(s, a) \left(\gamma \phi(s', a')^\top - \phi(s, a)^\top\right)$, that

$$\bar{\mathbf{A}}^{(c)}(\theta) = \mathbb{E}_{\substack{(s,a) \sim \mu_\theta \\ s' \sim P^{(c)}(\cdot|s,a) \\ a' \sim \pi_\theta(\cdot|s')}} \left[\mathbf{A}^{(c)}(s, a, s', a')\right] \ .$$

This gives

$$\begin{aligned}
\|\mathbf{A}^{(1)}(\theta) - \mathbf{A}^{(1)}(\theta_\star)\| &= \left\| \mathbb{E}_{\substack{(s,a) \sim \mu_\theta \\ s' \sim P^{(c)}(\cdot|s,a), \\ a' \sim \pi_\theta(\cdot|s')}} [\mathbf{A}^{(c)}(s, a, s', a')] - \mathbb{E}_{\substack{(s,a) \sim \mu_{\theta_\star} \\ s' \sim P^{(c)}(\cdot|s,a) \\ a' \sim \pi_{\theta_\star}(\cdot|s')}} [\mathbf{A}^{(c)}(s, a, s', a')] \right\| \\
&\leq \|\mu_\theta - \mu_{\theta_\star}\|_{\mathrm{TV}} \sup_{s,a,s',a'} \|\mathbf{A}^{(c)}(s, a, s', a')\| \ ,
\end{aligned}$$

and the result follows from the definition of C_A and Lemma B.1. The second inequality follows from similar derivations. □

B.2. Expression of difference of matrix products

In our derivations, we will use the follwing lemma, that allows to decompose the difference of two matrix products.

Lemma B.3. *For $c \in \{1, \ldots, N\}$, $t \geq 0$, $\eta_t \geq 0$, and $H \geq 0$, we have*

$$\Gamma^{(c)}_{t,1:H} - \mathrm{I} = \eta_t \sum_{h=1}^{H} \bar{\mathbf{A}}^{(c)}(\theta_\star) \Gamma^{(1)}_{t,1:h-1} \ , \tag{25}$$

$$\Gamma^{(c)}_{t,1:H} - \Gamma^{(avg)}_{1:H} = \eta_t \sum_{h=1}^{H} \Gamma^{(c)}_{t,h-1} (\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)) \Gamma^{(avg)}_{h+1:H} \ . \tag{26}$$

Proof. Remark that for all $h \geq 0$, we have

$$\Gamma^{(c)}_{t,1:h+1} - \mathrm{I} = (\mathrm{I} + \eta_t \bar{\mathbf{A}}^{(c)}(\theta_\star)) \Gamma^{(1)}_{t,1:h} - \mathrm{I} = \left(\Gamma^{(c)}_{t,1:h} - \mathrm{I}\right) + \eta_t \bar{\mathbf{A}}^{(c)}(\theta_\star) \Gamma^{(1)}_{t,1:h} \ ,$$

and the first identity follows. The second one follows from similar computations. □

C. Bound on terms of the decomposition

C.1. Decomposition of the error

Claim 1 (Restated). *Let $t \geq 0$. The updates of SARSA at block t can be written as*

$$\theta_{t,H}^{(1)} - \theta_\star = \Gamma_{t,1:H}^{(1)}(\theta_t - \theta_\star) + \textstyle\sum_{h=1}^{H} \eta_t \Gamma_{t,h+1:H}^{(1)} \varphi_{t,h-1}^{(1)} + \sum_{h=1}^{H} \eta_t \Gamma_{t,h+1:H}^{(1)} \varepsilon_{t,h-1}^{(1)}(Z_{t,h}^{(1)}) \ . \tag{10}$$

Proof. We have, by adding and removing $\eta_t \left(\bar{\mathbf{A}}^{(1)}(\theta_\star)\theta_{t,h}^{(1)} + \bar{\mathbf{b}}^{(1)}(\theta_\star) \right)$ to the update of $\theta_{t,h+1}^{(1)}$,

$$\begin{aligned}
\theta_{t,h+1}^{(1)} - \theta_\star &= \theta_{t,h}^{(1)} - \theta_\star + \eta_t \left(\mathbf{A}^{(1)}(Z_{t,h+1}^{(1)})\theta_{t,h}^{(1)} + \mathbf{b}^{(1)}(Z_{t,h+1}^{(1)}) \right) \\
&= \theta_{t,h}^{(1)} - \theta_\star + \eta_t \left(\bar{\mathbf{A}}^{(1)}(\theta_\star)\theta_{t,h}^{(1)} + \bar{\mathbf{b}}^{(1)}(\theta_\star) \right) \\
&\quad + \eta_t \left((\mathbf{A}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{A}}^{(1)}(\theta_\star))\theta_{t,h}^{(1)} + \mathbf{b}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{b}}^{(1)}(\theta_\star) \right) \ .
\end{aligned}$$

Since $\bar{\mathbf{b}}^{(1)}(\theta_\star) = -\bar{\mathbf{A}}^{(1)}(\theta_\star)\theta_\star$ in the single-agent case, we have

$$\begin{aligned}
\theta_{t,h+1}^{(1)} - \theta_\star &= \theta_{t,h}^{(1)} - \theta_\star + \eta_t \bar{\mathbf{A}}^{(1)}(\theta_\star) \left(\theta_{t,h}^{(1)} - \theta_\star \right) \\
&\quad + \eta_t \left((\mathbf{A}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{A}}^{(1)}(\theta_t))\theta_{t,h}^{(1)} + \mathbf{b}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{b}}^{(1)}(\theta_t) \right) \\
&\quad + \eta_t \left((\bar{\mathbf{A}}^{(1)}(\theta_t) - \bar{\mathbf{A}}^{(1)}(\theta_\star))\theta_{t,h}^{(1)} + \bar{\mathbf{b}}^{(1)}(\theta_t) - \bar{\mathbf{b}}^{(1)}(\theta_\star) \right) \ .
\end{aligned}$$

We then replace $\varphi_{t,h}^{(1)}$ and $\varepsilon_{t,h}^{(1)}(Z_{t,h+1}^{(1)})$ by their definitions

$$\begin{aligned}
\varphi_{t,h}^{(1)} &= (\bar{\mathbf{A}}^{(1)}(\theta_t) - \bar{\mathbf{A}}^{(1)}(\theta_\star))\theta_{t,h}^{(1)} + \bar{\mathbf{b}}^{(1)}(\theta_t) - \bar{\mathbf{b}}^{(1)}(\theta_\star) \ , \\
\varepsilon_{t,h}^{(1)}(Z_{t,h+1}^{(1)}) &= (\mathbf{A}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{A}}^{(1)}(\theta_t))\theta_{t,h}^{(1)} + \mathbf{b}^{(1)}(Z_{t,h+1}^{(1)}) - \bar{\mathbf{b}}^{(1)}(\theta_t) \ ,
\end{aligned}$$

and the result follows after unrolling the recursion. □

C.2. Bounds on iterates' norm and variance

First, we provide some bounds on the differences between the global and local iterates.

Lemma C.1 (Bound on the local iterates). *Assume A1. Let $c \in \{1, \dots, N\}$, $t, h \geq 0$, and $\theta_t \in \mathbb{R}^d$, and $\theta_{t,h}^{(c)} \in \mathbb{R}^d$ be the parameters obtained after h local TD updates started from θ_t. Then, whenever $\eta_t H \mathrm{C}_A \leq 1$ it holds that*

(a) the distance between last policy improvement and current estimate is bounded

$$\|\theta_t - \theta_{t,h}^{(c)}\| \leq 3\eta_t H(\mathrm{C}_A \mathrm{C}_{\text{proj}} + \mathrm{C}_b) \leq 3(\mathrm{C}_{\text{proj}} + 1) \ ,$$

(b) the norm of current iterate is bounded as

$$\|\theta_{t,h}^{(c)}\| \leq \widetilde{\mathrm{C}}_{\text{proj}} \stackrel{\Delta}{=} 4(\mathrm{C}_{\text{proj}} + 1) \ ,$$

(c) if $\eta_t H \mathrm{C}_A \leq 1/6$, the variance, conditionally on $\mathcal{F}_t$, of the local updates is bounded as

$$\mathbb{E}[\|\theta_{t,h}^{(c)} - \mathbb{E}[\theta_{t,h}^{(c)} \,|\mathcal{F}_t]\|^2 \,|\mathcal{F}_t] \leq 55\eta_t^2 H \tau_{\text{mix}} \mathrm{G}^2 \ ,$$

Proof. *Proof of C.1-(a).* Triangle inequality gives

$$\|\theta_t - \theta_{t,h+1}^{(c)}\| = \|\theta_t - \theta_{t,h}^{(c)} + \theta_{t,h}^{(c)} - \theta_{t,h+1}^{(c)}\| \leq \|\theta_t - \theta_{t,h}^{(c)}\| + \|\theta_{t,h}^{(c)} - \theta_{t,h+1}^{(c)}\| \ .$$

Plugging in the update, we have

$$\begin{aligned}\|\theta_t - \theta_{t,h+1}^{(c)}\| &\leq \|\theta_t - \theta_{t,h}^{(c)}\| + \eta_t \|\mathbf{A}^{(c)}(Z_{t,h+1}^{(c)})\theta_{t,h}^{(c)} + \mathbf{b}^{(c)}(Z_{t,h+1}^{(c)})\| \\ &\leq \|\theta_t - \theta_{t,h}^{(c)}\| + \eta_t \|\mathbf{A}^{(c)}(Z_{t,h+1}^{(c)})\|\,\|\theta_{t,h}^{(c)}\| + \eta_t \|\mathbf{b}^{(c)}(Z_{t,h+1}^{(c)})\| \enspace .\end{aligned}$$

Bounding $\|\theta_{t,h}^{(c)}\| \leq \|\theta_{t,h}^{(c)} - \theta_t\| + \|\theta_t\|$ and using the bounds from A1, we obtain

$$\begin{aligned}\|\theta_t - \theta_{t,h+1}^{(c)}\| &\leq \|\theta_t - \theta_{t,h}^{(c)}\| + \eta_t \mathrm{C}_A \|\theta_{t,h}^{(c)} - \theta_t\| + \eta_t \mathrm{C}_A \|\theta_t\| + \eta_t \mathrm{C}_b \\ &\leq (1 + \eta_t \mathrm{C}_A)\|\theta_t - \theta_{t,h}^{(c)}\| + \eta_t \mathrm{C}_A \mathrm{C}_{\text{proj}} + \eta_t \mathrm{C}_b \enspace ,\end{aligned}$$

where the second inequality comes from $\|\theta_t\| \leq \mathrm{C}_{\text{proj}}$. Unrolling the recursion, we obtain

$$\|\theta_t - \theta_{t,h}^{(c)}\| \leq \eta_t \sum_{\ell=0}^{h-1} (1 + \eta_t \mathrm{C}_A)^\ell (\mathrm{C}_A \mathrm{C}_{\text{proj}} + \mathrm{C}_b) \leq 3\eta_t H (\mathrm{C}_A \mathrm{C}_{\text{proj}} + \mathrm{C}_b) \enspace ,$$

where the last inequality comes from the fact that, for $\ell \leq H$ and $\eta_t \mathrm{C}_A \leq 1/H$, it holds that $(1+\eta_t \mathrm{C}_A)^\ell \leq (1+1/H)^H \leq 3$. The bound follows from $\eta_t H \leq 1/\mathrm{C}_A$ and $\mathrm{C}_b/\mathrm{C}_A \leq 1$.

Proof of C.1-(b). The second bound follows from $\|\theta_{t,h}^{(c)}\| \leq \|\theta_{t,h}^{(c)} - \theta_t\| + \|\theta_t\|$ and C.1-(a).

Proof of C.1-(c). First, remark that $\mathbb{E}[\|\theta_{t,0}^{(c)} - \mathbb{E}[\theta_{t,0}^{(c)} \,|\mathcal{F}_t]\|^2 \,|\mathcal{F}_t] = 0$, thus the result holds for $h = 0$. Let $h \geq 0$, and assume that the result holds for all $\ell \leq h$. Then, we have

$$\begin{aligned}&\mathbb{E}[\|\theta_{t,h+1}^{(c)} - \mathbb{E}[\theta_{t,h+1}^{(c)} \,|\mathcal{F}_t]\|^2 \,|\mathcal{F}_t] \\ &= \mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} + \mathbf{b}^{(c)}(Z_{\ell+1}^{(c)}) - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} + \mathbf{b}^{(c)}(Z_{\ell+1}^{(c)}) \,|\mathcal{F}_t]\Big\|^2 \,\Big|\mathcal{F}_t\Big] \\ &\leq 2\mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big\|^2 \,\Big|\mathcal{F}_t\Big] \\ &\quad + 2\mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{b}^{(c)}(Z_{\ell+1}^{(c)}) - \mathbb{E}[\mathbf{b}^{(c)}(Z_{\ell+1}^{(c)}) \,|\mathcal{F}_t]\Big\|^2 \,\Big|\mathcal{F}_t\Big] \enspace .\end{aligned}$$

We decompose

$$\begin{aligned}&\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t] \\ &= \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\Big(\theta_{t,\ell}^{(c)} - \mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big) \\ &\quad + \Big(\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)}) - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)}) \,|\mathcal{F}_t]\Big)\mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t] + \mathbb{E}\Big[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\Big(\theta_{t,\ell}^{(c)} - \mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big) \,\Big|\mathcal{F}_t\Big] \enspace ,\end{aligned}$$

which gives, using $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$,

$$\begin{aligned}&\mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big\|^2 \,\Big|\mathcal{F}_t\Big] \\ &\leq 6\mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\Big(\theta_{t,\ell}^{(c)} - \mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big)\Big\|^2 \,\Big|\mathcal{F}_t\Big] \\ &\quad + 3\mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \Big(\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)}) - \mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)}) \,|\mathcal{F}_t]\Big)\mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big\|^2 \,\Big|\mathcal{F}_t\Big] \enspace ,\end{aligned}$$

where we also used Jensen's inequality to bound

$$\Big\|\eta_t \sum_{\ell=0}^{h} \mathbb{E}\Big[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\Big(\theta_{t,\ell}^{(c)} - \mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big) \,\Big|\mathcal{F}_t\Big]\Big\|^2 \leq \mathbb{E}\Big[\Big\|\eta_t \sum_{\ell=0}^{h} \mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\Big(\theta_{t,\ell}^{(c)} - \mathbb{E}[\theta_{t,\ell}^{(c)} \,|\mathcal{F}_t]\Big)\Big\|^2 \,\Big|\mathcal{F}_r\Big] \enspace .$$

Since the $\mathbf{A}^{(c)}(\cdot)$ are uniformly bounded, we can use the induction hypothesis, Lemma A.8, and Lemma A.10, to obtain

$$\mathbb{E}\Big[\Big\|\eta_t\sum_{\ell=0}^{h}\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)}-\mathbb{E}[\mathbf{A}^{(c)}(Z_{\ell+1}^{(c)})\theta_{t,\ell}^{(c)}|\mathcal{F}_t]\Big\|^2\Big|\mathcal{F}_t\Big]\leq 330\eta_t^4H^3\mathrm{C}_A^2\tau_{\mathrm{mix}}\mathrm{G}^2+45\eta_t^2H\tau_{\mathrm{mix}}\mathrm{C}_A^2\widetilde{\mathrm{C}}_{\mathrm{proj}}^2 \ .$$

Similarly, Lemma A.8 gives

$$\mathbb{E}\Big[\Big\|\eta_t\sum_{\ell=0}^{h}\mathbf{b}^{(c)}(Z_{\ell+1}^{(c)})-\mathbb{E}[\mathbf{b}^{(c)}(Z_{\ell+1}^{(c)})\,|\mathcal{F}_t]\Big\|^2\Big|\mathcal{F}_t\Big]\leq 15\eta_t^2H\tau_{\mathrm{mix}}\mathrm{C}_b^2 \ ,$$

and the result follows from $\eta_t H\mathrm{C}_A\leq 1/6$. □

Proposition C.2. *Assume A1. Let $t\geq 0$, and $\eta_t\leq 1/\mathrm{C}_A$. Then, for any vector $u\in\mathbb{R}^d$, and $k\leq h$, we have*

$$\|\Gamma_{t,k:h}^{(1)}u\|^2\leq(1-\eta_t a)^{k-h+1}\|u\|^2 \ .$$

Proof. The result follows directly from A1. □

Lemma C.3. *Assume A1-5. Let $t\geq 0$ and assume that $\eta_t H\mathrm{C}_A\leq 1$, then for $c\in\{1,\dots,N\}$, we have*

$$\Big\|\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Z_{t,h}^{(c)})\Big]\Big\|\leq\frac{8\mathrm{G}}{3}\big(\eta_t\mathrm{C}_AH\tau_{\mathrm{mix}}+\delta\tau_{\mathrm{mix}}\big) \ ,$$

where $\delta=0$ if $Z_{t,0}^{(c)}$ is sampled from the stationary distribution ν_{θ_t} and $\delta=1$ otherwise.

Proof. If $\delta=1$, we use Lemma A.3 to construct a Markov chain $Y_{t,h}^{(c)}$ such that $Y_{t,h}^{(c)}\sim\nu_{\theta_t}$ starts from the stationary distribution of the policy π_{θ_t} and

$$\Big\|\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Z_{t,h}^{(c)})\Big]\Big\|\leq\Big\|\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Y_{t,h}^{(c)})\Big]\Big\|+\frac{8(\mathrm{C}_A\widetilde{\mathrm{C}}_{\mathrm{proj}}+\mathrm{C}_b)}{3}\delta\tau_{\mathrm{mix}} \ , \tag{27}$$

which also holds when $\delta=0$. Then, we write

$$\begin{aligned}\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Y_{t,h}^{(c)})\Big]&=\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\mathbb{E}\Big[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\theta_{t,h-1}^{(c)}+\widetilde{\mathbf{b}}^{(c)}(Y_{t,h}^{(c)})\Big]\\&=-\eta_t\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma_{t,h+1:H}^{(c)}\mathbb{E}\Big[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\big(\mathbf{A}^{(c)}(Y_{t,\ell}^{(c)})\theta_{t,\ell-1}^{(c)}+\mathbf{b}^{(c)}(Y_{t,\ell}^{(c)})\big)\Big] \ ,\end{aligned} \tag{28}$$

where we recall that $\widetilde{\mathbf{A}}^{(c)}(z)=\mathbf{A}^{(c)}(z)-\bar{\mathbf{A}}^{(c)}(\theta_t)$ and we used the fact that $\mathbb{E}[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\theta_t]=0$. Now, remark that, by A1 and Jensen's inequality,

$$\begin{aligned}\Big\|\mathbb{E}\Big[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})(\mathbf{A}^{(c)}(Y_{t,\ell}^{(c)})\theta_{t,\ell-1}^{(c)}+\mathbf{b}^{(c)}(Y_{t,\ell}^{(c)})\Big]\Big\|&=\Big\|\mathbb{E}\Big[\mathbb{E}[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\,|Y_{t,\ell}^{(c)}]\big(\mathbf{A}^{(c)}(Y_{t,\ell}^{(c)})\theta_{t,\ell-1}^{(c)}+\mathbf{b}^{(c)}(Y_{t,\ell}^{(c)})\big)\Big]\Big\|\\&\leq\mathbb{E}\Big[\Big\|\mathbb{E}[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\,|Y_{t,\ell}^{(c)}]\Big\|\Big](\mathrm{C}_A\widetilde{\mathrm{C}}_{\mathrm{proj}}+\mathrm{C}_b) \ .\end{aligned}$$

Then, A4 gives the bound $\Big\|\mathbb{E}[\widetilde{\mathbf{A}}^{(c)}(Y_{t,h}^{(c)})\,|Y_{t,\ell}^{(c)}]\Big\|\leq 2\mathrm{C}_A(1/4)^{\lfloor(h-\ell)/\tau_{\mathrm{mix}}\rfloor}$, and we obtain

$$\Big\|\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Y_{t,h}^{(c)})\Big]\Big\|\leq 2\eta_t\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\mathrm{C}_A(\mathrm{C}_A\widetilde{\mathrm{C}}_{\mathrm{proj}}+\mathrm{C}_b)(1/4)^{\lfloor(h-\ell)/\tau_{\mathrm{mix}}\rfloor} \ .$$

Bounding the inner sum by the sum of the series, plugging the result in (28) gives

$$\Big\|\mathbb{E}\Big[\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h}^{(c)}(Y_{t,h}^{(c)})\Big]\Big\|\leq\frac{8\eta_t\mathrm{C}_A(\mathrm{C}_A\widetilde{\mathrm{C}}_{\mathrm{proj}}+\mathrm{C}_b)H\tau_{\mathrm{mix}}}{3} \ , \tag{29}$$

and the result follows from plugging (29) in (27). □

Lemma C.4. *Assume A1-5. Let $t \geq 0$, and assume the step size satisfies $\eta_t H \mathrm{C}_A \leq 1$, then*

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\varepsilon^{(c)}_{t,h}(Z^{(c)}_{t,h})\Big\|^2\Big] \leq \frac{136 H \tau_{\mathrm{mix}}\mathrm{G}^2}{N} + 5504\eta_t^2\mathrm{C}_A^2\mathrm{G}^2H^2\tau_{\mathrm{mix}}^2 \ ,$$

where $\delta = 1$ if $Z^{(c)}_{t,0} \sim \nu_{\theta_t}$ and $\delta = 0$ otherwise.

Proof. We start from

$$\begin{aligned}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h})\Big\|^2\Big] &\leq 2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\theta^{(c)}_{t,h-1}\Big\|^2 \\ &\quad + 2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{b}}^{(c)}(Z^{(c)}_{t,h})\Big\|^2 \ . \end{aligned} \tag{30}$$

First, the second term of (30) can be bounded by Corollary A.9, which gives

$$2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{b}}^{(c)}(Y^{(c)}_{t,h})\Big\|^2 \leq \frac{34 H \tau_{\mathrm{mix}}\mathrm{C}_b^2}{N} \ . \tag{31}$$

Then, to bound the first term of (30), we recall that $\theta^{(c)}_{t,h} = \theta_t - \eta_t \sum_{\ell=1}^{h}\mathbf{A}^{(c)}(Y^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1} + \mathbf{b}^{(c)}(Y^{(c)}_{t,\ell})$, which gives the following decomposition, using $\theta^{(c)}_{t,h} = \theta^{(c)}_{t,h} - \mathbb{E}[\theta^{(c)}_{t,h}\,|\mathcal{F}_t] + \mathbb{E}[\theta^{(c)}_{t,h}\,|\mathcal{F}_t]$,

$$\begin{aligned}&2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\theta^{(c)}_{t,h-1}\Big\|^2\Big] \\ &\leq 4\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\mathbb{E}[\theta^{(c)}_{t,h-1}\,|\mathcal{F}_t]\Big\|^2 \\ &\quad + 4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\Big(\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t]\Big)\Big\|^2 \ . \end{aligned} \tag{32}$$

Since the $\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})$ are centered and independent from one agent to another, we can use Corollary A.9 to bound the first term as

$$4\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\mathbb{E}[\theta^{(c)}_{t,h}]\Big\|^2 \leq \frac{136 H \tau_{\mathrm{mix}}\mathrm{C}_A^2\widetilde{\mathrm{C}}^2_{\mathrm{proj}}}{N} \ , \tag{33}$$

To bound the second term, we use the decomposition

$$\begin{aligned}\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t] &= \mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\big(\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t]\big) \\ &\quad + \big(\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell}) - \mathbb{E}[\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})]\big)\mathbb{E}[\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t] \\ &\quad + \mathbb{E}[\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\big(\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t]\big)\,|\mathcal{F}_t] \ . \end{aligned}$$

Using $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ and Jensen's inequality, we have

$$\begin{aligned}&4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\Big(\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t]\Big)\Big\|^2 \\ &\leq 6\cdot 4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma^{(c)}_{t,h+1:H}\widetilde{\mathbf{A}}^{(c)}(Z^{(c)}_{t,h})\mathbf{A}^{(c)}(Z^{(c)}_{t,\ell})\Big(\theta^{(c)}_{t,\ell-1} - \mathbb{E}[\theta^{(c)}_{t,\ell-1}\,|\mathcal{F}_t]\Big)\Big\|^2\Big] \end{aligned}$$

Algorithm 2 SARSA: Single-Agent State-Action-Reward-State-Action

Input: step sizes $\eta_t > 0$, initial parameters θ_0, projection set $\mathcal{W}$, number of local steps $H > 0$, number of communications $T > 0$, initial distribution ϱ over states
Initialize first state $s_{-1,H} \sim \varrho$ and initial policy $\pi_{\theta_0} = \mathrm{Imp}_\beta(Q_{\theta_0}$
for step $t = 0$ to $T-1$ **do**
  Initialize $\theta_{t,0}^{(1)} = \theta_t$, take first action $a_{t,0}^{(1)} \sim \pi_{\theta_t}(\cdot|s_{t-1,H}^{(1)})$
  for step $h = 0$ to $H-1$ **do**
    Take action $a_{t,h+1}^{(1)} \sim \pi_{\theta_t}(\cdot|s_{t,h}^{(1)})$, observe reward $r^{(1)}(s_{t,h}^{(1)}, a_{t,h}^{(1)})$, next state $s_{t,h+1}^{(1)}$
    Compute $\delta_{t,h}^{(1)} = r^{(1)}(s_{t,h+1}^{(1)}, a_{t,h+1}^{(1)}) + \gamma\phi(s_{t,h+1}^{(1)}, a_{t,h+1}^{(1)})^\top\theta_{t,h}^{(1)} - \phi(s_{t,h}^{(1)}, a_{t,h}^{(1)})^\top\theta_{t,h}^{(1)}$
    Update $\theta_{t,h+1}^{(1)} = \theta_{t,h}^{(1)} + \eta_t\delta_{t,h}^{(1)}\phi(s_{t,h}^{(1)}, a_{t,h}^{(1)})$
  end for
  Update global parameter $\theta_{t+1} = \Pi\left(\theta_{t,H}^{(1)}\right)$
  Policy improvement $\pi_{\theta_{t+1}} = \mathrm{Imp}_\beta(Q_{\theta_{t+1}})$
end for
Return: θ_T

$$+ 3\cdot 4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma_{t,h+1:H}^{(c)}\widetilde{\mathbf{A}}^{(c)}(Z_{t,h}^{(c)})\mathring{\mathbf{A}}^{(c)}(Z_{t,\ell}^{(c)})\mathbb{E}[\theta_{t,\ell-1}^{(c)}\,|\mathcal{F}_t]\Big\|^2\Big]\ ,$$

where we defined $\mathring{\mathbf{A}}^{(c)}(Z_{t,\ell}^{(c)}) = \mathbf{A}^{(c)}(Z_{t,\ell}^{(c)}) - \mathbb{E}[\mathbf{A}^{(c)}(Z_{t,\ell}^{(c)})\,|\mathcal{F}_t]$. The second term can be bounded using Corollary A.13, which gives

$$3\cdot 4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h-1}\Gamma_{t,h+1:H}^{(c)}\widetilde{\mathbf{A}}^{(c)}(Z_{t,h}^{(c)})\widetilde{\mathbf{A}}^{(c)}(Z_{t,\ell}^{(c)})\mathbb{E}[\theta_{t,\ell-1}^{(c)}\,|\mathcal{F}_t]\Big\|^2\Big] \le 3744\eta_t^2H^2\tau_{\mathrm{mix}}^2\mathrm{C}_A^4\widetilde{\mathrm{C}}_{\mathrm{proj}}^2\ .$$

To bound the first one, we remark that, by Lemma C.1-(c), we have

$$\sup_{1\le h\le H}\mathbb{E}\Big[\Big\|\mathbf{A}^{(c)}(Z_{t,h}^{(c)})\Big(\theta_{t,h-1}^{(c)} - \mathbb{E}[\theta_{t,h-1}^{(c)}\,|\mathcal{F}_t]\Big)\Big\|^2\Big] \le 55\eta_t^2\mathrm{C}_A^2\mathrm{G}^2H\tau_{\mathrm{mix}}\ .$$

Thus, by Corollary A.11, we obtain

$$6\cdot 4\eta_t^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\sum_{\ell=1}^{h}\Gamma_{t,h+1:H}^{(c)}\widetilde{\mathbf{A}}^{(c)}(Z_{t,h}^{(c)})\mathbf{A}^{(c)}(Z_{t,\ell}^{(c)})\Big(\theta_{t,\ell-1}^{(c)} - \mathbb{E}[\theta_{t,\ell-1}^{(c)}]\Big)\Big\|^2\Big] \le 6\cdot 4\cdot 48\cdot 55\eta_t^4\mathrm{C}_A^4\mathrm{G}^2H^4\tau_{\mathrm{mix}}^2\ . \tag{34}$$

Plugging (33) and (34) in (32), together with (31), allows to upper bound (30) as

$$\begin{aligned}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\varepsilon_{t,h-1}^{(c)}(Y_{t,h}^{(c)})\Big\|^2\Big] &\le \frac{136H\tau_{\mathrm{mix}}\mathrm{G}^2}{N} + 63360\eta_t^4\mathrm{C}_A^4\mathrm{G}^2H^4\tau_{\mathrm{mix}}^2 + 3744\eta_t^2H^2\tau_{\mathrm{mix}}^2\mathrm{C}_A^4\widetilde{\mathrm{C}}_{\mathrm{proj}}^2\\ &\le \frac{136H\tau_{\mathrm{mix}}\mathrm{G}^2}{N} + 5504\eta_t^2\mathrm{C}_A^2\mathrm{G}^2H^2\tau_{\mathrm{mix}}^2\ ,\end{aligned}$$

where we also used $\eta_tH\mathrm{C}_A \le 1/6$, $\mathrm{C}_A^2\widetilde{\mathrm{C}}_{\mathrm{proj}}^2 \le \mathrm{G}^2$. □

D. Proof for single-agent SARSA— Proofs of Lemma 4.1 and Theorem 4.2

Lemma D.1 (Restated). *Assume A1–5. Let $t \ge 0$, assume that the step size satisfies $\eta_tH\mathrm{C}_A \le 1/6$. Then, it holds that*

$$\begin{aligned}\mathbb{E}[\|\theta_{t,H}^{(1)} - \theta_\star\|^2] \le (1-\tfrac{\eta_t aH}{4})\|\theta_t-\theta_\star\|^2 + 136\eta_t^2H\tau_{\mathrm{mix}}\mathrm{G}^2\\ + \delta\tfrac{58\eta_t\tau_{\mathrm{mix}}^2\mathrm{G}^2}{Ha} + \tfrac{976\eta_t^3H\tau_{\mathrm{mix}}^2\mathrm{G}^2\mathrm{C}_A^2}{a}\ ,\end{aligned}$$

where and $\delta = 0$ if episodes start in the stationary distribution and $\delta = 1$ otherwise.

Proof. We use a step size $\eta_{t,h} = \eta_t$, that verifies $\eta_t a \leq 1/2$, and remains constant during H steps and only depends on t. Expanding the expected square of Claim 1's error decomposition, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\theta_{t,H}^{(1)} - \theta_\star\|^2\right] \\
&= \|\Gamma_{t,1:H}^{(c)}(\theta_t - \theta_\star)\|^2 + 2\mathbb{E}\Big[\Big\langle \Gamma_{t,1:H}^{(c)}(\theta_t - \theta_\star)\,,\, \sum_{h=1}^{H} \eta_t \Gamma_{t,h+1:H}^{(c)} \left(\varepsilon_{t,h-1}^{(c)}(Z_{t,h}^{(c)}) + \varphi_{t,h-1}^{(c)}\right)\Big\rangle\Big] \\
&\quad + \mathbb{E}\left[\Big\| \sum_{h=1}^{H} \eta_t \Gamma_{t,h+1:H}^{(c)} \left(\varepsilon_{t,h-1}^{(c)}(Z_{t,h}^{(c)}) + \varphi_{t,h-1}^{(c)}\right)\Big\|^2\right] \\
&\leq \underbrace{\left[\|\Gamma_{t,1:H}^{(c)}\left(\theta_t^{(1)} - \theta_\star\right)\|^2\right]}_{\mathbf{T_1}} + \underbrace{2\eta_t\Big\langle \Gamma_{t,1:H}^{(c)}(\theta_t - \theta_\star)\,,\, \mathbb{E}\Big[\sum_{h=1}^{H} \Gamma_{t,h+1:H}^{(c)} \varepsilon_{t,h-1}^{(c)}(Z_{t,h}^{(c)})\Big]\Big\rangle}_{\mathbf{T_2}} \\
&\quad + \underbrace{2\eta_t\Big\langle \Gamma_{t,1:H}^{(c)}(\theta_t - \theta_\star)\,,\, \mathbb{E}\Big[\sum_{h=1}^{H} \Gamma_{t,h+1:H}^{(c)} \varphi_{t,h-1}^{(c)}\Big]\Big\rangle}_{\mathbf{T_3}} \\
&\quad + \underbrace{2\eta_t^2\mathbb{E}\left[\Big\| \sum_{h=1}^{H} \Gamma_{t,h+1:H}^{(c)} \varepsilon_{t,h-1}^{(c)}(Z_{t,h}^{(c)})\Big\|^2\right]}_{\mathbf{T_4}} + \underbrace{2\eta_t^2\mathbb{E}\left[\Big\| \sum_{h=1}^{H} \eta_t \Gamma_{t,h+1:H}^{(c)} \varphi_{t,h-1}^{(c)}\Big\|^2\right]}_{\mathbf{T_5}} \quad .
\end{aligned}
$$

In this decomposition, $\mathbf{T_1}$ is an optimization error, representing progress towards the solution $\theta_\star$; $\mathbf{T_2}$ is a sampling bias error due to the Markovian noise; $\mathbf{T_3}$ and $\mathbf{T_5}$ are error terms due to sub-optimality of the current policy; and $\mathbf{T_4}$ is a variance term. Next, we bound each term of this decomposition.

Bound on $\mathbf{T_1}$. The first term is the contraction term, and can be bounded using Proposition C.2

$$
\mathbf{T_1} \leq (1 - \eta_t a)^H \|\theta_t - \theta_\star\|^2 \quad . \tag{35}
$$

Bound on $\mathbf{T_2}$. We bound $\mathbf{T_2}$ using Cauchy-Schwarz inequality and Young's inequality

$$
\begin{aligned}
\mathbf{T_2} &\leq 2\eta_t \|\Gamma_{1:H}^{(1)}\| \|\theta_t - \theta_\star\| \Big\|\mathbb{E}\Big[\sum_{h=1}^{H} \Gamma_{h+1:H}^{(1)} \varepsilon^{(1)}(Y_{t,h}^{(1)})\Big]\Big\| \\
&\leq \frac{\eta_t H a}{8}\|\theta_t - \theta_\star\|^2 + \frac{8\eta_t}{Ha}\Big\|\mathbb{E}\Big[\sum_{h=1}^{H} \Gamma_{h+1:H}^{(1)} \varepsilon^{(1)}(Y_{t,h}^{(1)})\Big]\Big\|^2 \quad ,
\end{aligned}
$$

where we also used the fact that $\|\Gamma_{1:H}^{(1)}\| \leq 1$. Then, by Lemma C.3, we have

$$
\begin{aligned}
\mathbf{T_2} &\leq \frac{\eta_t H a}{8}\|\theta_t - \theta_\star\|^2 + \frac{8\eta_t}{Ha}\Big(\frac{8\mathrm{G}}{3}\big(\eta_t \mathrm{C}_A H \tau_{\mathrm{mix}} + \delta \tau_{\mathrm{mix}}\big)\Big)^2 \\
&\leq \frac{\eta_t H a}{8}\|\theta_t - \theta_\star\|^2 + \frac{58\eta_t^3 H \tau_{\mathrm{mix}}^2 \mathrm{C}_A^2 \mathrm{G}^2}{a} + \frac{58\delta\eta_t \mathrm{G}^2 \tau_{\mathrm{mix}}^2}{Ha} \quad .
\end{aligned} \tag{36}
$$

Bound on $\mathbf{T_3}$. The term $\mathbf{T_3}$ can be bounded using Cauchy-Schwarz inequality and Corollary B.2,

$$
\mathbf{T_3} \leq 2\eta_t \|\Gamma_{t,1:H}^{(c)}\| \|\theta_t - \theta_\star\| \Big\|\mathbb{E}\Big[\sum_{h=1}^{H} \Gamma_{t,h+1:H}^{(c)} \varphi_{t,h-1}^{(c)}\Big]\Big\| \leq 2\eta_t H \mathrm{C}_\mu \mathrm{G} \|\theta_t - \theta_\star\|^2 \quad , \tag{37}
$$

where we also used $\|\mathbb{E}\varphi_{t,h}^{(1)}\| \leq \|(\bar{\mathbf{A}}^{(1)}(\theta_t) - \bar{\mathbf{A}}^{(1)}(\theta_\star))\|\widetilde{\mathrm{C}}_{\mathrm{proj}} + \|\bar{\mathbf{b}}^{(1)}(\theta_t) - \bar{\mathbf{b}}^{(1)}(\theta_\star)\| \leq \mathrm{C}_\mu \mathrm{G}\|\theta_t - \theta_\star\|$.

Bound on $\mathbf{T_4}$. The term $\mathbf{T_4}$ can be bounded using Lemma C.4 with $N = 1$, which gives

$$
\mathbf{T_4} \leq 136\eta_t^2 H \tau_{\mathrm{mix}} \mathrm{G}^2 + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\mathrm{mix}}^2 \quad . \tag{38}
$$

Bound on $\mathbf{T_5}$. Finally, the last bound is also a consequence of Corollary B.2, which gives

$$\mathbf{T_5} \leq 2\eta_t^2 H \sum_{h=1}^{H} \mathbb{E}\left[\left\|\Gamma_{h+1:H}^{(1)} \varphi_{t,h-1}^{(1)}\right\|^2\right] \leq 8\eta_t^2 H^2 \mathrm{G}^2 \mathrm{C}_\mu^2 \|\theta_t - \theta_\star\|^2 \ . \tag{39}$$

Full error bound. Plugging (35), (36), (37), (38) and (39) in the above decomposition, we have

$$\begin{aligned}\mathbb{E}[\|\theta_{t,H}^{(1)} - \theta_\star\|^2] \leq (1-\eta a)^H \|\theta_t - \theta_\star\|^2 + \frac{\eta_t H a}{8}\|\theta_t - \theta_\star\|^2 + \frac{58\eta_t^3 H \tau_{\mathrm{mix}}^2 \mathrm{C}_A^2 \mathrm{G}^2}{a} + \frac{58\delta\eta_t \mathrm{G}^2 \tau_{\mathrm{mix}}^2}{Ha} \\ + 2\eta_t H \mathrm{C}_\mu \mathrm{G}\|\theta_t - \theta_\star\|^2 + 136\eta_t^2 H \tau_{\mathrm{mix}} \mathrm{G}^2 + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\mathrm{mix}}^2 + 8\eta_t^2 H^2 \mathrm{G}^2 \mathrm{C}_\mu^2 \|\theta_t - \theta_\star\|^2 \ .\end{aligned}$$

After reorganizing the terms, we obtain

$$\begin{aligned}\mathbb{E}[\|\theta_{t,H}^{(1)} - \theta_\star\|^2] \leq \left(1 - \eta a H/2 + \eta_t H a/8 + 2\eta_t H \mathrm{C}_\mu \mathrm{G} + 8\eta_t^2 H^2 \mathrm{G}^2 \mathrm{C}_\mu^2\right)\|\theta_t - \theta_\star\|^2 \\ + 136\eta_t^2 H \tau_{\mathrm{mix}} \mathrm{G}^2 + \frac{58\delta\eta_t \mathrm{G}^2 \tau_{\mathrm{mix}}^2}{Ha} + \frac{58\eta_t^3 H \tau_{\mathrm{mix}}^2 \mathrm{C}_A^2 \mathrm{G}^2}{a} + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\mathrm{mix}}^2 \ .\end{aligned}$$

Finally, remark that $16\mathrm{G}\mathrm{C}_\mu \leq a$, which implies that

$$1 - \eta a H/2 + \eta_t a H/8 + 2\eta_t H \mathrm{C}_\mu \mathrm{G} + 8\eta_t^2 H^2 \mathrm{G}^2 \mathrm{C}_\mu^2 \leq 1 - \eta a H/4 \ ,$$

and the result of the lemma follows by bounding $\frac{58\eta_t^3 H \tau_{\mathrm{mix}}^2 \mathrm{G}^2 \mathrm{C}_A^2}{a} + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\mathrm{mix}}^2 \leq \frac{976\eta_t^3 \mathrm{G}^2 \mathrm{C}_A^2 H \tau_{\mathrm{mix}}^2}{a}$. □

Theorem 4.2 (Restated). *Assume A1–5. Assume that the step size $\eta_t = \eta$ is constant and satisfies $\eta H \mathrm{C}_A \leq 1/5$ and that $H \geq \tau_{\mathrm{mix}}$. Then it holds that*

$$\begin{aligned}\mathbb{E}[\|\theta_T - \theta_\star\|^2] \leq (1 - \tfrac{\eta a H}{4})^T \|\theta_0 - \theta_\star\|^2 + \tfrac{544\eta\tau_{\mathrm{mix}}\mathrm{G}^2}{a} \\ + \delta\tfrac{232\tau_{\mathrm{mix}}^2\mathrm{G}^2}{H^2a^2} + \tfrac{3904\eta^2\tau_{\mathrm{mix}}^2\mathrm{G}^2\mathrm{C}_A^2}{a^2} \ ,\end{aligned}$$

where δ is defined in Lemma 4.1.

Proof. Since projections on convex sets are contractions, and $\theta_\star$ is within the set on which we project, we have

$$\|\theta_{t+1} - \theta_\star\|^2 \leq \|\bar{\theta}_{t+1} - \theta_\star\|^2 \ .$$

Applying Lemma 4.1 and unrolling the recursion gives the result. □

We now prove the corollary for SARSA's sample complexity.

Corollary D.2 (Restated). *Assume A1–5. Let $\epsilon > 0$, set $\eta \approx \min\left(\frac{1}{\mathrm{C}_A}, \frac{a\epsilon^2}{\mathrm{G}^2\tau_{\mathrm{mix}}}, \frac{a\epsilon}{\mathrm{G}\mathrm{C}_A\tau_{\mathrm{mix}}}\right)$ and $H \approx \max\left(1, \frac{\mathrm{G}\tau_{\mathrm{mix}}}{a\epsilon}\right)$, then SARSA reaches $\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \epsilon^2$ with*

$$TH \approx \max\left(\tfrac{\mathrm{C}_A}{a}, \tfrac{\mathrm{G}^2\tau_{\mathrm{mix}}}{a^2\epsilon^2}, \tfrac{\mathrm{C}_A\mathrm{G}\tau_{\mathrm{mix}}}{a^2\epsilon}\right)\log\left(\tfrac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$$

samples and $T \gtrsim \frac{\mathrm{C}_A}{a}\log\left(\frac{\|\theta_0 - \theta_\star\|^2}{\epsilon}\right)$ policy updates.

Proof. Let $\epsilon > 0$. From Theorem 4.2, we have

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \leq (1 - \tfrac{\eta a H}{4})^T \|\theta_0 - \theta_\star\|^2 + \frac{544\eta\tau_{\mathrm{mix}}\mathrm{G}^2}{a} + \delta\tfrac{232\tau_{\mathrm{mix}}^2\mathrm{G}^2}{H^2a^2} + \tfrac{3904\eta^2\tau_{\mathrm{mix}}^2\mathrm{G}^2\mathrm{C}_A^2}{a^2} \ ,$$

To obtain an overall mean squared error smaller than ϵ^2, each term has to be smaller than ϵ^2, which gives the conditions of η and H. The bounds on T and TH follow from bounding the exponentially decreasing term.

□

E. Proofs for federated SARSA

E.1. Convergence of FedSARSA— Proofs of Proposition 5.1 and Proposition 5.2

Proposition E.1 (Restated). *Assume A1–5. There exists a unique parameter $\theta_\star \in \mathcal{W}$ such that*

$$\tfrac{1}{N}\sum_{c=1}^N \bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star + \tfrac{1}{N}\sum_{c=1}^N \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \ .$$

Proof. The proof follows ideas similar to De Farias & Van Roy (2000)'s Theorem 5.1. For $c \in \{1, \dots, N\}$, we define the function $s^{(c)}$ as

$$s^{(c)}(\theta) = \bar{\mathbf{A}}^{(c)}(\theta)\theta + \bar{\mathbf{b}}^{(c)}(\theta) \ ,$$

as well as the map $F_\eta^{(c)} : \theta \mapsto \theta + \eta s^{(c)}(\theta)$, with $\eta > 0$. We show that $\frac{1}{N}\sum_{c=1}^N F_\eta^{(c)}$ has a fixed point.

Remark that $\frac{1}{N}\sum_{c=1}^N F_\eta^{(c)}$ is a continuous function, and that the set $\mathcal{C} := \{\theta \mid \|\theta\| \leq \frac{(1+\xi)\bar{R}}{1-\xi}\}$ is closed under $\frac{1}{N}\sum_{c=1}^N F_\eta^{(c)}$ for a well-chosen $\xi \in (0,1)$ (the same as for the individual $F_\eta^{(c)}$, whose explicit expression is provided in De Farias & Van Roy (2000)). Indeed, we have that

$$\left\| \frac{1}{N}\sum_{c=1}^N F_\eta^{(c)}(\theta) \right\| \leq \frac{1}{N}\sum_{c=1}^N \|F_\eta^{(c)}(\theta)\| \leq \frac{1}{N}\sum_{c=1}^N (\xi\|\theta\| + (1+\xi)\bar{R}) = \xi\|\theta\| + (1+\xi)\bar{R} \ .$$

By Brouwer's fixed-point theorem, this gives the existence of some $\theta_\star \in \mathbb{R}^d$ such that:

$$\begin{aligned}\frac{1}{N}\sum_{c=1}^N F_\eta^{(c)}\theta_\star = \theta_\star &\iff \frac{1}{N}\sum_{c=1}^N s^{(c)}(\theta_\star) = 0 \\ &\iff \frac{1}{N}\sum_{c=1}^N \bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star + \frac{1}{N}\sum_{c=1}^N \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \ ,\end{aligned}$$

which proves the existence of a fixed point. To prove unicity, let $\theta_*^1, \theta_*^2 \in \mathcal{W}$ be two fixed points of FedSARSA. Then

$$\theta_*^1 - \theta_*^2 = \theta_*^1 - \theta_*^2 - \frac{\eta}{N}\sum_{c=1}^N \bar{\mathbf{A}}^{(c)}(\theta_*^1)\theta_*^1 - \bar{\mathbf{A}}^{(c)}(\theta_*^2)\theta_*^2 + \bar{\mathbf{b}}^{(c)}(\theta_*^1) - \bar{\mathbf{b}}^{(c)}(\theta_*^2) \ .$$

This yields, denoting $\bar{\mathbf{A}} = 1/N\sum_c \bar{\mathbf{A}}^{(c)}$,

$$\theta_*^1 - \theta_*^2 = (I - \eta\bar{\mathbf{A}}(\theta_*^1))(\theta_*^1 - \theta_*^2) - \frac{\eta}{N}\sum_{c=1}^N (\bar{\mathbf{A}}^{(c)}(\theta_*^1) - \bar{\mathbf{A}}^{(c)}(\theta_*^2))\theta_*^2 + \bar{\mathbf{b}}^{(c)}(\theta_*^1) - \bar{\mathbf{b}}^{(c)}(\theta_*^2) \ .$$

Taking the norm, using triangle inequality, using Corollary B.2 to bound $\|\bar{\mathbf{A}}^{(c)}(\theta_*^1) - \bar{\mathbf{A}}^{(c)}(\theta_*^2)\| \leq \mathrm{C}_\mu \mathrm{C}_A$, $\|\bar{\mathbf{b}}^{(c)}(\theta_*^1) - \bar{\mathbf{b}}^{(c)}(\theta_*^2)\| \leq \mathrm{C}_\mu \mathrm{C}_b$, and using the bound $\|\theta_*^2\| \leq \mathrm{C}_{\text{proj}}$, we obtain

$$\|\theta_*^1 - \theta_*^2\| \leq (1 - \eta a + 2\eta \mathrm{G}\mathrm{C}_\mu)\|\theta_*^1 - \theta_*^2\| \leq (1 - \eta a/2)\|\theta_*^1 - \theta_*^2\|$$

where the first inequality comes from Lipschitzness of the policy improvement and the second one comes from A5. This proves that $\|\theta_*^1 - \theta_*^2\| = 0$, guaranteeing the uniqueness of FedSARSA's limit point.

□

Proposition E.2 (Restated). *Assume A1–5. For any $c \in \{1, \dots, N\}$, assume that $\theta_\star^{(c)} \in \mathcal{W}$, then the local optimum $\theta_\star^{(c)}$ (defined analogously to (6)) satisfies*

$$\|\theta_\star^{(c)} - \theta_\star\| \leq \frac{480}{79}(1 + \tau_{\text{mix}})(\epsilon_p\|\theta_\star\| + \epsilon_r) \ ,$$

where ϵ_p and ϵ_r are defined in (1).

Proof. From Proposition 5.3, we obtain

$$\|\bar{\mathbf{A}}^{(c)}(\theta_\star)(\vartheta_\star^{(c)} - \theta_\star)\| \leq \zeta_{\theta_\star} \ .$$

Using A2, we then obtain

$$\|\vartheta_\star^{(c)} - \theta_\star\| \leq \frac{\zeta_{\theta_\star}}{a} \ .$$

Now we need to bound $\|\vartheta_\star^{(c)} - \theta_\star^{(c)}\|$. They are respectively defined by:

$$\bar{\mathbf{A}}^{(c)}(\theta_\star)\vartheta_\star^{(c)} + \bar{\mathbf{b}}^{(c)}(\theta_\star) = 0,$$
$$\bar{\mathbf{A}}^{(c)}(\theta_\star^{(c)})\theta_\star^{(c)} + \bar{\mathbf{b}}^{(c)}(\theta_\star^{(c)}) = 0 \ .$$

Subtracting the equalities we obtain

$$\left(\bar{\mathbf{A}}^{(c)}(\theta_\star^{(c)}) - \bar{\mathbf{A}}^{(c)}(\theta_\star)\right)\theta_\star^{(c)} + \bar{\mathbf{A}}^{(c)}(\theta_\star)\left(\theta_\star^{(c)} - \vartheta_\star^{(c)}\right) = \bar{\mathbf{b}}^{(c)}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star^{(c)}) \ .$$

Using A2 a second time yields

$$\|\theta_\star^{(c)} - \vartheta_\star^{(c)}\| \leq \frac{1}{a}\left(\left\|\left(\bar{\mathbf{A}}^{(c)}(\theta_\star^{(c)}) - \bar{\mathbf{A}}^{(c)}(\theta_\star)\right)\theta_\star^{(c)}\right\| + \left\|\bar{\mathbf{b}}^{(c)}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star^{(c)})\right\|\right) \ .$$

Then, Corollary B.2, gives

$$\|\theta_\star^{(c)} - \vartheta_\star^{(c)}\| \leq \frac{1}{a}\left(\mathrm{C}_\mu\mathrm{C}_A\|\theta_\star^{(c)}\| + \mathrm{C}_\mu\mathrm{C}_b\right)\|\theta_\star^{(c)} - \theta_\star\| \, .$$

Using the triangle inequality, we obtain:

$$\|\theta_\star^{(c)} - \theta_\star\| \leq \frac{\zeta_{\theta_\star}}{a} + \frac{1}{a}\left(\mathrm{C}_\mu\mathrm{C}_A\|\theta_\star^{(c)}\| + \mathrm{C}_\mu\mathrm{C}_b\right)\|\theta_\star^{(c)} - \theta_\star\| \, .$$

Since $\|\theta_\star^{(c)}\| \leq \mathrm{C}_{\text{proj}}$ and using A5, we have

$$\mathrm{C}_{\text{proj}}\mathrm{C}_\mu\mathrm{C}_A + \mathrm{C}_\mu\mathrm{C}_b \leq a/80,$$

then the result follows from the definition of $\zeta_{\theta_\star}$. □

E.2. Bound on heterogeneity drift — Proofs of Proposition 5.3

Proposition E.3 (Restated). *Assume A1, A2, A4, and A5. For $c \in \{1, \dots, N\}$, there exists $\zeta_{\bar{\mathbf{A}}}, \zeta_{\theta_\star} \geq 0$ such that*

$$\|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|^2 \leq \zeta_{\bar{\mathbf{A}}}^2 \ , \quad \|\bar{\mathbf{A}}^{(c)}(\theta_\star)(\vartheta_\star^{(c)} - \theta_\star)\|^2 \leq \zeta_{\theta_\star}^2 \ ,$$

where we introduced the constants $\zeta_{\bar{\mathbf{A}}} \triangleq 4\mathrm{C}_A(1 + \tau_{\text{mix}})\epsilon_p$ and $\zeta_{\theta_\star} \triangleq 6(1 + \tau_{\text{mix}})(\epsilon_p\|\theta_\star\| + \epsilon_r)$.

Proof. This proofs is inspired by the proof of Zhang et al. (2024)'s Theorem 1.

(Heterogeneity of the $\bar{\mathbf{A}}^{(c)}$'s.) By Mitrophanov (2005)'s Theorem 3.1, we have

$$\|\mu_{\theta_\star}^{(c)} - \mu_{\theta_\star}^{(c)}\|_{\text{TV}} \leq 4(1 + \tau_{\text{mix}}) \sup_{\varrho \sim \mathcal{P}(\mathcal{S})} \|\varrho P_{\theta_\star}^{(c)} - \varrho P_{\theta_\star}^{(c')}\|_{TV} \ , \tag{40}$$

Thus, we have, for any $c \in \{1, \dots, N\}$

$$\frac{1}{N}\sum_{c'=1}^{N} \|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}^{(c')}(\theta_\star)\|^2 \leq \frac{\mathrm{C}_A^2}{N}\sum_{c'=1}^{N} \|\mu_{\theta_\star}^{(c)} - \mu_{\theta_\star}^{(c')}\|_{\text{TV}}^2 \leq 16\mathrm{C}_A^2(1 + \tau_{\text{mix}})^2\epsilon_p^2 \ .$$

(Heterogeneity of the $\vartheta_\star^{(c)}$'s.) By definition of $\theta_\star$ and $\vartheta_\star^{(c)}$, we have $\bar{\mathbf{A}}(\theta_\star)\theta_\star = \bar{\mathbf{b}}(\theta_\star)$ and $\bar{\mathbf{A}}^{(c)}(\theta_\star)\vartheta_\star^{(c)} = \bar{\mathbf{b}}^{(c)}(\theta_\star)$. This gives the identity

$$\bar{\mathbf{A}}(\theta_\star)\theta_\star - \bar{\mathbf{A}}^{(c)}(\theta_\star)\vartheta_\star^{(c)} = \bar{\mathbf{b}}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star) \ .$$

Adding and subtracting $\bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star$ on the left side, we obtain

$$(\bar{\mathbf{A}}(\theta_\star) - \bar{\mathbf{A}}^{(c)}(\theta_\star))\theta_\star + \bar{\mathbf{A}}^{(c)}(\theta_\star)(\theta_\star - \vartheta_\star^{(c)}) = \bar{\mathbf{b}}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star) \ .$$

Reorganizing the terms, we obtain

$$\bar{\mathbf{A}}^{(c)}(\theta_\star)(\theta_\star - \vartheta_\star^{(c)}) = (\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star))\theta_\star + \bar{\mathbf{b}}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star) \ ,$$

which gives, by taking the norm, using the triangle inequality and A2, that

$$\|\bar{\mathbf{A}}^{(c)}(\theta_\star)(\theta_\star - \vartheta_\star^{(c)})\| \leq \|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|\|\theta_\star\| + \|\bar{\mathbf{b}}(\theta_\star) - \bar{\mathbf{b}}^{(c)}(\theta_\star)\| \ .$$

Averaging over all $c \in \{1, \dots, N\}$ and using (40), we obtain

$$\frac{1}{N}\sum_{c=1}^{N} \|\theta_\star - \vartheta_\star^{(c)}\|^2 \leq 32\mathrm{C}_A^2(1+\tau_{\mathrm{mix}})^2\epsilon_p^2\|\theta_\star\|^2 + 32\mathrm{C}_b^2(1+\tau_{\mathrm{mix}})^2\epsilon_r^2 \ ,$$

and the result follows. □

Lemma E.4. *Let η_t such that $\eta_t H \mathrm{C}_A \leq 1$. Then, it holds that*

$$\|\Delta_{1:H}\| \leq 2\eta_t \mathrm{C}_A \widetilde{\mathrm{C}}_{\mathrm{proj}} \ .$$

Proof. We have

$$\begin{aligned}\|\Delta_{1:H}\| &= \Big\| \frac{1}{N}\sum_{c=1}^{N} \left(\mathrm{I} - \Gamma_{t,1:H}^{(c)}\right)\left(\vartheta_\star^{(c)} - \theta_\star\right) \Big\| \\ &\leq 2\|\mathrm{I} - \Gamma_{t,1:H}^{(c)}\|\widetilde{\mathrm{C}}_{\mathrm{proj}} \ ,\end{aligned}$$

where we used $\|\theta_\star\| \leq \widetilde{\mathrm{C}}_{\mathrm{proj}}$ and $\|\vartheta_\star^{(c)}\| \leq \widetilde{\mathrm{C}}_{\mathrm{proj}}$. The result comes from $\|\mathrm{I} - \Gamma_{t,1:H}^{(c)}\| \leq \eta_t H \mathrm{C}_A$. □

Lemma E.5. *Let η_t such that $\eta_t H \mathrm{C}_A \leq 1$. Then, it holds that*

$$\|\Delta_{1:H}\|^2 \leq \frac{\eta_t^4 H^2 (H-1)^2}{4}\zeta_{\bar{\mathbf{A}}}^2\zeta_{\theta_\star}^2 \ .$$

Proof. Using Lemma B.3, we have

$$\frac{1}{N}\sum_{c=1}^{N}\left(\mathrm{I} - \Gamma_{t,1:H}^{(c)}\right)\left(\vartheta_\star^{(c)} - \theta_\star\right) = -\frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma_{t,h+1:H}^{(c)}\bar{\mathbf{A}}^{(c)}\left(\vartheta_\star^{(c)} - \theta_\star\right) \ .$$

Then, we remark that

$$\sum_{c=1}^{N}\bar{\mathbf{A}}^{(c)}(\theta_\star)\vartheta_\star^{(c)} - \sum_{c=1}^{N}\bar{\mathbf{A}}^{(c)}(\theta_\star)\theta_\star = \sum_{c=1}^{N}\bar{\mathbf{b}}^{(c)}(\theta_\star) - \sum_{c=1}^{N}\bar{\mathbf{b}}^{(c)}(\theta_\star) = 0 \ .$$

Thus, using the notation $\Gamma_{h+1:H}^{(\mathrm{avg})} = \frac{1}{N}\sum_{c=1}^{N}\Gamma_{t,h+1:H}^{(c)}$, we have

$$\Delta_{1:H} = \frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\left(\bar{\Gamma}_{t,h+1:H}^{(c)} - \bar{\Gamma}_{t,h+1:H}^{(\mathrm{avg})}\right)\bar{\mathbf{A}}^{(c)}(\theta_\star)\left(\vartheta_\star^{(c)} - \theta_\star\right)$$

$$= \frac{\eta_t^2}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \bar{\Gamma}^{(c)}_{t,1:\ell-1} (\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)) \bar{\Gamma}^{(\mathrm{avg})}_{t,\ell+h+1:H} \bar{\mathbf{A}}^{(c)}(\theta_\star) \left(\vartheta_\star^{(c)} - \theta_\star\right) \ .$$

Consequently, we have, by the triangle inequality,

$$\begin{aligned}
\|\Delta_{1:H}\| &\leq \frac{\eta_t^2}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \|\bar{\Gamma}^{(c)}_{t,1:\ell-1} (\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)) \bar{\Gamma}^{(\mathrm{avg})}_{t,\ell+h+1:H} \bar{\mathbf{A}}^{(c)}(\theta_\star) \left(\vartheta_\star^{(c)} - \theta_\star\right) \| \\
&\leq \frac{\eta_t^2}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \sum_{\ell=1}^{h-1} \|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|^2 \|\bar{\mathbf{A}}^{(c)}(\theta_\star) \left(\vartheta_\star^{(c)} - \theta_\star\right) \| \\
&= \frac{\eta_t^2 H(H-1)}{2N} \sum_{c=1}^{N} \|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|^2 \|\bar{\mathbf{A}}^{(c)}(\theta_\star) \left(\vartheta_\star^{(c)} - \theta_\star\right) \| \ .
\end{aligned}$$

Using Cauchy-Schwarz inequality, we obtain

$$\|\Delta_{1:H}\|^2 \leq \frac{\eta_t^4 H^2 (H-1)^2}{4} \Big(\frac{1}{N} \sum_{c=1}^{N} \|\bar{\mathbf{A}}^{(c)}(\theta_\star) - \bar{\mathbf{A}}(\theta_\star)\|^2 \Big) \Big(\frac{1}{N} \sum_{c=1}^{N} \|\bar{\mathbf{A}}^{(c)}(\theta_\star) \left(\vartheta_\star^{(c)} - \theta_\star\right) \|^2 \Big) \ ,$$

which is the result. □

E.3. Convergence rate — Proofs of Lemma 5.4 and Theorem 5.5

Lemma E.6 (Restated). *Assume A1, A2, A4, and A5. Let $t \geq 0$, assume that the step size satisfies $\eta_t H \mathrm{C}_A \leq 1/6$. Then, it holds that,*

$$\begin{aligned}
\mathbb{E}[\|\bar{\theta}_{t+1} - \theta_\star\|^2] \leq (1 - \tfrac{\eta_t a H}{8}) \|\theta_t - \theta_\star\|^2 + \tfrac{9\eta_t^3 H(H-1)^2}{a} \zeta_{\bar{\mathbf{A}}}^2 \zeta_{\theta_\star}^2 \\
+ \tfrac{136 \eta_t^2 H \tau_{\mathrm{mix}} \mathrm{G}^2}{N} + \tfrac{58 \delta \eta_t \mathrm{G}^2 \tau_{\mathrm{mix}}^2}{Ha} + \tfrac{976 \eta_t^3 \mathrm{G}^2 \mathrm{C}_A^2 H \tau_{\mathrm{mix}}^2}{a} \ ,
\end{aligned}$$

where $\delta = 0$ if the $Z_{t,0}^{(c)}$ are sampled from the stationary distribution $\nu_{\theta_t}^{(c)}$ and $\delta = 1$ otherwise.

Proof. Starting from Claim 2, and expanding the square, we have

$$\begin{aligned}
\mathbb{E}\Big[\|\bar{\theta}_{t+1} - \theta_\star\|^2\Big] &= \Big\| \frac{1}{N} \sum_{c=1}^{N} \Gamma^{(c)}_{t,1:H} \left(\theta_t - \theta_\star\right) \Big\|^2 \\
&+ 2 \Big\langle \frac{1}{N} \sum_{c=1}^{N} \Gamma^{(c)}_{t,1:H} \left(\theta_t - \theta_\star\right) \ , \ \Delta_{1:H} + \frac{\eta_t}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \Gamma^{(c)}_{t,h+1:H} \mathbb{E}\Big[\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h}) + \varphi^{(c)}_{t,h-1} \Big| \mathcal{F}_t\Big] \Big\rangle \\
&+ \mathbb{E}\Big[\Big\| \Delta_{1:H} + \frac{\eta_t}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \Gamma^{(c)}_{t,h+1:H} \Big(\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h}) + \varphi^{(c)}_{t,h-1}\Big) \Big\|^2 \Big| \mathcal{F}_t \Big] \ .
\end{aligned}$$

This gives the decomposition

$$\begin{aligned}
&\mathbb{E}\Big[\|\bar{\theta}_{t+1} - \theta_\star\|^2\Big] \\
&= \underbrace{\Big\| \frac{1}{N} \sum_{c=1}^{N} \Gamma^{(c)}_{t,1:H} (\theta_t - \theta_\star) \Big\|^2}_{\mathbf{U_1}} + \underbrace{2 \Big\langle \frac{1}{N} \sum_{c=1}^{N} \Gamma^{(c)}_{t,1:H} (\theta_t - \theta_\star), \ \frac{\eta_t}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \Gamma^{(c)}_{t,h+1:H} \mathbb{E}\Big[\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h}) \Big| \mathcal{F}_t\Big] \Big\rangle}_{\mathbf{U_2}} \\
&+ \underbrace{2 \Big\langle \frac{1}{N} \sum_{c=1}^{N} \Gamma^{(c)}_{t,1:H} \left(\theta_t - \theta_\star\right) \ , \ \frac{\eta_t}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \Gamma^{(c)}_{t,h+1:H} \mathbb{E}\Big[\varphi^{(c)}_{t,h-1} \Big| \mathcal{F}_t\Big] \Big\rangle}_{\mathbf{U_3}}
\end{aligned}$$

$$+ \underbrace{3\mathbb{E}\Big[\Big\|\frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h})\Big\|^2\Big|\mathcal{F}_t\Big]}_{\mathbf{U_4}} + \underbrace{3\mathbb{E}\Big[\Big\|\frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\varphi^{(c)}_{t,h-1}\Big\|^2\Big|\mathcal{F}_t\Big]}_{\mathbf{U_5}}$$

$$+ \underbrace{3\|\Delta_{1:H}\|^2 + 2\Big\langle\frac{1}{N}\sum_{c=1}^{N}\Gamma^{(c)}_{t,1:H}\left(\theta_t-\theta_\star\right),\Delta_{1:H}\Big\rangle}_{\mathbf{U_6}} \ .$$

The terms $\mathbf{U_1}$ to $\mathbf{U_5}$ are analogous to the terms $\mathbf{T_1}$ to $\mathbf{T_5}$ in the single-agent setting, although with a variance term $\mathbf{U_4}$ whose leading term will scale in $1/N$. The term $\mathbf{U_6}$ is due to heterogeneity, and accounts for the differences in local updates from one agent to another.

Bound on $\mathbf{U_1}$. We have, using Jensen's inequality and Proposition C.2,

$$\mathbf{U_1} \leq \frac{1}{N}\sum_{c=1}^{N}\|\Gamma^{(c)}_{t,1:H}\left(\theta_t-\theta_\star\right)\|^2 \leq (1-\eta a)^H\|\theta_t-\theta_\star\|^2 \ . \tag{41}$$

Bound on $\mathbf{U_2}$. First, we use Lemma A.3 to construct a Markov chain $\{Y^{(c)}_{t,h}\}_{h\geq 0}$ initialized in the stationary distribution alike in the single-agent case. We then have

$$\begin{aligned}\mathbf{U_2} &\leq \frac{\eta_t aH}{8}\Big\|\frac{1}{N}\sum_{c=1}^{N}\Gamma^{(c)}_{t,1:H}\left(\theta_t-\theta_\star\right)\Big\|^2 + \frac{8}{\eta_t aH}\Big\|\frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\mathbb{E}\Big[\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h})\Big]\Big\|^2 \\ &\leq \frac{\eta_t aH}{8}\|\theta_t-\theta_\star\|^2 + \frac{8}{\eta_t aH}\Big\|\frac{\eta_t}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\Gamma^{(c)}_{t,h+1:H}\mathbb{E}\Big[\varepsilon^{(c)}_{t,h-1}(Z^{(c)}_{t,h})\Big]\Big\|^2 \ .\end{aligned}$$

Using Lemma C.3, we obtain

$$\mathbf{U_2} \leq \frac{\eta_t Ha}{8}\|\theta_t-\theta_\star\|^2 + \frac{58\eta_t^3 H\tau^2_{\text{mix}}\mathrm{C}_A^2\mathrm{G}^2}{a} + \frac{58\delta\eta_t\mathrm{G}^2\tau^2_{\text{mix}}}{Ha} \ . \tag{42}$$

Bound on $\mathbf{U_3}$. First, we use the triangle inequality to split the mean. Then, similarly to the bound on $\mathbf{T_3}$, we use Cauchy-Schwarz inequality and Corollary B.2 to obtain

$$\mathbf{U_3} \leq 2\eta_t H\mathrm{C}_\mu\mathrm{G}\|\theta_t-\theta_\star\|^2 \ . \tag{43}$$

Bound on $\mathbf{U_4}$. Using Lemma C.4, we have the bound

$$\mathbf{U_4} \leq \frac{136\eta_t^2 H\tau_{\text{mix}}\mathrm{G}^2}{N} + 5504\eta_t^4\mathrm{C}_A^2\mathrm{G}^2H^2\tau^2_{\text{mix}} \tag{44}$$

Bound on $\mathbf{U_5}$. Alike $\mathbf{T_5}$, the bound on $\mathbf{U_5}$ is a consequence of Corollary B.2,

$$\mathbf{U_5} \leq \frac{3\eta_t^2 H}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\mathbb{E}\Big[\Big\|\Gamma^{(1)}_{h+1:H}\varphi^{(c)}_{t,h-1}\Big\|^2\Big] \leq 12\eta_t^2H^2\mathrm{G}^2\mathrm{C}_\mu^2\|\theta_t-\theta_\star\|^2 \ . \tag{45}$$

Bound on $\mathbf{U_6}$. This term is due to heterogeneity. First, we split the second term using Young's inequality, then use Lemma E.5, which gives the bound

$$\begin{aligned}\mathbf{U_6} &\leq \frac{\eta_t a}{8}\Big\|\frac{1}{N}\sum_{c=1}^{N}\Gamma^{(c)}_{t,1:H}\left(\theta_t-\theta_\star\right)\Big\|^2 + \Big(3+\frac{8}{\eta_t a}\Big)\|\Delta_{1:H}\|^2 \\ &\leq \frac{\eta_t aH}{8}\Big\|\frac{1}{N}\sum_{c=1}^{N}\Gamma^{(c)}_{t,1:H}\left(\theta_t-\theta_\star\right)\Big\|^2 + \Big(3+\frac{8}{\eta_t aH}\Big)\eta_t^4H^2(H-1)^2\zeta^2_{\mathbf{A}}\zeta^2_{\theta_\star} \ ,\end{aligned}$$

which gives, using $3 \leq 1/(\eta_t a H)$,

$$\mathbf{U_6} \leq \frac{\eta_t a H}{8}\|\theta_t - \theta_\star\|^2 + \frac{9\eta_t^3 H^3 (H-1)^2}{a}\zeta_{\mathbf{A}}^2 \zeta_{\theta_\star}^2 \ . \tag{46}$$

Final bound. Plugging (41), (42), (43), (44), (45), and (46) the error decomposition above, we obtain

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_{t+1} - \theta_\star\|^2] \leq (1-\eta a)^H \|\theta_t - \theta_\star\|^2 + \frac{\eta_t H a}{8}\|\theta_t - \theta_\star\|^2 + \frac{58\eta_t^3 H \tau_{\text{mix}}^2 \mathrm{C}_A^2 \mathrm{G}^2}{a} + \frac{58\delta\eta_t \mathrm{G}^2 \tau_{\text{mix}}^2}{Ha} \\ + 2\eta_t H \mathrm{C}_\mu \mathrm{G}\|\theta_t - \theta_\star\|^2 + \frac{136\eta_t^2 H \tau_{\text{mix}} \mathrm{G}^2}{N} + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\text{mix}}^2 \\ + 12\eta_t^2 H^2 \mathrm{G}^2 \mathrm{C}_\mu^2 \|\theta_t - \theta_\star\|^2 + \frac{9\eta_t^3 H^3 (H-1)^2}{a}\zeta_{\mathbf{A}}^2 \zeta_{\theta_\star}^2 \ .\end{aligned}$$

Simplifying, we obtain

$$\begin{aligned}\mathbb{E}\Big[\|\bar{\theta}_{t+1} - \theta_\star\|^2\Big] \leq (1-\eta a/8)^H \|\theta_t - \theta_\star\|^2 + \frac{9\eta_t^3 H (H-1)^2}{a}\zeta_{\mathbf{A}}^2 \zeta_{\theta_\star}^2 + \frac{136\eta_t^2 H \tau_{\text{mix}} \mathrm{G}^2}{N} \\ + \frac{58\eta_t^3 H \tau_{\text{mix}}^2 \mathrm{C}_A^2 \mathrm{G}^2}{a} + \frac{58\delta\eta_t \mathrm{G}^2 \tau_{\text{mix}}^2}{Ha} + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\text{mix}}^2 \ ,\end{aligned}$$

which gives the result of the lemma follows by bounding $\frac{58\eta_t^3 H \tau_{\text{mix}}^2 \mathrm{G}^2 \mathrm{C}_A^2}{a} + 5504\eta_t^4 \mathrm{C}_A^2 \mathrm{G}^2 H^2 \tau_{\text{mix}}^2 \leq \frac{976\eta_t^3 \mathrm{G}^2 \mathrm{C}_A^2 H \tau_{\text{mix}}^2}{a}$. □

Theorem 5.5 (Restated)**.** *Assume A1, A2, A4, and A5, that the step size $\eta_t = \eta$ is constant and satisfies $\eta H \mathrm{C}_A \leq 1/5$ and that $H \geq \tau_{\text{mix}}$. Then it holds that*

$$\begin{aligned}\mathbb{E}[\|\theta_T - \theta_\star\|^2] \leq (1-\tfrac{\eta a H}{8})^T \|\theta_0 - \theta_\star\|^2 + \tfrac{72\eta^2 (H-1)^2}{a^2}\zeta_{\mathbf{A}}^2 \zeta_{\theta_\star}^2 \\ + \tfrac{1088\eta\tau_{\text{mix}}\mathrm{G}^2}{Na} + \tfrac{464\delta \mathrm{G}^2 \tau_{\text{mix}}^2}{H^2 a^2} + \tfrac{7808\eta^2 \mathrm{G}^2 \mathrm{C}_A^2 \tau_{\text{mix}}^2}{a^2} \ ,\end{aligned}$$

where $\delta = 0$ if the $Z_{t,0}^{(c)}$ are sampled from the stationary distribution $\nu_{\theta_t}^{(c)}$ and $\delta = 1$ otherwise.

Proof. The proof follows the same arguments as the proof of Theorem 4.2. □

Corollary E.7 (Restated)**.** *Assume A 1–5. Let $\epsilon > 0$. Set $\eta \approx \min\big(\frac{1}{\mathrm{C}_A}, \frac{Na\epsilon^2}{\mathrm{G}^2\tau_{\text{mix}}}, \frac{a\epsilon}{\mathrm{G}\mathrm{C}_A\tau_{\text{mix}}}\big)$, and H such that $H \lesssim \frac{\tau_{\text{mix}}\mathrm{G}}{\zeta_{\mathbf{A}}\zeta_{\theta_\star}} \max\big(\frac{\mathrm{G}}{N\epsilon}, \mathrm{C}_A\big)$ and $H \gtrsim \frac{\delta\tau_{\text{mix}}\mathrm{G}}{a\epsilon}$, then FedSARSA reaches $\mathbb{E}[\|\theta_T - \theta_\star\|^2] \lesssim \epsilon^2$ with $T \gtrsim \max\big(\frac{\mathrm{C}_A}{a}, \frac{\zeta_{\mathbf{A}}\zeta_{\theta_\star}}{a^2\epsilon}\big)\log\big(\frac{\|\theta_0-\theta_\star\|^2}{\epsilon}\big)$ communications, and*

$$TH \approx \max\big(\tfrac{\mathrm{C}_A}{a}, \tfrac{\mathrm{G}^2\tau_{\text{mix}}}{Na^2\epsilon^2}, \tfrac{\mathrm{G}\mathrm{C}_A\tau_{\text{mix}}}{a^2\epsilon}\big)\log\big(\tfrac{\|\theta_0-\theta_\star\|^2}{\epsilon}\big) \tag{12}$$

samples per agent.

Proof. Let $\epsilon > 0$. From Theorem 5.5, we have

$$\mathbb{E}[\|\theta_T - \theta_\star\|^2] \leq (1 - \tfrac{\eta a H}{8})\|\theta_t - \theta_\star\|^2 + \tfrac{72\eta^2(H-1)^2}{a^2}\zeta_{\mathbf{A}}^2\zeta_{\theta_\star}^2 + \tfrac{1088\eta\tau_{\text{mix}}\mathrm{G}^2}{Na} + \tfrac{464\delta\mathrm{G}^2\tau_{\text{mix}}^2}{H^2a^2} + \tfrac{7808\eta^2\mathrm{G}^2\mathrm{C}_A^2\tau_{\text{mix}}^2}{a^2} \ .$$

To obtain an overall mean squared error smaller than ϵ^2, each term has to be smaller than ϵ^2, which gives the conditions on η and H. For the exponential term to be small, we require $T \approx \frac{1}{\eta H a}\log(\frac{\|\theta_0-\theta_\star\|^2}{\epsilon^2})$ and $TH \approx \frac{1}{\eta a}\log(\frac{\|\theta_0-\theta_\star\|^2}{\epsilon^2})$, which give the bounds on T and TH. □