SINGLE-TIMESCALE PERSONALIZED AVERAGE-REWARD FEDERATED TD LEARNING ACROSS HETEROGENEOUS ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study personalized federated TD learning, in which a collection of agents interact with different environments and jointly learn their respective value functions. Inspired by the recent success of personalized federated learning (PFL), we focus on the setting where there exists a shared linear representation and the agents' optimal weights collectively lie in an unknown linear subspace. Under canonical single-timescale updates with average-reward objectives and Markovian sampling, we propose a natural personalized federated TD algorithm, named **pFedAR-TD**, in which agents jointly estimate a common subspace while updating agent-specific weights and mean rewards. We show that this decomposition can filter out conflicting signals, effectively mitigating the negative impacts of "misaligned" signals arising from environmental heterogeneity, and hence achieve linear speedup. Compared with the discounted-reward setting, the combination of personalization and average-reward objectives significantly complicates the convergence analysis. In particular, the additional need to learn local mean rewards results in three tightly coupled dynamics rather than two, calling for a careful balance between fine-grained characterization of error perturbations and the overall tractability of the system-level analysis. Our analysis exploits geometric properties induced by key algorithmic components such as projection, innovation filtration, and QR decomposition. A large subspace error early in training can amplify perturbations in the coupled iterates and jeopardize convergence. To address this, we further introduce an initialization phase and establish a maximal-deviation bound to keep the coupled iterates in a region where the convergence argument applies. Experiments are provided to show the benefits of learning via a shared structure to the more general control problem.

1 INTRODUCTION

In many real-world applications, ranging from simple assistive robotics to autonomous vehicles, autonomous agents operate in different local environments. For example, consider robot vacuums shown in Fig.1. Their interacting environments are mostly shaped by household layouts, furnishings, and obstacles. Similarly, autonomous vehicles deployed across different regions of a metropolitan area may encounter substantial variability in road conditions and traffic patterns (Huang et al., 2018; FHA, 2020; Caesar et al., 2020). Such environmental heterogeneity can lead to misaligned learning signals across agents, which may significantly hinder convergence and impair joint learning. Nevertheless, applying standard single-agent reinforcement learning (RL) solutions may lead to overall intensive, yet possibly redundant, computation and sample collection. This is because the cumulative knowledge obtained by one agent may partially overlap with others. In Fig. 1, vacuums may still encounter similar objects and face recurring navigation challenges. Likewise, autonomous vehicles in different regions share some features of roads and traffic interactions. These two competing intuitions raise a natural question:

Question: Can agents learn together while adapting to their different environments?

Existing works of Federated RL (FedRL) often overlook either environmental heterogeneity or local environment adapatation in training a shared policy or value-/Q-function across all agents (Jin et al., 2022; Wang et al., 2023; Xie & Song, 2023; Zhang et al., 2023a). Though this approach is practically

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

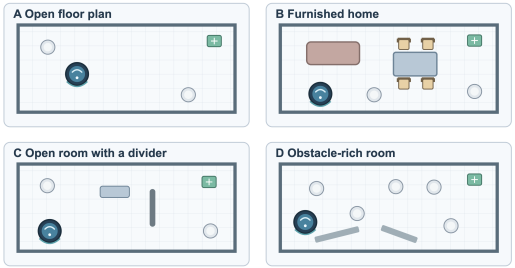


Figure 1: The environments that robot vacuums differ in obstacle shape, layouts, and surface smoothness.

effective, particularly in settings with mild environmental heterogeneity, in many applications (such as on-device recommender systems (Maeng et al., 2022) and smart healthcare (Nguyen et al., 2021)), heterogeneity is often moderate or even severe, which can significantly degrade the performance of a common policy. In a parallel line of research on multi-agent Markov games or competitive MARL, learning personalized (or localized) policies or value-/Q-functions has recently garnered considerable attention (Filar & Vrieze, 2012; Zhang et al., 2018; Qu et al., 2022; Zhang et al., 2023b; Daskalakis et al., 2023). In their setting, agents interact with one another within a shared environment, and the rewards received by individual agents are determined by the joint actions of all agents. It is easy to see that in such a setup, the more agents there are, the more complicated an agent's decision problem becomes. In fact, this line of work often suffers from the so-called curse of multi-agents—the sample complexity increases linearly or even exponentially (Song et al., 2022; Jin et al., 2024; Zhang et al., 2018) with respect to the number of agents.

In this paper, we take a step toward addressing the above question by studying cooperative RL. In particular, we study joint temporal-difference (TD) learning, which enables us to examine how agents can share information across heterogeneous environments while learning their own value functions, without the additional complexity of policy optimization (Sutton & Barto, 2018). We adopt a shared-subspace model studied extensively in personalized federated learning (PFL) and multi-task learning: under a common feature map, agents have distinct value-function weights that lie in a low-dimensional subspace (Tripuraneni et al., 2021; Collins et al., 2021; Thekumparampil et al., 2021; Duchi et al., 2022; Du et al., 2021; Tian et al., 2025; Niu et al., 2024). This established model provides a starting point for understanding how collaboration and personalization interact in federated RL. We propose a personalized federated TD algorithm with *single-timescale* updates and *average-reward* objectives. Our analysis goes beyond combining existing results for federated TD learning and subspace-based PFL. Agents have different transition kernels, so arguments that rely on a common kernel do not directly control their Markovian noise (Woo et al., 2023; Khodadadian et al., 2022); likewise, concentration arguments for fresh i.i.d. samples do not directly apply (Collins et al., 2021; Niu et al., 2024). The shared subspace, local value-function weights, and agent-specific mean rewards must instead be analyzed jointly.

Our contributions can be summarized as follows.

- We propose and analyze a cooperative yet personalized average-reward TD method, named **pFedAR-TD**, in which agents jointly estimate a common subspace while updating agent-specific weights and mean rewards. We introduce a Lyapunov function to capture joint subspace and local head estimation errors, and show that it converges to 0 at a rate $\widetilde{\mathcal{O}}(r/(TK))$, where r is the subspace dimension and K is the number of agents, and $\widetilde{\mathcal{O}}$ hides polylogarithmic factors and problem-dependent constants. Moreover, the reward estimation error decays at a rate of $\widetilde{\mathcal{O}}(1/(LT))$, where L is the length of the local trajectory in a round. A similar algorithm was considered in a recent work (Xiong et al., 2025) but under the simpler *two-timescale* and *discounted reward* setting. Moreover, the resulting rate for the joint estimator error is $\widetilde{\mathcal{O}}((TK)^{-2/3})$, which is significantly slower than our guarantee (Table 1), highlighting the improved efficiency enabled by our design and analysis.
- Our analysis extends beyond existing results in both RL and PFL. We study three coupled estimation processes—the shared subspace, local value-function weights, and local mean rewards—under Markovian sampling and heterogeneous transition kernels. In the PFL analysis of Collins et al. (2021), perturbations can be controlled by increasing the number of fresh samples; that argument does not directly apply here. Instead, we bound how perturbations propagate through the projected updates and QR decomposition, using an innovation filtration to handle the dependent noise. An initialization phase and a maximal-deviation bound then keep the coupled iterates within the region where the convergence analysis applies.
- Another technical novelty is on the convergence of subspace, which does not admit a contraction per se. Unlike classical TD learning that directly tracks parameter error, the subspace error is only

revealed through the local heads. We develop a geometric argument showing that projection onto the constraint set is non-expansive, while basis misalignment is second-order in the subspace error. Crucially, we prove that aggregate local head error is lower bounded by the subspace error, which converts local TD progress into principal angle contraction and yields a unified Lyapunov analysis.

- We validate the rate and the speedup in K on synthetic MDPs, and show on Gym environments that the personalized framework outperforms single-agent, universal-policy and two-timescale approaches in convergence speed and stability.

2 Related Work

Due to space limitations, we focus on FedRL under *heterogeneous environments* in the related work section. Additional references covering the broader scope can be found in Appendix B.

Fed RL under heterogeneous environments. There is an emerging interest in mathematically understanding the role of environmental heterogeneity in the performance of the federated versions of classic reinforcement learning algorithms (Jin et al., 2022; Wang et al., 2023; Xie & Song, 2023; Zhang et al., 2024), yet most existing studies focus on training a common policy or approximating a common function. Jin et al. (2022) studied federated Q-learning and policy gradient methods, assuming known transition kernels. Wang et al. (2025) studied federated Q-learning with unknown transition kernels, and found that, in the presence of environmental heterogeneity, the eventual convergence rate may depend solely on the number of communication rounds – multiple updates within each round cannot accelerate convergence. Wang et al. (2023) proposed FedTD(0) with linear function approximation, and linear convergence speedup in a low-heterogeneity regime. Xie & Song (2023) used the KL-divergence to penalize the deviation of local update from the global policy, and proved that under the setting of heterogeneous environments, the local update is beneficial for global convergence using their method. Yet, no convergence rate is provided. Zhang et al. (2024) proposed FedSARSA and proved that the algorithm can converge to a near-optimal solution. Mangold et al. (2024) proposed SCAFFLSA, which applies SCAFFOLD-style correction to heterogeneous federated SA/TD to mitigate client drift. Labbi et al. (2024) studied a model-based, tabular federated RL method under restricted heterogeneity, without personalization. Beikmohammadi et al. (2025) studied FedTD(0) under model mismatch, relying on the assumption that the averaged local models converge to the true model.

Personalization. Training personalized policies in FedRL in the presence of environment heterogeneity remains largely underexplored. Yang et al. (2024) studied the training of personalized policies in the context where each agent has a private reward function but shares a common transition kernel for the environment. The analysis focused solely on the convergence of the global policy, defined as the average of the local policies. Jin et al. (2022) proposed a heuristic deep FedRL method, where the policy network can be decomposed into a shared subnetwork and environment embedding layer. Xiong et al. (2025) studied personalized TD learning with two timescales and discounted rewards. Furthermore, the resulting rate for the joint estimator error is $\widetilde{\mathcal{O}}((TK)^{-2/3})$, which is significantly slower than our guarantee, highlighting the improved efficiency enabled by our design and analysis.

3 Problem Setup and Preliminaries

Federated Reinforcement Learning. We consider a multi-agent system that consists of one parameter server and K agents. The agents are modelled as Markov Decision Processes (MDP) $\{\mathcal{M}^k \mid \mathcal{M}^k = \langle P^k, \mathcal{S}, \mathcal{A}, R\rangle \text{ for } k = 1, 2, ..., K\}$, where $\mathcal{S}$ is the (possibly infinite) state space, $\mathcal{A}$ is a finite action space, P^k's are the transition kernels, and R is the reward function: $\mathcal{S}\times\mathcal{A} \to [-U_r, U_r]$. The environments P^k with which the agents interact are heterogeneous. The parameter server can orchestrate the learning at the agents through iteratively aggregating the updates at the agents.

	Timescale	Reward	Rate
Xiong et al. (2025)	two	discounted	$\widetilde{\mathcal{O}}\big((TK)^{-2/3}\big)$
pFedAR-TD (ours)	single	average	$\widetilde{\mathcal{O}}\big(\frac{r}{TK}\big)$

Table 1: Comparison with Xiong et al. (2025): K agents, T rounds.

We study average-reward reinforcement learning (Tsitsiklis & Van Roy, 1999), which provides a principled framework for maximizing the mean reward. Even in the single-agent setting, unlike discounted and finite-horizon MDPs, the sample complexity of average-reward settings remains largely unexplored (Jiao et al., 2026).

For a given policy π, we assume that the limiting state distribution induced by the transition kernel P^k and policy π exists for each agent $k \in [K]$; we denote this distribution by μ^k. The time-average expected reward at agent k is defined as:

$$J^k(\pi) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\big[\sum_{t=0}^{T-1} R(s_t^k, a_t^k)\big] = \mathop{\mathbb{E}}_{(s,a)\sim\mu^k\otimes\pi}[R(s,a)], \tag{1}$$

where (s_t^k, a_t^k) is the state-action pair of agent k at iteration t, for $t = 0, 1, \cdots$. The first expectation in Eq. (1) is taken over the state-action trajectories generated under the Markov chain, while the second is taken over their limiting distribution. Here, the notation $\mu^k \otimes \pi$ denotes sampling $s \sim \mu^k$ and $a \sim \pi(\cdot \mid s)$. The value function is defined as

$$V^{k,\pi}(s) = \mathbb{E}\big[\sum_{t=0}^{\infty}\big(R(s_t^k, a_t^k) - J^k(\pi)\big) \mid s_0^k = s\big]. \tag{2}$$

For ease of exposition, we drop the index and function argument in π. That is, for each k, $J^k(\pi)$ is written as J^k, and $V^{k,\pi}(s)$ as $V^k(s)$ for short.

Single-agent TD Learning. Let V^k belong to a function space with feature map $\phi : \mathcal{S} \to L^2(\mathbb{N})$. Denote the coefficients of the value V^k with respect to ϕ as $z^{k,*}$, where $z^{k,*} \in L^2(\mathbb{N})$. For simplicity, we assume ϕ is known and has finite dimension d, i.e., $\phi : \mathcal{S} \to \mathbb{R}^d$. Consequently, $V^k(s) = \langle \phi(s), z^{k,*} \rangle$, $\forall\, s \in \mathcal{S}$, and $k \in [K]$.

We consider the following linear function approximation: $\widehat{V}(s; z^k) = \phi(s)^\top z^k$. To drive the approximation $\widehat{V}(s; z^k)$ to V^k, single-agent TD learning (particularly TD(0)) update is applied to an agent to learn its value function:

$$\begin{aligned} z_{t+1}^k &= z_t^k + \beta\Big(R(s_t^k, a_t^k) - \eta_t^k + \big(\phi(s_{t+1}^k) - \phi(s_t^k)\big)^\top z_t^k\Big)\phi(s_t^k), \\ \eta_{t+1}^k &= \eta_t^k + \gamma\left(R(s_t^k, a_t^k) - \eta_t^k\right), \end{aligned} \tag{3}$$

where β and γ are stepsizes, and η_t^k is the reward estimator. In literature, $\delta_t^k := R(s_t^k, a_t^k) - \eta_t^k + \big(\phi(s_{t+1}^k) - \phi(s_t^k)\big)^\top z_t^k$ is often referred to as TD(0) error. It is well-known that under some standard technical assumptions $z_t^k \to z^{k,*}$ (Tsitsiklis & Van Roy, 1999; Sutton & Barto, 2018).

Federated TD Learning: Commonality and Heterogeneity. In multi-agent TD learning, the K agents aim to collaboratively learn their respective value functions under the given π based on their collective state-action trajectories. Learning with shared representation is widely recognized as an effective way to separate commonalities from heterogeneity across various heterogeneous data (Caruana, 1997; Ando et al., 2005; Bengio et al., 2013; LeCun et al., 2015; Niu et al., 2024).

Inspired by the literature of personalized FL and multi-task learning, following Tripuraneni et al. (2021); Collins et al. (2021); Thekumparampil et al. (2021); Duchi et al. (2022); Du et al. (2021); Tian et al. (2025); Niu et al. (2024), we assume that $\{z^{k,*} : k = 1, \cdots, K\}$ belong to and fully span an r-dimensional subspace of $\mathbb{R}^d$. That is, there exists an orthonormal matrix $\mathbf{B}^* \in \mathbb{R}^{d\times r}$ such that

$$z^{k,*} = \mathbf{B}^*\omega^{k,*}, \quad \text{where } \omega^{k,*} \in \mathbb{R}^r. \tag{4}$$

4 Algorithm: pFedAR-TD

We propose a personalized average-reward TD method, named **pFedAR-TD** based on joint linear approximation, in which agents iteratively estimate the underlying truth subspace $\mathbf{B}^*$ and their respective local "heads" $\omega^{k,*}$. Specifically, $\{z^k = \mathbf{B}\omega^k\}_{k\in[K]}$ are decomposed into the product of the common subspace estimate $\mathbf{B} \in \mathbb{R}^{d\times r}$, and an agent-specific head $\omega^k \in \mathbb{R}^r$.

A formal description can be found in Algorithm 1. At a high level, in each iteration, each agent rolls out a length-L trajectory, and uses the one-step TD errors $\{\delta_{t,\ell}^k\}_{\ell=0}^{L-1}$ – formally defined in Eq. (6) – and rewards to update estimates of its local head ω^k, common subspace $\mathbf{B}$, and local reward η^k. A similar scheme was also considered in (Xiong et al., 2025, Algorithm 2). The updates of ω^k and $\mathbf{B}$ can be interpreted as performing a perturbed stochastic gradient on the following joint linear approximation:

$$\min_{\mathbf{B},\omega^1,\cdots,\omega^K} \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\mu^k}[(\widehat{V}^k(s;\omega^k,\mathbf{B})-V^k(s))^2], \text{where } \widehat{V}^k(s;\omega^k,\mathbf{B})=\phi(s)^\top\mathbf{B}\omega^k. \tag{5}$$

The natural stochastic gradient descent updates of $\mathbf{B}$ and ω^k for Eq. (5) are:

$$\omega_{t+1}^k=\omega_t^k+\beta\left[V^k(s_{t,0}^k)-\phi(s_{t,0}^k)^\top\mathbf{B}_t\omega_t^k\right]\mathbf{B}_t{}^\top\phi(s_{t,0}^k),$$
$$\mathbf{B}_{t+1}^k=\mathbf{B}_t+\zeta\left[V^k(s_{t,0}^k)-\phi(s_{t,0}^k)^\top\mathbf{B}_t\omega_t^k\right]\phi(s_{t,0}^k)(\omega_t^k)^\top,$$

where β and ζ are stepsizes of ω^k and $\mathbf{B}$, respectively. However, V^k is unknown for each k. We follow batched semi-gradient TD(0): at each state $s_{t,\ell}^k$ of the length-L rollout, $V^k(s_{t,\ell}^k)$ is replaced by the bootstrapped target $R(s_{t,\ell}^k,a_{t,\ell}^k)-\eta_t^k+\phi(s_{t,\ell+1}^k)^\top\mathbf{B}_t\omega_t^k$, and the resulting L updates are averaged. The one-step TD errors along the rollout are

$$\delta_{t,\ell}^k=r_{t,\ell}^k-\eta_t^k+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k,\quad \ell=0,\dots,L-1, \tag{6}$$

where $r_{t,\ell}^k=R(s_{t,\ell}^k,a_{t,\ell}^k)$. Thus,

$$\widetilde{\omega}_{t+1}^k=\omega_t^k+\frac{\beta}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t{}^\top\phi(s_{t,\ell}^k), \tag{7}$$

$$\mathbf{B}_{t+1}^k=\mathbf{B}_t+\frac{\zeta}{L}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)(\omega_t^k)^\top, \tag{8}$$

where, for a matrix with orthonormal columns such as $\mathbf{B}_t$, the subscript $\perp$ denotes an orthonormal basis of the orthogonal complement of its column space.

Algorithmic Innovations. Although natural, Eq. (7) and Eq. (8) do not constitute the final updates used to obtain ω_{t+1}^k and $\mathbf{B}_{t+1}$. Several key algorithmic components are further incorporated to ensure favorable geometric properties of the error and perturbation terms. As can be seen in our analysis, these properties ultimately enable fine-grained control of the error dynamics, and enforce contraction of the principal-angle distance between $\mathbf{B}_t$ and $\mathbf{B}^*$—one of the major technical challenges.

(A) *Projection operation on local heads* Π_{U_ω}: This operation projects ω onto a convex ball with radius U_ω. In our analysis, we use both the coarse bound given by the projection as well as fine-grained characterization on the error dynamics similar to Khodadadian et al. (2022); Srikant & Ying (2019); Mitra (2024). Unlike Khodadadian et al. (2022); Srikant & Ying (2019); Mitra (2024) that relied on fine-grained analyses on the error dynamics to remove the need of projection of ω, in our problem, the combination of personalization and average-reward objectives significantly complicates the convergence analysis. In particular, the additional need to learn agent-specific mean rewards

Algorithm 1: pFedAR-TD

Input: Initial heads $\{\omega_0^k\}_{k=1}^K$; initial subspace $\mathbf{B}_0\in\mathbb{R}^{d\times r}$, which is orthonormal; three stepsizes: β for local heads, ζ for subspace, and γ for reward estimator; total iteration T; rollout length L; projection upper bound U_ω such that $\|\omega^{k,*}\|\le U_\omega$ for $k\in[K]$.
for $k=1,2,\dots,K$ **do**
  $\eta_0^k=0$
  Draw $s_{0,0}^k$ arbitrarily from $\mathcal{S}$
end for
for $t=0,1,2,\dots,T-1$ **do**
  for $k=1,2,\dots,K$ **do**
    Roll out L steps from $s_{t,0}^k$ under π and P^k: $(a_{t,\ell}^k,r_{t,\ell}^k,s_{t,\ell+1}^k)_{\ell=0}^{L-1}$.
    $\delta_{t,\ell}^k=r_{t,\ell}^k-\eta_t^k+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k$ for $\ell=0,\dots,L-1$.
    $\omega_{t+1}^k=\Pi_{U_\omega}\big(\omega_t^k+\frac{\beta}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)\big)$.
    $\mathbf{B}_{t+1}^k=\mathbf{B}_t+\frac{\zeta}{L}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\times\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)(\omega_t^k)^\top$.
    $\eta_{t+1}^k=\eta_t^k+\gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{t,\ell}^k-\eta_t^k\right)$.
    $s_{t+1,0}^k=s_{t,L}^k$.
  end for
  $\bar{\mathbf{B}}_{t+1}=\frac{1}{K}\sum_{k=1}^K\mathbf{B}_{t+1}^k$.
  $\mathbf{B}_{t+1},\mathbf{R}_{t+1}=\mathbb{QR}(\bar{\mathbf{B}}_{t+1})$.
  Server broadcasts $\mathbf{B}_{t+1}$ to all agents.
end for

results in three tightly coupled dynamics rather than two, calling for a careful balance between fine-grained characterization of error perturbations and the overall tractability of the system-level analysis. Without relying on any absolute bound on ω^k, characterizing the error dynamics quickly becomes intractable. Despite this technical convenience, it introduces "not so small" perturbation that interplays with the errors of other dynamics nontrivially that can be harmful to the convergence and potential speedup.

(B) *Subspace projected innovation in Eq. (8)*: Instead of moving $\mathbf{B}_{t+1}^k$ towards the local innovation $\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)(\omega_t^k)^\top$, we move $\mathbf{B}$ in the direction of the "residual" innovation that lies outside of $\mathbf{B}_t$. This can be viewed as innovation filtration in the subspace learning. Such filtration is not necessary in analyzing PFL with shared representation (Collins et al., 2021). Standard PFL with shared representations typically relies on fresh i.i.d. local samples, so QR re-orthogonalization and stochastic noise can be controlled by concentration arguments. In our setting, agents observe Markovian trajectories under heterogeneous transition kernels, and the TD drift involves heterogeneous matrices and time-varying projections. Our innovation filtration significantly mitigates some otherwise amplified perturbations arising from multiple cross terms. For example, under this projection, the perturbation in the QR decomposition is only second order; see Lemma D.1.

(C) *QR:* With QR, $\mathbf{B}_t$ maintains orthonormality, which is a key structural property that contributes to the contraction of the principal-angle distance between $\mathbf{B}_t$ and $\mathbf{B}^*$ (Collins et al., 2021).

Initialization Phase. Despite these techniques, a large subspace error early in training can amplify perturbations in the coupled updates and jeopardize convergence. We therefore initialize Algorithm 1 by having each agent run projected TD learning locally for T' rounds to obtain a coarse estimate of $z^{k,*}$. We then apply SVD to the agents' estimates to construct $\mathbf{B}_0$. The procedure is detailed in Algorithm 2.

5 Convergence Analysis

We present some technical assumptions adopted in our convergence analysis. Assumptions 1, 2, and 3 are widely adopted in the RL literature. Assumption 4 quantifies how well spread the underlying truth $z^{k,*}$ is in covering the r-dimensional subspace of $\mathbf{B}^*$.

Assumption 1 (Exploration). The matrix $A^k = \mathbb{E}_{s\sim\mu^k, a\sim\pi, s'\sim P^k}[\phi(s)(\phi(s') - \phi(s))^\top]$ is negative definite and its largest eigenvalue is upper bounded by $-\lambda$.

This assumption is widely adopted in analyzing TD learning with linear function approximation (Chen & Zhao, 2024; Bhandari et al., 2018; Chen et al., 2021; Olshevsky & Gharesifard, 2023; Qiu et al., 2021; Wu et al., 2020; Zou et al., 2019). Intuitively, A^k captures the one-step exploration of the policy π under the transition kernel P^k. For the tabular setting, Assumption 1 holds when the policy π explores all state-action pairs.

Assumption 2 (Uniform ergodicity). Let $\mathbb{P}_{0:\tau}^k(\cdot|s_0^k = s)$ denote the state distribution after τ steps given $s_0^k = s$ for agent k. There exist $m > 0$ and $\rho \in (0, 1)$, such that at all $k \in [K]$,

$$d_{TV}(\mathbb{P}_{0:\tau}^k(\cdot|s_0^k = s), \mu^k(\cdot)) \leq m\rho^\tau,\ \forall \tau \geq 0, s \in \mathcal{S}, \tag{9}$$

where d_{TV} is the total variation distance.

In our analysis, with a little abuse of notation, we also use $\mathbb{P}_{0:\tau}^k(\cdot,\cdot|s_0^k = s)$, which denotes the state-action pair at time τ given $s_0^k = s$. Recall that μ^k is the limiting distribution induced by policy π and the transition kernel $P^k, \forall k$. Uniform ergodicity is often assumed to characterize the noise induced by Markovian sampling (Chen & Zhao, 2024; Olshevsky & Gharesifard, 2023; Zou et al., 2019; Wu et al., 2020). Furthermore, we assume bounded features.

Assumption 3. $\|\phi(s)\| \leq 1$ for each $s \in \mathcal{S}$.

This assumption is standard in RL literature (Sutton et al., 1999; Sutton & Barto, 2018; Chen & Zhao, 2024; Olshevsky & Gharesifard, 2023; Zou et al., 2019; Wu et al., 2020). The existence of a shared common subspace helps to reduce the local learning problem from a d-dimensional problem to an r-dimensional problem. In PFL, when $r \ll d$, these benefits are well characterized and are substantial in derived bounds (Collins et al., 2021; Niu et al., 2024). However, the dependence on d is masked under Assumption 3, potentially obscuring the dimension reduction gains. More fine-

grained assumptions may therefore lead to significantly sharper results. Developing such refinements, however, would require a fundamental rethinking of TD learning and is left for future work.

$$\text{Let} \qquad \mathbf{Z}^* \in \mathbb{R}^{d\times K}, \qquad \text{with } z^{k,*} \text{ as the } k\text{-th column.} \tag{10}$$

By Eq. (4), we know that the rank of $\mathbf{Z}^*\mathbf{Z}^{*\top}$ is r. We require that each of the r dimensions of $\mathbf{B}^*$ is well-covered by $\{z^{k,*}\}_{k=1}^K$, formally stated in Assumption 4.

Assumption 4. $\frac{1}{K}\lambda^+_{\min}(\mathbf{Z}^*\mathbf{Z}^{*\top}) \geq \alpha$ for some $\alpha > 0$.

In our analysis, we also use $\lambda^+_{\min} = \lambda^+_{\min}(\mathbf{Z}^*\mathbf{Z}^{*\top})$ to denote the smallest nonzero eigenvalue of $\mathbf{Z}^*\mathbf{Z}^{*\top}$.

5.1 Main Convergence Results

In Algorithm 1, three groups of variables are updated: the common subspace estimate $\mathbf{B}_t$, the local heads ω_t^k, and estimates of local time-averaged rewards (defined in Eq. (1)). In this subsection, we characterize their evolutions over time. For ease of exposition, we introduce a set of notation:

$$x_t^k = \mathbf{B}_t\omega_t^k - z^{k,*},\ m_t = \mathbf{B}_\perp^{*\top}\mathbf{B}_t,\ y_t^k = J^k - \eta_t^k. \tag{11}$$

Define $M_t = \mathbb{E}\|m_t\|_F^2$, $\bar{X}_t = \frac{1}{K}\sum_{k=1}^K \mathbb{E}\|x_t^k\|^2$, $\bar{Y}_t = \frac{1}{K}\sum_{k=1}^K \mathbb{E}|y_t^k|^2$, together with the corresponding conditional expectations, which are the objects our one-step recursions are actually stated for. Fix an integer $\tau \geq 0$, and let $\mathbb{E}_i[\cdot]$ condition on the trajectories up to time i, as defined in Appendix C:

$$\widehat{M}_t = \mathbb{E}_{t-2\tau}\|m_t\|_F^2, \quad \widehat{X}_t = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{t-2\tau}\|x_t^k\|^2, \quad \widehat{Y}_t = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{t-2\tau}|y_t^k|^2, \tag{12}$$

so that $\bar{X}_t = \mathbb{E}[\widehat{X}_t]$, $M_t = \mathbb{E}[\widehat{M}_t]$ and $\bar{Y}_t = \mathbb{E}[\widehat{Y}_t]$ by the tower property. We keep the total expectation to the very end: Lemma 5.1 and Lemma 5.2 are stated for the conditional expectations, and $\mathbb{E}$ is taken only when the two are combined in Appendix H.

Let $\beta = c\zeta$ for a constant $c > 0$, and write $\mathcal{V}_t = \bar{X}_t + cM_t$.

Lemma 5.1 ((informal) Upper bound of local head errors)**.** *Let $\mathbf{P}^* = \mathbf{B}^*(\mathbf{B}^*)^\top$ and $\mathbf{P}_t = \mathbf{B}_t\mathbf{B}_t^\top$ for each t, with $\mathbf{P}_\perp^* = \mathbf{I} - \mathbf{P}^*$ and $\mathbf{P}_{t,\perp} = \mathbf{I} - \mathbf{P}_t$. Under the conditions collected in Section I.5, for any positive constants c_4, c_5, c_6,*

$$\begin{aligned}\widehat{X}_{t+1} \leq \widehat{X}_t + \frac{2c\zeta}{K}\sum_{k=1}^K \mathbb{E}_{t-2\tau}\left[(x_t^k)^\top\mathbf{P}_t A^k x_t^k\right] + \left(c_4^2 c + c_5^2 + c_6^2 + \frac{4U_\omega^4}{c_5^2}\right)\zeta\widehat{X}_t \\ + \mathcal{C}_{X,1}(\tau^2)\zeta^2\widehat{X}_t + \left[\left(\frac{c}{c_4^2} + \frac{U_\omega^4}{c_6^2}\right)\zeta + \mathcal{C}_{X,3}(\tau^2)\zeta^2\right]\widehat{Y}_t \\ + \mathcal{C}_{X,5}(\tau^2)\frac{\zeta^2}{L} + \mathcal{C}_{X,6}(\tau^2)\frac{\zeta^2}{K} + O(\zeta^3) + O(\gamma^2\zeta^2),\end{aligned}$$

where $\mathcal{C}_{X,1}(\tau^2)$, $\mathcal{C}_{X,3}(\tau^2)$, $\mathcal{C}_{X,5}(\tau^2)$, and $\mathcal{C}_{X,6}(\tau^2)$ are on the order of τ^2, formally defined in Appendix E.1.

Deriving Lemma 5.1 is highly nontrivial and requires a fundamental detour from the existing analysis (Chen & Zhao, 2024). For our problem, the main drift of the decay in $\|x_t^k\|$ may arise from $\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\rangle$. However, the environmental heterogeneity significantly complicates the characterization of this term. Specifically,

$$\left\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\right\rangle = \langle x_t^k, \mathbf{P}_t A^k x_t^k\rangle + \text{“perturbation”}.$$

Zooming into this expression, the main negative drift arises from the term $\langle x_t^k, \mathbf{P}_t A^k x_t^k\rangle$. In a single-agent or homogeneous environment it reduces to $\langle x_t^k, A^k x_t^k\rangle$, which Assumption 1 signs directly. Under environmental heterogeneity it does not, because of $\mathbf{P}_t$ and its interplay with the heterogeneous A^k: (1) $\mathbf{P}_t$ is not of full rank, (2) the local head error is distorted by the product $\mathbf{P}_t A^k$, and (3) $\mathbf{P}_t$ varies over time.

Lemma 5.2 (Principal angle distance analysis). *Under the conditions collected in Section I.5, for any positive c_1,*

$$\widehat{M}_{t+1} \le \widehat{M}_t + \frac{2\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[\left(\mathbf{P}_{t,\perp}\mathbf{P}^*_\perp x_t^k\right)^\top A^k x_t^k\right] + c_1^2\zeta\widehat{M}_t + \mathcal{C}_{M,4}(\tau^2)\zeta^2\widehat{X}_t$$
$$+ \left(\mathcal{C}_{M,5}(\tau^2)\zeta^2 + \frac{U_\omega^2}{c_1^2}\zeta\right)\widehat{Y}_t + 2C_{25}(\tau^2)\frac{\zeta^2}{L} + \mathcal{C}_{M,1}\zeta^4$$
$$+ 2C_{28}(\tau^4)\gamma^2\zeta^2 + \mathcal{C}_{M,2}(\tau^2)\frac{\zeta^2}{K} + \mathcal{C}_{M,3}\zeta^6 + \mathcal{C}_{M,6}\zeta^3,$$

where the constants are defined in Appendix F.

The analysis tracks the weighted sum $\widehat{X}_t + c\widehat{M}_t$. Adding the recursion of Lemma 5.1 to c times that of Lemma 5.2, the two cross terms combine: by $\mathbf{P}_t + \mathbf{P}_{t,\perp} = \mathbf{I}$ and $\mathbf{P}^*_\perp = \mathbf{I} - \mathbf{P}^*$,

$$(x_t^k)^\top\mathbf{P}_t A^k x_t^k + (\mathbf{P}_{t,\perp}\mathbf{P}^*_\perp x_t^k)^\top A^k x_t^k = (x_t^k)^\top A^k x_t^k - (\mathbf{P}_{t,\perp}\mathbf{P}^* x_t^k)^\top A^k x_t^k,$$

where Assumption 1 bounds the first term by $-\lambda\|x_t^k\|^2$, and $\|A^k\|_2 \le 2$ with $\|\mathbf{P}_{t,\perp}\mathbf{P}^*\|_2 \le \|m_t\|_2$ bound the second by $2\|m_t\|_2\|x_t^k\|^2$. The weighted sum therefore has a negative drift in $\widehat{X}_t$ while $\|m_t\|_2 \le \lambda/4$.

The subspace error is revealed only through the local heads. Lemma 5.3 bounds it by the aggregate head error, $\alpha M_t \le \bar{X}_t$, so that $\mathcal{V}_t \le (1 + c/\alpha)\bar{X}_t$: a drift in $\bar{X}_t$ becomes a drift in $\mathcal{V}_t$.

Lemma 5.3 (Lower bounding local head errors w.r.t. pincipal angle distance). *It holds that*

$$\frac{1}{K}\sum_{k=1}^{K}\|x_t^k\|^2 \ge \frac{\lambda^+_{\min}(\mathbf{Z}^*\mathbf{Z}^{*\top})}{K}\|m_t\|_F^2 \ge \alpha\|m_t\|_F^2.$$

Notably, Lemma 5.3 is not a problem-specific lower bound, which is often used to characterize the fundamental limits of a family of problems. Instead, it serves as a key intermediate result that is specific to our algorithm. As can be seen from our analysis, it says that under Algorithm 1, M_t and X_t can be learned and will be learned simultaneously.

Lemma 5.4 (Reward analysis). *Let $\gamma = \frac{C\log(T)}{T}$. Under the conditions collected in Section I.5,*

$$\mathbb{E}(y_T^k)^2 \le \frac{28U_r^2}{T^2} + \frac{\gamma}{L}\left(4U_r^2 + \frac{8U_r^2 m\rho}{1-\rho}\right) + \frac{8U_r^2 m}{L(1-\rho)}\left(\frac{\gamma^2}{\rho}\right)^L + \frac{8U_r^2\tau\gamma}{(1-\rho)^2L^2} + \widetilde{\mathcal{O}}\left(\frac{1}{T^2}\right),$$

which implies that for sufficiently large T, $\bar{Y}_T = \widetilde{\mathcal{O}}\left(\frac{1}{LT}\right)$.

Proposition 1 (Conditional convergence (informal)). *Consider Algorithm 1 with*

$$\beta = c\zeta, \qquad \zeta = \frac{8(\alpha+c)\log(TK)}{c\lambda\alpha(T-2\tau+1)}, \qquad \gamma = \frac{C\log(T)}{T},$$

$L = K$, $\tau = \lceil 2\log_\rho(\zeta)\rceil$, and the conditions collected in Section I.5. Suppose that $\|m_t\|_2 \le \lambda/4$ on the considered time interval. Then $\mathcal{V}_{T+1} = \widetilde{\mathcal{O}}\left(\frac{1}{TK}\right)$. Here $\widetilde{\mathcal{O}}$ hides polylogarithmic and problem-specific constants.

Proposition 1 assumes that $\|m_t\|_2 \le \lambda/4$, i.e., the subspace error stays bounded from above along the whole run. We will show that this holds with high probability in two steps. First, we will show that, with the initialization phase (formally described in Alg. 2), $\|\mathbf{B}_\perp^{*\top}\mathbf{B}_0\|_2 = \|m_0\|_2 \le \widetilde{t}_0$ is below a small threshold with high probability. Second, we split the error into two parts, according to whether the process has stayed below $\widetilde{t}_b$, where $\widetilde{t}_b > \widetilde{t}_0$. For ease of exposition, we refer to $\widetilde{t}_b$ as the barrier. With no upcrossing of $\widetilde{t}_b$, the conditions of Proposition 1 are satisfied, so its recursion can be unrolled on this part (Lemma 5.5). Above $\widetilde{t}_b$, which occurs only with low probability (Lemma 5.6), the error is bounded. Theorem 1 combines the two parts, as well as initialization, with Proposition 1.

Formally, we track the pathwise companion of $\mathcal{V}_t$, $n_t := \frac{1}{K}\sum_{k=1}^K\|x_t^k\|^2 + c\|m_t\|_F^2$, so that, $\mathcal{V}_t = \mathbb{E}[n_t]$. Since $n_t \ge c\|m_t\|_F^2 \ge c\|m_t\|_2^2$, $n_t \le c\lambda^2/16$ implies $\|m_t\|_2 \le \lambda/4$. Let $\widetilde{t}_b = c\lambda^2/32$, and let $T_{\text{exit}} := \min\left\{t \ge 0 : n_t > \frac{c\lambda^2}{32}\right\}$, $n_t^{\text{in}} := \mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}}\, n_t$. In the next two lemmas, $a := \frac{c\lambda\alpha}{4(\alpha+c)}\zeta$, and $U_n, \mathcal{C}_r, \mathcal{C}_{\epsilon,1}, \mathcal{C}_{\epsilon,2}$ are the constants of Lemma I.1, Lemma G.1 and Section I.3.

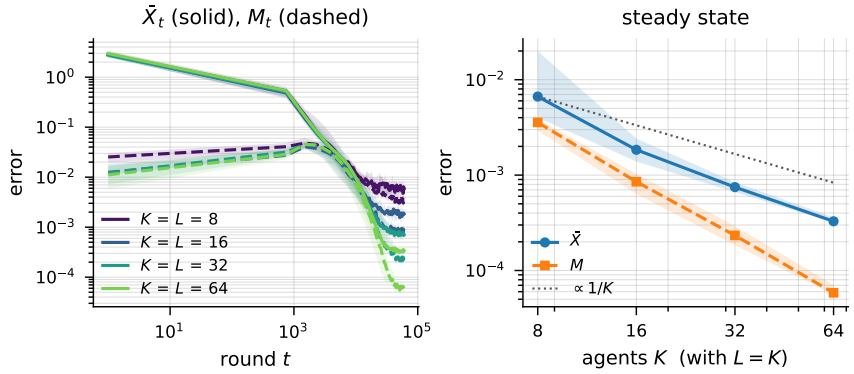


Figure 2: Synthetic MDPs, twelve seeds (median, interquartile range). Left: $\bar{X}_t$ (solid) and M_t (dashed) along the run. Right: their steady states against K at $L = K$, with a $1/K$ reference.

Lemma 5.5. *Under the conditions collected in Section I.5,*

$$\mathbb{E}\big[n^{\text{in}}_{T+1}\big] \le (1-a)^{T+1-2\tau}\big(4U_\omega^2 + cr\big) + 2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\,\zeta\,T\max\big\{e^{-2\gamma T}, e^{-a(T-2\tau)}\big\} + \frac{2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\,\zeta}{a}\Big(\frac{\gamma}{L} + \tau^2\gamma^2\Big) + \frac{\mathcal{C}_{\epsilon,2}\,\zeta^2}{a}\Big(\frac{1}{L} + \frac{1}{K}\Big),$$

which is $\widetilde{\mathcal{O}}(1/(TK))$ at $L = K \le T$.

Lemma 5.6 (High probability bound of stopping time)**.** *Let $\delta \in (0,1)$. Prepend the initialization phase of Algorithm 2 and assume in addition the conditions collected in Section I.5. Then*

$$\Pr\big[T_{\text{exit}} \le T\big] \;\le\; \delta + \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big),$$

which is at most $2/(TK)$ at $\delta = 1/(TK)$ under the lower bound on T in Theorem 2.

Theorem 1 (Global convergence (informal))**.** *Run Algorithm 2 under the conditions of Theorem 2, which require T and on K to be sufficiently large. Then, with no hypothesis on the trajectory,*

$$\mathcal{V}_{T+1} = \widetilde{\mathcal{O}}\Big(\frac{r}{TK}\Big).$$

The proofs of Lemma 5.5, Lemma 5.6 and Theorem 1 are deferred to Appendix I (Corollary 1, Lemma I.7 and Theorem 2).

6 Experiments

Synthetic MDPs. We take $n = 40$, $d = 20$, $r = 3$, and $K \in \{8, 16, 32, 64\}$ agents. The s-th row of $\Phi \in \mathbb{R}^{n\times d}$ is $\phi(s)$, with i.i.d. Gaussian entries rescaled by one constant so that $\max_s \|\phi(s)\| = 1$; $\mathbf{B}^* \in \mathbb{R}^{d\times r}$ is the Q factor of a Gaussian matrix, $\omega^{k,*}$ is Gaussian, $z^{k,*} = \mathbf{B}^*\omega^{k,*}$. Each agent has its own transition kernel P^k, and $r^k = V^k - P^kV^k + J^k$ with $V^k = \Phi z^{k,*}$ and $J^k = 0$. We run $T = 6\times 10^4$ rounds at $L = K$, $\beta = 0.08$, $\zeta = 0.05$, $\gamma = 0.05$, $\omega_0^k = 0$ and $\eta_0^k = 0$, preceded by an initialization phase of $T' = 10^4$ rounds at $L' = 4$. One 64-agent instance is drawn per seed and each K uses its first K agents; we report medians over twelve seeds with the interquartile range, the steady state being the average over the last $T/5$ rounds. The two parts of $\mathcal{V}_t$ fall from $\bar{X} = 6.7\times 10^{-3}$ and $M = 3.6\times 10^{-3}$ at $K = 8$ to 3.3×10^{-4} and 5.8×10^{-5} at $K = 64$, showing the speedup in K. Appendix L compares these stepsizes with a two-timescale schedule.

Gym environments. We evaluate the personalization framework on heterogeneous Acrobot, CartPole and Hopper: **pFedAR-TD** against **Single TD** and **FedTD-Uniform** for value-function approximation under a fixed policy, against a two-timescale scheme, and extended to control through the actor-critic **pFed-AC**. Appendix M gives the configurations and the results.

AI USE STATEMENT

We used large language models (LLMs) to aid and polish the writing of this paper, including improving grammar, wording, and readability. All LLM-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the content of this paper.

REFERENCES

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

Ali Beikmohammadi, Sarit Khirirat, Peter Richtárik, and Sindri Magnússon. 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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, nill, 2020. nill.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34: 25294–25307, 2021.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36, 2024.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4180–4234. PMLR, 2023.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

John C Duchi, Vitaly Feldman, Lunjia Hu, and Kunal Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. *Advances in Neural Information Processing Systems*, 35:5854–5866, 2022.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

FHA. Traffic analysis tools: Next generation simulation, 2020. URL https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm. [Accessed: Sep. 13, 2024].

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 954–960, nill, 2018. nill.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Yuchen Jiao, Jiin Woo, Gen Li, Gauri Joshi, and Yuejie Chi. Sample complexity of average-reward q-learning: From single-agent to federated reinforcement learning. *arXiv preprint arXiv:2601.13642*, 2026.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, 49(4): 2295–2322, 2024.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022.

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

Safwan Labbi, Daniil Tiapkin, Lorenzo Mancini, Paul Mangold, and Eric Moulines. Federated ucbvi: Communication-efficient federated regret minimization with heterogeneous agents. *arXiv preprint arXiv:2410.22908*, 2024.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Kiwan Maeng, Haiyu Lu, Luca Melis, John Nguyen, Mike Rabbat, and Carole-Jean Wu. Towards fair federated recommendation learning: Characterizing the inter-dependence of system and data heterogeneity. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 156–167, 2022.

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

Aritra Mitra. A simple finite-time analysis of td learning with linear function approximation. *IEEE Transactions on Automatic Control*, 70(2):1388–1394, 2024.

Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE communications surveys & tutorials*, 23(3):1622–1658, 2021.

Xiaochun Niu, Lili Su, Jiaming Xu, and Pengkun Yang. Collaborative learning with shared linear representations: Statistical rates and optimal algorithms. *arXiv preprint arXiv:2409.04919*, 2024.

Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022.

Sudeep Salgia and Yuejie Chi. The sample-communication complexity trade-off in federated q-learning. *Advances in Neural Information Processing Systems*, 37:39694–39747, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6MmiS0HUJHR.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on learning theory*, pp. 2803–2830. PMLR, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *Journal of Machine Learning Research*, 26(187):1–125, 2025.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35 (11):1799–1808, 1999.

Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

Muxing Wang, Pengkun Yang, and Lili Su. On the convergence rates of federated q-learning across heterogeneous environments. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EkLAG3gt3g.

Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37157–37216. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/woo23a.html.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Zhijie Xie and Shenghui Song. FedKL: Tackling Data Heterogeneity in Federated Reinforcement Learning by Penalizing KL Divergence. *IEEE Journal on Selected Areas in Communications*, 41 (4):1227–1242, April 2023. ISSN 1558-0008. doi: 10.1109/JSAC.2023.3242734. URL https://ieeexplore.ieee.org/abstract/document/10038492?casa_token=yGyMDlnL_FsAAAAA:hqNvzWEb6yVKwTZVdHKLVnorDg07AWx4uujDsLLTTY_7unjr1ew8Yv4_UUAWfCx3X1b9wHNySP8.

Guojun Xiong, Shufan Wang, Daniel Jiang, and Jian Li. On the linear speedup of personalized federated reinforcement learning with shared representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SXZr8aDKia.

Tong Yang, Shicong Cen, Yuting Wei, Yuxin Chen, and Yuejie Chi. Federated natural policy gradient and actor critic methods for multi-task reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DUFD6vsyF8.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *arXiv preprint arXiv:2401.15273*, 2024.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International conference on machine learning*, pp. 5872–5881. PMLR, 2018.

Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. On the convergence of sarsa with linear function approximation. In *International Conference on Machine Learning*, pp. 41613–41646. PMLR, 2023a.

Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023b.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

Appendices

A Limitations

First, while our convergence bound achieves $\widetilde{\mathcal{O}}(r/(TK))$, the dependence on the feature dimension d is masked under Assumption 3 (bounded features). Relaxing this assumption to obtain bounds that explicitly capture the benefit of a low-dimensional subspace ($r \ll d$) would require a more refined analysis and is left for future work. Second, our theoretical analysis focuses on TD-learning (i.e., policy evaluation under a fixed policy); extending the finite-time guarantees to the full actor-critic setting with policy updates remains an open problem.

B Additional Related Work

Fed RL under homogeneous environments. When the environments of all agents are homogeneous, it has been shown that the federated version of classic reinforcement learning algorithms can significantly alleviate the data collection burden on individual agents (Woo et al., 2023; Khodadadian et al., 2022). Specifically, the per-agent sample complexity decreases linearly in terms of the number of agents – an effect commonly referred to in existing literature as linear speedup. Recent work has made significant progress in understanding federated RL under homogeneous environments. Woo et al. (2023) studied federated Q-learning. In addition to a linear speedup, Woo et al. (2023) uncovered the blessing of heterogeneity in terms of state-action exploration – a completely different notion of heterogeneity from our focus. Salgia & Chi (2025) studied the sample-communication complexity trade-off, and developed a new algorithm that achieves order-optimal sample and communication complexities.

Model/Policy personalization. In FL, model personalization (often referred to as PFL) has garnered significant attention in recent years. The success of personalization techniques depends on balancing the bias introduced by using global knowledge that may not generalize to individual clients, and the variance inherent in learning from limited local datasets. Popular PFL techniques include regularized local objectives (T Dinh et al., 2020; Li et al., 2021), local-global parameter interpolation (Deng et al., 2020), meta-learning (Fallah et al., 2020; Jiang et al., 2019), and representation learning (Collins et al., 2021; Xu et al., 2023). Extending these ideas of PFL to RL introduces an even greater level of difficulty that arises from the inherent sample correlation induced by sequential decision-making, and the non-stationarity of sample quality as the policy evolves.

In the RL literature, training personalized (or localized) policies has recently garnered considerable attention in the context of multi-agent Markov games or competitive MARL (Filar & Vrieze, 2012; Zhang et al., 2018; Qu et al., 2022; Zhang et al., 2023b; Daskalakis et al., 2023), which depart fundamentally from our focus. Specifically, in their setting, agents interact with one another within a shared environment, and the rewards received by individual agents are determined by the joint actions of all agents. Each agent aims to learn a localized policy that contributes to a joint policy across all agents, with the goal of maximizing its own welfare. A commonly adopted objective among the participating agents is to learn some form of equilibrium, such as a Nash Equilibrium.

C Notations and Preliminaries

Let $v_{0:t}^k$ denote the Markovian trajectory up to time t generated under initial state distribution d_0, given policy π, and the local transition kernel P^k:

$$v_{0:t}^k = (s_0^k, a_0^k, s_1^k, ..., s_{t+1}^k),$$

and $v_{0:t}$ denote the collection of Markovian trajectories generated by all agents:

$$v_{0:t} = \{v_{0:t}^k\}_{k=1}^K.$$

Similarly, we define $v_{0:(t,\ell)}^k$ and $v_{0:(t,\ell)}$ as the Markovian trajectory generated up to the ℓ-th sample in time t.

Recall that $\mathbb{E}_j[X_t]$ denotes the conditional expectation of any random variable X_t given the collection of Markovian trajectories up to time j:

$$\mathbb{E}_j[X_t] = \mathbb{E}\left[X_t \mid v_{0:j}\right] = \mathbb{E}\left[X_t \mid \{v_{0:j}^k\}_{k=1}^K\right].$$

Write $\mathcal{F}_j := \sigma(v_{0:j})$ for the σ-algebra these trajectories generate, so that $\mathbb{E}_j[X_t] = \mathbb{E}[X_t \mid \mathcal{F}_j]$ and $\{\mathcal{F}_j\}_{j\geq 0}$ is a filtration.

Let

$$\widetilde{x}_{t+1}^k = \bar{\mathbf{B}}_{t+1}\widetilde{\omega}_{t+1}^k - z^{k,*}, \qquad \bar{m}_{t+1} = \mathbf{B}_\perp^{*\top}\bar{\mathbf{B}}_{t+1}, \tag{13}$$

where $\widetilde{\omega}_{t+1}^k$ is given in Eq. (7) and $\bar{\mathbf{B}}_{t+1}$ in Algorithm 1.

Denote

$$\widetilde{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k); \quad \bar{b}^k = \mathbb{E}_{s\sim\mu^k, a\sim\pi}\left[(R(s,a) - J^k)\phi(s)\right]. \tag{14}$$

In addition, let

$$\begin{aligned}
\xi_t^k =& \frac{1}{L}\sum_{\ell=0}^{L-1}(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k) \\
&+ \frac{1}{L}\sum_{\ell=0}^{L-1}(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}_t\omega_t^k\phi(s_{t,\ell}^k) \\
&- \frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}\left[(R(s^{(\ell)}, a^{(\ell)}) - J^k)\phi(s^{(\ell)})\right] \\
&- \frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}\left[(\phi(s^{(\ell+1)}) - \phi(s^{(\ell)}))^\top \mathbf{B}_t\omega_t^k\phi(s^{(\ell)})\right],
\end{aligned} \tag{15}$$

where the expectation is taken with respect to $s^{(\ell)} \sim \mu^k$, $a^{(\ell)} \sim \pi$, $s^{(\ell+1)} \sim P^k$. This can be viewed as the Markovian noise of the TD error of agent k at time t. Moreover, let

$$\begin{aligned}
\widetilde{A}_t^k &= \frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top, \\
\text{and } \mathbf{b}_t^k &= \frac{1}{L}\sum_{\ell=0}^{L-1}\big((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) \\
&\qquad - \phi(s_{t,\ell}^k))^\top \mathbf{B}^*\omega^{k,*}\big)\phi(s_{t,\ell}^k).
\end{aligned} \tag{16}$$

The following rewrite appears repeatedly in our analysis.

Proposition 2. *For any t and agent k, the averaged innovation $\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)$ can be rewritten as*

$$\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k) = \xi_t^k + A^k x_t^k + y_t^k\bar{\phi}_t^k, \tag{17}$$

and

$$\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k) = \widetilde{A}_t^k x_t^k + y_t^k\bar{\phi}_t^k + \mathbf{b}_t^k, \tag{18}$$

where $\bar{\phi}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)$.

The proof of Proposition 2 can be found in Appendix K.

D Bound on QR Decomposition and Perturbation

The results in this section hold for arbitrary $\mathbf{B}_0$ that has orthonormal columns.

Recall from Algorithm 1 the update of $\bar{\mathbf{B}}$:

$$\bar{\mathbf{B}}_{t+1} = \mathbf{B}_t + \frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top$$

For ease of exposition, let

$$\mathbf{Q}_t := \frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top. \tag{19}$$

The following couple of lemmas enable us to bound the "distortion" caused by the QR decomposition. Specifically, Lemma D.1 bounds the perturbation in terms of $\|\mathbf{Q_t}\|$, which is further bounded in Lemma D.2. It is worth noting that, in the traditional PFL (Collins et al., 2021), such distortion can be well controlled by the number of $i.i.d.$ local samples that are freshly drawn in each iteration. However, their analysis does not apply to our setting due to the Markov sampling. We fundamentally depart from the analysis in Collins et al. (2021) in deriving upper bounds on those distortions via constructing a fixed-point iteration.

Lemma D.1 (Perturbation of QR)**.** *For each $t \geq 1$, when $U_\delta U_\omega \zeta \leq 1/2$ (where U_δ is defined in Lemma J.2), it holds that*

$$\|\mathbf{R}_{t+1} - \mathbf{I}\|_2 \leq 2\|\mathbf{Q}_t\|_F^2, \tag{20}$$

$$\|\mathbf{R}_{t+1}^{-1}\| \leq 1, \tag{21}$$

$$\|\mathbf{R}_{t+1}^{-1} - \mathbf{I}\|_2 \leq 2\|\mathbf{Q}_t\|_F^2. \tag{22}$$

Proof. Recall that $\bar{\mathbf{B}}_{t+1} = \mathbf{B}_{t+1}\mathbf{R}_{t+1}$. We rewrite $\mathbf{R}_{t+1}^\top\mathbf{R}_{t+1}$ as

$$\begin{aligned}\mathbf{R}_{t+1}^\top\mathbf{R}_{t+1} = \bar{\mathbf{B}}_{t+1}^\top\bar{\mathbf{B}}_{t+1} &= \mathbf{B}_t^\top\mathbf{B}_t + \mathbf{B}_t^\top\mathbf{Q}_t + \mathbf{Q}_t^\top\mathbf{B}_t + \mathbf{Q}_t^\top\mathbf{Q}_t \\ &= \mathbf{I} + \mathbf{B}_t^\top\mathbf{Q}_t + \mathbf{Q}_t^\top\mathbf{B}_t + \mathbf{Q}_t^\top\mathbf{Q}_t \\ &= \mathbf{I} + \mathbf{Q}_t^\top\mathbf{Q}_t.\end{aligned}$$

The last equality holds because $\mathbf{Q}_t$ lives in span$(\mathbf{B}_{t,\perp})$.

It can also be rewritten as

$$\begin{aligned}\mathbf{R}_{t+1}^\top\mathbf{R}_{t+1} &= (\mathbf{I} + \mathbf{R}_{t+1} - \mathbf{I})^\top(\mathbf{I} + \mathbf{R}_{t+1} - \mathbf{I}) \\ &= \mathbf{I} + (\mathbf{R}_{t+1} - \mathbf{I})^\top + (\mathbf{R}_{t+1} - \mathbf{I}) + (\mathbf{R}_{t+1} - \mathbf{I})^\top(\mathbf{R}_{t+1} - \mathbf{I}).\end{aligned}$$

Thus, we get

$$(\mathbf{R}_{t+1} - \mathbf{I})^\top + (\mathbf{R}_{t+1} - \mathbf{I}) = \mathbf{Q}_t^\top\mathbf{Q}_t - (\mathbf{R}_{t+1} - \mathbf{I})^\top(\mathbf{R}_{t+1} - \mathbf{I}). \tag{23}$$

We construct a fixed-point iteration to "solve" $\mathbf{R}_{t+1} - \mathbf{I}$ in terms of $\mathbf{Q}_t^\top\mathbf{Q}_t$ under Eq. (23).

Toward this, we decompose the symmetric matrix $\mathbf{Q}_t^\top\mathbf{Q}_t$ into

$$\mathbf{Q}_t^\top\mathbf{Q}_t = \mathbf{R}_0 + \mathbf{R}_0^\top,$$

where $\mathbf{R}_0$ is upper triangular. It is easy to see that

$$\|\mathbf{R}_0\|_F \leq \left\|\mathbf{R}_0 + \mathbf{R}_0^\top\right\|_F \leq \|\mathbf{Q}_t\|_F^2 \leq \zeta^2 U_\delta^2 U_\omega^2.$$

Let $\Delta_{t+1} \triangleq \mathbf{R}_{t+1} - \mathbf{I} - \mathbf{R}_0$. Then Eq. (23) is equivalent to

$$\Delta_{t+1}^\top + \Delta_{t+1} = -(\Delta_{t+1}^\top\Delta_{t+1} + \Delta_{t+1}^\top\mathbf{R}_0 + \mathbf{R}_0^\top\Delta_{t+1} + \mathbf{R}_0^\top\mathbf{R}_0). \tag{24}$$

Next, we define a sequence of $\Delta_{t+1,i}$ for $i \geq 0$. Let $\Delta_{t+1,0} = 0$ for $i \geq 1$ and let $\Delta_{t+1,i}$ be an upper triangular matrix satisfying

$$\Delta_{t+1,i}^\top + \Delta_{t+1,i} = -(\Delta_{t+1,i-1}^\top \Delta_{t+1,i-1} + \Delta_{t+1,i-1}^\top \mathbf{R}_0 + \mathbf{R}_0^\top \Delta_{t+1,i-1} + \mathbf{R}_0^\top \mathbf{R}_0).$$

We first show the boundedness of the sequence. Note that

$$\|\Delta_{t+1,i}\|_F \leq \left\|\Delta_{t+1,i}^\top + \Delta_{t+1,i}\right\|_F \leq \|\Delta_{t+1,i-1}\|_F^2 + 2\|\mathbf{R}_0\|_F \|\Delta_{t+1,i-1}\|_F + \|\mathbf{R}_0\|_F^2, \quad (25)$$

where the second inequality applies the triangle inequality, and the first inequality holds because the absolute value of all entries in $\Delta_{t+1,i}^\top + \Delta_{t+1,i}$ is larger than that in $\Delta_{t+1,i}$.

Provided $\zeta^2 U_\delta^2 U_\omega^2 \leq 1/4$, we have $\|\mathbf{R}_0\|_F \leq 1/4$, and for any $\|\Delta_{t+1,i-1}\|_F \leq 1/4$, we have $\|\Delta_{t+1,i}\|_F \leq 1/4$ by Eq. (25). By induction, the entire sequence satisfies $\|\Delta_{t+1,i}\|_F \leq 1/4$ for all $i \geq 0$.

Next, we show the convergence of the sequence $\Delta_{t+1,i}$. By definition,

$$\begin{aligned}
&(\Delta_{t+1,i+1}^\top + \Delta_{t+1,i+1}) - (\Delta_{t+1,i}^\top + \Delta_{t+1,i}) \\
=& -(\Delta_{t+1,i}^\top(\Delta_{t+1,i} - \Delta_{t+1,i-1}) + (\Delta_{t+1,i} - \Delta_{t+1,i-1})^\top \Delta_{t+1,i-1} \\
&+ (\Delta_{t+1,i} - \Delta_{t+1,i-1})^\top \mathbf{R}_0 + \mathbf{R}_0^\top (\Delta_{t+1,i} - \Delta_{t+1,i-1})).
\end{aligned}$$

By taking the norm on both sides, we obtain

$$\begin{aligned}
\|\Delta_{t+1,i+1} - \Delta_{t+1,i}\|_F &\leq \left\|(\Delta_{t+1,i+1} - \Delta_{t+1,i})^\top + (\Delta_{t+1,i+1} - \Delta_{t+1,i})\right\|_F \\
&\leq \frac{1}{4}\|\Delta_{t+1,i} - \Delta_{t+1,i-1}\|_F + \frac{1}{4}\|\Delta_{t+1,i} - \Delta_{t+1,i-1}\|_F \\
&\quad + 2\|\mathbf{R}_0\|_F \|\Delta_{t+1,i} - \Delta_{t+1,i-1}\|_F .
\end{aligned}$$

Therefore, since $\|\mathbf{R}_0\|_F < \frac{1}{4}$, the difference $\|\Delta_{t+1,i+1} - \Delta_{t+1,i}\|_F$ geometrically converges to zero, which implies that the sequence $\Delta_{t+1,i}$ converges.

The limit of the sequence satisfies the fixed-point Eq. (24). Taking norm on both sides of Eq. (24) yields

$$\|\Delta_{t+1}\|_F \leq \left\|\Delta_{t+1}^\top + \Delta_{t+1}\right\|_F \leq \|\Delta_{t+1}\|_F^2 + 2\|\mathbf{R}_0\|_F \|\Delta_{t+1}\|_F + \|\mathbf{R}_0\|_F^2 .$$

Since $\|\Delta_{t+1}\|_F \leq 1/4$, we obtain that

$$\|\Delta_{t+1}\|_F \leq \frac{1 - 2\|\mathbf{R}_0\|_F - \sqrt{1 - 4\|\mathbf{R}_0\|_F}}{2} \leq 4\|\mathbf{R}_0\|_F^2 .$$

We conclude that

$$\|\mathbf{R}_{t+1} - \mathbf{I}\|_F = \|\mathbf{R}_0 + \Delta_{t+1}\|_F \leq \|\mathbf{R}_0\|_F + 4\|\mathbf{R}_0\|_F^2 \leq 2\|\mathbf{R}_0\|_F \leq 2\|\mathbf{Q}_t\|_F^2 .$$

For any $u \in \mathbb{R}^r$, Eq. (23) gives

$$\|\mathbf{R}_{t+1}u\|^2 = u^\top \mathbf{R}_{t+1}^\top \mathbf{R}_{t+1} u = \|u\|^2 + \|\mathbf{Q}_t u\|^2 \geq \|u\|^2.$$

Thus $\|\mathbf{R}_{t+1}^{-1}\| \leq 1$. Moreover,

$$\|\mathbf{R}_{t+1}^{-1} - \mathbf{I}\|_2 = \|\mathbf{R}_{t+1}^{-1}(\mathbf{I} - \mathbf{R}_{t+1})\|_2 \leq \|\mathbf{R}_{t+1}^{-1}\|_2 \|\mathbf{I} - \mathbf{R}_{t+1}\|_2.$$

Together with the bound above on $\mathbf{R}_{t+1} - \mathbf{I}$, we obtain

$$\begin{aligned}
&\|\mathbf{R}_{t+1} - \mathbf{I}\|_2 \leq \|\mathbf{R}_{t+1} - \mathbf{I}\|_F \leq 2\|\mathbf{Q}_t\|_F^2, \\
&\|\mathbf{R}_{t+1}^{-1}\| \leq 1, \\
&\|\mathbf{R}_{t+1}^{-1} - \mathbf{I}\|_2 \leq 2\|\mathbf{Q}_t\|_F^2.
\end{aligned}$$

□

Lemma D.2. *For any $t \geq \tau$,*

$$\mathbb{E}_{t-\tau}\|\mathbf{Q}_t\|_F^2 \leq 16U_\omega^2\zeta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}\|x_t^k\|^2 + 4\frac{U_\delta^2}{L}U_\omega^2\frac{\zeta^2}{K} + 16U_\omega^2U_\delta^2\zeta^2\frac{m^2\rho^{2\tau L}}{L^2(1-\rho)^2}$$
$$+ 4\tau^2U_\delta^4\zeta^2\beta^2 + \frac{4U_\omega^2\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}|y_t^k|^2. \tag{26}$$

Proof. For any $t \geq \tau$, we have

$$\|\mathbf{Q}_t\|_F^2 = \|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{P}_{t,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2 \leq \|\frac{\zeta}{KL}\sum_{k=1}^{K}\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2$$
$$= \frac{\zeta^2}{K^2}\|\sum_{k=1}^{K}\Big(\widetilde{A}_t^k x_t^k + y_t^k\bar{\phi}_t^k + \mathbf{b}_t^k\Big)(\omega_t^k)^\top\|_F^2 \quad \text{(by Proposition 2)}$$
$$\leq \underbrace{\frac{4\zeta^2}{K^2}\|\sum_{k=1}^{K}\widetilde{A}_t^k x_t^k(\omega_t^k)^\top\|_F^2}_{I_1} + \underbrace{\frac{4\zeta^2}{K^2}\|\sum_{k=1}^{K}\mathbf{b}_t^k(\omega_{t-\tau}^k)^\top\|_F^2}_{I_2}$$
$$+ \underbrace{\frac{4\zeta^2}{K^2}\|\sum_{k=1}^{K}\mathbf{b}_t^k(\omega_t^k-\omega_{t-\tau}^k)^\top\|_F^2}_{I_3} + \underbrace{\frac{4\zeta^2}{K^2}\|\sum_{k=1}^{K}y_t^k\bar{\phi}_t^k(\omega_t^k)^\top\|_F^2}_{I_4}.$$

Recall that $\widetilde{A}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top, \mathbf{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}\Big((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}^*\omega^{k,*}\Big)\phi(s_{t,\ell}^k)$. Then,

$$\|\widetilde{A}_t^k\| = \|\frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\| \leq 2. \tag{27}$$

For I_1, we have,

$$I_1 \leq \frac{4\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k x_t^k(\omega_t^k)^\top\|^2 \leq 16U_\omega^2\zeta^2\frac{1}{K}\sum_{k=1}^{K}\|x_t^k\|^2.$$

For I_2, the Markovian noise, we have under expectation that,

$$\frac{4\zeta^2}{K^2}\mathbb{E}_{t-\tau}\|\sum_{k=1}^{K}\mathbf{b}_t^k(\omega_{t-\tau}^k)^\top\|_F^2$$
$$\leq \frac{4\zeta^2}{K^2}\mathbb{E}_{t-\tau}\sum_{k=1}^{K}\|\mathbf{b}_t^k(\omega_{t-\tau}^k)^\top\|^2 + \frac{8\zeta^2}{K^2}\sum_{k=1}^{K}\sum_{k'=k+1}^{K}\mathbb{E}_{t-\tau}\Big\langle\mathbf{b}_t^k(\omega_{t-\tau}^k)^\top, \mathbf{b}_t^{k'}(\omega_{t-\tau}^{k'})^\top\Big\rangle$$
$$\leq 4\frac{U_\delta^2}{L}U_\omega^2\frac{\zeta^2}{K} + 16U_\omega^2U_\delta^2\zeta^2\frac{m^2\rho^{2\tau L}}{L^2(1-\rho)^2}. \quad \text{(by Lemma K.5, Lemma K.6, independence)}$$

For I_3,

$$\begin{aligned}
&\frac{4\zeta^2}{K^2}\|\sum_{k=1}^{K}\mathbf{b}_t^k(\omega_t^k-\omega_{t-\tau}^k)^\top\|_F^2\\
\leq&\frac{4\zeta^2}{K}\sum_{k=1}^{K}\|\mathbf{b}_t^k(\omega_t^k-\omega_{t-\tau}^k)^\top\|^2\\
\leq&\frac{4\zeta^2}{K}\sum_{k=1}^{K}\|\mathbf{b}_t^k\|^2\|\omega_t^k-\omega_{t-\tau}^k\|^2\\
\leq&4U_\delta^2\zeta^2\left(\tau U_\delta\beta\right)^2 \quad \text{(by Lemma K.1)}\\
=&4\tau^2U_\delta^4\zeta^2\beta^2.
\end{aligned}$$

Lastly, for I_4, we have

$$I_4\leq\frac{4\zeta^2}{K}\sum_{k=1}^{K}\|y_t^k\bar{\phi}_t^k(\omega_t^k)^\top\|^2\leq\frac{4U_\omega^2\zeta^2}{K}\sum_{k=1}^{K}|y_t^k|^2.$$

In summary,

$$\begin{aligned}
\mathbb{E}_{t-\tau}\|\mathbf{Q}_t\|_F^2\leq&16U_\omega^2\zeta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}\|x_t^k\|^2+4\frac{U_\delta^2}{L}U_\omega^2\frac{\zeta^2}{K}+16U_\omega^2U_\delta^2\zeta^2\frac{m^2\rho^{2\tau L}}{L^2(1-\rho)^2}\\
&+4\tau^2U_\delta^4\zeta^2\beta^2+\frac{4U_\omega^2\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}|y_t^k|^2.
\end{aligned}$$

□

E BOUNDS ON LOCAL HEAD ERRORS

E.1 UPPER BOUND: PROOF OF LEMMA 5.1

We first provide a formal statement of Lemma 5.1, which involves a set of constants whose formal definitions are deferred to the lemma proof.

Lemma E.1 (Upper bound of local head errors). *Let $\mathbf{P}^* = \mathbf{B}^*(\mathbf{B}^*)^\top$ and $\mathbf{P}_t = \mathbf{B}_t\mathbf{B}_t^\top$ for each t, with $\mathbf{P}_\perp^* = \mathbf{I} - \mathbf{P}^*$ and $\mathbf{P}_{t,\perp} = \mathbf{I} - \mathbf{P}_t$. Choose $\tau = \lceil 2\log_\rho(\zeta)\rceil$, $U_\delta U_\omega \zeta \le \frac{1}{2}$, and $\beta U_\delta \le U_\omega$, where $U_\delta = 2U_r + 2U_\omega$ and $\beta = c\zeta$ for arbitrary $c > 0$ that can be tuned. For any positive constants c_4, c_5, c_6,*

$$\begin{aligned}
\widehat{X}_{t+1} \le \widehat{X}_t &+ \frac{2c\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top \mathbf{P}_t A^k x_t^k\right] \\
&+ \left(c_4^2 c + c_5^2 + c_6^2 + \frac{4U_\omega^4}{c_5^2}\right)\zeta\widehat{X}_t + \mathcal{C}_{X,1}(\tau^2)\zeta^2\widehat{X}_t \\
&+ \left[\left(\frac{c}{c_4^2} + \frac{U_\omega^4}{c_6^2}\right)\zeta + \mathcal{C}_{X,3}(\tau^2)\zeta^2\right]\widehat{Y}_t \\
&+ \mathcal{C}_{X,5}(\tau^2)\frac{\zeta^2}{L} + \mathcal{C}_{X,6}(\tau^2)\frac{\zeta^2}{K} + \mathcal{C}_{X,7}\zeta^3 + \mathcal{C}_{X,8}\zeta^4 + \mathcal{C}_{X,9}\gamma^2\zeta^2 + \mathcal{C}_{X,13}\zeta^6,
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{C}_{X,1}(\tau^2) = \left(C_{31}c + (36 + 108\tau^2)c^2 + C_{33}c + \left(36U_\omega^4 + (2.5 + 3U_\omega)\,\tau^2 C_{18}\right) + 576U_\omega^4\right), \\
&\mathcal{C}_{X,3}(\tau^2) = \left[C_{32}c + (9 + 27\tau^2)c^2 + C_{34}c + 9U_\omega^4 + (2.5 + 3U_\omega)\,\tau^2 C_{23} + 144U_\omega^4\right], \\
&\mathcal{C}_{X,5}(\tau^2) = C_{39}(\tau^2), \qquad \mathcal{C}_{X,6}(\tau^2) = C_{41}(\tau^2) + 144\frac{U_\delta^2}{L}U_\omega^4, \\
&\mathcal{C}_{X,7} = C_{42} + C_{43}(\tau), \qquad \mathcal{C}_{X,8} = C_{40}(\tau^4) + 144U_\omega^2\tau^2U_\delta^4c^2, \\
&\mathcal{C}_{X,9} = C_{45}(\tau^4), \qquad \mathcal{C}_{X,13} = \left[576U_\omega^4U_\delta^2m^2 + C_{44}(\tau^2)\right].
\end{aligned}$$

The missing definitions of constants such as C_{31} can be found in the analysis.

Proof. Recall that

$$\omega_{t+1}^k = \Pi_{U_\omega}(\widetilde{\omega}_{t+1}^k), \qquad \mathbf{B}_{t+1} = \bar{\mathbf{B}}_{t+1}\mathbf{R}_{t+1}^{-1}, \qquad \bar{\mathbf{B}}_{t+1} = \mathbf{B}_{t+1}\mathbf{R}_{t+1}.$$

Since $\mathbf{B}_{t+1}$ has orthonormal columns, we have

$$\mathbf{P}_{t+1,\perp} = \mathbf{I} - \mathbf{B}_{t+1}\mathbf{B}_{t+1}^\top.$$

Therefore,

$$\begin{aligned}
\|x_{t+1}^k\|^2 &= \|\mathbf{B}_{t+1}\omega_{t+1}^k - z^{k,*}\|^2 \\
&= \left\|\mathbf{B}_{t+1}\left(\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\right) - \mathbf{P}_{t+1,\perp}z^{k,*}\right\|^2 \\
&= \|\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 + \|\mathbf{P}_{t+1,\perp}z^{k,*}\|^2 \\
&\quad - 2\underbrace{\left\langle \mathbf{B}_{t+1}\left(\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\right), \mathbf{P}_{t+1,\perp}z^{k,*}\right\rangle}_{=0 \text{ because } \mathbf{B}_{t+1}^\top\mathbf{P}_{t+1,\perp}=\mathbf{0}} \\
&= \|\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 + \|\mathbf{P}_{t+1,\perp}z^{k,*}\|^2.
\end{aligned}$$

Moreover,

$$\|\mathbf{B}_{t+1}^\top z^{k,*}\| \le \|z^{k,*}\| = \|\omega^{k,*}\| \le U_\omega.$$

Thus, $\mathbf{B}_{t+1}^\top z^{k,*}$ belongs to the ball onto which Π_{U_ω} projects. The Euclidean projection property gives

$$\left\langle \widetilde{\omega}_{t+1}^k - \omega_{t+1}^k, \mathbf{B}_{t+1}^\top z^{k,*} - \omega_{t+1}^k\right\rangle \le 0.$$

Consequently,

$$\begin{aligned}
&\|\widetilde{\omega}_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 \\
&= \|\widetilde{\omega}_{t+1}^k - \omega_{t+1}^k\|^2 + \|\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 \\
&\quad + 2\left\langle \widetilde{\omega}_{t+1}^k - \omega_{t+1}^k, \omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\right\rangle \\
&\geq \|\widetilde{\omega}_{t+1}^k - \omega_{t+1}^k\|^2 + \|\omega_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2.
\end{aligned}$$

Then, we have

$$\|x_{t+1}^k\|^2 \leq \|\widetilde{\omega}_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 + \|\mathbf{P}_{t+1,\perp} z^{k,*}\|^2.$$

Using the orthogonal decomposition once more,

$$\|\widetilde{\omega}_{t+1}^k - \mathbf{B}_{t+1}^\top z^{k,*}\|^2 + \|\mathbf{P}_{t+1,\perp} z^{k,*}\|^2 = \|\mathbf{B}_{t+1}\widetilde{\omega}_{t+1}^k - z^{k,*}\|^2.$$

Therefore,

$$\begin{aligned}
\|x_{t+1}^k\|^2 &\leq \|\mathbf{B}_{t+1}\widetilde{\omega}_{t+1}^k - z^{k,*}\|^2 \\
&= \|\widetilde{x}_{t+1}^k\|^2 + 2\left\langle \mathbf{B}_{t+1}(\mathbf{I}_r - \mathbf{R}_{t+1})\widetilde{\omega}_{t+1}^k, \widetilde{x}_{t+1}^k\right\rangle \\
&\quad + \|\mathbf{B}_{t+1}(\mathbf{I}_r - \mathbf{R}_{t+1})\widetilde{\omega}_{t+1}^k\|^2. \qquad (28)
\end{aligned}$$

Since $\|\mathbf{Q}_t\|_F^2 \leq \zeta^2 U_\delta^2 U_\omega^2 \leq 1/4$, $\|\mathbf{R}_{t+1}\| \leq 3/2$. Together with $\beta U_\delta \leq U_\omega$, this gives

$$\|\widetilde{\omega}_{t+1}^k\| \leq 2U_\omega, \qquad \|\widetilde{x}_{t+1}^k\| \leq \|\bar{\mathbf{B}}_{t+1}\|\|\widetilde{\omega}_{t+1}^k\| + \|z^{k,*}\| \leq 4U_\omega.$$

The QR correction therefore satisfies

$$\begin{aligned}
&2\left\langle \mathbf{B}_{t+1}(\mathbf{I}_r - \mathbf{R}_{t+1})\widetilde{\omega}_{t+1}^k, \widetilde{x}_{t+1}^k\right\rangle \\
&\leq 2\|\mathbf{I}_r - \mathbf{R}_{t+1}\|_F \|\widetilde{\omega}_{t+1}^k\|\|\widetilde{x}_{t+1}^k\| \\
&\leq 32U_\omega^2\|\mathbf{Q}_t\|_F^2,
\end{aligned}$$

whereas

$$\begin{aligned}
\|\mathbf{B}_{t+1}(\mathbf{I}_r - \mathbf{R}_{t+1})\widetilde{\omega}_{t+1}^k\|^2 &\leq \|\mathbf{I}_r - \mathbf{R}_{t+1}\|_F^2 \|\widetilde{\omega}_{t+1}^k\|^2 \\
&\leq 16U_\omega^2\|\mathbf{Q}_t\|_F^4 \\
&\leq 4U_\omega^2\|\mathbf{Q}_t\|_F^2.
\end{aligned}$$

Combining these bounds with Eq. (28), we obtain

$$\mathbb{E}_{t-\tau}\|x_{t+1}^k\|^2 \leq \mathbb{E}_{t-\tau}\|\widetilde{x}_{t+1}^k\|^2 + 36U_\omega^2\mathbb{E}_{t-\tau}\|\mathbf{Q}_t\|_F^2. \qquad (29)$$

We bound $\|\widetilde{x}_{t+1}^k\|$ as follows:

$$\begin{aligned}
\widetilde{x}_{t+1}^k &= \bar{\mathbf{B}}_{t+1}\widetilde{\omega}_{t+1}^k - z^{k,*} \\
&= \left(\mathbf{B}_t + \frac{\zeta}{K}\sum_{i=1}^K \mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top\right)\left(\omega_t^k + \beta\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)\right) - z^{k,*} \\
&= \mathbf{B}_t\omega_t^k - z^{k,*} + \beta\mathbf{B}_t(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)) + \frac{\zeta}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top\omega_t^k \\
&\quad + \zeta\beta(\frac{1}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)\right).
\end{aligned}$$

Squaring both sides, we have

$$\begin{aligned}\|\widetilde{x}_{t+1}^k\|^2 =&\|\bar{\mathbf{B}}_{t+1}\widetilde{\omega}_{t+1}^k - z^{k,*}\|^2\\ \leq&\|\mathbf{B}_t\omega_t^k - z^{k,*}\|^2 + \underbrace{2\beta\langle\mathbf{B}_t\omega_t^k - z^{k,*}, \mathbf{B}_t\mathbf{B}_t^\top\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\rangle}_{I_1}\\ &\underbrace{+2\zeta\langle\mathbf{B}_t\omega_t^k - z^{k,*}, \frac{1}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^i)\delta_{t,\ell}^i(\omega_t^i)^\top\omega_t^k\rangle}_{I_2}\\ &\underbrace{+2\zeta\beta\left\langle\mathbf{B}_t\omega_t^k - z^{k,*}, (\frac{1}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)\right)\right\rangle}_{I_3}\\ &\underbrace{+3\beta^2\|\mathbf{B}_t(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k))\|^2}_{I_4} + \underbrace{3\zeta^2\|\frac{1}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top\omega_t^k\|^2}_{I_5}\\ &\underbrace{+3\beta^2\zeta^2\|(\frac{1}{K}\sum_{i=1}^K\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^i\phi(s_{t,\ell}^i)(\omega_t^i)^\top)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\mathbf{B}_t^\top\phi(s_{t,\ell}^k)\right)\|^2}_{I_6}. \end{aligned}\tag{30}$$

We bound these terms separately.

Bounding I_1. Roughly speaking, the main drift of the decay in $\|x_t^k\|$ may arise from the head-innovation term I_1 in Eq. (30). However, the environmental heterogeneity significantly complicates the characterization of this term. Specifically, by Proposition 2, I_1 can be written as:

$$\begin{aligned}2\beta\left\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\right\rangle &= 2\beta\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top\xi_t^k\rangle + 2\beta\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top A^k x_t^k\rangle + 2\beta\langle x_t^k, \mathbf{B}_t\mathbf{B}_t^\top y_t^k\bar{\phi}_t^k\rangle\\ &= 2\beta\langle x_t^k, \mathbf{P}_t A^k x_t^k\rangle + \underbrace{2\beta\langle x_t^k, \mathbf{P}_t\xi_t^k\rangle}_{I_{12}} + \underbrace{2\beta\langle x_t^k, \mathbf{P}_t y_t^k\bar{\phi}_t^k\rangle}_{I_{13}},\end{aligned}\tag{31}$$

recalling that $\mathbf{P}_t = \mathbf{B}_t\mathbf{B}_t^\top$. The first term in this display is combined with the PAD term in the proof of Proposition 1.

For I_{13},we have

$$I_{13} \leq 2\beta\|x_t^k\||y_t^k|.\tag{32}$$

Recall

$$\begin{aligned}\widetilde{b}_t^k &= \frac{1}{L}\sum_{\ell=0}^{L-1}(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k);\\ \bar{b}^k &= \mathbb{E}_{s\sim\mu^k, a\sim\pi}\left[(R(s,a) - J^k)\phi(s)\right].\end{aligned}$$

For I_{12}, recall from Eq. (15) that,

$$\xi_t^k = \widetilde{b}_t^k - \bar{b}^k + (\widetilde{A}_t^k - A^k)\mathbf{B}_t\omega_t^k.\tag{33}$$

Then,

$$
\begin{aligned}
&2\beta\left\langle x_t^k, \mathbf{P}_t\xi_t^k\right\rangle\\
=&2\beta\left\langle x_t^k, \mathbf{P}_t\left[\widetilde{b}_t^k-\bar{b}^k+(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right]\right\rangle\\
=&2\beta\left\langle x_t^k, \mathbf{P}_t(\widetilde{b}_t^k-\bar{b}^k)\right\rangle+2\beta\left\langle x_t^k, \mathbf{P}_t(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right\rangle\\
=&\underbrace{2\beta\left\langle x_t^k-x_{t-\tau}^k, \mathbf{P}_t(\widetilde{b}_t^k-\bar{b}^k)\right\rangle}_{I_{121}}+\underbrace{2\beta\left\langle x_{t-\tau}^k, (\mathbf{P}_t-\mathbf{P}_{t-\tau})(\widetilde{b}_t^k-\bar{b}^k)\right\rangle}_{I_{122}}\\
&+\underbrace{2\beta\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}(\widetilde{b}_t^k-\bar{b}^k)\right\rangle}_{I_{123}}\\
&+\underbrace{2\beta\left\langle x_t^k-x_{t-\tau}^k, \mathbf{P}_t(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right\rangle}_{I_{124}}+\underbrace{2\beta\left\langle x_{t-\tau}^k, (\mathbf{P}_t-\mathbf{P}_{t-\tau})(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right\rangle}_{I_{125}}\\
&+\underbrace{2\beta\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}(\widetilde{A}_t^k-A^k)(\mathbf{B}_t\omega_t^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k)\right\rangle}_{I_{126}}\\
&+\underbrace{2\beta\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\right\rangle}_{I_{127}}.
\end{aligned}
$$

For I_{121}, we have under conditional expectation,

$$
\begin{aligned}
&2\beta\mathbb{E}_{t-\tau}\left\langle x_t^k-x_{t-\tau}^k, \mathbf{P}_t\big(\widetilde{b}_t^k-\bar{b}^k\big)\right\rangle\\
\leq&2\beta\mathbb{E}_{t-\tau}\|x_t^k-x_{t-\tau}^k\|\|\mathbf{P}_t\|\|\widetilde{b}_t^k-\bar{b}^k\|\\
\leq&2\beta\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_t^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\|\|\widetilde{b}_t^k-\bar{b}^k\|\\
\leq&\frac{1}{2}\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_t^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\|^2+2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k-\bar{b}^k\|^2\\
\leq&\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_t^k-\mathbf{B}_t\omega_{t-\tau}^k\|^2+\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_{t-\tau}^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\|^2+2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k-\bar{b}^k\|^2\\
\leq&\mathbb{E}_{t-\tau}\|\omega_t^k-\omega_{t-\tau}^k\|^2+U_\omega\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k-\bar{b}^k\|^2.
\end{aligned}
\tag{34}
$$

Similarly, for I_{122}, we have

$$
\begin{aligned}
&2\beta\mathbb{E}_{t-\tau}\left\langle x_{t-\tau}^k, (\mathbf{P}_t-\mathbf{P}_{t-\tau})\left[\widetilde{b}_t^k-\bar{b}^k\right]\right\rangle\\
\leq&2\beta\mathbb{E}_{t-\tau}\|x_{t-\tau}^k\|\|\mathbf{P}_t-\mathbf{P}_{t-\tau}\|\|\widetilde{b}_t^k-\bar{b}^k\| \quad \text{(by Cauchy Schwartz)}\\
\leq&8U_\omega\beta\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|\|\widetilde{b}_t^k-\bar{b}^k\|\\
\leq&\frac{1}{2}\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+32U_\omega^2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k-\bar{b}^k\|^2.
\end{aligned}
\tag{35}
$$

For I_{124},

$$
\begin{aligned}
&2\beta\mathbb{E}_{t-\tau}\left\langle x_t^k-x_{t-\tau}^k, \mathbf{P}_t(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right\rangle\\
\leq&2U_\omega\beta\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_t^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\|\|\widetilde{A}_t^k-A^k\|\\
\leq&\frac{1}{2}\mathbb{E}_{t-\tau}\|\mathbf{B}_t\omega_t^k-\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\|^2+2U_\omega^2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k-A^k\|^2\\
\leq&\mathbb{E}_{t-\tau}\|\omega_t^k-\omega_{t-\tau}^k\|^2+U_\omega\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+2U_\omega^2\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k-A^k\|^2.
\end{aligned}
\tag{36}
$$

For I_{125},

$$
\begin{aligned}
&2\beta\mathbb{E}_{t-\tau}\left\langle x_{t-\tau}^k, (\mathbf{P}_t-\mathbf{P}_{t-\tau})(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\right\rangle\\
\leq&8\beta U_\omega^2\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|\|\widetilde{A}_t^k-A^k\|\\
\leq&\frac{1}{2}\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+32U_\omega^4\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k-A^k\|^2.
\end{aligned}
\tag{37}
$$

For I_{126}, we have,

$$\begin{aligned}
&2\beta\mathbb{E}_{t-\tau}\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}(\widetilde{A}_t^k - A^k)(\mathbf{B}_t\omega_t^k - \mathbf{B}_{t-\tau}\omega_{t-\tau}^k)\right\rangle \\
\leq& 4\beta U_\omega \mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|\|\mathbf{B}_t\omega_t^k - \mathbf{B}_{t-\tau}\omega_{t-\tau}^k\| \\
\leq& 8\beta^2 U_\omega^2 \mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 + \mathbb{E}_{t-\tau}\|\omega_t^k - \omega_{t-\tau}^k\|^2 + U_\omega\mathbb{E}_{t-\tau}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2
\end{aligned} \tag{38}$$

Summing up the upper bounds of I_{121}(Eq. (34)), I_{122}(Eq. (35)), I_{124} (Eq. (36)), I_{125} (Eq. (37)) and I_{126} (Eq. (38)), conditioning on $v_{t-\tau}$, we get,

$$\begin{aligned}
&\mathbb{E}_{t-\tau}\left[I_{121} + I_{122} + I_{124} + I_{125} + I_{126}\right] \\
\leq& 3\mathbb{E}_{t-\tau}\|\omega_t^k - \omega_{t-\tau}^k\|^2 + (1+3U_\omega)\,\mathbb{E}_{t-\tau}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 \\
&+ \left(10U_\omega^2 + 32U_\omega^4\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 + \left(2 + 32U_\omega^2\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k - \bar{b}^k\|^2 \\
\leq& 3\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-\tau}\|\omega_{j+1}^k - \omega_j^k\|^2 + (1+3U_\omega)\,\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-\tau}\|\mathbf{B}_{j+1} - \mathbf{B}_j\|^2 \\
&+ \left(10U_\omega^2 + 32U_\omega^4\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 + \left(2 + 32U_\omega^2\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k - \bar{b}^k\|^2.
\end{aligned} \tag{39}$$

For I_{123}, by Assumption 2 and Lemma K.6, we have

$$\begin{aligned}
2\beta\mathbb{E}_{t-\tau}\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}\left[\widetilde{b}_t^k - \bar{b}^k\right]\right\rangle &\leq 2\beta\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}\mathbb{E}_{t-\tau}\left[\widetilde{b}_t^k - \bar{b}^k\right]\right\rangle \\
&\leq 2\beta\|x_{t-\tau}^k\|\|\mathbf{P}_{t-\tau}\|\|\mathbb{E}_{t-\tau}\left[\widetilde{b}_t^k - \bar{b}^k\right]\| \\
&\leq 2\beta\cdot 2U_\omega\cdot 4U_r\frac{m\rho^{\tau L}}{L(1-\rho)} = 16U_\omega U_r\beta\frac{m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned}$$

For I_{127}, similarly, by Assumption 2 and Lemma K.6, we have

$$2\beta\mathbb{E}_{t-\tau}\left\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau}(\widetilde{A}_t^k - A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k\right\rangle \leq 16U_\omega^2\beta\frac{m\rho^{\tau L}}{L(1-\rho)}.$$

Then, by the tower property, we have

$$\mathbb{E}_{t-\tau}[I_{123} + I_{127}] \leq 16U_\omega(U_r + U_\omega)\beta\frac{m\rho^{\tau L}}{L(1-\rho)}. \tag{40}$$

Putting Eq. (40) and Eq. (39) together, we get

$$\begin{aligned}
\mathbb{E}_{t-\tau}[I_{12}] \leq\ & 3\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-\tau}\|\omega_{j+1}^k - \omega_j^k\|^2 + (1+3U_\omega)\,\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-\tau}\|\mathbf{B}_{j+1} - \mathbf{B}_j\|^2 \\
&+ \left(10U_\omega^2 + 32U_\omega^4\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 \\
&+ \left(2 + 32U_\omega^2\right)\beta^2\mathbb{E}_{t-\tau}\|\widetilde{b}_t^k - \bar{b}^k\|^2 + 16U_\omega(U_r + U_\omega)\beta\frac{m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned} \tag{41}$$

Combining Eq. (41) and Eq. (32) and using the tower property, we obtain

$$\begin{aligned}
\mathbb{E}_{t-2\tau}[I_1] \leq\ & 2\beta\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top\mathbf{P}_t A^k x_t^k\right] + c_4^2\beta\mathbb{E}_{t-2\tau}\|x_t^k\|^2 + \frac{\beta}{c_4^2}\mathbb{E}_{t-2\tau}|y_t^k|^2 \\
&+ 3\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k - \omega_j^k\|^2 + (1+3U_\omega)\,\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-2\tau}\|\mathbf{B}_{j+1} - \mathbf{B}_j\|^2 \\
&+ \left(10U_\omega^2 + 32U_\omega^4\right)\beta^2\mathbb{E}_{t-2\tau}\|\widetilde{A}_t^k - A^k\|^2 + \left(2 + 32U_\omega^2\right)\beta^2\mathbb{E}_{t-2\tau}\|\widetilde{b}_t^k - \bar{b}^k\|^2 \\
&+ 16U_\omega(U_r + U_\omega)\beta\frac{m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned} \tag{42}$$

Bounding I_2. By Proposition 2, I_2 can be written as,

$$\begin{aligned}
&\frac{2\zeta}{KL}\langle x_t^k, \sum_{i=1}^K \mathbf{P}_{t,\perp}\Big(\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^i)\delta_{t,\ell}^i\Big)(\omega_t^i)^\top\omega_t^k\rangle \\
=&\frac{2\zeta}{K}\langle x_t^k, \sum_{i=1}^K \mathbf{P}_{t,\perp}\widetilde{A}_t^i x_t^i(\omega_t^i)^\top\omega_t^k\rangle + \frac{2\zeta}{K}\langle x_t^k, \sum_{i=1}^K \mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\rangle \\
&+ \frac{2\zeta}{K}\langle x_t^k, \sum_{i=1}^K \mathbf{P}_{t,\perp}y_t^i\bar{\phi}_t^i(\omega_t^i)^\top\omega_t^k\rangle \\
\leq&\frac{4U_\omega^2\zeta}{K}\sum_{i=1}^K\|x_t^k\|\|x_t^i\| + \underbrace{\frac{2\zeta}{K}\langle x_t^k, \sum_{i=1}^K \mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\rangle}_{I_{21}} + \frac{2U_\omega^2\zeta}{K}\sum_{i=1}^K\|x_t^k\|\|y_t^i\|.
\end{aligned} \tag{43}$$

For I_{21}, similar to the derivations of the upper bounds of terms from I_{121} to I_{126}, we have

$$\begin{aligned}
&\frac{2\zeta}{K}\sum_{i=1}^K\langle x_t^k, \mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\rangle \\
=&\frac{2\zeta}{K}\sum_{i=1}^K\langle x_t^k - x_{t-\tau}^k, \mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\rangle + \frac{2\zeta}{K}\sum_{i=1}^K\langle x_{t-\tau}^k, (\mathbf{P}_{t,\perp} - \mathbf{P}_{t-\tau,\perp})\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\rangle \\
&+ \frac{2\zeta}{K}\sum_{i=1}^K\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau,\perp}\mathbf{b}_t^i(\omega_t^i - \omega_{t-\tau}^i)^\top\omega_t^k\rangle \\
&+ \frac{2\zeta}{K}\sum_{i=1}^K\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau,\perp}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top(\omega_t^k - \omega_{t-\tau}^k)\rangle \\
&+ \frac{2\zeta}{K}\sum_{i=1}^K\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau,\perp}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\omega_{t-\tau}^k\rangle \\
\leq&\frac{3}{2}\|\omega_t^k - \omega_{t-\tau}^k\|^2 + \frac{3}{2}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 + \big(18U_\omega^4 + 32U_\omega^6\big)\frac{\zeta^2}{K}\sum_{i=1}^K\|\mathbf{b}_t^i\|^2 \\
&\frac{1}{2}\frac{1}{K}\sum_{i=1}^K\|\omega_t^i - \omega_{t-\tau}^i\|^2 + \frac{2\zeta}{K}\sum_{i=1}^K\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau,\perp}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\omega_{t-\tau}^k\rangle.
\end{aligned} \tag{44}$$

For the last term in Eq. (44), by Cauchy–Schwarz inequality and Lemma K.6, we have

$$\begin{aligned}
\frac{2\zeta}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\langle x_{t-\tau}^k, \mathbf{P}_{t-\tau,\perp}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\omega_{t-\tau}^k\rangle &\leq \frac{2\zeta}{K}\sum_{i=1}^K 2U_\omega^3\|\mathbb{E}_{t-\tau}[\mathbf{b}_t^i]\| \\
&\leq \frac{8U_\delta U_\omega^3\zeta m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned} \tag{45}$$

Then, putting Eq. (45) back to Eq. (44) and then putting the result back to Eq. (43) and using tower property, we have

$$\begin{aligned}
\mathbb{E}_{t-\tau}[I_2] \leq\ & 4U_\omega^2\frac{\zeta}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\|x_t^k\|\|x_t^i\| + 2U_\omega^2\frac{\zeta}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\|x_t^k\|\|y_t^i\| + \frac{3}{2}\mathbb{E}_{t-\tau}\|\omega_t^k - \omega_{t-\tau}^k\|^2 \\
&+ \frac{3}{2}\mathbb{E}_{t-\tau}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 + \big(18U_\omega^4 + 32U_\omega^6\big)\frac{\zeta^2}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\|\mathbf{b}_t^i\|^2 + \frac{1}{2}\frac{1}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\|\omega_t^i - \omega_{t-\tau}^i\|^2 \\
&+ \frac{8U_\delta U_\omega^3\zeta m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned} \tag{46}$$

Applying Young's inequality, for arbitrary $c_5 > 0$ and $c_6 > 0$, we have

$$4U_\omega^2 \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|x_t^k\| \|x_t^i\| \leq c_5^2 \zeta \mathbb{E}_{t-\tau} \|x_t^k\|^2 + \frac{4U_\omega^4}{c_5^2} \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|x_t^i\|^2,$$
$$2U_\omega^2 \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|x_t^k\| |y_t^i| \leq c_6^2 \zeta \mathbb{E}_{t-\tau} \|x_t^k\|^2 + \frac{U_\omega^4}{c_6^2} \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} |y_t^i|^2.$$

We get

$$\begin{aligned}
\mathbb{E}_{t-\tau}[I_2] \leq\ & c_5^2 \zeta \mathbb{E}_{t-\tau} \|x_t^k\|^2 + \frac{4U_\omega^4}{c_5^2} \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|x_t^i\|^2 + c_6^2 \zeta \mathbb{E}_{t-\tau} \|x_t^k\|^2 + \frac{U_\omega^4}{c_6^2} \frac{\zeta}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} |y_t^i|^2 \\
& + \frac{3}{2} \mathbb{E}_{t-\tau} \|\omega_t^k - \omega_{t-\tau}^k\|^2 + \frac{3}{2} \mathbb{E}_{t-\tau} \|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 + \left(18U_\omega^4 + 32U_\omega^6\right) \frac{\zeta^2}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|\mathbf{b}_t^i\|^2 \\
& + \frac{1}{2} \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{t-\tau} \|\omega_t^i - \omega_{t-\tau}^i\|^2 + \frac{8U_\delta U_\omega^3 \zeta m \rho^{\tau L}}{L(1-\rho)}.
\end{aligned} \tag{47}$$

For I_3, by Proposition 2, we have

$$\begin{aligned}
I_3 =& \underbrace{\left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \frac{1}{L} \sum_{\ell=0}^{L-1} \delta_{t,\ell}^i \phi(s_{t,\ell}^i)(\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \widetilde{A}_t^k x_t^k \right) \right\rangle}_{I_{31}} \\
& + \underbrace{\left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \frac{1}{L} \sum_{\ell=0}^{L-1} \delta_{t,\ell}^i \phi(s_{t,\ell}^i)(\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top y_t^k \bar{\phi}_t^k \right) \right\rangle}_{I_{32}} \\
& + \underbrace{\left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \frac{1}{L} \sum_{\ell=0}^{L-1} \delta_{t,\ell}^i \phi(s_{t,\ell}^i)(\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \mathbf{b}_t^k \right) \right\rangle}_{I_{33}}.
\end{aligned}$$

By Lemma J.2, I_{31} and I_{32} can be bounded as

$$I_{31} \leq 2U_\delta U_\omega \|x_t^k\|^2, \tag{48}$$
$$I_{32} \leq U_\delta U_\omega \|x_t^k\| |y_t^k|. \tag{49}$$

For I_{33}, we have,

$$\begin{aligned}
I_{33} =& \left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \widetilde{A}_t^i x_t^i (\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \mathbf{b}_t^k \right) \right\rangle \\
& + \left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} y_t^i \bar{\phi}_t^i (\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \mathbf{b}_t^k \right) \right\rangle \\
& + \left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \mathbf{b}_t^i (\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \mathbf{b}_t^k \right) \right\rangle \\
\leq & 2U_\delta U_\omega \frac{1}{K} \sum_{i=1}^{K} \|x_t^k\| \|x_t^i\| + U_\delta U_\omega \frac{1}{K} \sum_{i=1}^{K} \|x_t^k\| |y_t^i| \\
& + \left\langle x_t^k, \left(\frac{1}{K} \sum_{i=1}^{K} \mathbf{P}_{t,\perp} \mathbf{b}_t^i (\omega_t^i)^\top \right) \left(\mathbf{B}_t^\top \mathbf{b}_t^k \right) \right\rangle.
\end{aligned} \tag{50}$$

For the last term, we have,

$$
\begin{aligned}
&\mathbb{E}_{t-\tau}\left\langle x_t^k, \left(\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\right)\left(\mathbf{B}_t^\top\mathbf{b}_t^k\right)\right\rangle \\
=&\mathbb{E}_{t-\tau}(x_t^k - x_{t-\tau}^k)^\top\mathbf{P}_{t,\perp}\frac{1}{K}\sum_{i=1}^{K}\mathbf{b}_t^i(\omega_t^i)^\top\left(\mathbf{B}_t^\top\mathbf{b}_t^k\right) \\
&+\mathbb{E}_{t-\tau}(x_{t-\tau}^k)^\top(\mathbf{P}_{t,\perp}-\mathbf{P}_{t-\tau,\perp})\frac{1}{K}\sum_{i=1}^{K}\mathbf{b}_t^i(\omega_t^i)^\top\left(\mathbf{B}_t^\top\mathbf{b}_t^k\right) \\
&+\mathbb{E}_{t-\tau}(x_{t-\tau}^k)^\top\mathbf{P}_{t-\tau,\perp}\frac{1}{K}\sum_{i=1}^{K}\mathbf{b}_t^i(\omega_t^i-\omega_{t-\tau}^i)^\top\left(\mathbf{B}_t^\top\mathbf{b}_t^k\right) \\
&+\mathbb{E}_{t-\tau}(x_{t-\tau}^k)^\top\mathbf{P}_{t-\tau,\perp}\frac{1}{K}\sum_{i=1}^{K}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\left((\mathbf{B}_t-\mathbf{B}_{t-\tau})^\top\mathbf{b}_t^k\right) \\
&+\mathbb{E}_{t-\tau}(x_{t-\tau}^k)^\top\mathbf{P}_{t-\tau,\perp}\frac{1}{K}\sum_{i=1}^{K}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\left(\mathbf{B}_{t-\tau}^\top\mathbf{b}_t^k\right) \quad \text{(bridging)} \\
\leq&\mathbb{E}_{t-\tau}\|\omega_t^k-\omega_{t-\tau}^k\|U_\delta^2U_\omega+7U_\delta^2U_\omega^2\mathbb{E}_{t-\tau}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|+2U_\omega U_\delta^2\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|\omega_t^i-\omega_{t-\tau}^i\| \\
&+2U_\omega^2U_\delta^2\frac{1}{K}+\mathbb{E}_{t-\tau}(x_{t-\tau}^k)^\top\mathbf{P}_{t-\tau,\perp}\frac{1}{K}\sum_{i\neq k}\mathbf{b}_t^i(\omega_{t-\tau}^i)^\top\left(\mathbf{B}_{t-\tau}^\top\mathbf{b}_t^k\right) \\
\leq&3\tau U_\omega U_\delta^3\beta+14U_\delta^3U_\omega^3\tau\zeta+28U_\delta^4U_\omega^4\tau\zeta^2+2U_\omega^2U_\delta^2\frac{1}{K}+8U_\omega^2U_\delta^2m^2\frac{\rho^{2\tau L}}{L^2(1-\rho)^2},
\end{aligned}
$$

where the last inequality follows from (a) the independence across agents, (b) Assumption 2, and (c) Lemma K.1.

Putting this bound back to I_{33} (Eq. (50)), and taking conditional expectation, we get,

$$
\begin{aligned}
\mathbb{E}_{t-\tau}[I_{33}] \leq\ & 2U_\delta U_\omega\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|x_t^k\|\|x_t^i\|+U_\delta U_\omega\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|x_t^k\||y_t^i| \\
&+3\tau U_\omega U_\delta^3\beta+14U_\delta^3U_\omega^3\tau\zeta+28U_\delta^4U_\omega^4\tau\zeta^2+2U_\omega^2U_\delta^2\frac{1}{K}+8U_\omega^2U_\delta^2m^2\frac{\rho^{2\tau L}}{L^2(1-\rho)^2}. \qquad (51)
\end{aligned}
$$

Putting I_{31} (Eq. (48)), I_{32} (Eq. (49)), and I_{33} (Eq. (51)) back to I_3, and taking expectation, we have

$$\begin{aligned}
\mathbb{E}_{t-\tau}[I_3] \leq\ & 2\zeta\beta\Bigg(2U_\delta U_\omega \mathbb{E}_{t-\tau}\|x_t^k\|^2 + U_\delta U_\omega \mathbb{E}_{t-\tau}\|x_t^k\||y_t^k| \\
& + 2U_\delta U_\omega \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^k\|\|x_t^i\| + U_\delta U_\omega \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^k\||y_t^i| \\
& + 3\tau U_\omega U_\delta^3 \beta + 14 U_\delta^3 U_\omega^3 \tau\zeta + 28 U_\delta^4 U_\omega^4 \tau\zeta^2 + 2U_\omega^2 U_\delta^2 \frac{1}{K} + 8U_\omega^2 U_\delta^2 m^2 \frac{\rho^{2\tau L}}{L^2(1-\rho)^2}\Bigg) \\
\leq\ & 4U_\delta U_\omega \zeta\beta \mathbb{E}_{t-\tau}\|x_t^k\|^2 + 2U_\delta U_\omega \zeta\beta \mathbb{E}_{t-\tau}\|x_t^k\||y_t^k| \\
& + 4U_\delta U_\omega \zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^k\|\|x_t^i\| + 2U_\delta U_\omega \zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^k\||y_t^i| \\
& + 6\tau U_\omega U_\delta^3 \zeta\beta^2 + 28 U_\delta^3 U_\omega^3 \tau\zeta^2\beta + 56 U_\delta^4 U_\omega^4 \tau\zeta^3\beta \\
& + 4U_\delta^2 U_\omega^2 \frac{\zeta\beta}{K} + 16 U_\delta^2 U_\omega^2 m^2 \zeta\beta \frac{\rho^{2\tau L}}{L^2(1-\rho)^2} \\
\leq\ & \underbrace{8U_\delta U_\omega}_{:=C_{31}} \zeta\beta \mathbb{E}_{t-\tau}\|x_t^k\|^2 + \underbrace{U_\delta U_\omega}_{:=C_{32}} \zeta\beta \mathbb{E}_{t-\tau}|y_t^k|^2 \\
& + \underbrace{2U_\delta U_\omega}_{:=C_{33}} \zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^i\|^2 + \underbrace{U_\delta U_\omega}_{:=C_{34}} \zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}|y_t^i|^2 \\
& + \underbrace{6\tau U_\omega U_\delta^3}_{:=C_{35}(\tau)} \zeta\beta^2 + \underbrace{28 U_\delta^3 U_\omega^3 \tau}_{C_{36}(\tau)} \zeta^2\beta + \underbrace{56 U_\delta^4 U_\omega^4 \tau}_{C_{37}(\tau)} \zeta^3\beta + \underbrace{4U_\delta^2 U_\omega^2}_{C_{38}} \frac{\zeta\beta}{K} \\
& + 16 U_\delta^2 U_\omega^2 m^2 \zeta\beta \frac{\rho^{2\tau L}}{L^2(1-\rho)^2} \quad \text{(Young's inequality)} \\
=\ & C_{31}\zeta\beta \mathbb{E}_{t-\tau}\|x_t^k\|^2 + C_{32}\zeta\beta \mathbb{E}_{t-\tau}|y_t^k|^2 + C_{33}\zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}\|x_t^i\|^2 + C_{34}\zeta\beta \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{t-\tau}|y_t^i|^2 \\
& + C_{35}(\tau)\zeta\beta^2 + C_{36}(\tau)\zeta^2\beta + C_{37}(\tau)\zeta^3\beta + C_{38}\frac{\zeta\beta}{K} + 16U_\omega^2 U_\delta^2 m^2 \zeta\beta \frac{\rho^{2\tau L}}{L^2(1-\rho)^2}. \qquad (52)
\end{aligned}$$

For I_4, we have,

$$\begin{aligned}
& \|\mathbf{B}_t \frac{1}{L}\sum_{\ell=0}^{L-1} \delta_{t,\ell}^k \mathbf{B}_t^\top \phi(s_{t,\ell}^k)\|^2 \\
\leq\ & 3\|\mathbf{B}_t\mathbf{B}_t^\top(\widetilde{A}_t^k x_t^k)\|^2 + 3\|\mathbf{B}_t\mathbf{B}_t^\top y_t^k \bar{\phi}_t^k\|^2 + 3\|\mathbf{B}_t\mathbf{B}_t^\top \mathbf{b}_t^k\|^2 \quad \text{(by Proposition 2)} \\
\leq\ & 12\|x_t^k\|^2 + 3|y_t^k|^2 + 3\|\mathbf{b}_t^k\|^2. \qquad (53)
\end{aligned}$$

From Lemma K.5, we know

$$\mathbb{E}_{t-\tau}\|\mathbf{b}_t^k\|^2 \leq \frac{U_\delta^2}{L}. \qquad (54)$$

Then, taking conditional expectation on Eq. (53), we get

$$\mathbb{E}_{t-\tau}\|\mathbf{B}_t \frac{1}{L}\sum_{\ell=0}^{L-1} \delta_{t,\ell}^k \mathbf{B}_t^\top \phi(s_{t,\ell}^k)\|^2 \leq 12\mathbb{E}_{t-\tau}\|x_t^k\|^2 + 3\mathbb{E}_{t-\tau}|y_t^k|^2 + 3\frac{U_\delta^2}{L}. \qquad (55)$$

Therefore, we have

$$\mathbb{E}_{t-\tau}[I_4] \leq 36\beta^2 \mathbb{E}_{t-\tau}\|x_t^k\|^2 + 9\beta^2 \mathbb{E}_{t-\tau}|y_t^k|^2 + 9U_\delta^2 \frac{\beta^2}{L}. \qquad (56)$$

For I_5, similarly, under conditional expectation, we have

$$\begin{aligned}
&\mathbb{E}_{t-\tau}\|\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^{i}\phi(s_{t,\ell}^{i})(\omega_t^i)^\top\omega_t^k\|^2\\
\leq&3\mathbb{E}_{t-\tau}\|\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}\widetilde{A}_t^i x_t^i(\omega_t^i)^\top\omega_t^k\|^2+3\mathbb{E}_{t-\tau}\|\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}y_t^i\bar{\phi}_t^i(\omega_t^i)^\top\omega_t^k\|^2\\
&+3\mathbb{E}_{t-\tau}\|\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}\mathbf{b}_t^i(\omega_t^i)^\top\omega_t^k\|^2\\
\leq&12U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|x_t^i\|^2+3U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}|y_t^i|^2+3U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|\mathbf{b}_t^i\|^2\\
\leq&12U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|x_t^i\|^2+3U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}|y_t^i|^2+3U_\omega^4\frac{U_\delta^2}{L}. \quad \text{(by Eq. (54))}
\end{aligned}$$

Then, we have

$$\mathbb{E}_{t-\tau}[I_5]\leq 36U_\omega^4\zeta^2\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}\|x_t^i\|^2+9\zeta^2U_\omega^4\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-\tau}|y_t^i|^2+9U_\omega^4U_\delta^2\frac{\zeta^2}{L}. \tag{57}$$

For I_6, we have,

$$I_6\leq 3\beta^2\zeta^2\|(\frac{1}{K}\sum_{i=1}^{K}\mathbf{P}_{t,\perp}\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^{i}\phi(s_{t,\ell}^{i})(\omega_t^i)^\top)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^{k}\mathbf{B}_t^\top\phi(s_{t,\ell}^{k})\right)\|^2\leq 3U_\delta^4U_\omega^2\beta^2\zeta^2. \tag{58}$$

Applying Lemma K.5, Lemma K.2, Lemma K.3, using $\beta=c\zeta$ and the tower property, we further get,

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\widetilde{x}_{t+1}^k\|^2\leq 2\beta\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top\mathbf{P}_tA^kx_t^k\right]+\left(1+c_4^2c\zeta+c_5^2\zeta+c_6^2\zeta+C_{31}c\zeta^2+(36+108\tau^2)c^2\zeta^2\right)\mathbb{E}_{t-2\tau}\|x_t^k\|^2\\
&+\left(\frac{1}{c_4^2}c\zeta+C_{32}c\zeta^2+(9+27\tau^2)c^2\zeta^2\right)\mathbb{E}_{t-2\tau}|y_t^k|^2\\
&+\left(\frac{4U_\omega^4}{c_5^2}\zeta+C_{33}c\zeta^2+\left(36U_\omega^4+\left(\frac{5}{2}+3U_\omega\right)\tau^2C_{18}\right)\zeta^2\right)\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^i\|^2\\
&+\left(\frac{U_\omega^4}{c_6^2}\zeta+C_{34}c\zeta^2+\left(9U_\omega^4+\left(\frac{5}{2}+3U_\omega\right)\tau^2C_{23}\right)\zeta^2\right)\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{t-2\tau}|y_t^i|^2\\
&+C_{39}(\tau^2)\frac{\zeta^2}{L}+C_{41}(\tau^2)\frac{\zeta^2}{K}+C_{42}\zeta\frac{\rho^{\tau L}}{L(1-\rho)}+C_{43}(\tau)\zeta^3+C_{44}(\tau^2)\zeta^2\frac{\rho^{2\tau L}}{L^2(1-\rho)^2}+C_{40}(\tau^4)\zeta^4+C_{45}(\tau^4)\gamma^2\zeta^2,
\end{aligned} \tag{59}$$

where

$$C_{39}(\tau^2) = \Bigg[\left(10U_\omega^2 + 32U_\omega^4\right)\left(16 + 32\frac{m}{1-\rho}\right)c^2 + \left(2 + 32U_\omega^2\right)\left(16U_r^2 + 32U_r^2\frac{m}{1-\rho}\right)c^2 + \left(18U_\omega^4 + 32U_\omega^6\right)U_\delta^2 + 9U_\delta^2c^2 + 9U_\omega^4U_\delta^2 + 5\tau^2C_7c^2\Bigg],$$

$$C_{40}(\tau^4) = \Bigg[C_{37}(\tau)c + 3U_\delta^4U_\omega^2c^2 + 5\tau^2C_5(\tau^2)c^4 + 5\tau^2C_6(\tau^2)c^2 + \left(\frac{5}{2} + 3U_\omega\right)\tau^2\left(C_{16}(\tau^2)c^2 + C_{17}(\tau^2) + C_{19}(\tau^2)c^2 + C_{24}\right)\Bigg]$$

$$C_{41}(\tau^2) = \left[\left(\frac{5}{2} + 3U_\omega\right)\tau^2C_{20} + C_{38}c\right], C_{42} = \left[16U_\omega(U_r + U_\omega)mc + 8U_\delta U_\omega^3m\right],$$

$$C_{43}(\tau) = \left[C_{35}(\tau)c^2 + C_{36}(\tau)c\right], C_{44}(\tau^2) = \left[16U_\omega^2U_\delta^2m^2c + \left(\frac{5}{2} + 3U_\omega\right)\tau^2C_{21}m^2\right],$$

$$C_{45}(\tau^4) = \left[5\tau^2C_8(\tau^2)c^2 + \left(\frac{5}{2} + 3U_\omega\right)\tau^2C_{22}(\tau^2)\right].$$

Invoking Eq. (26) in Eq. (29), we obtain

$$\begin{aligned}\mathbb{E}_{t-\tau}\|x_{t+1}^k\|^2 \le \mathbb{E}_{t-\tau}\|\widetilde{x}_{t+1}^k\|^2 &+ 576U_\omega^4\zeta^2\frac{1}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}\|x_t^i\|^2 \\ &+ 144U_\omega^4\zeta^2\frac{1}{K}\sum_{i=1}^K\mathbb{E}_{t-\tau}|y_t^i|^2 \\ &+ 144\frac{U_\delta^2}{L}U_\omega^4\frac{\zeta^2}{K} + 576U_\omega^4U_\delta^2\zeta^2\frac{m^2\rho^{2\tau L}}{L^2(1-\rho)^2} \\ &+ 144U_\omega^2\tau^2U_\delta^4\zeta^2\beta^2. \end{aligned} \tag{60}$$

Applying $\mathbb{E}_{t-2\tau}$ to Eq. (60) and using the tower property to bring it to the same conditioning level as Eq. (59), then taking the average over agents, invoking Eq. (59), and using $\beta = c\zeta$ and $\tau = \lceil 2\log_\rho(\zeta)\rceil$, we obtain

$$\begin{aligned}\widehat{X}_{t+1} \le \widehat{X}_t &+ \frac{2c\zeta}{K}\sum_{k=1}^K\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top\mathbf{P}_tA^kx_t^k\right] \\ &+ \left(c_4^2c + c_5^2 + c_6^2 + \frac{4U_\omega^4}{c_5^2}\right)\zeta\widehat{X}_t + \mathcal{C}_{X,1}(\tau^2)\zeta^2\widehat{X}_t \\ &+ \left[\left(\frac{c}{c_4^2} + \frac{U_\omega^4}{c_6^2}\right)\zeta + \mathcal{C}_{X,3}(\tau^2)\zeta^2\right]\widehat{Y}_t \\ &+ \mathcal{C}_{X,5}(\tau^2)\frac{\zeta^2}{L} + \mathcal{C}_{X,6}(\tau^2)\frac{\zeta^2}{K} + \mathcal{C}_{X,7}\zeta^3 + \mathcal{C}_{X,8}\zeta^4 + \mathcal{C}_{X,9}\gamma^2\zeta^2 + \mathcal{C}_{X,13}\zeta^6, \end{aligned} \tag{61}$$

where

$$\begin{aligned}&\mathcal{C}_{X,1}(\tau^2) = C_{31}c + (36 + 108\tau^2)c^2 + C_{33}c + \left(36U_\omega^4 + (2.5 + 3U_\omega)\tau^2C_{18}\right) + 576U_\omega^4, \\ &\mathcal{C}_{X,3}(\tau^2) = C_{32}c + (9 + 27\tau^2)c^2 + C_{34}c + 9U_\omega^4 + (2.5 + 3U_\omega)\tau^2C_{23} + 144U_\omega^4, \\ &\mathcal{C}_{X,5}(\tau^2) = C_{39}(\tau^2), \mathcal{C}_{X,6}(\tau^2) = C_{41}(\tau^2) + 144\frac{U_\delta^2}{L}U_\omega^4, \\ &\mathcal{C}_{X,7} = C_{42} + C_{43}(\tau), \qquad \mathcal{C}_{X,8} = C_{40}(\tau^4) + 144U_\omega^2\tau^2U_\delta^4c^2, \\ &\mathcal{C}_{X,9} = C_{45}(\tau^4), \qquad \mathcal{C}_{X,13} = \left[576U_\omega^4U_\delta^2m^2 + C_{44}(\tau^2)\right].\end{aligned}$$

□

E.2 LOWER BOUND: PROOF OF LEMMA 5.3

Recall that $x_t^k = \mathbf{B}_t\omega_t^k - z^{k,*} = \mathbf{B}_t\omega_t^k - \mathbf{B}^*\omega^{k,*}$.

By definition, for each $k \in [K]$, $\mathbf{B}_t\omega_t^k$ lies in the subspace spanned by $\mathbf{B}_t$. As illustrated in Fig.3, the distance between $\mathbf{B}_t\omega_t^k$ and $z^{k,*}$ is no smaller than the distance between $z^{k,*}$ and the subspace spanned by $\mathbf{B}_t$. Hence,

$$\begin{aligned}\|x_t^k\| &= \|\mathbf{B}_t\omega_t^k - z^{k,*}\| \geq \|\mathbf{P}_{t,\perp} z^{k,*}\| \\ &= \|(\mathbf{I} - \mathbf{P}_t)z^{k,*}\| = \|(\mathbf{P}^* - \mathbf{P}_t)z^{k,*}\| \\ &= \|(\mathbf{P}_t - \mathbf{P}^*)z^{k,*}\|.\end{aligned}$$

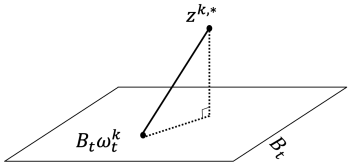


Figure 3: Geometry illustration.

Since the k-th column of $\mathbf{Z}^*$ is $z^{k,*} = \mathbf{B}^*\omega^{k,*}$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\|x_t^k\|^2 &\geq \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{P}_{t,\perp} z^{k,*}\|^2 \\ &= \frac{1}{K}\|\mathbf{P}_{t,\perp}\mathbf{Z}^*\|_F^2 \\ &= \frac{1}{K}\operatorname{tr}\left((\mathbf{P}_{t,\perp}\mathbf{Z}^*)^\top(\mathbf{P}_{t,\perp}\mathbf{Z}^*)\right) \\ &= \frac{1}{K}\operatorname{tr}\left(\mathbf{Z}^{*\top}\mathbf{P}_{t,\perp}\mathbf{Z}^*\right) \\ &= \operatorname{tr}\left(\mathbf{P}_{t,\perp}\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K}\right) \\ &= \operatorname{tr}\left[\mathbf{P}_{t,\perp}\left(\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K} - \alpha\mathbf{P}^*\right)\right] + \alpha\operatorname{tr}(\mathbf{P}_{t,\perp}\mathbf{P}^*).\end{aligned}$$

Moreover,

$$\begin{aligned}\operatorname{tr}(\mathbf{P}_{t,\perp}\mathbf{P}^*) &= \operatorname{tr}\left(\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\mathbf{B}^*\mathbf{B}^{*\top}\right) \\ &= \operatorname{tr}\left(\mathbf{B}^{*\top}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\mathbf{B}^*\right) \\ &= \|\mathbf{B}_\perp^{*\top}\mathbf{B}_t\|_F^2 \\ &= \|m_t\|_F^2.\end{aligned}$$

It remains to show that the first trace, $\operatorname{tr}\left[\mathbf{P}_{t,\perp}\left(\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K} - \alpha\mathbf{P}^*\right)\right]$, is nonnegative.

By the full-span setting,

$$\operatorname{range}(\mathbf{Z}^*) = \operatorname{range}(\mathbf{B}^*), \qquad \operatorname{rank}(\mathbf{Z}^*) = r.$$

Let $u_1, \ldots, u_r$ be orthonormal eigenvectors of $\mathbf{Z}^*\mathbf{Z}^{*\top}/K$ corresponding to its r positive eigenvalues $\lambda_1, \ldots, \lambda_r$. The coverage assumption gives $\lambda_i \geq \alpha$ for every $i \in [r]$. For every $v \in \mathbb{R}^d$,

$$\begin{aligned}v^\top\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K}v &= \sum_{i=1}^{r}\lambda_i(u_i^\top v)^2 \\ &\geq \alpha\sum_{i=1}^{r}(u_i^\top v)^2 \\ &= \alpha\|\mathbf{P}^*v\|^2 \\ &= v^\top(\alpha\mathbf{P}^*)v.\end{aligned}$$

Thus,

$$\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K} \succeq \alpha\mathbf{P}^*.$$

Let $q_1, \ldots, q_{d-r}$ be an orthonormal basis of $\mathrm{range}(\mathbf{P}_{t,\perp})$. Since $\mathbf{P}_{t,\perp}$ is the orthogonal projector onto this space,

$$\mathbf{P}_{t,\perp} = \sum_{i=1}^{d-r} q_i q_i^\top.$$

Therefore,

$$\begin{aligned}
&\mathrm{tr}\left[\mathbf{P}_{t,\perp}\left(\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K} - \alpha\mathbf{P}^*\right)\right] \\
&= \sum_{i=1}^{d-r} q_i^\top \left(\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K} - \alpha\mathbf{P}^*\right) q_i \\
&\geq 0.
\end{aligned}$$

Consequently,

$$\frac{1}{K}\sum_{k=1}^{K}\|x_t^k\|^2 \geq \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{P}_{t,\perp} z^{k,*}\|^2 \geq \mathrm{tr}\left(\mathbf{P}_{t,\perp}\frac{\mathbf{Z}^*\mathbf{Z}^{*\top}}{K}\right) \geq \alpha\|m_t\|_F^2.$$

F Proof of Lemma 5.2 (Bounds on Principal Angle Distance)

Recall from Algorithm 1 of the update of $\bar{\mathbf{B}}$ that $\bar{\mathbf{B}}_{t+1} = \mathbf{B}_t + \frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top$. Since $\mathbf{B}_{t+1} = \bar{\mathbf{B}}_{t+1}\mathbf{R}_{t+1}^{-1}$, we have

$$\|\mathbf{B}_\perp^{*\top}\mathbf{B}_{t+1}\|_F^2 = \|\mathbf{B}_\perp^{*\top}\bar{\mathbf{B}}_{t+1}\mathbf{R}_{t+1}^{-1}\|_F^2 \le \|\mathbf{B}_\perp^{*\top}\bar{\mathbf{B}}_{t+1}\|_F^2,$$

where the inequality follows from Eq. (21). That is,

$$\|m_{t+1}\|_F^2 \le \|\bar{m}_{t+1}\|_F^2.$$

For $\|\bar{m}_{t+1}\|_F^2$, we have

$$\begin{aligned}
\|\bar{m}_{t+1}\|_F^2 &= \|\mathbf{B}_\perp^{*\top}\bar{\mathbf{B}}_{t+1}\|_F^2 = \|\mathbf{B}_\perp^{*\top}\mathbf{B}_t + \frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2 \\
&= \|\mathbf{B}_\perp^{*\top}\mathbf{B}_t\|_F^2 + 2\langle\mathbf{B}_\perp^{*\top}\mathbf{B}_t, \frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\rangle \\
&\quad + \|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2 \\
&= \|m_t\|_F^2 + \underbrace{\|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2}_{I.1} + 2\langle\mathbf{B}_\perp^{*\top}\mathbf{B}_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top\rangle \\
&\quad \underbrace{+ 2\langle\mathbf{B}_\perp^{*\top}\mathbf{B}_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}y_t^k\bar{\phi}_t^k(\omega_t^k)^\top\rangle}_{I.2} + 2\left\langle\mathbf{B}_\perp^{*\top}\mathbf{B}_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}A^k x_t^k(\omega_t^k)^\top\right\rangle,
\end{aligned} \tag{62}$$

where the last equality follows from Proposition 2.

For I.1 in Eq. (62), recalling that $\|\mathbf{Q}_t\|_F^2 = \|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{P}_{t,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2$, we have

$$\begin{aligned}
I.1 &= \|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2 \\
&= \|\mathbf{B}_\perp^{*\top}\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{P}_{t,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big)(\omega_t^k)^\top\|_F^2 \\
&\le \|\mathbf{B}_\perp^{*\top}\|^2\|\mathbf{Q}_t\|_F^2 \\
&\le \|\mathbf{Q}_t\|_F^2.
\end{aligned}$$

Similar to the derivation in the proof of Lemma 5.1, I.2 in Eq. (62) can be bounded as

$$\begin{aligned}
I.2 &= 2\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}y_t^k\bar{\phi}_t^k(\omega_t^k)^\top\rangle \\
&\le 2\|m_t\|\frac{\zeta}{K}\sum_{k=1}^{K}\|\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}y_t^k\bar{\phi}_t^k(\omega_t^k)^\top\| \le 2U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\|m_t\||y_t^k|.
\end{aligned}$$

For the last inner product in Eq. (62), the trace identity gives

$$2\left\langle\mathbf{B}_\perp^{*\top}\mathbf{B}_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}A^k x_t^k(\omega_t^k)^\top\right\rangle = \frac{2\zeta}{K}\sum_{k=1}^{K}\left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k. \tag{63}$$

Here, we use

$$\left\langle AB, vw^\top \right\rangle = \text{trace}\left(A^\top vw^\top\right) = w^\top B^\top A^\top v$$

and $\mathbf{P}_{t,\perp}\mathbf{P}_\perp^*\mathbf{B}_t\omega_t^k = \mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k$. Combining the bounds of the first two inner products in Eq. (62) with Eq. (63), we have

$$\begin{aligned}\|\bar{m}_{t+1}\|_F^2 \le \|m_t\|_F^2 + \frac{2\zeta}{K}\sum_{k=1}^K \left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k + \|\mathbf{Q}_t\|_F^2 \\ + 2\left\langle m_t, \frac{\zeta}{K}\sum_{k=1}^K \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top \right\rangle + \frac{2U_\omega\zeta}{K}\sum_{k=1}^K \|m_t\||y_t^k|. \end{aligned} \tag{64}$$

By Eq. (21) and Eq. (64), we have

$$\begin{aligned}\|m_{t+1}\|_F^2 \le \|m_t\|_F^2 + \frac{2\zeta}{K}\sum_{k=1}^K \left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k + \|\mathbf{Q}_t\|_F^2 \\ + 2\left\langle m_t, \frac{\zeta}{K}\sum_{k=1}^K \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top \right\rangle + \frac{2U_\omega\zeta}{K}\sum_{k=1}^K \|m_t\||y_t^k|. \end{aligned} \tag{65}$$

By Young's inequality, we have for any $c_1 > 0$,

$$2U_\omega\frac{\zeta}{K}\sum_{k=1}^K \|m_t\||y_t^k| \le c_1^2\zeta\|m_t\|^2 + \frac{U_\omega^2}{c_1^2}\zeta\frac{1}{K}\sum_{k=1}^K |y_t^k|^2,$$

Putting this back to Eq. (65), we get,

$$\begin{aligned}\|m_{t+1}\|_F^2 \le (1+c_1^2\zeta)\|m_t\|_F^2 + \frac{2\zeta}{K}\sum_{k=1}^K \left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k + \|\mathbf{Q}_t\|_F^2 \\ + 2\left\langle m_t, \frac{\zeta}{K}\sum_{k=1}^K \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top \right\rangle + \frac{U_\omega^2\zeta}{c_1^2 K}\sum_{k=1}^K |y_t^k|^2. \end{aligned} \tag{66}$$

For the Markovian-noise term in Eq. (66), we have

$$
\begin{aligned}
&\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top\rangle \\
=&\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\big[\widetilde{b}_t^k-\bar{b}^k+(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k\big](\omega_t^k)^\top\rangle \quad \text{(by Eq. (33))} \\
=&\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\big[\widetilde{b}_t^k-\bar{b}^k\big](\omega_t^k)^\top\rangle+\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}(\widetilde{A}_t^k-A^k)\mathbf{B}_t\omega_t^k(\omega_t^k)^\top\rangle \\
=&\langle m_t, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}\big[\widetilde{b}_t^k-\bar{b}^k\big](\omega_t^k-\omega_{t-\tau}^k)^\top\rangle \\
&+\langle m_t-m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}\big[\widetilde{b}_t^k-\bar{b}^k\big](\omega_{t-\tau}^k)^\top\rangle \\
&+\langle m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}(\mathbf{P}_{t,\perp}-\mathbf{P}_{t-\tau,\perp})\frac{\zeta}{K}\sum_{k=1}^{K}\big[\widetilde{b}_t^k-\bar{b}^k\big](\omega_{t-\tau}^k)^\top\rangle \\
&\underbrace{+\langle m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}\big[\widetilde{b}_t^k-\bar{b}^k\big](\omega_{t-\tau}^k)^\top\rangle}_{I_1} \\
&+\langle m_t, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\left(\mathbf{B}_t-\mathbf{B}_{t-\tau}\right)\omega_t^k(\omega_t^k)^\top\rangle \\
&+\langle m_t, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}(\omega_t^k-\omega_{t-\tau}^k)(\omega_t^k)^\top\rangle \\
&+\langle m_t, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_t^k-\omega_{t-\tau}^k)^\top\rangle \\
&+\langle m_t-m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle \\
&+\langle m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}\left(\mathbf{P}_{t,\perp}-\mathbf{P}_{t-\tau,\perp}\right)\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle \\
&\underbrace{+\langle m_{t-\tau}, \mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle}_{I_2}. \qquad (67)
\end{aligned}
$$

Note that I_1 and I_2 in the above decomposition can be bounded under expectation conditioned on $v_{0:t-\tau}$, and other terms are in terms of $\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2$, $\|\omega_t^k-\omega_{t-\tau}^k\|^2$, $\left\|\widetilde{b}_t^k-\bar{b}^k\right\|^2$, and $\left\|\widetilde{A}_t^k-A^k\right\|^2$

by applying Cauchy-Schwartz. Specifically, for the first term, we have

$$
\begin{aligned}
&\langle m_t, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}\left[\widetilde{b}_t^k-\bar{b}^k\right](\omega_t^k-\omega_{t-\tau}^k)^\top\rangle\\
\leq&\frac{\zeta}{K}\sum_{k=1}^{K}\left\|\widetilde{b}_t^k-\bar{b}^k\right\|\left\|\omega_t^k-\omega_{t-\tau}^k\right\|\\
\leq&\frac{\zeta}{K}\sum_{k=1}^{K}\left(\frac{\zeta}{2}\left\|\widetilde{b}_t^k-\bar{b}^k\right\|^2+\frac{1}{2\zeta}\left\|\omega_t^k-\omega_{t-\tau}^k\right\|^2\right)\\
\leq&\frac{1}{2}\frac{\zeta^2}{K}\sum_{k=1}^{K}\left\|\widetilde{b}_t^k-\bar{b}^k\right\|^2+\frac{1}{2K}\sum_{k=1}^{K}\left\|\omega_t^k-\omega_{t-\tau}^k\right\|^2.
\end{aligned}
$$

Applying similar analysis to all other terms in Eq. (67), we obtain

$$
\begin{aligned}
&\langle m_t-m_{t-\tau}, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}\left[\widetilde{b}_t^k-\bar{b}^k\right](\omega_{t-\tau}^k)^\top\rangle\\
\leq&\frac{1}{2}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+\frac{U_\omega^2}{2}\frac{\zeta^2}{K}\sum_{k=1}^{K}\left\|\widetilde{b}_t^k-\bar{b}^k\right\|^2,\\
&\langle m_{t-\tau}, \mathbf{B}_\perp^{*\top}(\mathbf{P}_{t,\perp}-\mathbf{P}_{t-\tau,\perp})\frac{\zeta}{K}\sum_{k=1}^{K}\left[\widetilde{b}_t^k-\bar{b}^k\right](\omega_{t-\tau}^k)^\top\rangle\\
\leq&2\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2+\frac{1}{2}U_\omega^2\frac{\zeta^2}{K}\sum_{k=1}^{K}\left\|\widetilde{b}_t^k-\bar{b}^k\right\|^2,\\
&\langle m_t, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\left(\mathbf{B}_t-\mathbf{B}_{t-\tau}\right)\omega_t^k(\omega_t^k)^\top\rangle\\
\leq&\frac{1}{2}U_\omega^4\frac{\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k-A^k\|^2+\frac{1}{2}\|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|^2,\\
&\langle m_t, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}(\omega_t^k-\omega_{t-\tau}^k)(\omega_t^k)^\top\rangle\\
&\leq\frac{U_\omega}{2}\frac{\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k-A^k\|^2+\frac{U_\omega}{2}\frac{1}{K}\sum_{k=1}^{K}\|\omega_t^k-\omega_{t-\tau}^k\|^2,\\
&\langle m_t, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_t^k-\omega_{t-\tau}^k)^\top\rangle\\
&\leq\frac{U_\omega}{2}\frac{\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k-A^k\|^2+\frac{U_\omega}{2}\frac{1}{K}\sum_{k=1}^{K}\|\omega_t^k-\omega_{t-\tau}^k\|^2,\\
&\langle m_t-m_{t-\tau}, \mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k-A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle
\end{aligned}
$$

$$\leq \frac{1}{2}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 + \frac{1}{2}U_\omega^4 \frac{\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k - A^k\|^2,$$

$$\langle m_{t-\tau}, \mathbf{B}_\perp^{*\top}\left(\mathbf{P}_{t,\perp} - \mathbf{P}_{t,\perp}\right)\frac{\zeta}{K}\sum_{k=1}^{K}(\widetilde{A}_t^k - A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle$$

$$\leq 2\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 + \frac{U_\omega^4}{2}\frac{\zeta^2}{K}\sum_{k=1}^{K}\|\widetilde{A}_t^k - A^k\|^2$$

Combining the above bounds and pulgging them back to Eq. (67), and taking expectation conditioned on $v_{0:t-\tau}$, we get

$$\begin{aligned}
&\mathbb{E}_{t-\tau}\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top\rangle \\
\leq&\left(\frac{1}{2}+U_\omega^2\right)\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}\left\|\widetilde{b}_t^k - \bar{b}^k\right\|^2 + \left(\frac{1}{2}+U_\omega\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}\left\|\omega_t^k - \omega_{t-\tau}^k\right\|^2 \\
&+\left(U_\omega + \frac{3}{2}U_\omega^4\right)\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 + 8\mathbb{E}_{t-\tau}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 \\
&\underbrace{+\langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t-\tau,\perp}\mathbb{E}_{t-\tau}\left[\widetilde{b}_t^k - \bar{b}^k\right](\omega_{t-\tau}^k)^\top\rangle}_{I_1} \\
&\underbrace{+\langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_\perp^{*\top}\mathbf{P}_{t-\tau,\perp}\mathbb{E}_{t-\tau}[\widetilde{A}_t^k - A^k]\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle}_{I_2}. \qquad (68)
\end{aligned}$$

Note that Lemma K.5 holds for any $t \geq 0$ and any state $s_{t,k}^0$. Applying Lemma K.5, we have

$$\mathbb{E}_{t-\tau}\left\|\widetilde{b}_t^k - \bar{b}^k\right\|^2 \leq \frac{1}{L}\left(16U_r^2 + 32U_r^2\frac{m}{1-\rho}\right). \qquad (69)$$

Similarly, we have,

$$\mathbb{E}_{t-\tau}\|\widetilde{A}_t^k - A^k\|^2 \leq \frac{1}{L}\left(16 + 32\frac{m}{1-\rho}\right). \quad (\text{Lemma K.5, Eq. (153)}) \qquad (70)$$

Applying Lemma K.2 and Lemma K.3, we have the following,

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\omega_t^k - \omega_{t-\tau}^k\|^2 \\
\leq&\tau\sum_{j=t-\tau}^{t-1}\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k - \omega_j^k\|^2 \\
\leq&\tau^2 C_5(\tau^2)\beta^4 + \tau^2 C_6(\tau^2)\zeta^2\beta^2 + 24\tau^2\beta^2\mathbb{E}_{t-2\tau}\|x_t^k\|^2 \\
&+\tau^2 C_7\frac{\beta^2}{L} + \tau^2 C_8(\tau^2)\gamma^2\beta^2 + 2\tau^2\beta^2\mathbb{E}_{t-2\tau}|y_t^k|^2, \qquad (71)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\|^2 \\
\leq&\tau^2 C_{16}(\tau^2)\zeta^2\beta^2 + \tau^2 C_{17}(\tau^2)\zeta^4 + \tau^2 C_{18}\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2 + \tau^2 C_{19}(\tau^2)\zeta^2\beta^2 + \tau^2 C_{20}\frac{\zeta^2}{K} \\
&+\tau^2 C_{21}\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2} + \tau^2 C_{22}(\tau^2)\gamma^2\zeta^2 + \tau^2 C_{23}\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k|^2 + \tau^2 C_{24}\zeta^4. \qquad (72)
\end{aligned}$$

For I_1, by Lemma K.6, we have

$$\begin{aligned}\mathbb{E}_{t-\tau}[I_1] &= \mathbb{E}_{t-\tau}\langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}\left[\widetilde{b}_t^k - \bar{b}^k\right](\omega_{t-\tau}^k)^\top\rangle \\ &= \langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}\mathbb{E}_{t-\tau}\left[\widetilde{b}_t^k - \bar{b}^k\right](\omega_{t-\tau}^k)^\top\rangle \\ &\leq \frac{4U_rU_\omega\zeta m\rho^{\tau L}}{L(1-\rho)}.\end{aligned} \tag{73}$$

Similarly, for I_2, by Lemma K.6, we have

$$\begin{aligned}\mathbb{E}_{t-\tau}[I_2] &= \mathbb{E}_{t-\tau}\langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}(\widetilde{A}_t^k - A^k)\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle \\ &\leq \langle m_{t-\tau}, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t-\tau,\perp}\mathbb{E}_{t-\tau}\left[\widetilde{A}_t^k - A^k\right]\mathbf{B}_{t-\tau}\omega_{t-\tau}^k(\omega_{t-\tau}^k)^\top\rangle \\ &\leq \frac{4U_\omega^2\zeta m\rho^{\tau L}}{L(1-\rho)}.\end{aligned} \tag{74}$$

Putting Eq. (69), Eq. (70), Eq. (71), Eq. (72), Eq. (73) and Eq. (74) back to Eq. (68), using the tower property and $\beta = c\zeta$, we get

$$\begin{aligned}&\mathbb{E}_{t-2\tau}\langle m_t, \frac{\zeta}{K}\sum_{k=1}^{K}\mathbf{B}_{\perp}^{*\top}\mathbf{P}_{t,\perp}\xi_t^k(\omega_t^k)^\top\rangle \\ \leq&\left(\frac{1}{2}+U_\omega^2\right)\left(16U_r^2+32U_r^2\frac{m}{1-\rho}\right)\frac{\zeta^2}{L} \\ &+\left(\frac{1}{2}+U_\omega\right)\Bigg(\tau^2C_5(\tau^2)c^4\zeta^4+\tau^2C_6(\tau^2)c^2\zeta^4+24\tau^2c^2\zeta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2 \\ &\qquad+\tau^2C_7c^2\frac{\zeta^2}{L}+\tau^2C_8(\tau^2)c^2\gamma^2\zeta^2+2\tau^2c^2\zeta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k|^2\Bigg) \\ &+\left(U_\omega+\frac{3}{2}U_\omega^4\right)\left(16+32\frac{m}{1-\rho}\right)\frac{\zeta^2}{L} \\ &+8\Bigg(\tau^2C_{16}(\tau^2)c^2\zeta^4+\tau^2C_{17}(\tau^2)\zeta^4+\tau^2C_{18}\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2+\tau^2C_{19}(\tau^2)c^2\zeta^4 \\ &\qquad+\tau^2C_{20}\frac{\zeta^2}{K}+\tau^2C_{21}\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}+\tau^2C_{22}(\tau^2)\gamma^2\zeta^2+\tau^2C_{23}\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k|^2+\tau^2C_{24}\zeta^4\Bigg) \\ &+\frac{2U_rU_\omega\zeta m\rho^{\tau L}}{L(1-\rho)}+\frac{2U_\omega^2\zeta m\rho^{\tau L}}{L(1-\rho)} \\ \leq&C_{25}(\tau^2)\frac{\zeta^2}{L}+C_{26}(\tau^4)\zeta^4+C_{27}(\tau^4)\zeta^4+C_{28}(\tau^4)\gamma^2\zeta^2 \\ &+C_{29}(\tau^2)\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2+C_{30}(\tau^2)\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k|^2 \\ &+8\tau^2C_{20}\frac{\zeta^2}{K}+8\tau^2C_{21}\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}+\frac{2U_rU_\omega\zeta m\rho^{\tau L}}{L(1-\rho)}+\frac{2U_\omega^2\zeta m\rho^{\tau L}}{L(1-\rho)}.\end{aligned} \tag{75}$$

where

$$C_{25}(\tau^2) = \left(\frac{1}{2} + U_\omega^2\right)\left(16U_r^2 + 32U_r^2\frac{m}{1-\rho}\right) + \left(U_\omega + \frac{3}{2}U_\omega^4\right)\left(16 + 32\frac{m}{1-\rho}\right) + \left(\frac{1}{2} + U_\omega\right)\tau^2 C_7 c^2,$$

$$C_{26}(\tau^4) = \left[\left(\frac{1}{2} + U_\omega\right)\left(\tau^2 C_5(\tau^2)c^4 + \tau^2 C_6(\tau^2)c^2\right) + 8\left(\tau^2 C_{16}(\tau^2)c^2 + \tau^2 C_{17}(\tau^2)\right)\right],$$

$$C_{27}(\tau^4) = \left[8\tau^2 C_{19}(\tau^2)c^2 + 8\tau^2 C_{24}\right], C_{28}(\tau^4) = \left[\left(\frac{1}{2} + U_\omega\right)\tau^2 C_8(\tau^2)c^2 + 8\tau^2 C_{22}(\tau^2)\right],$$

$$C_{29}(\tau^2) = \left[24\tau^2 c^2\left(\frac{1}{2} + U_\omega\right) + 8\tau^2 C_{18}\right], C_{30}(\tau^2) = \left[2\tau^2 c^2\left(\frac{1}{2} + U_\omega\right) + 8\tau^2 C_{23}\right].$$

Choosing $\tau = \lceil 2\log_\rho(\zeta)\rceil$ and applying $\mathbb{E}_{t-2\tau}$ (the tower property brings the $\mathbb{E}_{t-\tau}$ terms to the same level), we get,

$$\begin{aligned}\widehat{M}_{t+1} \leq \widehat{M}_t + \frac{2\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[\left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k\right] + c_1^2\zeta\widehat{M}_t + \mathcal{C}_{M,4}(\tau^2)\zeta^2\widehat{X}_t \\ + \left(\mathcal{C}_{M,5}(\tau^2)\zeta^2 + \frac{U_\omega^2}{c_1^2}\zeta\right)\widehat{Y}_t + 2C_{25}(\tau^2)\frac{\zeta^2}{L} + \mathcal{C}_{M,1}\zeta^4 \\ + 2C_{28}(\tau^4)\gamma^2\zeta^2 + \mathcal{C}_{M,2}(\tau^2)\frac{\zeta^2}{K} + \mathcal{C}_{M,3}\zeta^6 + \mathcal{C}_{M,6}\zeta^3, \end{aligned} \tag{76}$$

where

$$\mathcal{C}_{M,1} = \left(2C_{27}(\tau^4) + 4\tau^2 U_\delta^4 c^2 + 2C_{26}(\tau^4)\right), \mathcal{C}_{M,2}(\tau^2) = \left(16\tau^2 C_{20} + 4U_\delta^2 U_\omega^2\right),$$

$$\mathcal{C}_{M,3} = \left(16\tau^2 C_{21} + 4U_\omega^2 U_\delta^2\right)\frac{4m^2}{L^2(1-\rho)^2}, \mathcal{C}_{M,4}(\tau^2) = \left[16U_\omega^2 + 2C_{29}(\tau^2)\right],$$

$$\mathcal{C}_{M,5}(\tau^2) = \left[2C_{30}(\tau^2) + 4U_\omega^2\right], \mathcal{C}_{M,6} = \left[\frac{4U_r U_\omega m}{L(1-\rho)} + \frac{4U_\omega^2 m}{L(1-\rho)}\right].$$

G PROOF OF LEMMA 5.4

By Eq. (11) and the update of η^k in Algorithm 1, we have

$$\begin{aligned}
y_{t+1}^k =& J^k - \eta_{t+1}^k = J^k - \eta_t^k - \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - \eta_t^k\right) = y_t^k - \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - \eta_t^k\right) \\
=& y_t^k - \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - J^k + J^k - \eta_t^k\right) \\
=& y_t^k - \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - J^k + y_t^k\right) \\
=& (1-\gamma)y_t^k - \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - J^k\right) \\
=& (1-\gamma)^{t+1} y_0^k - \gamma\sum_{i=0}^{t}(1-\gamma)^{t-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right).
\end{aligned}$$

That is, for every $t \geq 0$,

$$y_t^k = (1-\gamma)^t y_0^k - \gamma\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right), \tag{77}$$

which holds at any stepsize and for any η_0^k, since it is only the update of η^k unrolled.

Then,

$$\begin{aligned}
(y_t^k)^2 =& (1-\gamma)^{2t}(y_0^k)^2 + \gamma^2\left(\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right)\right)^2 \\
& - 2(1-\gamma)^t y_0^k \gamma\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right).
\end{aligned}$$

Taking expectation, we get

$$\begin{aligned}
\mathbb{E}(y_t^k)^2 =& \underbrace{(1-\gamma)^{2t}(y_0^k)^2}_{I_1} + \underbrace{\gamma^2\mathbb{E}\left(\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right)\right)^2}_{I_2} \\
& \underbrace{- 2(1-\gamma)^t y_0^k \gamma\mathbb{E}\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right)}_{I_3}.
\end{aligned} \tag{78}$$

For I_1 in Eq. (78), using the fact that $\ln(1-\gamma) \leq -\gamma$ for $\gamma \in (0,1)$, we have

$$I_1 = (1-\gamma)^{2t}(y_0^k)^2 \leq (y_0^k)^2 e^{-2\gamma t}.$$

For I_3 in Eq. (78), we have,

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\\
=&\mathbb{E}\sum_{i=0}^{\tau-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right) + \mathbb{E}\sum_{i=\tau}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\\
\leq& 2U_r\frac{(1-\gamma)^{t-\tau}(1-(1-\gamma)^{\tau})}{1-(1-\gamma)} + \sum_{i=\tau}^{t-1}(1-\gamma)^{t-1-i}\mathbb{E}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\\
\leq& 2U_r\tau(1-\gamma)^{t-\tau} + \frac{2U_r}{L}\sum_{i=\tau}^{t-1}(1-\gamma)^{t-1-i}\sum_{\ell=0}^{L-1}\sup_{s\in\mathcal{S}} d_{TV}\left(\mathbb{P}^k(s_{i,\ell}^k \mid s_0^k = s), \mu^k\right)\\
\leq& 2U_r\tau(1-\gamma)^{t-\tau} + \frac{2U_r m}{L}\sum_{i=\tau}^{t-1}(1-\gamma)^{t-1-i}\sum_{\ell=0}^{L-1}\rho^{iL+\ell}\\
\leq& 2U_r\tau(1-\gamma)^{t-\tau} + \frac{2U_r m\rho^{\tau L}}{L(1-\rho)\gamma}.
\end{aligned}$$

where the first inequality follows because $\left|\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right| \leq 2U_r$ and $1-(1-\gamma)^{\tau} \leq \tau\gamma$, the second inequality follows since $\mathbb{E}(r_{i,\ell}^k - J^k) \leq 2U_r \sup_{s\in\mathcal{S}} d_{TV}\left(\mathbb{P}^k(s_{i,\ell}^k \mid s_0^k = s), \mu^k\right)$, and the third inequality follows from Assumption 2. Thus,

$$\begin{aligned}
|I_3| \leq& 2(1-\gamma)^t|y_0^k|\left(2U_r\tau\gamma(1-\gamma)^{t-\tau} + \frac{2U_r m\rho^{\tau L}}{L(1-\rho)}\right)\\
\leq& 4|y_0^k|U_r\tau\gamma(1-\gamma)^{2t-\tau} + \frac{4U_r(1-\gamma)^t|y_0^k|m\rho^{\tau L}}{L(1-\rho)}.
\end{aligned}$$

Lastly, for I_2 in Eq. (78), we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\right)^2\\
=&\mathbb{E}\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)^2\\
&+ 2\mathbb{E}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{j,\ell}^k - J^k\right)\\
\leq&\mathbb{E}\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)^2\\
&\underbrace{+ 2\mathbb{E}\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{j,\ell}^k - J^k\right)}_{I_{21}}\\
&\underbrace{+ 2\mathbb{E}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{j,\ell}^k - J^k\right)}_{I_{22}}\\
&\underbrace{+ 2\mathbb{E}\sum_{i=t-\tau}^{t-1}\sum_{j=i+1}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k - J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{j,\ell}^k - J^k\right)}_{I_{23}}. \qquad (79)
\end{aligned}$$

Recall $v_{0:t}^k$ is the Markovian trajectory at agent k:

$$v_{0:t}^k = (s_0^k, a_0^k, s_1^k, a_1^k, ..., s_{t-\tau+1}^k, a_{t-\tau+1}^k, ..., s_t^k, a_t^k, s_{t+1}^k). \tag{80}$$

For the first term, we have

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right)^2 \\
=&\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\frac{1}{L^2}\left(\sum_{\ell=0}^{L-1}\mathbb{E}(r_{i,\ell}^k - J^k)^2 + 2\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\mathbb{E}(r_{i,\ell}^k - J^k)(r_{i,\ell'}^k - J^k)\right) \\
\leq&\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\frac{1}{L^2}\left(4U_r^2 L + 8U_r^2 m\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\rho^{\ell'-\ell}\right) \\
\leq&\sum_{i=0}^{t-1}(1-\gamma)^{2t-2-2i}\frac{1}{L^2}\left(4U_r^2 L + 8U_r^2 mL\frac{\rho}{1-\rho}\right) \\
\leq&\frac{1}{1-(1-\gamma)^2}\frac{1}{L}\left(4U_r^2 + 8U_r^2 m\frac{\rho}{1-\rho}\right) \\
\leq&\frac{1}{\gamma L}\left(4U_r^2 + 8U_r^2 m\frac{\rho}{1-\rho}\right).
\end{aligned}$$

For I_{21} in Eq. (79), we have

$$\begin{aligned}
I_{21} =&2\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{i,\ell}^k - J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{j,\ell}^k - J^k\right)|v_{0:i}^k\right]\right] \\
\leq&2\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}2U_r\mathbb{E}\left[\left|\mathbb{E}\left[\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{j,\ell}^k - J^k\right)|v_{0:i}^k\right]\right|\right] \\
\leq&\frac{2}{L}\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}2U_r\sum_{\ell=0}^{L-1}2U_r m\rho^{(j-i-1)L+\ell} \\
=&\frac{8U_r^2 m}{L}\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\rho^{(j-i-1)L}\frac{1-\rho^L}{1-\rho} \\
\leq&\frac{8U_r^2 m}{L(1-\rho)}\rho^{(\tau-1)L}\sum_{i=0}^{t-\tau-1}\sum_{j=i+\tau}^{t-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j} \\
\leq&\frac{8U_r^2 m}{L(1-\rho)}\rho^{(\tau-1)L}\sum_{i=0}^{t-\tau-1}(1-\gamma)^{t-1-i}\frac{1-(1-\gamma)^{t-i-\tau}}{\gamma} \\
\leq&\frac{8U_r^2 m}{L(1-\rho)}\rho^{(\tau-1)L}\frac{(1-\gamma)^\tau}{\gamma^2}.
\end{aligned}$$

For I_{22}, we have

$$
\begin{aligned}
&2\mathbb{E}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{i,\ell}^k-J^k\right)\left(\frac{1}{L}\sum_{\ell=0}^{L-1}r_{j,\ell}^k-J^k\right)\\
\leq&\frac{2}{L^2}\mathbb{E}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\sum_{\ell=0}^{L-1}\sum_{\ell'=0}^{L-1}(r_{i,\ell}^k-J^k)(r_{j,\ell'}^k-J^k)\\
\leq&\frac{8mU_r^2}{L^2}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\sum_{\ell=0}^{L-1}\sum_{\ell'=0}^{L-1}\rho^{(j-i)L+(\ell'-\ell)}\\
=&\frac{8mU_r^2}{L^2}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\rho^{(j-i-1)L+1}\left(\frac{1-\rho^L}{1-\rho}\right)^2\\
\leq&\frac{8mU_r^2}{L^2}\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\frac{1}{(1-\rho)^2}.
\end{aligned}
$$

Noting that

$$
\begin{aligned}
&\sum_{i=0}^{t-\tau-1}\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-i}(1-\gamma)^{t-1-j}\\
=&\sum_{i=0}^{t-\tau-1}(1-\gamma)^{t-1-i}\left(\sum_{j=i+1}^{i+\tau-1}(1-\gamma)^{t-1-j}\right)\\
=&\sum_{i=0}^{t-\tau-1}(1-\gamma)^{t-1-i}\frac{(1-\gamma)^{t-i-\tau}(1-(1-\gamma)^{\tau-1})}{\gamma}\\
=&\frac{(1-(1-\gamma)^{\tau-1})}{\gamma}\sum_{i=0}^{t-\tau-1}(1-\gamma)^{2t-1-2i-\tau}.
\end{aligned}
$$

Since $(1-\gamma)^\tau \geq 1-\tau\gamma$, we have

$$
\frac{(1-(1-\gamma)^\tau)}{\gamma}\leq\frac{1-(1-\tau\gamma)}{\gamma}=\tau.
$$

In addition,

$$
\sum_{i=0}^{t-\tau-1}(1-\gamma)^{2t-1-2i-\tau}\leq\frac{(1-\gamma)^{\tau+1}}{1-(1-\gamma)^2}\leq\frac{(1-\gamma)^\tau}{\gamma}.
$$

Thus, I_{22} in Eq. (79) can be bounded as

$$
I_{22}\leq\tau\frac{(1-\gamma)^\tau}{\gamma}\frac{8U_r^2}{L^2}\frac{1}{(1-\rho)^2}.
$$

For I_{23} in Eq. (79), we have

$$
I_{23}\leq 2\sum_{i=t-\tau}^{t-1}\sum_{j=i+1}^{t-1}4U_r^2\leq 8U_r^2\frac{\tau(\tau-1)}{2}.
$$

Now, we can bound I_2 in Eq. (78) as

$$
\begin{aligned}
I_2&\leq\gamma^2\left(\frac{1}{\gamma L}\left(4U_r^2+8U_r^2m\frac{\rho}{1-\rho}\right)+\frac{8U_r^2m}{L(1-\rho)}\rho^{(\tau-1)L}\frac{(1-\gamma)^\tau}{\gamma^2}+\tau\frac{(1-\gamma)^\tau}{\gamma}\frac{8U_r^2}{L^2}\frac{1}{(1-\rho)^2}+8U_r^2\frac{\tau(\tau-1)}{2}\right)\\
&\leq\frac{\gamma}{L}\left(4U_r^2+8U_r^2m\frac{\rho}{1-\rho}\right)+\frac{8U_r^2m}{L(1-\rho)}\rho^{(\tau-1)L}(1-\gamma)^\tau+\frac{\gamma}{L^2}\tau(1-\gamma)^\tau\frac{8U_r^2}{(1-\rho)^2}+8U_r^2\frac{\tau(\tau-1)}{2}\gamma^2.
\end{aligned}
$$

Combining the three bounds gives the form of the result that does not depend on T, on the value of γ, or on η_0^k, and that is therefore the one reused elsewhere.

Lemma G.1 (Reward analysis, general form)**.** *Let $\gamma \le 1/2$ and let τ satisfy $\gamma\tau \le 1/2$ and $\rho^{\tau L} \le \gamma$. There is a*

$$\mathcal{C}_r = \mathcal{O}\Big(U_r^2\Big(1 + \frac{m}{\rho(1-\rho)} + \frac{\tau}{(1-\rho)^2 L}\Big)\Big)$$

such that, for every $t \ge 0$ and every k,

$$\mathbb{E}(y_t^k)^2 \;\le\; \mathcal{C}_r\Big(e^{-2\gamma t} + \frac{\gamma}{L} + \tau^2\gamma^2\Big). \tag{81}$$

Proof. Throughout, $|y_0^k| = |J^k - \eta_0^k| \le 2U_r$, which holds because η_t^k is a convex combination of rewards. Use $\mathbb{E}(y_t^k)^2 = I_1 + I_2 - I_3 \le I_1 + I_2 + |I_3|$, since $I_1, I_2 \ge 0$. By $\gamma\tau \le 1/2$, $(1-\gamma)^{-\tau} \le e^{\gamma\tau/(1-\gamma)} \le e^{2\gamma\tau} \le e \le 3$, so $(1-\gamma)^{2t-\tau} \le 3e^{-2\gamma t}$, and $I_1 = (1-\gamma)^{2t}(y_0^k)^2 \le 4U_r^2 e^{-2\gamma t}$.

For the first term of $|I_3|$, $\tau\gamma \le 1/2$ gives

$$4|y_0^k|U_r\tau\gamma(1-\gamma)^{2t-\tau} \le 24U_r^2\tau\gamma e^{-2\gamma t} \le 12U_r^2 e^{-2\gamma t}.$$

For its second term, $(1-\gamma)^t \le 1$ and $\rho^{\tau L} \le \gamma$ give

$$\frac{4U_r(1-\gamma)^t|y_0^k|m\rho^{\tau L}}{L(1-\rho)} \le \frac{8U_r^2 m}{1-\rho}\cdot\frac{\gamma}{L}.$$

In I_2, the second term is at most $\frac{8U_r^2 m}{\rho(1-\rho)}\frac{\gamma}{L}$ by $\rho^{(\tau-1)L} \le \rho^{\tau L}/\rho^L \le \gamma/\rho$, the third is $\frac{8U_r^2\tau}{(1-\rho)^2 L}\cdot\frac{\gamma}{L}$, and the fourth is at most $4U_r^2\tau^2\gamma^2$. Collecting the coefficients of $e^{-2\gamma t}$, of γ/L and of $\tau^2\gamma^2$ gives $\mathcal{C}_r$. □

Lemma 5.4 is Eq. (81) at $t = T$ with $\gamma = C\log(T)/T$ and τ of Lemma 5.1; the remainder of this section tracks the individual terms instead of collecting them into $\mathcal{C}_r$.

Putting I_1, I_2 and I_3 together, and using $\mathbb{E}(y_T^k)^2 = I_1 + I_2 - I_3 \le I_1 + I_2 + |I_3|$ (since $I_1, I_2 \ge 0$), and choosing $\gamma = \frac{C\log(T)}{T}$, we get

$$\begin{aligned}
\mathbb{E}(y_T^k)^2 &\le (y_0^k)^2 e^{-2\gamma T} + \frac{\gamma}{L}\left(4U_r^2 + 8U_r^2 m\frac{\rho}{1-\rho}\right) + \frac{8U_r^2 m}{L(1-\rho)}\rho^{(\tau-1)L}(1-\gamma)^\tau \\
&+ \frac{\gamma}{L^2}\tau(1-\gamma)^\tau\frac{8U_r^2}{(1-\rho)^2} + 8U_r^2\frac{\tau(\tau-1)}{2}\gamma^2 + 4|y_0^k|U_r\tau\gamma e^{\gamma\tau}e^{-\gamma 2T} + \frac{4U_r(1-\gamma)^T|y_0^k|m\rho^{\tau L}}{L(1-\rho)} \\
&\le 4U_r^2\frac{1}{T^2} + \frac{C\log(T)}{LT}\left(4U_r^2 + 8U_r^2 m\frac{\rho}{1-\rho}\right) + \frac{8U_r^2 m}{L(1-\rho)}\rho^{(\tau-1)L} + \frac{8U_r^2\tau}{(1-\rho)^2}\frac{C\log(T)}{TL^2} \\
&+ 4U_r^2\tau(\tau-1)\frac{C^2\log^2(T)}{T^2} + 8U_r^2\tau\gamma e^{\gamma\tau}\frac{1}{T^2} + \frac{8U_r^2 m\rho^{\tau L}}{L(1-\rho)T}.
\end{aligned}$$

By $\tau = \lceil 2\log_\rho(\zeta)\rceil$ and $\zeta \le \gamma$, $\rho^\tau \le \zeta^2 \le \gamma^2$; with $\gamma\tau \le 1/2$, it holds that $e^{\gamma\tau} \le e^{1/2} \le 3$, $\rho^{\tau L} \le \gamma^{2L}$, and $\rho^{(\tau-1)L} \le (\gamma^2/\rho)^L$. Thus, the upper bound of $\mathbb{E}(y_T^k)^2$ can be simplified as

$$\begin{aligned}
\mathbb{E}(y_T^k)^2 &\le 28U_r^2\frac{1}{T^2} + \frac{C\log(T)}{LT}\left(4U_r^2 + 8U_r^2 m\frac{\rho}{1-\rho}\right) + \frac{8U_r^2 m}{L(1-\rho)}\left(\frac{C^2\log^2(T)}{\rho T^2}\right)^L + \frac{8U_r^2\tau}{(1-\rho)^2}\frac{C\log(T)}{TL^2} \\
&+ 4U_r^2\tau(\tau-1)\frac{C^2\log^2(T)}{T^2} + \frac{8U_r^2 m}{L(1-\rho)T}\left(\frac{C^2\log^2(T)}{T^2}\right)^L,
\end{aligned}$$

which trivially implies that for sufficiently large T,

$$\bar{Y}_T = \widetilde{\mathcal{O}}\left(\frac{1}{LT}\right).$$

H Proof of Proposition 1

Recall from Eq. (12) the conditional quantities $\widehat{X}_t, \widehat{M}_t, \widehat{Y}_t$, whose total expectations are $\bar{X}_t, M_t, \bar{Y}_t$. Lemma 5.1 and Lemma 5.2 give the two one-step recursions

$$\widehat{X}_{t+1} \le \widehat{X}_t + \frac{2c\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top \mathbf{P}_t A^k x_t^k\right] + \left(c_4^2 c + c_5^2 + c_6^2 + \frac{4U_\omega^4}{c_5^2}\right)\zeta\widehat{X}_t$$
$$+ \mathcal{C}_{X,1}(\tau^2)\zeta^2\widehat{X}_t + \mathcal{O}(\zeta\widehat{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right),$$

$$\widehat{M}_{t+1} \le \widehat{M}_t + \frac{2\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[\left(\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k\right)^\top A^k x_t^k\right] + c_1^2\zeta\widehat{M}_t$$
$$+ \mathcal{C}_{M,4}(\tau^2)\zeta^2\widehat{X}_t + \mathcal{O}(\zeta\widehat{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right).$$

Set

$$\widehat{\mathcal{V}}_t = \widehat{X}_t + c\widehat{M}_t.$$

Adding the first recursion to c times the second, we obtain

$$\widehat{\mathcal{V}}_{t+1} \le \widehat{\mathcal{V}}_t + \frac{2c\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top \mathbf{P}_t A^k x_t^k + (\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k)^\top A^k x_t^k\right]$$
$$+ \left(c_4^2 c + c_5^2 + c_6^2 + \frac{4U_\omega^4}{c_5^2}\right)\zeta\widehat{X}_t + cc_1^2\zeta\widehat{M}_t + [\mathcal{C}_{X,1}(\tau^2) + c\,\mathcal{C}_{M,4}(\tau^2)]\zeta^2\widehat{X}_t$$
$$+ \mathcal{O}(\zeta\widehat{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right).$$

For every $k \in [K]$,

$$(x_t^k)^\top \mathbf{P}_t A^k x_t^k + (\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k)^\top A^k x_t^k = (x_t^k)^\top A^k x_t^k - (\mathbf{P}_{t,\perp}\mathbf{P}^* x_t^k)^\top A^k x_t^k.$$

Since $\|A^k\|_2 \le 2$ and $\|\mathbf{P}_{t,\perp}\mathbf{P}^*\|_2 \le \|m_t\|_2$, while $\|m_t\|_2 \le \lambda/4$, the two displayed terms satisfy

$$\frac{2c\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[(x_t^k)^\top \mathbf{P}_t A^k x_t^k + (\mathbf{P}_{t,\perp}\mathbf{P}_\perp^* x_t^k)^\top A^k x_t^k\right]$$
$$\le -2c\lambda\zeta\widehat{X}_t + \frac{4c\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\left[\|m_t\|_2\|x_t^k\|^2\right] \le -c\lambda\zeta\widehat{X}_t.$$

Choose

$$c_4^2 = \frac{\lambda}{30}, \qquad c_1^2 = \frac{\lambda\alpha}{6}, \qquad c_5^2 = c_6^2 = 2U_\omega^2.$$

The coverage lower bound $\widehat{X}_t \ge \alpha\widehat{M}_t$ gives $cc_1^2\zeta\widehat{M}_t \le c\lambda\zeta\widehat{X}_t/6$. Thus, if $c\lambda \ge 20U_\omega^2$,

$$\widehat{\mathcal{V}}_{t+1} \le \widehat{\mathcal{V}}_t - \frac{c\lambda}{2}\zeta\widehat{X}_t + [\mathcal{C}_{X,1}(\tau^2) + c\,\mathcal{C}_{M,4}(\tau^2)]\zeta^2\widehat{X}_t + \mathcal{O}(\zeta\widehat{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right).$$

Finally, $\widehat{\mathcal{V}}_t \le (1 + c/\alpha)\widehat{X}_t$, so for $[\mathcal{C}_{X,1}(\tau^2) + c\,\mathcal{C}_{M,4}(\tau^2)]\zeta \le \frac{c\lambda}{4}$,

$$\widehat{\mathcal{V}}_{t+1} \le \left(1 - \frac{c\lambda\alpha}{4(\alpha+c)}\zeta\right)\widehat{\mathcal{V}}_t + \mathcal{O}(\zeta\widehat{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right), \qquad t \ge 2\tau. \tag{82}$$

Taking the total expectation in Eq. (82) and using $\mathbb{E}[\widehat{\mathcal{V}}_t] = \mathcal{V}_t$ and $\mathbb{E}[\widehat{Y}_t] = \bar{Y}_t$ gives the unconditional recursion

$$\mathcal{V}_{t+1} \le \left(1 - \frac{c\lambda\alpha}{4(\alpha+c)}\zeta\right)\mathcal{V}_t + \mathcal{O}(\zeta\bar{Y}_t) + \mathcal{O}\left(\frac{\zeta^2}{L} + \frac{\zeta^2}{K}\right), \qquad t \ge 2\tau. \tag{83}$$

For $0 \le t \le 2\tau$, $\mathcal{V}_t \le 4U_\omega^2 + cr$ by $\|x_t^k\| \le \|\mathbf{B}_t\omega_t^k\| + \|z^{k,*}\| \le 2U_\omega$ and $\|m_t\|_F^2 \le r$. Unrolling Eq. (83) from $t = 2\tau$ with the stepsize in Proposition 1, using Lemma 5.4, and setting $L = K$, gives $\mathcal{V}_{T+1} = \widetilde{\mathcal{O}}(1/(TK))$.

I Removing the Trajectory Hypothesis

Proposition 1 assumes $\|m_t\|_2 \le \lambda/4$, and in this section, we replace this condition with a high-probability statement. We will first run an initialization phase to ensure that the starting point of Algorithm 1 satisfies the condition. Then, we show with high probability the sequence stays within the condition.

I.1 Pathwise Lyapunov function and the exit time

All quantities in Appendix H are in unconditional expectation: $\mathcal{V}_t = \bar{X}_t + cM_t$. We define the *pathwise* quantity

$$n_t := \frac{1}{K}\sum_{k=1}^{K} \|x_t^k\|^2 + c\|m_t\|_F^2 \ge 0, \qquad \text{so that} \qquad \mathcal{V}_t = \mathbb{E}[n_t]. \tag{84}$$

Since $\|m_t\|_2 \le \|m_t\|_F$ and $n_t \ge c\|m_t\|_F^2$,

$$n_t \le \frac{c\lambda^2}{16} \quad \Longrightarrow \quad \|m_t\|_2^2 \le \frac{n_t}{c} \le \frac{\lambda^2}{16} \quad \Longrightarrow \quad \|m_t\|_2 \le \frac{\lambda}{4}, \tag{85}$$

so a barrier on n_t delivers the hypothesis that Appendix H consumes. Define the *first exit time*

$$T_{\text{exit}} := \min\Big\{t \ge 0 :\ n_t > \frac{c\lambda^2}{32}\Big\} \qquad (= \infty \text{ if no such } t \le T), \tag{86}$$

a stopping time with respect to $\{\mathcal{F}_t\}$.

Lemma I.1 (Deterministic one-step increment). *Suppose $U_\delta U_\omega \zeta \le 1/2$ and $\beta = c\zeta$. Then, for every t and on every sample path,*

$$\big|n_{t+1} - n_t\big| \le U_n \zeta, \qquad U_n := U_\delta U_\omega \big[4c + 8U_\omega^2 + 4cr\big].$$

Proof. By Lemma K.1 with $\tau = 1$,

$$\big\|\omega_{t+1}^k - \omega_t^k\big\| \le c\zeta U_\delta, \qquad \big\|\mathbf{B}_{t+1} - \mathbf{B}_t\big\|_2 \le \zeta U_\delta U_\omega + 2\zeta^2 U_\delta^2 U_\omega^2 \le 2\zeta U_\delta U_\omega.$$

Since $\|\mathbf{B}_{t+1}\|_2 = 1$ and $\|\omega_t^k\| \le U_\omega$,

$$\big\|x_{t+1}^k - x_t^k\big\| \le \big\|\omega_{t+1}^k - \omega_t^k\big\| + \big\|\mathbf{B}_{t+1} - \mathbf{B}_t\big\|_2 \big\|\omega_t^k\big\| \le \zeta U_\delta\big(c + 2U_\omega^2\big),$$

and $\|m_{t+1} - m_t\|_F = \big\|\mathbf{B}_\perp^{*\top}(\mathbf{B}_{t+1} - \mathbf{B}_t)\big\|_F \le \sqrt{r}\,\|\mathbf{B}_{t+1} - \mathbf{B}_t\|_2 \le 2\sqrt{r}\,\zeta U_\delta U_\omega$. With $\big|\|u\|^2 - \|v\|^2\big| \le \|u - v\|(\|u\| + \|v\|)$, $\|x_t^k\| \le \|\mathbf{B}_t\omega_t^k\| + \|z^{k,*}\| \le 2U_\omega$ and $\|m_t\|_F \le \|\mathbf{B}_t\|_F = \sqrt{r}$,

$$\big|\|x_{t+1}^k\|^2 - \|x_t^k\|^2\big| \le 4U_\omega U_\delta\big(c + 2U_\omega^2\big)\zeta, \qquad \big|\|m_{t+1}\|_F^2 - \|m_t\|_F^2\big| \le 4r\,\zeta U_\delta U_\omega,$$

and Eq. (84) gives the claim. □

I.2 Initialization phase

Setting. In Algorithm 1, replace r by d, and $\mathbf{B}_0 =$ by $\mathbf{I}_d$. Then $\mathbf{B}_{t,\perp} \in \mathbb{R}^{d\times 0}$, so $\mathbf{B}_{t,\perp}\mathbf{B}_{t,\perp}^\top = 0$, and the update of $\mathbf{B}^k$ in Algorithm 1 gives $\mathbf{Q}_t \equiv 0$, $\mathbf{R}_{t+1} = \mathbf{I}$, and

$$\mathbf{B}_t \equiv \mathbf{I}_d, \qquad \mathbf{P}_t = \mathbf{I}_d, \qquad \mathbf{P}_{t,\perp} = 0, \qquad m_t = 0.$$

The initialization phase runs for T' rounds, with head stepsize β', reward stepsize γ', and its own mixing index $\tau' := \lceil 2\log_\rho(\gamma')\rceil$, so that $\rho^{\tau'} \le \gamma'^2$. Writing $z_t^k := \omega_t^k$ and $x_t^k = z_t^k - z^{k,*}$, the head update is

$$z_{t+1}^k = \Pi_{U_\omega}\Big(z_t^k + \frac{\beta'}{L}\sum_{\ell=0}^{L-1} \delta_{t,\ell}^k \phi(s_{t,\ell}^k)\Big).$$

After T' rounds each agent sends $z_{T'}^k$; the server sets $\mathbf{B}_0$ to the top-r eigenvectors of $\frac{1}{K}\sum_k z_{T'}^k (z_{T'}^k)^\top$ and broadcasts it, and each agent sets $\omega_0^k = \mathbf{B}_0^\top z_{T'}^k$. Algorithm 1 then starts from $(\mathbf{B}_0, \{\omega_0^k\}_k)$.

Algorithm 2 pFedAR-TD with initialization

Input: stepsizes β', γ' and horizon T' for the initialization; stepsizes β, ζ, γ and horizon T for Algorithm 1; rank r; projection bound U_ω.
For $k = 1, \dots, K$: $\eta_0^k = 0$, z_0^k arbitrary with $\|z_0^k\| \le U_\omega$, $s_{0,0}^k$ drawn arbitrarily from $\mathcal{S}$.
INITIALIZATION. $r = d$, $\mathbf{B}_t \equiv \mathbf{I}_d$, no server step
for $t = 0, 1, \dots, T' - 1$ **do**
  for $k = 1, \dots, K$ **do**
    $\delta_{t,\ell}^k = r_{t,\ell}^k - \eta_t^k + \big(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k)\big)^\top z_t^k$ for $\ell = 0, \dots, L-1$.
    $z_{t+1}^k = \Pi_{U_\omega}\Big(z_t^k + \frac{\beta'}{L}\sum_{\ell=0}^{L-1} \delta_{t,\ell}^k \phi(s_{t,\ell}^k)\Big)$.
    $\eta_{t+1}^k = \eta_t^k + \gamma'\Big(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - \eta_t^k\Big)$.
    $s_{t+1,0}^k = s_{t,L}^k$.
  end for
end for
HANDOFF.
Each agent sends $z_{T'}^k$.
Server sets $\mathbf{B}_0$ to the top-r eigenvectors of $\frac{1}{K}\sum_k z_{T'}^k (z_{T'}^k)^\top$ and broadcasts it.
Each agent sets $\omega_0^k = \mathbf{B}_0^\top z_{T'}^k$.
Each agent keeps $s_{0,0}^k \leftarrow s_{T',L}^k$; the reward estimator restarts, $\eta_0^k = 0$.
Run Algorithm 1 from $\big(\mathbf{B}_0, \{\omega_0^k\}\big)$ with stepsizes β, ζ and γ satisfying Eq. (118), horizon T.

Reduction. In Eq. (30) the terms I_2, I_3, I_5, I_6 each carry a factor $\mathbf{P}_{t,\perp} = 0$, and in Eq. (60) every perturbation term carries $\|\mathbf{Q}_t\|_F = 0$. With $\mathbf{B}_t\mathbf{B}_t^\top = \mathbf{I}$, and Π_{U_ω} non-expansive with $\|z^{k,*}\| \le U_\omega$,

$$\big\|x_{t+1}^k\big\|^2 \le \big\|x_t^k\big\|^2 + 2\beta'\Big\langle x_t^k,\ \frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big\rangle + 3\beta'^2\Big\|\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big\|^2. \tag{87}$$

Lemma I.2 (Initialization recursion and mean-square bound)**.** *For β' with $36(1+\tau'^2)\beta' \le \lambda/2$, $t > 2\tau'$ and every k,*

$$\mathbb{E}_{t-2\tau'}\big[\|x_{t+1}^k\|^2\big] \le (1-\lambda\beta')\,\mathbb{E}_{t-2\tau'}\big[\|x_t^k\|^2\big] + \epsilon_t',$$

where

$$\epsilon_t' = \Big(\frac{2\beta'}{\lambda} + 9\beta'^2 + 3\tau'^2\beta'^2\Big)\mathbb{E}_{t-2\tau'}|y_t^k|^2 + \frac{\mathcal{C}'\beta'^2}{L} + 72\tau'^4U_\delta^2\beta'^4 + 36\tau'^4U_r^2\gamma'^2\beta'^2 + 16U_\omega(U_r+U_\omega)\beta'\frac{m\rho^{\tau' L}}{L(1-\rho)},$$

$$\mathcal{C}' := \Big(9 + \tfrac{9}{2}\tau'^2\Big)U_\delta^2 + 10U_\omega^2\Big(16 + \frac{32m}{1-\rho}\Big) + 2\Big(16U_r^2 + \frac{32U_r^2 m}{1-\rho}\Big).$$

Consequently, for $T' \ge 2\tau'$ and every k,

$$\mathbb{E}\big\|x_{T'}^k\big\|^2 \le (1-\lambda\beta')^{T'-2\tau'}\,4U_\omega^2 + \sum_{s=2\tau'}^{T'-1}(1-\lambda\beta')^{T'-1-s}\,\mathbb{E}[\epsilon_s']. \tag{88}$$

Proof. Apply $\mathbb{E}_{t-2\tau'}$ to Eq. (87). For the second term, Eq. (31) at $\beta = \beta'$ and $\mathbf{P}_t = \mathbf{I}$ splits it into $2\beta'\langle x_t^k, A^k x_t^k\rangle + I_{12} + I_{13}$, and $2\beta'\langle x_t^k, A^k x_t^k\rangle \le -2\lambda\beta'\|x_t^k\|^2$ by Assumption 1, while $I_{13} \le 2\beta'\|x_t^k\||y_t^k| \le \frac{\lambda\beta'}{2}\|x_t^k\|^2 + \frac{2\beta'}{\lambda}|y_t^k|^2$.

In the seven-way split of I_{12} preceding Eq. (34), I_{122} and I_{125} carry a factor $\mathbf{P}_t - \mathbf{P}_{t-\tau'} = 0$, and $\mathbf{B}_t = \mathbf{B}_{t-\tau'} = \mathbf{I}_d$ gives $\|x_t^k - x_{t-\tau'}^k\| = \|\omega_t^k - \omega_{t-\tau'}^k\|$. Hence, by Cauchy–Schwarz and $2ab \le \frac{1}{2}a^2 + 2b^2$,

$$I_{121} \le 2\beta'\|\omega_t^k - \omega_{t-\tau'}^k\|\|\widetilde{b}_t^k - \bar{b}^k\| \le \tfrac{1}{2}\|\omega_t^k - \omega_{t-\tau'}^k\|^2 + 2\beta'^2\|\widetilde{b}_t^k - \bar{b}^k\|^2,$$

$$I_{124} \le 2\beta' U_\omega\|\omega_t^k - \omega_{t-\tau'}^k\|\|\widetilde{A}_t^k - A^k\| \le \tfrac{1}{2}\|\omega_t^k - \omega_{t-\tau'}^k\|^2 + 2U_\omega^2\beta'^2\|\widetilde{A}_t^k - A^k\|^2,$$

$$I_{126} \le 4\beta' U_\omega\|\widetilde{A}_t^k - A^k\|\|\omega_t^k - \omega_{t-\tau'}^k\| \le \tfrac{1}{2}\|\omega_t^k - \omega_{t-\tau'}^k\|^2 + 8U_\omega^2\beta'^2\|\widetilde{A}_t^k - A^k\|^2,$$

and $\mathbb{E}_{t-\tau'}[I_{123}+I_{127}] \le 16U_\omega(U_r+U_\omega)\beta' \frac{m\rho^{\tau' L}}{L(1-\rho)}$ by Eq. (40). By Cauchy–Schwarz and Lemma K.2, whose ζ-terms enter through $\|\mathbf{B}_j - \mathbf{B}_t\|$ in Lemma K.4 and vanish since $\mathbf{B}_j \equiv \mathbf{I}_d$,

$$\begin{aligned}\frac{3}{2}\mathbb{E}_{t-2\tau'}\|\omega_t^k - \omega_{t-\tau'}^k\|^2 &\le \frac{3}{2}\tau' \sum_{j=t-\tau'}^{t-1} \mathbb{E}_{t-2\tau'}\|\omega_{j+1}^k - \omega_j^k\|^2 \\ &\le 36\tau'^2\beta'^2\mathbb{E}_{t-2\tau'}\|x_t^k\|^2 + 3\tau'^2\beta'^2\mathbb{E}_{t-2\tau'}|y_t^k|^2 + \frac{9}{2}\tau'^2 U_\delta^2\frac{\beta'^2}{L} \\ &\quad + 72\tau'^4 U_\delta^2\beta'^4 + 36\tau'^4 U_r^2\gamma'^2\beta'^2.\end{aligned}$$

Summing, and using Eq. (153), Eq. (151),

$$\begin{aligned}\mathbb{E}_{t-2\tau'}[I_{12}] &\le 36\tau'^2\beta'^2\mathbb{E}_{t-2\tau'}\|x_t^k\|^2 + 3\tau'^2\beta'^2\mathbb{E}_{t-2\tau'}|y_t^k|^2 \\ &\quad + \frac{\beta'^2}{L}\Big[\tfrac{9}{2}\tau'^2U_\delta^2 + 10U_\omega^2\Big(16 + \frac{32m}{1-\rho}\Big) + 2\Big(16U_r^2 + \frac{32U_r^2 m}{1-\rho}\Big)\Big] \\ &\quad + 72\tau'^4U_\delta^2\beta'^4 + 36\tau'^4U_r^2\gamma'^2\beta'^2 + 16U_\omega(U_r+U_\omega)\beta'\frac{m\rho^{\tau' L}}{L(1-\rho)}.\end{aligned}$$

For the third term of Eq. (87), Eq. (56) at $\beta = \beta'$ gives

$$3\beta'^2\,\mathbb{E}_{t-2\tau'}\Big\|\frac{1}{L}\sum_\ell \delta_{t,\ell}^k\phi(s_{t,\ell}^k)\Big\|^2 \le 36\beta'^2\mathbb{E}_{t-2\tau'}\|x_t^k\|^2 + 9\beta'^2\mathbb{E}_{t-2\tau'}|y_t^k|^2 + 9U_\delta^2\frac{\beta'^2}{L}.$$

Adding and collecting the $\mathbb{E}_{t-2\tau'}\|x_t^k\|^2$ coefficients,

$$\mathbb{E}_{t-2\tau'}\big[\|x_{t+1}^k\|^2\big] \le \Big(1 - \frac{3\lambda\beta'}{2} + 36(1+\tau'^2)\beta'^2\Big)\mathbb{E}_{t-2\tau'}\big[\|x_t^k\|^2\big] + \epsilon_t',$$

and $36(1+\tau'^2)\beta' \le \lambda/2$ gives $1 - \frac{3\lambda\beta'}{2} + 36(1+\tau'^2)\beta'^2 \le 1 - \lambda\beta'$. For Eq. (88), take the total expectation, unroll from $t = 2\tau'$, and use $\|x_{2\tau'}^k\| \le \|\mathbf{B}_{2\tau'}\omega_{2\tau'}^k\| + \|z^{k,*}\| \le 2U_\omega$. □

Lemma I.3 (Initialization high-probability bound)**.** *Write $\varepsilon'^2 := \frac{1}{K}\sum_{k=1}^K \|z_{T'}^k - z^{k,*}\|^2$, and let B denote the right-hand side of Eq. (88). For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\varepsilon'^2 \;\le\; B + 2U_\omega\sqrt{\frac{2B\log(1/\delta)}{K}} + \frac{8U_\omega^2\log(1/\delta)}{3K}.$$

Proof. Put $X_k := \|x_{T'}^k\|^2$, so $\varepsilon'^2 = \frac{1}{K}\sum_k X_k$. During this phase, agents exchange nothing before T', so $X_1, \dots, X_K$ are independent. By Π_{U_ω} and $\|z^{k,*}\| \le U_\omega$, $0 \le X_k \le 4U_\omega^2$, so

$$\mathrm{Var}(X_k) \le \mathbb{E}[X_k^2] \le 4U_\omega^2\,\mathbb{E}[X_k] \le 4U_\omega^2 B$$

by Eq. (88). Bernstein's inequality gives, with probability at least $1-\delta$,

$$\frac{1}{K}\sum_k X_k \le \mathbb{E}[X_1] + \sqrt{\frac{2\,\mathrm{Var}(X_1)\log(1/\delta)}{K}} + \frac{2\cdot 4U_\omega^2\log(1/\delta)}{3K},$$

and substituting $\mathbb{E}[X_1] \le B$ and $\mathrm{Var}(X_1) \le 4U_\omega^2 B$, both of which the right-hand side is increasing in, gives the claim. □

Lemma I.4. *With $\mathbf{B}_0, \omega_0^k$ as above,*

$$\frac{1}{K}\sum_{k=1}^K \|x_0^k\|^2 \le 4\varepsilon'^2, \qquad \|m_0\|_F^2 \le \frac{4\varepsilon'^2}{\alpha}, \qquad n_0 \le 4\varepsilon'^2\Big(1 + \frac{c}{\alpha}\Big).$$

Proof. $\mathbf{B}_0$ minimises $\frac{1}{K}\sum_k \|\mathbf{P}_\perp z_{T'}^k\|^2$ over rank-r orthogonal projections, so comparing with $\mathbf{B}^*$ and using $\mathbf{P}_\perp^* z^{k,*} = 0$,

$$\frac{1}{K}\sum_k \|\mathbf{P}_{0,\perp} z_{T'}^k\|^2 \le \frac{1}{K}\sum_k \|\mathbf{P}_\perp^* z_{T'}^k\|^2 = \frac{1}{K}\sum_k \|\mathbf{P}_\perp^*(z_{T'}^k - z^{k,*})\|^2 \le \varepsilon'^2.$$

By $\mathbf{P}_0 + \mathbf{P}_{0,\perp} = \mathbf{I}$, $x_0^k = \mathbf{P}_0 z_{T'}^k - z^{k,*} = (z_{T'}^k - z^{k,*}) - \mathbf{P}_{0,\perp} z_{T'}^k$, and the triangle inequality applied to the concatenation of the K blocks in $\mathbb{R}^{dK}$ gives

$$\sqrt{\sum_k \|x_0^k\|^2} \le \sqrt{\sum_k \|z_{T'}^k - z^{k,*}\|^2} + \sqrt{\sum_k \|\mathbf{P}_{0,\perp} z_{T'}^k\|^2} \le \sqrt{K}\,\varepsilon' + \sqrt{K}\,\varepsilon',$$

so $\frac{1}{K}\sum_k \|x_0^k\|^2 \le 4\varepsilon'^2$. The second bound is Lemma 5.3, which gives $\alpha\|m_0\|_F^2 \le \frac{1}{K}\sum_k \|x_0^k\|^2$; the third follows by substituting both into Eq. (84). □

Lemma I.5. *γ', β', T' can be chosen, independently of T and K, so that*

$$B \;\le\; \frac{c\lambda^2}{1024\big(1+\frac{c}{\alpha}\big)}.$$

Proof. By Lemma I.2,

$$B = (1-\lambda\beta')^{T'-2\tau'} 4U_\omega^2 + \sum_{s=2\tau'}^{T'-1} (1-\lambda\beta')^{T'-1-s}\Big[\Big(\frac{2\beta'}{\lambda} + 9\beta'^2 + 3\tau'^2\beta'^2\Big)\mathbb{E}|y_s^k|^2 + \frac{\mathcal{C}'\beta'^2}{L} + 72\tau'^4 U_\delta^2 \beta'^4 + 36\tau'^4 U_r^2 \gamma'^2 \beta'^2 + 16U_\omega(U_r+U_\omega)\beta' \frac{m\rho^{\tau' L}}{L(1-\rho)}\Big]. \tag{89}$$

Take $\lambda\beta' \le \gamma'/2$ and $\gamma'\tau' \le \frac{1}{2}$, the latter secured by the choice of γ' below and giving $\gamma' \le \frac{1}{2}$ by $\tau' \ge 1$.

The terms inside the sum of Eq. (89) other than the one containing y do not depend on s, so $\sum_s (1-\lambda\beta')^{T'-1-s} \le \frac{1}{\lambda\beta'}$ bounds their contribution by

$$\frac{\mathcal{C}'\beta'}{\lambda L} + \frac{72\tau'^4 U_\delta^2 \beta'^3}{\lambda} + \frac{36\tau'^4 U_r^2 \gamma'^2 \beta'}{\lambda} + \frac{16U_\omega(U_r+U_\omega) m\rho^{\tau' L}}{\lambda L(1-\rho)}. \tag{90}$$

For the term containing y, $\tau' = \lceil 2\log_\rho(\gamma')\rceil$ gives $\rho^{\tau' L} \le \gamma'^{2L} \le \gamma'$, so Lemma G.1 applies at stepsize γ' and index τ'; with $\sum_s (1-\lambda\beta')^{T'-1-s} \le \frac{1}{\lambda\beta'}$ it gives

$$\begin{aligned}
&\sum_{s=2\tau'}^{T'-1} (1-\lambda\beta')^{T'-1-s}\Big(\frac{2\beta'}{\lambda} + (9+3\tau'^2)\beta'^2\Big)\mathbb{E}|y_s^k|^2 \\
&\quad\le \Big(\frac{2}{\lambda^2} + \frac{(9+3\tau'^2)\beta'}{\lambda}\Big)\mathcal{C}_r\Big(\frac{\gamma'}{L} + \tau'^2\gamma'^2\Big) \\
&\qquad + \Big(\frac{2\beta'}{\lambda} + (9+3\tau'^2)\beta'^2\Big)\mathcal{C}_r \sum_{s=2\tau'}^{T'-1} (1-\lambda\beta')^{T'-1-s} e^{-2\gamma' s},
\end{aligned}$$

Reindexing the last sum by $j = T' - 1 - s$,

$$\sum_{s=2\tau'}^{T'-1} (1-\lambda\beta')^{T'-1-s} e^{-2\gamma' s} = e^{-2\gamma'(T'-1)} \sum_{j=0}^{T'-1-2\tau'} \big((1-\lambda\beta')e^{2\gamma'}\big)^j,$$

whose ratio exceeds $1+\gamma'$:

$$(1-\lambda\beta')e^{2\gamma'} \;\ge\; \Big(1-\frac{\gamma'}{2}\Big)e^{2\gamma'} \;\ge\; 1+\gamma',$$

the first step by $\lambda\beta' \le \gamma'/2$ and the second because $(1-\frac{\gamma'}{2})e^{2\gamma'} - 1 - \gamma'$ vanishes at $\gamma' = 0$ and has derivative $e^{2\gamma'}(\frac{3}{2} - \gamma') - 1 \ge 0$ on $[0, \frac{1}{2}]$. Hence

$$\begin{aligned}
&e^{-2\gamma'(T'-1)} \sum_{j=0}^{T'-1-2\tau'} \big((1-\lambda\beta')e^{2\gamma'}\big)^j \\
&\quad\le e^{-2\gamma'(T'-1)} \frac{\big((1-\lambda\beta')e^{2\gamma'}\big)^{T'-2\tau'}}{\gamma'} = \frac{(1-\lambda\beta')^{T'-2\tau'} e^{-2\gamma'(2\tau'-1)}}{\gamma'} \\
&\quad\le \frac{e^{-\lambda\beta'(T'-2\tau')}}{\gamma'},
\end{aligned}$$

the last step by $\tau' \geq 1$ and $0 \leq 1-\lambda\beta' \leq e^{-\lambda\beta'}$. The latter also bounds the first term of Eq. (89) by $4U_\omega^2 e^{-\lambda\beta'(T'-2\tau')}$, so

$$\begin{aligned} B \;\leq\; & \Big(\frac{2}{\lambda^2}+\frac{(9+3\tau'^2)\beta'}{\lambda}\Big)\mathcal{C}_r\Big(\frac{\gamma'}{L}+\tau'^2\gamma'^2\Big) \\ & + Eq.\,(90) \\ & + \Big(4U_\omega^2+\frac{\mathcal{C}_r}{\gamma'}\Big(\frac{2\beta'}{\lambda}+(9+3\tau'^2)\beta'^2\Big)\Big)e^{-\lambda\beta'(T'-2\tau')}. \end{aligned} \tag{91}$$

Each of the four terms below is made at most $\frac{c\lambda^2}{4096\left(1+\frac{c}{\alpha}\right)}$, a quarter of the target.

First term of Eq. (91). By $\beta' \leq \gamma'/(2\lambda) \leq 1/\lambda$ its bracket is at most $\frac{2}{\lambda^2}+\frac{9+3\tau'^2}{\lambda^2}=\frac{11+3\tau'^2}{\lambda^2}$, so it suffices that

$$\frac{(11+3\tau'^2)\mathcal{C}_r}{\lambda^2}\Big(\frac{\gamma'}{L}+\tau'^2\gamma'^2\Big) \;\leq\; \frac{c\lambda^2}{4096\big(1+\frac{c}{\alpha}\big)}. \tag{92}$$

Last term of Eq. (90). By $\rho^{\tau' L} \leq \gamma'^{2L} \leq \gamma'^2$ it suffices that

$$\gamma'^2 \;\leq\; \frac{c\lambda^3 L(1-\rho)}{65536\,U_\omega(U_r+U_\omega)m\big(1+\frac{c}{\alpha}\big)}. \tag{93}$$

Since $\tau' = \lceil 2\log_\rho(\gamma')\rceil = \mathcal{O}(\log\frac{1}{\gamma'})$ grows only logarithmically as $\gamma' \downarrow 0$, we can pick a γ' small enough that Eq. (92), Eq. (93) and $\gamma'\tau' \leq \frac{1}{2}$ hold together, and this γ' in turn fixes τ'.

First three terms of Eq. (90). By $\beta'^3 \leq \beta'\gamma'^2/\lambda^2$ they are at most $\beta'\big[\frac{\mathcal{C}'}{\lambda L}+\frac{72\tau'^4U_\delta^2\gamma'^2}{\lambda^3}+\frac{36\tau'^4U_r^2\gamma'^2}{\lambda}\big]$, so it suffices that

$$\beta' \;\leq\; \min\Big\{\frac{\gamma'}{2\lambda},\;\frac{\lambda}{72(1+\tau'^2)},\;\frac{c\lambda^2}{4096\big(1+\frac{c}{\alpha}\big)\big[\frac{\mathcal{C}'}{\lambda L}+\frac{72\tau'^4U_\delta^2\gamma'^2}{\lambda^3}+\frac{36\tau'^4U_r^2\gamma'^2}{\lambda}\big]}\Big\}, \tag{94}$$

the first two entries being $\lambda\beta' \leq \gamma'/2$ and $36(1+\tau'^2)\beta' \leq \lambda/2$. Now that γ' and τ' are fixed, the right-hand side of Eq. (94) is a number, and β' can be chosen to satisfy it.

Third term of Eq. (91). Solving for T', it suffices that

$$\begin{aligned} T' \;\geq\; & 2\tau' + \frac{1}{\lambda\beta'}\log\Big[\frac{4096\big(1+\frac{c}{\alpha}\big)}{c\lambda^2} \\ & \times\Big(4U_\omega^2+\frac{\mathcal{C}_r}{\gamma'}\Big(\frac{2\beta'}{\lambda}+(9+3\tau'^2)\beta'^2\Big)\Big)\Big]. \end{aligned} \tag{95}$$

With γ', τ' and β' all fixed, the right-hand side of Eq. (95) is a number as well, and T' can be chosen to satisfy it.

Adding the four, $B \leq \frac{c\lambda^2}{1024(1+\frac{c}{\alpha})}$. None of γ', β', T' involves T or K. □

Proposition 3. *Run the initialization phase with γ', β', T' chosen as in Lemma I.5, and let $K \geq 8192\,U_\omega^2(1+c/\alpha)\log(2/\delta)/(c\lambda^2)$. Then, with probability at least $1-\delta/2$,*

$$n_0 \;\leq\; \frac{c\lambda^2}{64}.$$

Proof. By Lemma I.4 it suffices that $\varepsilon'^2 \leq \frac{c\lambda^2}{256(1+c/\alpha)}$. Apply Lemma I.3 at $\delta/2$, with $B \leq \frac{c\lambda^2}{1024(1+c/\alpha)}$ from Lemma I.5; the condition on K reads $\frac{2\log(2/\delta)}{K} \leq \frac{c\lambda^2}{4096U_\omega^2(1+c/\alpha)}$, so the second term of Lemma I.3 is

$$2U_\omega\sqrt{B\cdot\frac{2\log(2/\delta)}{K}} \leq 2U_\omega\sqrt{\frac{c\lambda^2}{1024(1+\frac{c}{\alpha})}\cdot\frac{c\lambda^2}{4096U_\omega^2(1+\frac{c}{\alpha})}} = \frac{c\lambda^2}{1024\big(1+\frac{c}{\alpha}\big)},$$

and the third is $\frac{8U_\omega^2 \log(2/\delta)}{3K} = \frac{4}{3}U_\omega^2 \cdot \frac{2\log(2/\delta)}{K} \le \frac{c\lambda^2}{3072(1+\frac{c}{\alpha})}$. Adding B,

$$\varepsilon'^2 \le \Big(1+1+\tfrac{1}{3}\Big)\frac{c\lambda^2}{1024\big(1+\frac{c}{\alpha}\big)} < \frac{c\lambda^2}{256\big(1+\frac{c}{\alpha}\big)}.$$

□

I.3 Below the barrier

Write $a := \frac{c\lambda\alpha}{4(\alpha+c)}\zeta$. Eq. (82) is derived under the hypothesis $\|m_t\|_2 \le \lambda/4$ of Proposition 1, which by Eq. (85) holds pointwise on $\{n_t \le c\lambda^2/16\}$; and $\widehat{\mathcal{V}}_t = \mathbb{E}_{t-2\tau}[n_t]$ by Eq. (12) and Eq. (84). So Eq. (82) reads

$$\text{on } \Big\{n_t \le \frac{c\lambda^2}{16}\Big\}: \qquad \mathbb{E}_{t-2\tau}\big[n_{t+1}\big] \le (1-a)\,\mathbb{E}_{t-2\tau}\big[n_t\big] + \epsilon_t, \qquad t \ge 2\tau, \tag{96}$$

where ϵ_t is a *random variable*. Collecting the coefficients of $\widehat{Y}_t$ in Eq. (61) and Eq. (76),

$$\epsilon_t \;\le\; 2\mathcal{C}_{\epsilon,1}\,\zeta\,\widehat{Y}_t + \mathcal{C}_{\epsilon,2}\,\zeta^2\Big(\frac{1}{L}+\frac{1}{K}\Big), \quad \mathcal{C}_{\epsilon,1} := \frac{c}{c_4^2} + \frac{U_\omega^4}{c_6^2} + \frac{cU_\omega^2}{c_1^2} = \frac{30c}{\lambda} + \frac{U_\omega^2}{2} + \frac{6cU_\omega^2}{\lambda\alpha}, \tag{97}$$

where $\mathcal{C}_{\epsilon,2}$ collects $\mathcal{C}_{X,5}, \mathcal{C}_{X,6}, c\mathcal{C}_{M,2}, 2cC_{25}$; the factor 2 covers $[\mathcal{C}_{X,3}(\tau^2) + c\mathcal{C}_{M,5}(\tau^2)]\zeta \le \mathcal{C}_{\epsilon,1}$, and the terms in $\zeta^3, \zeta^4, \gamma^2\zeta^2, \zeta^6$ are absorbed into $\mathcal{C}_{\epsilon,2}$ once $\zeta \le 1/L$ and $\gamma^2 \le 1/K$.

Recall that

$$n_t^{\text{in}} = \mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}}\, n_t. \tag{98}$$

Lemma I.6. *Assume the conditions of Eq. (97) and $2\tau U_n \zeta \le c\lambda^2/32$. Then for $t \ge 2\tau$,* unconditionally,

$$\mathbb{E}\big[n_{t+1}^{\text{in}}\big] \le (1-a)\,\mathbb{E}\big[n_t^{\text{in}}\big] + \mathbb{E}[\epsilon_t].$$

Proof. By definition, $n_{t+1}^{\text{in}} := \mathbf{1}_{\{T_{\text{exit}} > t+1-2\tau\}}\, n_{t+1}$. Since $\{T_{\text{exit}} > t+1-2\tau\} \subseteq \{T_{\text{exit}} > t-2\tau\}$ and $n_{t+1} \ge 0$, we get

$$n_{t+1}^{\text{in}} \le \mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}}\, n_{t+1},$$

which implies

$$\begin{aligned}
\mathbb{E}_{t-2\tau}[n_{t+1}^{\text{in}}] \le& \mathbb{E}_{t-2\tau}\left[\mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}}\, n_{t+1}\right] \\
=& \mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}} \mathbb{E}_{t-2\tau}\left[n_{t+1}\right] \\
\overset{(a)}{\le}& \mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}} (1-a)\,\mathbb{E}_{t-2\tau}\big[n_t\big] + \epsilon_t \\
=& (1-a)\,\mathbb{E}_{t-2\tau}\big[\mathbf{1}_{\{T_{\text{exit}} > t-2\tau\}} n_t\big] + \epsilon_t \\
=& (1-a)\,\mathbb{E}_{t-2\tau}\big[n_t^{\text{in}}\big] + \epsilon_t.
\end{aligned}$$

For (a), on $\{T_{\text{exit}} > t-2\tau\}$ we have $n_{t-2\tau} \le \frac{c\lambda^2}{32}$ by Eq. (86), so

$$n_t \le n_{t-2\tau} + 2\tau U_n \zeta \le \frac{c\lambda^2}{32} + \frac{c\lambda^2}{32} = \frac{c\lambda^2}{16},$$

and Eq. (96) applies. Off that event the indicator vanishes and (a) reads $0 \le \epsilon_t$, so (a) holds everywhere. Taking the total expectation and using the tower property gives

$$\mathbb{E}\big[n_{t+1}^{\text{in}}\big] \le (1-a)\,\mathbb{E}\big[n_t^{\text{in}}\big] + \mathbb{E}[\epsilon_t].$$

□

Corollary 1. *Assume the conditions of Lemma I.6 and of Lemma G.1, together with $\zeta^2 \le \gamma$ and $a \le 1$. Then*

$$\begin{aligned}
\mathbb{E}\big[n_{T+1}^{\text{in}}\big] \;\le& (1-a)^{T+1-2\tau}\big(4U_\omega^2 + cr\big) + 2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\,\zeta\,T\max\big\{e^{-2\gamma T}, e^{-a(T-2\tau)}\big\} \\
&+ \frac{2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\,\zeta}{a}\Big(\frac{\gamma}{L} + \tau^2\gamma^2\Big) + \frac{\mathcal{C}_{\epsilon,2}\,\zeta^2}{a}\Big(\frac{1}{L}+\frac{1}{K}\Big).
\end{aligned} \tag{99}$$

Proof. Applying Lemma I.6 at $t = T$, then at $t = T-1$, and so on down to $t = 2\tau$,

$$\begin{aligned}\mathbb{E}\big[n_{T+1}^{\text{in}}\big] &\le (1-a)\,\mathbb{E}\big[n_T^{\text{in}}\big] + \mathbb{E}[\epsilon_T] \\ &\le (1-a)^2\,\mathbb{E}\big[n_{T-1}^{\text{in}}\big] + (1-a)\,\mathbb{E}[\epsilon_{T-1}] + \mathbb{E}[\epsilon_T] \\ &\;\;\vdots \\ &\le (1-a)^{T+1-2\tau}\,\mathbb{E}\big[n_{2\tau}^{\text{in}}\big] + \sum_{s=2\tau}^{T}(1-a)^{T-s}\,\mathbb{E}[\epsilon_s].\end{aligned}$$

Here $n_{2\tau}^{\text{in}} \le n_{2\tau} \le 4U_\omega^2 + cr$ by $\|x_t^k\| \le 2U_\omega$ and $\|m_t\|_F^2 \le r$; the total expectation in Eq. (97) with $\mathbb{E}[\widehat{Y}_s] = \bar{Y}_s$ gives

$$\mathbb{E}[\epsilon_s] \;\le\; 2\mathcal{C}_{\epsilon,1}\,\zeta\,\bar{Y}_s + \mathcal{C}_{\epsilon,2}\,\zeta^2\Big(\frac{1}{L} + \frac{1}{K}\Big);$$

and with $\tau = \lceil 2\log_\rho(\zeta)\rceil$ we have $\rho^{\tau L} \le \zeta^{2L} \le \zeta^2 \le \gamma$, so Lemma G.1 applies and, averaged over k, $\bar{Y}_s \le \mathcal{C}_r\big(e^{-2\gamma s} + \frac{\gamma}{L} + \tau^2\gamma^2\big)$. Substituting the three,

$$\begin{aligned}\mathbb{E}\big[n_{T+1}^{\text{in}}\big] \le (1-a)^{T+1-2\tau}\big(4U_\omega^2 + cr\big) + 2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\zeta\sum_{s=2\tau}^{T}(1-a)^{T-s}e^{-2\gamma s} \\ + \Big[2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\zeta\Big(\frac{\gamma}{L} + \tau^2\gamma^2\Big) + \mathcal{C}_{\epsilon,2}\zeta^2\Big(\frac{1}{L} + \frac{1}{K}\Big)\Big]\sum_{s=2\tau}^{T}(1-a)^{T-s}.\end{aligned}$$

For the third term, $\sum_{s=2\tau}^{T}(1-a)^{T-s} = \sum_{j=0}^{T-2\tau}(1-a)^j \le \frac{1}{a}$. For the second, use $(1-a)^{T-s} \le e^{-a(T-s)}$; the exponent of $e^{-a(T-s)}e^{-2\gamma s}$ is affine in s, so its maximum over $2\tau \le s \le T$ is at an endpoint, namely $e^{-2\gamma T}$ at $s = T$ and at most $e^{-a(T-2\tau)}$ at $s = 2\tau$, and there are at most T summands. This gives Eq. (99). □

I.4 Stopped process

Lemma I.7 (Exit probability). *Let $\delta \in (0,1)$. Under the conditions collected in Section I.5, with the initialization phase as in Proposition 3 and the reward bound of Lemma I.8, each holding with probability at least $1 - \delta/2$,*

$$\Pr\big[T_{\text{exit}} \le T\big] \;\le\; \delta + \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big).$$

Proof. Splitting at the event that the initialization phase hands over $n_0 \le c\lambda^2/64$,

$$\big\{T_{\text{exit}} \le T\big\} = \Big(\big\{T_{\text{exit}} \le T\big\} \cap \Big\{n_0 \le \frac{c\lambda^2}{64}\Big\}\Big) \cup \Big(\big\{T_{\text{exit}} \le T\big\} \cap \Big\{n_0 > \frac{c\lambda^2}{64}\Big\}\Big),$$

so, by $\Pr[A \cup B] \le \Pr[A] + \Pr[B]$ and $A \cap B \subseteq B$,

$$\Pr\big[T_{\text{exit}} \le T\big] \;\le\; \Pr\Big[\big\{T_{\text{exit}} \le T\big\} \cap \Big\{n_0 \le \frac{c\lambda^2}{64}\Big\}\Big] + \frac{\delta}{2}, \tag{100}$$

the last term by Proposition 3.

By the definition of T_{exit},

$$\big\{T_{\text{exit}} \le T\big\} = \Big\{\max_{0\le t\le T} n_t > \frac{c\lambda^2}{32}\Big\}. \tag{101}$$

On $\{n_0 \le c\lambda^2/64\}$, for $t \le 2\tau$, Lemma I.1 applied t times gives $n_t \le n_0 + 2\tau U_n\zeta \le c\lambda^2/64 + 2\tau U_n\zeta \le c\lambda^2/32$, the last step by $2\tau U_n\zeta \le c\lambda^2/64$, so no $t \le 2\tau$ crosses the barrier $c\lambda^2/32$ and

$$\Big\{\max_{0\le t\le T} n_t > \frac{c\lambda^2}{32}\Big\} = \Big\{\max_{2\tau\le t\le T} n_t > \frac{c\lambda^2}{32}\Big\}. \tag{102}$$

The drift Eq. (96) is a statement about $\mathbb{E}_{t-2\tau}[n_t]$, not about n_t, so n_t should be replaced by it. Since n_t is not $\mathcal{F}_{t-2\tau}$-measurable, Lemma I.1 is used twice: $n_{t-2\tau}$ is $\mathcal{F}_{t-2\tau}$-measurable, so for $t \geq 2\tau$

$$\left|\mathbb{E}_{t-2\tau}[n_t] - n_t\right| \leq \left|\mathbb{E}_{t-2\tau}\left[n_t - n_{t-2\tau}\right]\right| + \left|n_{t-2\tau} - n_t\right| \;\leq\; 4\tau U_n \zeta, \tag{103}$$

pointwise. Hence $n_t \leq \mathbb{E}_{t-2\tau}[n_t] + 4\tau U_n\zeta$, and taking the maximum,

$$\left\{\max_{2\tau\leq t\leq T} n_t > \frac{c\lambda^2}{32}\right\} \subseteq \left\{\max_{2\tau\leq t\leq T} \mathbb{E}_{t-2\tau}[n_t] > \frac{c\lambda^2}{32} - 4\tau U_n\zeta\right\}. \tag{104}$$

Finally, Eq. (96) holds only on $\{n_t \leq c\lambda^2/16\}$, so $\mathbb{E}_{t-2\tau}[n_t]$ must be stopped at its own first crossing

$$\sigma := \min\left\{t \geq 2\tau :\; \mathbb{E}_{t-2\tau}[n_t] > \frac{c\lambda^2}{32} - 4\tau U_n\zeta\right\} \qquad (= \infty \text{ if no such } t),$$

a stopping time for $\{\mathcal{F}_{t-2\tau}\}$. The stopped process $\mathbb{E}_{(t\wedge\sigma)-2\tau}\left[n_{t\wedge\sigma}\right]$ exceeds the threshold $\frac{c\lambda^2}{32} - 4\tau U_n\zeta$ of σ somewhere in $[2\tau, T]$ exactly when $\mathbb{E}_{t-2\tau}[n_t]$ does. If $\mathbb{E}_{u-2\tau}[n_u]$ exceeds the threshold at some $2\tau \leq u \leq T$, then $\sigma \leq u$, so the stopped process at u equals $\mathbb{E}_{\sigma-2\tau}[n_\sigma]$, which exceeds the threshold by the definition of σ. If the stopped process exceeds the threshold at some u, then $\sigma \leq u \leq T$, and $\mathbb{E}_{t-2\tau}[n_t]$ exceeds the threshold at $t = \sigma$. So

$$\left\{\max_{2\tau\leq t\leq T} \mathbb{E}_{t-2\tau}[n_t] > \frac{c\lambda^2}{32} - 4\tau U_n\zeta\right\} = \left\{\max_{2\tau\leq t\leq T} \mathbb{E}_{(t\wedge\sigma)-2\tau}\left[n_{t\wedge\sigma}\right] > \frac{c\lambda^2}{32} - 4\tau U_n\zeta\right\}. \tag{105}$$

It remains to bound the probability of the right-hand event. For $2\tau \leq t \leq T$

$$\begin{aligned}\mathbb{E}_{(t\wedge\sigma)-2\tau}\left[n_{t\wedge\sigma}\right] &= \mathbb{E}_0\left[n_{2\tau}\right] \\ &\quad + \sum_{s=2\tau}^{t-1}\Big(\mathbb{E}_{((s+1)\wedge\sigma)-2\tau}\left[n_{(s+1)\wedge\sigma}\right] \\ &\qquad - \mathbb{E}_{(s\wedge\sigma)-2\tau}\left[n_{s\wedge\sigma}\right]\Big),\end{aligned} \tag{106}$$

the first term because $\sigma \geq 2\tau$. So the maximum in Eq. (105) is bounded once the starting value $\mathbb{E}_0[n_{2\tau}]$ and the increment are.

The starting value. By Lemma I.1 applied 2τ times, $\left|\mathbb{E}_0[n_{2\tau}] - n_0\right| \leq 2\tau U_n\zeta$, so on $\{n_0 \leq c\lambda^2/64\}$,

$$\mathbb{E}_0\left[n_{2\tau}\right] \;\leq\; \frac{c\lambda^2}{64} + 2\tau U_n\zeta.$$

The increment. Its two terms coincide when $\sigma \leq s$, so

$$\begin{aligned}&\mathbb{E}_{((s+1)\wedge\sigma)-2\tau}\left[n_{(s+1)\wedge\sigma}\right] \\ &\qquad - \mathbb{E}_{(s\wedge\sigma)-2\tau}\left[n_{s\wedge\sigma}\right] \\ &= \mathbf{1}_{\{\sigma>s\}}\left(\mathbb{E}_{s+1-2\tau}[n_{s+1}] - \mathbb{E}_{s-2\tau}[n_s]\right),\end{aligned} \tag{107}$$

and $\mathbf{1}_{\{\sigma>s\}}$ is $\mathcal{F}_{s-2\tau}$-measurable because σ is a stopping time for $\{\mathcal{F}_{t-2\tau}\}$.

The decomposition. Split the right-hand side of Eq. (107):

$$\begin{aligned}&\mathbf{1}_{\{\sigma>s\}}\left(\mathbb{E}_{s+1-2\tau}[n_{s+1}] - \mathbb{E}_{s-2\tau}[n_s]\right) \\ &\quad = \underbrace{\mathbf{1}_{\{\sigma>s\}}\left(\mathbb{E}_{s-2\tau}[n_{s+1}] - \mathbb{E}_{s-2\tau}[n_s]\right)}_{\text{compensator}} + \underbrace{\mathbf{1}_{\{\sigma>s\}}\left(\mathbb{E}_{s+1-2\tau}[n_{s+1}] - \mathbb{E}_{s-2\tau}[n_{s+1}]\right)}_{=:\,\Xi_{s+1}},\end{aligned} \tag{108}$$

one term with the filtration fixed and the index advanced, the other with the index fixed and the filtration advanced.

The compensator. On $\{\sigma > s\}$ the definition of σ gives $\mathbb{E}_{s-2\tau}[n_s] \leq c\lambda^2/32 - 4\tau U_n\zeta$, so Eq. (103) gives $n_s \leq \mathbb{E}_{s-2\tau}[n_s] + 4\tau U_n\zeta \leq c\lambda^2/32 \leq c\lambda^2/16$ and Eq. (96) applies there:

$$\mathbb{E}_{s-2\tau}\left[n_{s+1}\right] \;\leq\; (1-a)\,\mathbb{E}_{s-2\tau}[n_s] + \epsilon_s, \qquad a = \frac{c\lambda\alpha}{4(\alpha+c)}\zeta. \tag{109}$$

With $\mathbb{E}_{s-2\tau}[n_s] \geq 0$ and $\epsilon_s \geq 0$,

$$\mathbf{1}_{\{\sigma>s\}}\left(\mathbb{E}_{s-2\tau}[n_{s+1}] - \mathbb{E}_{s-2\tau}[n_s]\right) \;\leq\; \mathbf{1}_{\{\sigma>s\}}\left(-a\,\mathbb{E}_{s-2\tau}[n_s] + \epsilon_s\right) \;\leq\; \epsilon_s, \tag{110}$$

the last inequality also covering $\{\sigma \leq s\}$, where the left side is 0.

The term Ξ. Its three factors $\mathbb{E}_{s+1-2\tau}[n_{s+1}]$, $\mathbb{E}_{s-2\tau}[n_{s+1}]$ and $\mathbf{1}_{\{\sigma>s\}}$ are measurable with respect to $\mathcal{F}_{s+1-2\tau}$, $\mathcal{F}_{s-2\tau}$ and $\mathcal{F}_{s-2\tau}$ respectively, and the last two σ-algebras are contained in the first, so

$$\Xi_{s+1} \text{ is } \mathcal{F}_{s+1-2\tau}\text{-measurable.} \tag{111}$$

Since $\mathcal{F}_{s-2\tau} \subseteq \mathcal{F}_{s+1-2\tau}$ and $\mathbf{1}_{\{\sigma>s\}}$ is $\mathcal{F}_{s-2\tau}$-measurable, the tower property gives

$$\mathbb{E}\big[\Xi_{s+1} \,\big|\, \mathcal{F}_{s-2\tau}\big] = \mathbf{1}_{\{\sigma>s\}}\Big(\mathbb{E}\big[\mathbb{E}_{s+1-2\tau}[n_{s+1}] \,\big|\, \mathcal{F}_{s-2\tau}\big] - \mathbb{E}_{s-2\tau}\big[n_{s+1}\big]\Big) = 0. \tag{112}$$

The first term of Ξ_{s+1} is Eq. (103) at $t = s+1$, the second the same argument across $2\tau+1$ steps, so

$$\big|\Xi_{s+1}\big| \;\le\; 4\tau U_n\zeta + (4\tau+2)U_n\zeta \;=\; 2(4\tau+1)U_n\zeta. \tag{113}$$

The exit event. By Eq. (108) and Eq. (110), the summand of Eq. (106) is at most $\epsilon_s + \Xi_{s+1}$, so for every $2\tau \le t \le T$

$$\mathbb{E}_{(t\wedge\sigma)-2\tau}\big[n_{t\wedge\sigma}\big] \;\le\; \mathbb{E}_0\big[n_{2\tau}\big] + \sum_{s=2\tau}^{t-1}\epsilon_s + \sum_{s=2\tau}^{t-1}\Xi_{s+1}. \tag{114}$$

Taking the maximum in Eq. (114), and using $\epsilon_s \ge 0$ to make $\sum_{s=2\tau}^{t-1}\epsilon_s$ increasing in t and so pull it out of the maximum,

$$\max_{2\tau\le t\le T}\mathbb{E}_{(t\wedge\sigma)-2\tau}\big[n_{t\wedge\sigma}\big] \;\le\; \frac{c\lambda^2}{64} + 2\tau U_n\zeta + \sum_{s=2\tau}^{T-1}\epsilon_s + \max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1}.$$

With Eq. (101)–Eq. (105),

$$\begin{aligned}&\Big\{T_{\text{exit}} \le T\Big\} \cap \Big\{n_0 \le \frac{c\lambda^2}{64}\Big\}\\ &\qquad\subseteq \Big\{\sum_{s=2\tau}^{T-1}\epsilon_s + \max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1} > \frac{c\lambda^2}{64} - 6\tau U_n\zeta\Big\}.\end{aligned}$$

The drift budget $\sum_{s=2\tau}^{T-1}\epsilon_s$. The threshold Azuma is given must be deterministic and bounded below by a positive constant, and $\sum_{s=2\tau}^{T-1}\epsilon_s$ is neither. By Eq. (97),

$$\sum_{s=2\tau}^{T-1}\epsilon_s \;\le\; 2\mathcal{C}_{\epsilon,1}\,\zeta\sum_{s=2\tau}^{T-1}\widehat{Y}_s + \mathcal{C}_{\epsilon,2}\,(T-2\tau)\,\zeta^2\Big(\frac{1}{L}+\frac{1}{K}\Big),$$

the second term deterministic and the first random. Cut the first out at $c\lambda^2/128$,

$$\begin{aligned}&\Big\{\sum_{s=2\tau}^{T-1}\epsilon_s + \max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1} > \frac{c\lambda^2}{64} - 6\tau U_n\zeta\Big\}\\ &\qquad\subseteq \Big\{2\mathcal{C}_{\epsilon,1}\,\zeta\sum_{s=2\tau}^{T-1}\widehat{Y}_s > \frac{c\lambda^2}{128}\Big\} \cup \Big\{\max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1} > u\Big\},\end{aligned} \tag{115}$$

where $u := \frac{c\lambda^2}{128} - 6\tau U_n\zeta - \mathcal{C}_{\epsilon,2}(T-2\tau)\zeta^2(\frac{1}{L}+\frac{1}{K})$.

Second event. Apply Lemma I.11 with $\mathcal{G}_s = \mathcal{F}_{s-2\tau}$, $D_{s+1} = \Xi_{s+1}$ for $s = 2\tau, \dots, T-1$, and $N = T-2\tau$. Its three hypotheses are Eq. (111), Eq. (112) and Eq. (113). Hence

$$\begin{aligned}\Pr\Big[\max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1} \ge u\Big] \;&\le\; \exp\Big(-\frac{u^2}{2(T-2\tau)\big(2(4\tau+1)U_n\zeta\big)^2}\Big)\\ &\le\; \exp\Big(-\frac{u^2}{8T\big((4\tau+1)U_n\zeta\big)^2}\Big).\end{aligned}$$

For the threshold, the two quantities subtracted in u are handled by requiring

$$6\tau U_n\zeta \;\le\; \frac{c\lambda^2}{512}, \qquad \mathcal{C}_{\epsilon,2}(T-2\tau)\zeta^2\Big(\frac{1}{L}+\frac{1}{K}\Big) \;\le\; \frac{c\lambda^2}{512}, \tag{116}$$

which give $u \ge \frac{c\lambda^2}{128} - \frac{c\lambda^2}{256} = \frac{c\lambda^2}{256}$, and hence

$$\Pr\Big[\max_{2\tau\le t\le T}\sum_{s=2\tau}^{t-1}\Xi_{s+1} \ge u\Big] \;\le\; \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big).$$

First event. Recall $\widehat{Y}_s = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{s-2\tau}|y_s^k|^2$ and $y_s^k = J^k - \eta_s^k$. By Eq. (77) at stepsize γ, $y_s^k = (1-\gamma)^s y_0^k - \nu_s^k$ with

$$\nu_s^k = \frac{\gamma}{L}\sum_{i=0}^{s-1}(1-\gamma)^{s-1-i}\sum_{\ell=0}^{L-1}\big(r_{i,\ell}^k - J^k\big),$$

so $(a+b)^2 \le 2a^2+2b^2$ together with $|y_0^k| = |J^k - \eta_0^k| \le 2U_r$ gives

$$\widehat{Y}_s \;\le\; 8U_r^2(1-\gamma)^{2s} + \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}_{s-2\tau}\big[(\nu_s^k)^2\big]. \tag{117}$$

First part. $\sum_{s\ge 0}(1-\gamma)^{2s} = \frac{1}{\gamma(2-\gamma)} \le \frac{1}{\gamma}$, so

$$2\mathcal{C}_{\epsilon,1}\,\zeta\sum_{s=2\tau}^{T-1}8U_r^2(1-\gamma)^{2s} \;\le\; \frac{16\,\mathcal{C}_{\epsilon,1}U_r^2\,\zeta}{\gamma} \;\le\; \frac{c\lambda^2}{256},$$

the last step holding as soon as

$$\gamma \;\ge\; \frac{4096\,\mathcal{C}_{\epsilon,1}\,U_r^2}{c\lambda^2}\,\zeta, \tag{118}$$

a lower bound on γ in terms of ζ.

Second part. What is needed is a pointwise bound on the noise,

$$\big|\nu_t^k\big| \;\le\; \sqrt{\mathcal{C}_\nu}, \qquad \mathcal{C}_\nu := \frac{c\lambda^2}{4096\,\mathcal{C}_{\epsilon,1}\,\zeta(T-2\tau)}, \tag{119}$$

for every k and every $0 \le t \le T$, which Lemma I.8 supplies with probability at least $1-\delta/2$.

Split ν_s^k at $\mathcal{F}_{s-2\tau}$,

$$\nu_s^k = (1-\gamma)^{2\tau}\nu_{s-2\tau}^k + \frac{\gamma}{L}\sum_{i=s-2\tau}^{s-1}(1-\gamma)^{s-1-i}\sum_{\ell=0}^{L-1}\big(r_{i,\ell}^k - J^k\big).$$

The second term has only 2τ summands, each at most $2U_r$ in absolute value, so it is at most $4U_r\tau\gamma$ everywhere and

$$\big|\nu_s^k\big| \;\le\; (1-\gamma)^{2\tau}\big|\nu_{s-2\tau}^k\big| + 4U_r\tau\gamma$$

pointwise. The right-hand side is $\mathcal{F}_{s-2\tau}$-measurable, so taking $\mathbb{E}_{s-2\tau}$ of the square and using monotonicity,

$$\mathbb{E}_{s-2\tau}\big[(\nu_s^k)^2\big] \;\le\; \Big((1-\gamma)^{2\tau}\big|\nu_{s-2\tau}^k\big| + 4U_r\tau\gamma\Big)^2. \tag{120}$$

Now substitute Eq. (119) at $t = s-2\tau$, together with the stepsize condition

$$4U_r\tau\gamma \;\le\; \sqrt{\mathcal{C}_\nu}\,: \tag{121}$$

the right-hand side of Eq. (120) is at most $4\mathcal{C}_\nu$ on the event of Eq. (119), and there

$$\frac{4\mathcal{C}_{\epsilon,1}\,\zeta}{K}\sum_{s=2\tau}^{T-1}\sum_{k=1}^{K}\mathbb{E}_{s-2\tau}\big[(\nu_s^k)^2\big] \;\le\; \frac{c\lambda^2}{256}.$$

Adding the two parts, $2\mathcal{C}_{\epsilon,1}\zeta\sum_{s=2\tau}^{T-1}\widehat{Y}_s \le \frac{c\lambda^2}{256} + \frac{c\lambda^2}{256} = \frac{c\lambda^2}{128}$ whenever Eq. (119) holds, so the first event of Eq. (115) is contained in the complement of Eq. (119).

The two events of Eq. (115), bounded by Lemma I.8 and by the display above, give

$$\Pr\Big[\big\{T_{\text{exit}} \le T\big\} \cap \Big\{n_0 \le \frac{c\lambda^2}{64}\Big\}\Big] \;\le\; \frac{\delta}{2} + \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big),$$

and substituting this into Eq. (100),

$$\Pr\big[T_{\text{exit}} \le T\big] \;\le\; \frac{\delta}{2} + \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big) + \frac{\delta}{2}$$
$$= \;\delta + \exp\Big(-\frac{c^2\lambda^4}{524288\,T\big((4\tau+1)U_n\zeta\big)^2}\Big).$$

□

Lemma I.8 (Pointwise bound on the reward error)**.** *Let $\delta \in (0,1)$. Under the conditions on γ collected in Section I.5, Eq. (119) holds with probability at least $1-\delta/2$.*

Proof. By Eq. (77), $y_t^k = (1-\gamma)^t y_0^k - \nu_t^k$ with

$$\nu_t^k = \frac{\gamma}{L}\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\sum_{\ell=0}^{L-1}\big(r_{i,\ell}^k - J^k\big).$$

We can write

$$r_{i,\ell}^k - J^k = \underbrace{\big[V^k(s_{i,\ell}^k) - V^k(s_{i,\ell+1}^k)\big]}_{\text{telescopes}} + D_{i,\ell}^k, \qquad D_{i,\ell}^k := r_{i,\ell}^k - J^k + V^k(s_{i,\ell+1}^k) - V^k(s_{i,\ell}^k), \tag{122}$$

with V^k the value function of Eq. (2).

Martingale difference. $D_{i,\ell}^k$ is a function of $(s_{i,\ell}^k, a_{i,\ell}^k, s_{i,\ell+1}^k)$, hence $v_{0:(i,\ell)}$-measurable, and

$$\mathbb{E}\big[D_{i,\ell}^k \;\big|\; v_{0:(i,\ell-1)}\big] = \mathbb{E}_{(a,s')\sim\pi\otimes P^k}\big[R(s,a) - J^k + V^k(s') - V^k(s) \;\big|\; s = s_{i,\ell}^k\big] = 0, \tag{123}$$

by the Markov property, $(a_{i,\ell}^k, s_{i,\ell+1}^k) \sim \pi \otimes P^k$ given $s_{i,\ell}^k$, and the Bellman equation for V^k.

Boundedness. What is needed is a pointwise bound on $D_{i,\ell}^k$. By Eq. (122), $|r_{i,\ell}^k| \le U_r$ and $|J^k| \le U_r$,

$$\big|D_{i,\ell}^k\big| \;\le\; \big|r_{i,\ell}^k - J^k\big| + \big|V^k(s_{i,\ell+1}^k)\big| + \big|V^k(s_{i,\ell}^k)\big| \;\le\; 2U_r + 2\|V^k\|_\infty,$$

so $\|V^k\|_\infty$ is what is left. By Eq. (2), with $\mathbb{E}_s$ the expectation along the chain started at $s_0^k = s$,

$$\begin{aligned}\big|V^k(s)\big| \;&\le\; \sum_{t\ge0}\Big|\mathbb{E}_s\big[R(s_t^k, a_t^k)\big] - J^k\Big| \\ &=\; \sum_{t\ge0}\Big|\int_{\mathcal{S}}\sum_{a'\in\mathcal{A}}\pi(a'\mid s')R(s',a')\,\big(\mathbb{P}_{0:t}^k(ds'\mid s) - \mu^k(ds')\big)\Big| \\ &\le\; 2U_r\sum_{t\ge0} d_{TV}\big(\mathbb{P}_{0:t}^k(\cdot\mid s), \mu^k\big) \;\le\; 2U_r m\sum_{t\ge0}\rho^t \;=\; \frac{2U_r m}{1-\rho}\end{aligned}$$

at every s, the last two steps by $|R| \le U_r$ and Assumption 2, so that

$$\|V^k\|_\infty \;\le\; \frac{2U_r m}{1-\rho}, \qquad \big|D_{i,\ell}^k\big| \;\le\; 2U_r\Big(1 + \frac{2m}{1-\rho}\Big) \quad \text{pointwise.} \tag{124}$$

Substituting Eq. (122) into ν_t^k, we get

$$\underbrace{\frac{\gamma}{L}\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\big(V^k(s_{i,0}^k) - V^k(s_{i+1,0}^k)\big)}_{\text{(I)}} + \underbrace{\frac{\gamma}{L}\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}\sum_{\ell=0}^{L-1}D_{i,\ell}^k}_{\text{(II)}}, \tag{125}$$

where we used that the inner sum telescopes to $V^k(s_{i,0}^k) - V^k(s_{i,L}^k)$ and that $s_{i,L}^k = s_{i+1,0}^k$ in Algorithm 2.

(I) is bounded pointwise. Summing by parts, we get

$$\begin{aligned}
\big|(\mathrm{I})\big| &= \frac{\gamma}{L}\Big|\sum_{i=0}^{t-1}(1-\gamma)^{t-1-i}V^k(s^k_{i,0}) - \sum_{i=1}^{t}(1-\gamma)^{t-i}V^k(s^k_{i,0})\Big| \\
&= \frac{\gamma}{L}\Big|(1-\gamma)^{t-1}V^k(s^k_{0,0}) - V^k(s^k_{t,0}) + \sum_{i=1}^{t-1}\big((1-\gamma)^{t-1-i} - (1-\gamma)^{t-i}\big)V^k(s^k_{i,0})\Big| \\
&\le \frac{\gamma}{L}\|V^k\|_\infty\Big((1-\gamma)^{t-1} + 1 + \sum_{i=1}^{t-1}\big|(1-\gamma)^{t-1-i} - (1-\gamma)^{t-i}\big|\Big) \\
&= \frac{\gamma}{L}\|V^k\|_\infty\Big((1-\gamma)^{t-1} + 1 + 1 - (1-\gamma)^{t-1}\Big) \;=\; \frac{2\gamma}{L}\,\|V^k\|_\infty \\
&\le \frac{4\gamma U_r m}{L(1-\rho)} \;\le\; \frac{\sqrt{\mathcal{C}_\nu}}{2},
\end{aligned}$$

where we shifted the index in the second sum, then used the triangle inequality with $|V^k(\cdot)| \le \|V^k\|_\infty$, then dropped the absolute values by $(1-\gamma)^{t-1-i} \ge (1-\gamma)^{t-i}$ and telescoped, then Eq. (124), and finally the stepsize condition $\gamma \le \frac{L\sqrt{\mathcal{C}_\nu}(1-\rho)}{8U_r m}$.

(II) is a sum of martingale differences: by Eq. (123) and Eq. (124), its summands $\frac{\gamma}{L}(1-\gamma)^{t-1-i}D^k_{i,\ell}$ satisfy Eq. (130) for $\{v_{0:(i,\ell)}\}$ with $a_{i,\ell} = \frac{2\gamma U_r}{L}(1-\gamma)^{t-1-i}\big(1+\frac{2m}{1-\rho}\big)$. We get

$$\begin{aligned}
\sum_{i=0}^{t-1}\sum_{\ell=0}^{L-1} a_{i,\ell}^2 &= \frac{4\gamma^2 U_r^2}{L}\Big(1+\frac{2m}{1-\rho}\Big)^2 \sum_{i=0}^{t-1}(1-\gamma)^{2(t-1-i)} \\
&\le \frac{4\gamma^2 U_r^2}{L}\Big(1+\frac{2m}{1-\rho}\Big)^2 \cdot \frac{1}{\gamma(2-\gamma)} \;\le\; \frac{4\gamma U_r^2}{L}\Big(1+\frac{2m}{1-\rho}\Big)^2. \qquad (126)
\end{aligned}$$

Since $|(\mathrm{I})| \le \sqrt{\mathcal{C}_\nu}/2$ on every path, Eq. (119) follows once $|(\mathrm{II})| \le \sqrt{\mathcal{C}_\nu}/2$ for every $t \le T$ and every k. Fix one such pair and apply Lemma I.10 at $u = \sqrt{\mathcal{C}_\nu}/2$, once to (II) and once to $-(\mathrm{II})$:

$$\begin{aligned}
\Pr\Big[\big|(\mathrm{II})\big| > \frac{\sqrt{\mathcal{C}_\nu}}{2}\Big] &\le 2\exp\Big(-\frac{u^2}{2\sum_{i,\ell} a_{i,\ell}^2}\Big) \;\le\; 2\exp\Big(-\frac{\mathcal{C}_\nu/4}{\frac{8\gamma U_r^2}{L}\big(1+\frac{2m}{1-\rho}\big)^2}\Big) \\
&= 2\exp\Big(-\frac{\mathcal{C}_\nu\, L}{32\, U_r^2\big(1+\frac{2m}{1-\rho}\big)^2\gamma}\Big),
\end{aligned}$$

by $u^2 = \mathcal{C}_\nu/4$ and Eq. (126). At $t = 0$ the sum defining ν_0^k is empty, so $\nu_0^k = 0$ and Eq. (119) holds there. For $1 \le t \le T$ we get

$$\begin{aligned}
\Pr\Big[\max_{1\le t\le T,\,1\le k\le K}|\nu_t^k| > \sqrt{\mathcal{C}_\nu}\Big] &= \Pr\Big[\bigcup_{t=1}^{T}\bigcup_{k=1}^{K}\{|\nu_t^k| > \sqrt{\mathcal{C}_\nu}\}\Big] \\
&\le \Pr\Big[\bigcup_{t=1}^{T}\bigcup_{k=1}^{K}\Big\{\big|(\mathrm{II})\big| > \frac{\sqrt{\mathcal{C}_\nu}}{2}\Big\}\Big] \\
&\le \sum_{t=1}^{T}\sum_{k=1}^{K}\Pr\Big[\big|(\mathrm{II})\big| > \frac{\sqrt{\mathcal{C}_\nu}}{2}\Big] \\
&\le 2TK\exp\Big(-\frac{\mathcal{C}_\nu\, L}{32\, U_r^2\big(1+\frac{2m}{1-\rho}\big)^2\gamma}\Big).
\end{aligned}$$

This is at most $\delta/2$ exactly when

$$\exp\Big(-\frac{\mathcal{C}_\nu\, L}{32\, U_r^2\big(1+\frac{2m}{1-\rho}\big)^2\gamma}\Big) \;\le\; \frac{\delta}{4TK},$$

that is, taking logarithms,

$$\frac{\mathcal{C}_\nu\, L}{32\, U_r^2 \big(1+\frac{2m}{1-\rho}\big)^2 \gamma} \;\geq\; \log \frac{4TK}{\delta}. \tag{127}$$

□

I.5 Global convergence

The conditions. Collected from the four subsections above, in three groups.

The stepsizes ζ, β and the index τ.

$\beta = c\zeta, \qquad \tau = \lceil 2\log_\rho(\zeta)\rceil$	Lemmas 5.1, 5.2, I.1
$U_\delta U_\omega \zeta \leq \frac{1}{\sqrt{6}}, \qquad U_\omega \zeta \leq \frac{1}{2}, \qquad \beta U_\delta \leq U_\omega$	Lemmas 5.1, 5.2, Appendix K
$\zeta \leq \frac{1}{L}, \qquad [\mathcal{C}_{X,3}(\tau^2) + c\mathcal{C}_{M,5}(\tau^2)]\zeta \leq \mathcal{C}_{\epsilon,1}$	*Eq.* (97)
$2\tau U_n \zeta \leq \frac{c\lambda^2}{64}$	*Eq.* (102)
$6\tau U_n \zeta \leq \frac{c\lambda^2}{512}, \qquad \mathcal{C}_{\epsilon,2}(T-2\tau)\zeta^2\big(\frac{1}{L}+\frac{1}{K}\big) \leq \frac{c\lambda^2}{512}$	*Eq.* (116)
$a = \frac{c\lambda\alpha}{4(\alpha+c)}\zeta \leq 1$	Corollary 1
$[\mathcal{C}_{X,1}(\tau^2) + c\,\mathcal{C}_{M,4}(\tau^2)]\zeta \leq \frac{c\lambda}{4}$	*Eq.* (82)

The first entry of the second line covers both the $U_\delta U_\omega \zeta \leq \frac{1}{2}$ of Lemma 5.1 and the $U_\delta^2 U_\omega^2 \zeta^2 \leq \frac{1}{6}$ of Lemma 5.2; the fourth line keeps the threshold of σ positive and implies the $2\tau U_n \zeta \leq c\lambda^2/32$ of Lemma I.6.

The stepsize γ.

$\gamma\tau \leq \frac{1}{2}, \qquad \rho^{\tau L} \leq \gamma, \qquad \gamma^2 \leq \frac{1}{K}$		Lemma G.1, *Eq.* (97)
$\zeta^2 \leq \gamma$		Corollary 1
$\zeta \leq \gamma$		Lemma 5.4
$\gamma \geq \frac{4096\mathcal{C}_{\epsilon,1}U_r^2}{c\lambda^2}\zeta$		*Eq.* (118)
$\gamma \leq \frac{L\sqrt{\mathcal{C}_\nu}(1-\rho)}{8U_r m}, \qquad$	$\frac{\mathcal{C}_\nu L}{32U_r^2(1+2m/(1-\rho))^2\gamma} \geq \log\frac{4TK}{\delta}$	*Eq.* (127)
$4U_r\tau\gamma \leq \sqrt{\mathcal{C}_\nu}$		*Eq.* (121)

The constants c, C and the horizons.

$c\lambda \geq 20U_\omega^2, \qquad C \geq 1, \qquad T \geq 2\tau$	Proposition 1
$K \geq \frac{8192U_\omega^2(1+c/\alpha)\log(2/\delta)}{c\lambda^2}$	Proposition 3

The choice. Take the stepsizes of Proposition 1 together with $\delta = 1/(TK)$:

$$\beta = c\zeta, \quad \tau = \lceil 2\log_\rho(\zeta)\rceil, \quad \zeta = \frac{8(\alpha+c)\log(TK)}{c\lambda\alpha(T-2\tau+1)}, \quad \gamma = \frac{C\log(T)}{T}, \quad \delta = \frac{1}{TK}, \tag{128}$$

with $T \geq 4\tau$ and $K \leq T$, giving $T - 2\tau + 1 \geq T/2$ and $\log(TK) \leq 2\log T$, and

$$C = \max\Big\{1,\; \frac{131072\,\mathcal{C}_{\epsilon,1}U_r^2(\alpha+c)}{c^2\lambda^3\alpha},\; \frac{32(\alpha+c)}{c\lambda\alpha}\Big\},$$

which meets Eq. (118) and $\zeta \le \gamma$. The three groups then require

$$
\begin{aligned}
T \;&\ge\; \frac{16(\alpha+c)\log(TK)}{c\lambda\alpha}\max\Big\{\sqrt{6}\,U_\delta U_\omega,\; 2U_\omega,\; \frac{cU_\delta}{U_\omega},\; L,\; \frac{c\lambda\alpha}{4(\alpha+c)},\\
&\qquad\qquad \frac{\mathcal{C}_{X,3}(\tau^2)+c\mathcal{C}_{M,5}(\tau^2)}{\mathcal{C}_{\epsilon,1}},\; \frac{4[\mathcal{C}_{X,1}(\tau^2)+c\,\mathcal{C}_{M,4}(\tau^2)]}{c\lambda}\Big\},\\
T \;&\ge\; \frac{49152\,(\alpha+c)\,\tau U_n\log(TK)}{c^2\lambda^3\alpha},\\
T \;&\ge\; \frac{65536\,\mathcal{C}_{\epsilon,2}(\alpha+c)^2\log^2(TK)}{c^3\lambda^4\alpha^2}\Big(\frac{1}{L}+\frac{1}{K}\Big),\\
T \;&\ge\; C\log(T)\max\Big\{2\tau,\; \sqrt{K},\; \frac{4U_r\tau}{\sqrt{\mathcal{C}_\nu}},\; \frac{8U_r m}{L\sqrt{\mathcal{C}_\nu}\,(1-\rho)}\Big\},\\
T \;&\ge\; \frac{32U_r^2\big(1+\frac{2m}{1-\rho}\big)^2 C\log(T)}{\mathcal{C}_\nu\,L}\,\log\frac{4TK}{\delta},
\end{aligned}
$$

and that

$$
K \;\ge\; \frac{8192\,U_\omega^2(1+c/\alpha)\log(2TK)}{c\lambda^2}.
$$

Theorem 2 (Global convergence, formal version of Theorem 1)**.** *Run Algorithm 2 with Eq. (128), let $L = K \le T$, and assume the conditions above and*

$$
T \;\ge\; \frac{134217728\,(4\tau+1)^2U_n^2(\alpha+c)^2\log^3(TK)}{c^4\lambda^6\alpha^2}. \tag{129}
$$

Then, with no hypothesis on the trajectory,

$$
\mathcal{V}_{T+1} \;=\; \widetilde{\mathcal{O}}\Big(\frac{r}{TK}\Big).
$$

Proof. By Eq. (84), $\mathcal{V}_{T+1} = \mathbb{E}[n_{T+1}]$, so

$$
\begin{aligned}
\mathcal{V}_{T+1} &= \mathbb{E}\big[n_{T+1}\mathbf{1}_{\{T_{\text{exit}}>T\}}\big] + \mathbb{E}\big[n_{T+1}\mathbf{1}_{\{T_{\text{exit}}\le T\}}\big]\\
&\le \mathbb{E}\big[n_{T+1}^{\text{in}}\big] + \big(4U_\omega^2+cr\big)\Pr\big[T_{\text{exit}}\le T\big],
\end{aligned}
$$

where for the first term $\tau \ge 1$ gives $\{T_{\text{exit}} > T\} \subseteq \{T_{\text{exit}} > T+1-2\tau\}$, which with $n_{T+1} \ge 0$ and Eq. (98) bounds it by $\mathbb{E}[n_{T+1}^{\text{in}}]$, and for the second $\|x_t^k\| \le 2U_\omega$ and $\|m_t\|_F \le \|\mathbf{B}_t\|_F = \sqrt{r}$ give $n_{T+1} \le 4U_\omega^2 + cr$ pointwise.

Corollary 1 bounds $\mathbb{E}[n_{T+1}^{\text{in}}]$ by four terms. Under Eq. (128), $a = \frac{2\log(TK)}{T-2\tau+1}$ and they are, in order,

$$
(1-a)^{T+1-2\tau}\big(4U_\omega^2+cr\big) \;\le\; e^{-2\log(TK)}\big(4U_\omega^2+cr\big) = \frac{4U_\omega^2+cr}{(TK)^2},
$$

$$
2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\zeta\,T\max\big\{e^{-2\gamma T}, e^{-a(T-2\tau)}\big\} \;\le\; \frac{2e\mathcal{C}_{\epsilon,1}\mathcal{C}_r\zeta}{T} \;=\; \widetilde{\mathcal{O}}\Big(\frac{1}{T^2}\Big),
$$

by $e^{-2\gamma T} = T^{-2C} \le T^{-2}$ and $e^{-a(T-2\tau)} = e^a(TK)^{-2} \le e\,T^{-2}$, and

$$
\frac{2\mathcal{C}_{\epsilon,1}\mathcal{C}_r\zeta}{a}\Big(\frac{\gamma}{L}+\tau^2\gamma^2\Big) = \widetilde{\mathcal{O}}\Big(\frac{1}{TL}\Big), \qquad \frac{\mathcal{C}_{\epsilon,2}\zeta^2}{a}\Big(\frac{1}{L}+\frac{1}{K}\Big) = \widetilde{\mathcal{O}}\Big(\frac{1}{TK}\Big),
$$

by $\zeta/a = \frac{4(\alpha+c)}{c\lambda\alpha}$. At $L = K \le T$ all four are $\widetilde{\mathcal{O}}(1/(TK))$.

Lemma I.7 gives $\Pr[T_{\text{exit}} \le T] \le \delta + \exp\big(-\frac{c^2\lambda^4}{524288\,T((4\tau+1)U_n\zeta)^2}\big)$. By $T - 2\tau + 1 \ge T/2$, $\zeta \le \frac{16(\alpha+c)\log(TK)}{c\lambda\alpha T}$, so the exponent is at least $\frac{c^4\lambda^6\alpha^2\,T}{134217728\,(4\tau+1)^2U_n^2(\alpha+c)^2\log^2(TK)}$, which is at least $\log(TK)$ by Eq. (129). At $\delta = 1/(TK)$ we get $\Pr[T_{\text{exit}} \le T] \le 2/(TK)$, so $(4U_\omega^2+cr)\Pr[T_{\text{exit}} \le T] = \mathcal{O}(r/(TK))$. □

I.6 TOOLS

Stated in the form in which they are used. All three are standard.

Lemma I.9 (Doob's maximal inequality). *Let $\{M_n\}_{n=0}^N$ be a non-negative submartingale for a filtration $\{\mathcal{G}_n\}$: $M_n \geq 0$, M_n is $\mathcal{G}_n$-measurable, and $\mathbb{E}[M_{n+1} \mid \mathcal{G}_n] \geq M_n$. Then for every $x > 0$,*

$$\Pr\Big[\max_{0\leq n\leq N} M_n \;\geq\; x\Big] \;\leq\; \frac{\mathbb{E}[M_N]}{x}.$$

Lemma I.10 (Azuma–Hoeffding). *Let $\{\mathcal{G}_n\}$ be a filtration and let $\{D_n\}_{n=1}^N$ satisfy*

$$D_n \text{ is } \mathcal{G}_n\text{-measurable}, \qquad \mathbb{E}[D_n \mid \mathcal{G}_{n-1}] = 0, \qquad |D_n| \leq a_n \text{ pointwise}, \tag{130}$$

with a_n constants. Then for every $u > 0$,

$$\Pr\Big[\sum_{j=1}^N D_j \;\geq\; u\Big] \;\leq\; \exp\Big(-\frac{u^2}{2\sum_{j=1}^N a_j^2}\Big).$$

Lemma I.11 (Azuma–Hoeffding, maximal form). *Under the hypotheses of Lemma I.10, for every $u > 0$,*

$$\Pr\Big[\max_{1\leq n\leq N} \sum_{j=1}^n D_j \;\geq\; u\Big] \;\leq\; \exp\Big(-\frac{u^2}{2\sum_{j=1}^N a_j^2}\Big),$$

the same bound as for the fixed endpoint.

Proof. Write $S_n = \sum_{j\leq n} D_j$, fix $\theta > 0$, and put $M_n = e^{\theta S_n}$ with $M_0 = 1$.

Step 1: M_n is a non-negative submartingale. It is positive, and $\mathcal{G}_n$-measurable because each D_j with $j \leq n$ is $\mathcal{G}_j \subseteq \mathcal{G}_n$-measurable. Since S_n is $\mathcal{G}_n$-measurable,

$$\mathbb{E}\big[M_{n+1} \;\big|\; \mathcal{G}_n\big] = M_n\, \mathbb{E}\big[e^{\theta D_{n+1}} \;\big|\; \mathcal{G}_n\big] \;\geq\; M_n\, e^{\theta\, \mathbb{E}[D_{n+1}|\mathcal{G}_n]} = M_n,$$

by conditional Jensen and $\mathbb{E}[D_{n+1} \mid \mathcal{G}_n] = 0$.

Step 2: Doob. The map $y \mapsto e^{\theta y}$ is increasing, so $\{S_n \geq u\} = \{M_n \geq e^{\theta u}\}$ for each n, and therefore $\{\max_n S_n \geq u\} = \{\max_n M_n \geq e^{\theta u}\}$. Lemma I.9 at $x = e^{\theta u}$ gives

$$\Pr\Big[\max_{n\leq N} S_n \geq u\Big] \;\leq\; e^{-\theta u}\, \mathbb{E}[M_N]. \tag{131}$$

Step 3: the moment generating function. Hoeffding's lemma states that a random variable X with $\mathbb{E}X = 0$ and $|X| \leq a$ satisfies $\mathbb{E}e^{\theta X} \leq e^{\theta^2 a^2/2}$. Applied conditionally on $\mathcal{G}_{n-1}$, where D_n has conditional mean 0 and is still bounded by a_n pointwise,

$$\mathbb{E}\big[e^{\theta D_n} \;\big|\; \mathcal{G}_{n-1}\big] \;\leq\; e^{\theta^2 a_n^2/2}.$$

Peeling one step at a time, using that $e^{\theta S_{n-1}}$ is $\mathcal{G}_{n-1}$-measurable and the tower property,

$$\mathbb{E}\big[e^{\theta S_n}\big] = \mathbb{E}\Big[e^{\theta S_{n-1}}\, \mathbb{E}\big[e^{\theta D_n} \;\big|\; \mathcal{G}_{n-1}\big]\Big] \;\leq\; e^{\theta^2 a_n^2/2}\, \mathbb{E}\big[e^{\theta S_{n-1}}\big],$$

so $\mathbb{E}[M_N] \leq \exp\big(\frac{\theta^2}{2}\sum_{j\leq N} a_j^2\big)$.

Step 4: optimising. Substituting into Eq. (131),

$$\Pr\Big[\max_{n\leq N} S_n \geq u\Big] \;\leq\; \exp\Big(-\theta u + \tfrac{\theta^2}{2}\sum_{j\leq N} a_j^2\Big),$$

and $\theta = u/\sum_j a_j^2$ minimises the exponent, giving the stated bound. □

J Supporting Propositions and Lemmas

J.1 Boundedness

Lemma J.1. *For any $t \geq 0, L \geq 0, k \in [K]$, it holds that $\eta_t^k \in [-U_r, U_r]$.*

Proof. We prove Lemma J.1 by induction. The algorithm initializes $\eta_0^k = 0$. Suppose that $\eta_{t'}^k \in [-U_r, U_r]$ for $0 \leq t' \leq t-1$ and for any $k \in [K]$. We have

$$\eta_t^k = \eta_{t-1}^k + \gamma\left(\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t-1,\ell}^k - \eta_{t-1}^k\right) = (1-\gamma)\eta_{t-1}^k + \frac{\gamma}{L}\sum_{\ell=0}^{L-1} r_{t-1,\ell}^k \leq (1-\gamma)U_r + \frac{\gamma}{L}L \cdot U_r = U_r.$$

Similarly, it can be shown that $\eta_t^k \geq -U_r$.

□

Lemma J.2. *Suppose that Assumption 3 holds. Define $U_\delta := 2U_r + 2U_\omega$. For any $t \geq 0, L \geq 0, k \in [K]$, and $\ell = 0, \ldots, L-1$, it holds that*

$$|\delta_{t,\ell}^k| \leq U_\delta, \quad \textit{and} \quad \|\mathbf{b}_t^k\| \leq U_\delta. \tag{132}$$

Proof. From Eq. (6),

$$\begin{aligned}
|\delta_{t,\ell}^k| &= |(r_{t,\ell}^k - \eta_t^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}_t \omega_t^k| \\
&\leq |r_{t,\ell}^k - \eta_t^k| + |(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}_t \omega_t^k| \\
&\leq 2U_r + 2U_\omega,
\end{aligned}$$

where the last inequality holds due to Lemma J.1 ($|r_{t,\ell}^k - \eta_t^k| \leq 2U_r$), $\|\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k)\| \leq 2$, and $\|\mathbf{B}_t \omega_t^k\| \leq U_\omega$.

Recall from Eq. (16) that $\mathbf{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1} g_\ell^k$, where $g_\ell^k = \left((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}^* \omega^{k,*}\right)\phi(s_{t,\ell}^k)$. Since $|r_{t,\ell}^k - J^k| \leq 2U_r$, $\|(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}^* \omega^{k,*}\| \leq 2U_\omega$, and $\|\phi(s_{t,\ell}^k)\| \leq 1$, each $\|g_\ell^k\| \leq 2U_r + 2U_\omega = U_\delta$, so

$$\|\mathbf{b}_t^k\| = \left\|\frac{1}{L}\sum_{\ell=0}^{L-1} g_\ell^k\right\| \leq \frac{1}{L}\sum_{\ell=0}^{L-1}\|g_\ell^k\| \leq U_\delta.$$

□

K Other Supporting Results

Proof of Proposition 2. It follows from Eq. (6) that for any t and L:

$$\begin{aligned}
&\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k) = \frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k - \eta_t^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k) \\
=&\frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k) + \frac{1}{L}\sum_{\ell=0}^{L-1}(J^k - \eta_t^k)\phi(s_{t,\ell}^k) \\
=&\underbrace{\frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top x_t^k}_{\widetilde{A}_t^k} \qquad \text{(recall that } x_t^k = \mathbf{B}_t\omega_t^k - \mathbf{B}^*\omega^{k,*}\text{)} \\
&+\underbrace{\frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}^*\omega^{k,*}\right)\phi(s_{t,\ell}^k)}_{\mathbf{b}_t^k} + y_t^k\bar{\phi}_t^k \\
=&\widetilde{A}_t^k x_t^k + \mathbf{b}_t^k + y_t^k\bar{\phi}_t^k,
\end{aligned}$$

proving Eq. (18). In the remaining proof, we focus on proving the first decomposition in Eq. (17). We have,

$$\begin{aligned}
&\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k) = \frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k-\eta_t^k)+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k)\\
=&\frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k-J^k)+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k)+\frac{1}{L}\sum_{\ell=0}^{L-1}(J^k-\eta_t^k)\phi(s_{t,\ell}^k)\\
=&\frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k-J^k)+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k)\\
&-\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top\mathbf{B}_t\omega_t^k\right)\phi(s^{(\ell)})\\
&+\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top\mathbf{B}_t\omega_t^k\right)\phi(s^{(\ell)})\\
&+y_t^k\bar{\phi}_t^k.
\end{aligned}$$

By the definition of the one-step TD fixed point $z^{k,*}$ under the steady-state distribution μ^k, we have

$$\mathbb{E}_{s\sim\mu^k,a\sim\pi,s'\sim P^k}\left((R(s,a)-J^k)+(\phi(s')-\phi(s))^\top z^{k,*}\right)\phi(s)=0. \tag{133}$$

Then, since each summand is at steady state,

$$\begin{aligned}
&\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top\mathbf{B}_t\omega_t^k\right)\phi(s^{(\ell)})\\
=&\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top\mathbf{B}_t\omega_t^k\right)\phi(s^{(\ell)})\\
&-\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top z^{k,*}\right)\phi(s^{(\ell)})\\
=&\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\phi(s^{(\ell)})(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top x_t^k\\
=&\underbrace{\mathbb{E}_{s\sim\mu^k,a\sim\pi,s'\sim P^k}\left[\phi(s)(\phi(s')-\phi(s))^\top\right]}_{A^k}x_t^k\\
=&A^kx_t^k.
\end{aligned}$$

Then,

$$\frac{1}{L}\sum_{\ell=0}^{L-1}\delta_{t,\ell}^k\phi(s_{t,\ell}^k)=\xi_t^k+y_t^k\bar{\phi}_t^k+A^kx_t^k,$$

where

$$\begin{aligned}
\xi_t^k=&\frac{1}{L}\sum_{\ell=0}^{L-1}\left((r_{t,\ell}^k-J^k)+(\phi(s_{t,\ell+1}^k)-\phi(s_{t,\ell}^k))^\top\mathbf{B}_t\omega_t^k\right)\phi(s_{t,\ell}^k)\\
&-\frac{1}{L}\sum_{\ell=0}^{L-1}\mathbb{E}_{s^{(\ell)}\sim\mu^k,\,a^{(\ell)}\sim\pi,\,s^{(\ell+1)}\sim P^k}\left((R(s^{(\ell)},a^{(\ell)})-J^k)+(\phi(s^{(\ell+1)})-\phi(s^{(\ell)}))^\top\mathbf{B}_t\omega_t^k\right)\phi(s^{(\ell)}).
\end{aligned}$$

□

Lemma K.1. *Fix $\tau\in\mathbb{N}$. When $\zeta U_\delta U_\omega\le 1/2$, for any $t\ge\tau>0$ and $k\in[K]$, it holds that*

$$\|\omega_t^k-\omega_{t-\tau}^k\|\le\tau U_\delta\beta,\ \ \textit{and}\ \ \|\mathbf{B}_t-\mathbf{B}_{t-\tau}\|\le\tau U_\delta U_\omega\zeta+2\tau U_\delta^2U_\omega^2\zeta^2.$$

Proof. For any $t \geq \tau > 0$ and $k \in [K]$, we have

$$\begin{aligned}
\|\omega_t^k - \omega_{t-\tau}^k\| \leq & \sum_{i=t-\tau}^{t-1} \|\omega_{i+1}^k - \omega_i^k\| \\
= & \sum_{i=t-\tau}^{t-1} \Big\| \Pi_{U_\omega}\Big(\omega_i^k + \frac{\beta}{L}\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \mathbf{B}_i^\top \phi(s_{i,\ell}^k)\Big) - \Pi_{U_\omega}(\omega_i^k)\Big\| \\
\leq & \sum_{i=t-\tau}^{t-1} \Big\| \frac{\beta}{L}\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \mathbf{B}_i^\top \phi(s_{i,\ell}^k)\Big\| \\
\leq & \tau U_\delta \beta,
\end{aligned}$$

where the last inequality follows from Lemma J.2, since $\big\|\frac{\beta}{L}\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \mathbf{B}_i^\top \phi(s_{i,\ell}^k)\big\| \leq \frac{\beta}{L}\sum_{\ell=0}^{L-1} |\delta_{i,\ell}^k| \leq \beta U_\delta$. In addition,

$$\begin{aligned}
&\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\| \leq \sum_{i=t-\tau}^{t-1} \|\mathbf{B}_{i+1} - \mathbf{B}_i\| \\
&= \sum_{i=t-\tau}^{t-1} \|(\mathbf{B}_i + \frac{\zeta}{KL}\sum_{k=1}^{K} \mathbf{B}_{i,\perp}\mathbf{B}_{i,\perp}^\top\Big(\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \phi(s_{i,\ell}^k)\Big)(\omega_i^k)^\top)\mathbf{R}_{i+1}^{-1} - \mathbf{B}_i\| \\
&= \sum_{i=t-\tau}^{t-1} \|(\mathbf{B}_i + \frac{\zeta}{KL}\sum_{k=1}^{K} \mathbf{B}_{i,\perp}\mathbf{B}_{i,\perp}^\top\Big(\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \phi(s_{i,\ell}^k)\Big)(\omega_i^k)^\top)\mathbf{R}_{i+1}^{-1} - \mathbf{B}_i\mathbf{R}_{i+1}^{-1} + \mathbf{B}_i\mathbf{R}_{i+1}^{-1} - \mathbf{B}_i\| \\
&\leq \sum_{i=t-\tau}^{t-1} \|\frac{\zeta}{KL}\sum_{k=1}^{K} \mathbf{B}_{i,\perp}\mathbf{B}_{i,\perp}^\top\Big(\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \phi(s_{i,\ell}^k)\Big)(\omega_i^k)^\top\|\|\mathbf{R}_{i+1}^{-1}\| + \sum_{i=t-\tau}^{t-1} \|\mathbf{R}_{i+1}^{-1} - \mathbf{I}\|\|\mathbf{B}_i\| \\
&\leq \zeta U_\delta U_\omega \sum_{i=t-\tau}^{t-1} \|\mathbf{R}_{i+1}^{-1}\| + \sum_{i=t-\tau}^{t-1} \|\mathbf{R}_{i+1}^{-1} - \mathbf{I}\| \\
&\leq \zeta U_\delta U_\omega \sum_{i=t-\tau}^{t-1} 1 + \sum_{i=t-\tau}^{t-1} 2\|\mathbf{Q}_i\|_F^2,
\end{aligned}$$

where the last inequality follows from Lemma D.1, and $\mathbf{Q}_i$ is defined in Eq. (19) as

$$\mathbf{Q}_i = \frac{\zeta}{KL}\sum_{k=1}^{K} \mathbf{B}_{i,\perp}\mathbf{B}_{i,\perp}^\top\Big(\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \phi(s_{i,\ell}^k)\Big)(\omega_i^k)^\top.$$

We have

$$\begin{aligned}
\|\mathbf{Q}_i\|_F &\leq \frac{\zeta}{KL}\sum_{k=1}^{K} \Big\|\mathbf{P}_{i,\perp}\Big(\sum_{\ell=0}^{L-1} \delta_{i,\ell}^k \phi(s_{i,\ell}^k)\Big)(\omega_i^k)^\top\Big\|_F \\
&\leq \frac{\zeta}{KL}\sum_{k=1}^{K} \Big(\sum_{\ell=0}^{L-1} |\delta_{i,\ell}^k|\Big)\|\mathbf{P}_{i,\perp}\|\|\omega_i^k\| \\
&\leq \zeta U_\delta U_\omega. \qquad (134)
\end{aligned}$$

Therefore, we have

$$\|\mathbf{B}_t - \mathbf{B}_{t-\tau}\| \leq \tau\zeta U_\delta U_\omega + 2\tau\zeta^2 U_\delta^2 U_\omega^2.$$

□

Lemma K.2. *Fix $\tau \in \mathbb{N}$. For any $k \in [K]$ and any $j \geq t - \tau > \tau$, where $t \geq 2\tau$, it holds that*

$$\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k - \omega_j^k\| \leq C_1(\tau)\beta^2 + C_2(\tau)\zeta\beta + C_3(\tau)\gamma\beta + C_4\frac{\beta}{\sqrt{L}} + 2\beta\mathbb{E}_{t-2\tau}\|x_t^k\| + \beta\mathbb{E}_{t-2\tau}|y_t^k|,$$

$$\begin{aligned}
\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k - \omega_j^k\|^2 \leq & C_5(\tau^2)\beta^4 + C_6(\tau^2)\zeta^2\beta^2 + 24\beta^2\mathbb{E}_{t-2\tau}\|x_t^k\|^2 \\
& + C_7\frac{\beta^2}{L} + C_8(\tau^2)\gamma^2\beta^2 + 2\beta^2\mathbb{E}_{t-2\tau}|y_t^k|^2,
\end{aligned}$$

where

$$C_1(\tau)=2\tau U_\delta, C_2(\tau)=4\tau U_\delta U_\omega^2, C_3(\tau)=2\tau U_r, C_4=U_\delta,$$
$$C_5(\tau^2)=48\tau^2U_\delta^2, C_6(\tau^2)=192\tau^2U_\omega^3U_\delta^2, C_7=3U_\delta^2, C_8(\tau^2)=24\tau^2U_r^2. \tag{135}$$

Proof. For any j, one has

$$\begin{aligned}
&\|\omega_{j+1}^k-\omega_j^k\|=\|\Pi_{U_\omega}(\omega_j^k+\frac{\beta}{L}\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\mathbf{B}_j^\top\phi(s_{j,\ell}^k))-\Pi_{U_\omega}(\omega_j^k)\|\\
\le&\Big\|\frac{\beta}{L}\sum_{\ell=0}^{L-1}\mathbf{B}_j^\top\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big\| \qquad \text{(by non-expansive property of projection)}\\
=&\beta\|\mathbf{B}_j^\top\left(\widetilde{A}_j^kx_j^k+y_j^k\bar{\phi}_j^k+\mathbf{b}_j^k\right)\| \qquad \text{(by Proposition 2)}\\
\le&\beta\|\mathbf{B}_j^\top\widetilde{A}_j^kx_j^k\|+\beta\|\mathbf{B}_j^\top\mathbf{b}_j^k\|+\beta\|\mathbf{B}_j^\top y_j^k\bar{\phi}_j^k\|\\
\le&2\beta\|x_j^k\|+\beta\|\mathbf{b}_j^k\|+\beta|y_j^k|.
\end{aligned}\tag{136}$$

By Lemma K.5, and by the tower property and the notation defined in Section C,

$$\mathbb{E}_{t-2\tau}\|\mathbf{b}_j^k\|\le\frac{U_\delta}{\sqrt{L}},$$

where $U_\delta=2U_r+2U_\omega$ as defined in Lemma K.5 (recall $\mathbf{b}_j^k$ is a martingale term, so its bound carries no mixing factor). Then, taking expectation up to $t-2\tau$, Eq. (136) can be written as,

$$\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k-\omega_j^k\|\le2\beta\mathbb{E}_{t-2\tau}\|x_j^k\|+\frac{\beta U_\delta}{\sqrt{L}}+\beta\mathbb{E}_{t-2\tau}|y_j^k|. \tag{137}$$

By Lemma K.4, Eq. (137) can be written as,

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k-\omega_j^k\|\le2\beta\mathbb{E}_{t-2\tau}\|x_j^k\|+\frac{\beta U_\delta}{\sqrt{L}}+\beta\mathbb{E}_{t-2\tau}|y_j^k|\\
\le&2\beta\left(\tau U_\delta\beta+2\tau U_\delta U_\omega^2\zeta+\mathbb{E}_{t-2\tau}\|x_t^k\|\right)+\frac{\beta U_\delta}{\sqrt{L}}\\
&+\beta\left(2\tau U_r\gamma+\mathbb{E}_{t-2\tau}|y_t^k|\right)\\
=&\underbrace{2\tau U_\delta}_{:=C_1(\tau)}\beta^2+\underbrace{4\tau U_\delta U_\omega^2}_{:=C_2(\tau)}\zeta\beta+\underbrace{2\tau U_r}_{:=C_3(\tau)}\gamma\beta+\frac{\beta}{\sqrt{L}}\underbrace{U_\delta}_{:=C_4}+2\beta\mathbb{E}_{t-2\tau}\|x_t^k\|+\beta\mathbb{E}_{t-2\tau}|y_t^k|\\
\le&C_1(\tau)\beta^2+C_2(\tau)\zeta\beta+C_3(\tau)\gamma\beta+C_4\frac{\beta}{\sqrt{L}}+2\beta\mathbb{E}_{t-2\tau}\|x_t^k\|+\beta\mathbb{E}_{t-2\tau}|y_t^k|.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k-\omega_j^k\|^2=\mathbb{E}_{t-2\tau}\|\Pi_{U_\omega}(\omega_j^k+\frac{\beta}{L}\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\mathbf{B}_j^\top\phi(s_{j,\ell}^k))-\Pi_{U_\omega}(\omega_j^k)\|^2\\
\le&\mathbb{E}_{t-2\tau}\Big\|\frac{\beta}{L}\sum_{\ell=0}^{L-1}\mathbf{B}_j^\top\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big\|^2\\
=&\beta^2\mathbb{E}_{t-2\tau}\|\mathbf{B}_j^\top\left(\widetilde{A}_j^kx_j^k+y_j^k\bar{\phi}_j^k+\mathbf{b}_j^k\right)\|^2 \qquad \text{(by Proposition 2)}\\
\le&3\beta^2\mathbb{E}_{t-2\tau}\|\mathbf{B}_j^\top\widetilde{A}_j^kx_j^k\|^2+3\beta^2\mathbb{E}_{t-2\tau}\|\mathbf{B}_j^\top\mathbf{b}_j^k\|^2+3\beta^2\mathbb{E}_{t-2\tau}\|\mathbf{B}_j^\top y_j^k\bar{\phi}_j^k\|^2\\
\le&12\beta^2\mathbb{E}_{t-2\tau}\|x_j^k\|^2+3\beta^2\mathbb{E}_{t-2\tau}\|\mathbf{b}_j^k\|^2+3\beta^2\mathbb{E}_{t-2\tau}|y_j^k|^2.
\end{aligned}\tag{138}$$

By Lemma K.5 and the tower property, we get,

$$\mathbb{E}_{t-2\tau}\|\mathbf{b}_t^k\|^2\le\frac{U_\delta^2}{L}.$$

By Lemma K.4, Eq. (138) can be written as,

$$\begin{aligned}&\mathbb{E}_{t-2\tau}\|\omega_{j+1}^k-\omega_j^k\|^2\le 3\left(4\beta^2\mathbb{E}_{t-2\tau}\|x_j^k\|^2+\frac{U_\delta^2}{L}\beta^2+\beta^2\mathbb{E}_{t-2\tau}|y_j^k|^2\right)\\ \le&48\tau^2U_\delta^2\beta^4+192\tau^2U_\omega^3U_\delta^2\zeta^2\beta^2+24\beta^2\mathbb{E}_{t-2\tau}\|x_t^k\|^2\\ &+3U_\delta^2\frac{\beta^2}{L}+24\tau^2U_r^2\gamma^2\beta^2+6\beta^2\mathbb{E}_{t-2\tau}|y_t^k|^2.\end{aligned}$$

□

Lemma K.3. *Fix $\tau>0$. Choosing ζ so that $\zeta U_\delta U_\omega\le 1/2$ and $U_\omega\zeta\le 1/2$, for any $j\ge t-\tau\ge\tau$, we have*

$$\begin{aligned}\mathbb{E}_{t-2\tau}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\|]\le&C_9(\tau)\zeta\beta+C_{10}(\tau)\zeta^2+2U_\omega\frac{\zeta}{K}\sum_{k=1}^K\mathbb{E}_{t-2\tau}\|x_t^k\|+C_{11}(\tau)\zeta\beta+C_{12}\frac{\zeta}{\sqrt{K}}\\ &+C_{13}\zeta\frac{m\rho^{\tau L}}{L(1-\rho)}+C_{14}(\tau)\gamma\zeta+U_\omega\frac{\zeta}{K}\sum_{k=1}^K\mathbb{E}_{t-2\tau}|y_t^k|+C_{15}\zeta^2,\\ \mathbb{E}_{t-2\tau}\|\mathbf{B}_{j+1}-\mathbf{B}_j\|^2\le&C_{16}(\tau^2)\zeta^2\beta^2+C_{17}(\tau^2)\zeta^4+C_{18}\frac{\zeta^2}{K}\sum_{k=1}^K\mathbb{E}_{t-2\tau}\|x_t^k\|^2+C_{19}(\tau^2)\zeta^2\beta^2\\ &+C_{20}\frac{\zeta^2}{K}+C_{21}\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}+C_{22}(\tau^2)\gamma^2\zeta^2+C_{23}\frac{\zeta^2}{K}\sum_{k=1}^K\mathbb{E}_{t-2\tau}|y_t^k|^2+C_{24}\zeta^4,\end{aligned}$$

where

$$\begin{aligned}&C_9(\tau)=2U_\omega\tau U_\delta, C_{10}(\tau)=4\tau U_\delta U_\omega^3, C_{11}(\tau)=U_\delta^2\tau, C_{12}=U_\omega U_\delta,\\ &C_{13}=2U_\delta U_\omega, C_{14}(\tau)=2U_\omega\tau U_r, C_{15}=2U_\delta^2U_\omega^2,\\ &C_{16}(\tau^2)=96\tau^2U_\omega^2U_\delta^2, C_{17}(\tau^2)=384\tau^2U_\omega^5U_\delta^2, C_{18}=48U_\omega^2, C_{19}=12U_\delta^4\tau^2,\\ &C_{20}=12U_\omega^2U_\delta^2, C_{21}=12U_\omega^2U_\delta^2, C_{22}(\tau^2)=48U_\omega^2\tau^2U_r^2, C_{23}=12U_\omega^2, C_{24}=8U_\delta^4U_\omega^4.\end{aligned}$$

Proof. Fix $\tau>0$. For any $j\ge t-\tau$, the one-step perturbation of $\mathbf{B}_j$ can be bounded as

$$\begin{aligned}&\|\mathbf{B}_{j+1}-\mathbf{B}_j\|\\ =&\|(\mathbf{B}_j+\frac{\zeta}{KL}\sum_{k=1}^K\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top)\mathbf{R}_{j+1}^{-1}-\mathbf{B}_j\|\\ =&\|(\mathbf{B}_j+\frac{\zeta}{KL}\sum_{k=1}^K\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top)\mathbf{R}_{j+1}^{-1}-\mathbf{B}_j\mathbf{R}_{j+1}^{-1}+\mathbf{B}_j\mathbf{R}_{j+1}^{-1}-\mathbf{B}_j\|\\ \le&\left\|\frac{\zeta}{KL}\sum_{k=1}^K\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top\right\|\|\mathbf{R}_{j+1}^{-1}\|+\|\mathbf{R}_{j+1}^{-1}-\mathbf{I}\|\|\mathbf{B}_j\|\\ \le&\left\|\frac{\zeta}{KL}\sum_{k=1}^K\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top\right\|+2\|\mathbf{Q}_j\|_F^2,\end{aligned}\tag{139}$$

where the last inequality follows from Lemma D.1 and the fact that $\|\mathbf{B}_j\|\le 1$. From Eq. (19), we know $\|\mathbf{Q}_j\|_F\le U_\delta U_\omega\zeta$. Thus, it remains to bound $\left\|\frac{\zeta}{KL}\sum_{k=1}^K\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top\right\|$,

for which we have

$$
\begin{aligned}
&\left\| \frac{\zeta}{KL} \sum_{k=1}^{K} \mathbf{P}_{j,\perp} \Big(\sum_{\ell=0}^{L-1} \delta_{j,\ell}^k \phi(s_{j,\ell}^k) \Big) (\omega_j^k)^\top \right\| \\
=&\frac{\zeta}{KL} \left\| \sum_{k=1}^{K} \mathbf{P}_{j,\perp} \Big(\sum_{\ell=0}^{L-1} \delta_{j,\ell}^k \phi(s_{j,\ell}^k) \Big) (\omega_j^k)^\top \right\| \\
\leq&\frac{\zeta}{KL} \left\| \sum_{k=1}^{K} \Big(\sum_{\ell=0}^{L-1} \delta_{j,\ell}^k \phi(s_{j,\ell}^k) \Big) (\omega_j^k)^\top \right\| \\
=&\frac{\zeta}{K} \left\| \sum_{k=1}^{K} \left(\widetilde{A}_j^k x_j^k + y_j^k \bar{\phi}_j^k + \mathbf{b}_j^k \right) (\omega_j^k)^\top \right\| \\
\leq&\frac{\zeta}{K} \left\| \sum_{k=1}^{K} \widetilde{A}_j^k x_j^k (\omega_j^k)^\top \right\| + \frac{\zeta}{K} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_j^k)^\top \right\| + \frac{\zeta}{K} \left\| \sum_{k=1}^{K} y_j^k \bar{\phi}_j^k (\omega_j^k)^\top \right\| \\
\leq&2U_\omega \frac{\zeta}{K} \sum_{k=1}^{K} \|x_j^k\| + \underbrace{\frac{\zeta}{K} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_j^k - \omega_{j-\tau}^k)^\top \right\|}_{I_1} + \underbrace{\frac{\zeta}{K} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_{j-\tau}^k)^\top \right\|}_{I_2} + U_\omega \frac{\zeta}{K} \sum_{k=1}^{K} |y_j^k|.
\end{aligned}
\tag{140}
$$

For I_1 in Eq. (140),

$$
\frac{\zeta}{K} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_j^k - \omega_{j-\tau}^k)^\top \right\| \leq U_\delta^2 \tau \zeta \beta. \quad \text{(by Lemma K.1)} \tag{141}
$$

For I_2 in Eq. (140), taking conditional expectation with respect to the Markovian trajectories up to time $j-\tau$, we have

$$
\begin{aligned}
&\frac{\zeta}{K} \mathbb{E}_{j-\tau} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_{j-\tau}^k)^\top \right\| \\
\leq&\frac{\zeta}{K} \left(\mathbb{E}_{j-\tau} \left\| \sum_{k=1}^{K} \mathbf{b}_j^k (\omega_{j-\tau}^k)^\top \right\|^2 \right)^{\frac{1}{2}} \quad \text{(by Jensen's inequality for expectation)} \\
=&\frac{\zeta}{K} \left(\mathbb{E}_{j-\tau} \sum_{k=1}^{K} \left\| \mathbf{b}_j^k (\omega_{j-\tau}^k)^\top \right\|^2 + 2 \sum_{k=1}^{K} \sum_{k'=k+1}^{K} \mathbb{E}_{j-\tau} \left\langle \mathbf{b}_j^k (\omega_{j-\tau}^k)^\top, \mathbf{b}_j^{k'} (\omega_{j-\tau}^{k'})^\top \right\rangle \right)^{\frac{1}{2}} \\
\overset{(a)}{\leq}&\frac{\zeta}{K} \left(K U_\omega^2 U_\delta^2 + 2 \sum_{k=1}^{K} \sum_{k'=k+1}^{K} \mathbb{E}_{j-\tau} \left[\mathbf{b}_j^k (\omega_{j-\tau}^k)^\top \right] \mathbb{E}_{j-\tau} \left[\mathbf{b}_j^{k'} (\omega_{j-\tau}^{k'})^\top \right] \right)^{\frac{1}{2}} \\
\leq&\frac{\zeta}{K} \left(K U_\omega^2 U_\delta^2 + K^2 U_\delta^2 U_\omega^2 \frac{4m^2 \rho^{2\tau L}}{L^2 (1-\rho)^2} \right)^{\frac{1}{2}} \\
\leq&U_\omega U_\delta \frac{\zeta}{\sqrt{K}} + 2U_\delta U_\omega \zeta \frac{m \rho^{\tau L}}{L(1-\rho)}.
\end{aligned}
\tag{142}
$$

where inequality (a) follows from Lemma J.2 and conditional independence, and the mixing term uses $\|\mathbb{E}_{j-\tau}[\mathbf{b}_j^k]\| \leq 2U_\delta \frac{m\rho^{\tau L}}{L(1-\rho)}$ from Lemma K.6.

Plugging Eq. (141) and Eq. (142) back to Eq. (140) and Eq. (139), we get, under expectation,

$$\begin{aligned}\mathbb{E}_{j-\tau}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\|] &= \mathbb{E}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\| \mid v_{0:j-\tau}] \\ &\le 2U_\omega \frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{j-\tau}\|x_j^k\| + U_\delta^2\tau\zeta\beta + U_\omega U_\delta\frac{\zeta}{\sqrt{K}} \\ &\quad + 2U_\delta U_\omega\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{j-\tau}|y_j^k| + 2U_\delta^2U_\omega^2\zeta^2. \end{aligned} \tag{143}$$

Finally,

$$\begin{aligned}&\mathbb{E}_{t-2\tau}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\|] = \mathbb{E}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\| \mid v_{0:t-2\tau}] \\ =&\mathbb{E}[\,\mathbb{E}[\|\mathbf{B}_{j+1}-\mathbf{B}_j\| \mid v_{0:j-\tau}] \mid v_{0:t-2\tau}] \qquad \text{(by the tower property of conditional expecation)} \\ \le&2U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}[\mathbb{E}_{j-\tau}\|x_j^k\|] + U_\delta^2\tau\zeta\beta + U_\omega U_\delta\frac{\zeta}{\sqrt{K}} \qquad \text{(by applying Eq. (143))} \\ &+ 2U_\delta U_\omega\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}[\mathbb{E}_{j-\tau}|y_j^k|] + 2U_\delta^2U_\omega^2\zeta^2 \\ =&2U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}[\|x_j^k\|] + U_\delta^2\tau\zeta\beta + U_\omega U_\delta\frac{\zeta}{\sqrt{K}} \\ &+ 2U_\delta U_\omega\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}[|y_j^k|] + 2U_\delta^2U_\omega^2\zeta^2 \\ \le&2U_\omega\zeta\left(\tau U_\delta\beta + 2\tau U_\delta U_\omega^2\zeta + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|\right) + U_\delta^2\tau\zeta\beta + U_\omega U_\delta\frac{\zeta}{\sqrt{K}} \\ &+ 2U_\delta U_\omega\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\left(2\tau U_r\gamma + \mathbb{E}_{t-2\tau}|y_t^k|\right) + 2U_\delta^2U_\omega^2\zeta^2 \quad \text{(applying Lemma K.4)} \\ =&\underbrace{2U_\omega\tau U_\delta}_{:=C_9(\tau)}\zeta\beta + \underbrace{4\tau U_\delta U_\omega^3}_{:=C_{10}(\tau)}\zeta^2 + 2U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\| + \underbrace{U_\delta^2\tau}_{:=C_{11}(\tau)}\zeta\beta + \underbrace{U_\omega U_\delta}_{:=C_{12}}\frac{\zeta}{\sqrt{K}} \\ &+ \underbrace{2U_\delta U_\omega}_{:=C_{13}}\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + \underbrace{2U_\omega\tau U_r}_{:=C_{14}(\tau)}\gamma\zeta + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k| + \underbrace{2U_\delta^2U_\omega^2}_{:=C_{15}}\zeta^2 \\ =&C_9(\tau)\zeta\beta + C_{10}(\tau)\zeta^2 + 2U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\| + C_{11}(\tau)\zeta\beta + C_{12}\frac{\zeta}{\sqrt{K}} \\ &+ C_{13}\zeta\frac{m\rho^{\tau L}}{L(1-\rho)} + C_{14}(\tau)\gamma\zeta + U_\omega\frac{\zeta}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k| + C_{15}\zeta^2. \end{aligned}$$

Analogously,

$$\begin{aligned}&\|\mathbf{B}_{j+1}-\mathbf{B}_j\|^2 \\ \le&2\left\|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^k\phi(s_{j,\ell}^k)\Big)(\omega_j^k)^\top\right\|^2 + 8\|\mathbf{Q}_j\|_F^4. \end{aligned} \tag{144}$$

Similar to Eq. (140), the first term can be decomposed and bounded as,

$$
\begin{aligned}
&2\left\|\frac{\zeta}{KL}\sum_{k=1}^{K}\mathbf{P}_{j,\perp}\Big(\sum_{\ell=0}^{L-1}\delta_{j,\ell}^{k}\phi(s_{j,\ell}^{k})\Big)(\omega_j^k)^\top\right\|^2\\
\leq&\frac{6\zeta^2}{K^2}\left\|\sum_{k=1}^{K}\widetilde{A}_j^k x_j^k(\omega_j^k)^\top\right\|^2+\frac{6\zeta^2}{K^2}\left\|\sum_{k=1}^{K}\mathbf{b}_j^k(\omega_j^k)^\top\right\|^2+\frac{6\zeta^2}{K^2}\left\|\sum_{k=1}^{K}y_j^k\bar{\phi}_j^k(\omega_j^k)^\top\right\|^2\\
\leq&24U_\omega^2\frac{\zeta^2}{K}\sum_{k=1}^{K}\|x_j^k\|^2+\underbrace{\frac{12\zeta^2}{K^2}\left\|\sum_{k=1}^{K}\mathbf{b}_j^k(\omega_j^k-\omega_{j-\tau}^k)^\top\right\|^2}_{I_3}\\
&+\underbrace{\frac{12\zeta^2}{K^2}\left\|\sum_{k=1}^{K}\mathbf{b}_j^k(\omega_{j-\tau}^k)^\top\right\|^2}_{I_4}+U_\omega^2\frac{6\zeta^2}{K}\sum_{k=1}^{K}|y_j^k|^2.
\end{aligned}
\tag{145}
$$

For I_3 in Eq. (145), by Lemma J.2 and Lemma K.1, we have

$$
\frac{12\zeta^2}{K^2}\left\|\sum_{k=1}^{K}\mathbf{b}_j^k(\omega_j^k-\omega_{j-\tau}^k)^\top\right\|^2\leq 12U_\delta^4\tau^2\zeta^2\beta^2. \tag{146}
$$

For I_4 in Eq. (145), we take the conditional expectation with respect to the Markovian trajectories up to time $j-\tau$. Similar to Eq. (142), we obtain

$$
\begin{aligned}
\frac{12\zeta^2}{K^2}\mathbb{E}_{j-\tau}\left\|\sum_{k=1}^{K}\mathbf{b}_j^k(\omega_{j-\tau}^k)^\top\right\|^2&\leq\frac{12\zeta^2}{K^2}\left(KU_\omega^2U_\delta^2+K^2U_\delta^2U_\omega^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}\right)\\
&\leq 12U_\omega^2U_\delta^2\frac{\zeta^2}{K}+12U_\omega^2U_\delta^2\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}.
\end{aligned}
\tag{147}
$$

Plugging Eq. (146) and Eq. (147) back to Eq. (145) and Eq. (144), and applying the tower property, we get

$$
\begin{aligned}
&\mathbb{E}_{t-2\tau}\|\mathbf{B}_{j+1}-\mathbf{B}_j\|^2=\mathbb{E}_{t-2\tau}\left[\mathbb{E}_{j-\tau}\|\mathbf{B}_{j+1}-\mathbf{B}_j\|^2\right]\\
\leq&24U_\omega^2\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_j^k\|^2+12U_\delta^4\tau^2\zeta^2\beta^2+12U_\omega^2U_\delta^2\frac{\zeta^2}{K}\\
&+12U_\omega^2U_\delta^2\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}+U_\omega^2\frac{6\zeta^2}{K^2}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_j^k|^2+8(U_\delta U_\omega\zeta)^4\\
\leq&24U_\omega^2\zeta^2\left(4\tau^2U_\delta^2\beta^2+16\tau^2U_\omega^3U_\delta^2\zeta^2+2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2\right)\\
&+12U_\delta^4\tau^2\zeta^2\beta^2+12U_\omega^2U_\delta^2\frac{\zeta^2}{K}+12U_\omega^2U_\delta^2\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}\\
&+U_\omega^2\frac{6\zeta^2}{K^2}\sum_{k=1}^{K}\left(8\tau^2U_r^2\gamma^2+2\mathbb{E}_{t-2\tau}|y_t^k|^2\right)+8(U_\delta U_\omega\zeta)^4\quad\text{(applying Lemma K.4)}\\
\leq&96\tau^2U_\omega^2U_\delta^2\zeta^2\beta^2+384\tau^2U_\omega^5U_\delta^2\zeta^4+48U_\omega^2\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}\|x_t^k\|^2\\
&+12U_\delta^4\tau^2\zeta^2\beta^2+12U_\omega^2U_\delta^2\frac{\zeta^2}{K}+12U_\omega^2U_\delta^2\zeta^2\frac{4m^2\rho^{2\tau L}}{L^2(1-\rho)^2}+48U_\omega^2\tau^2U_r^2\gamma^2\zeta^2\\
&+12U_\omega^2\frac{\zeta^2}{K}\sum_{k=1}^{K}\mathbb{E}_{t-2\tau}|y_t^k|^2+8U_\delta^4U_\omega^4\zeta^4.
\end{aligned}
$$

□

Lemma K.4. *For $0 \le t-\tau \le j \le t-1$, we have for any $k \in [K]$,*

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|x_j^k\| \le \tau U_\delta \beta + 2\tau U_\delta U_\omega^2 \zeta + \mathbb{E}_{t-2\tau}\|x_t^k\|,\\
&\mathbb{E}_{t-2\tau}\|x_j^k\|^2 \le 4\tau^2 U_\delta^2 \beta^2 + 16\tau^2 U_\omega^3 U_\delta^2 \zeta^2 + 2\mathbb{E}_{t-2\tau}\|x_t^k\|^2,\\
&\mathbb{E}_{t-2\tau}|y_j^k| \le 2\tau U_r \gamma + \mathbb{E}_{t-2\tau}|y_t^k|,\\
&\mathbb{E}_{t-2\tau}|y_j^k|^2 \le 8\tau^2 U_r^2 \gamma^2 + 2\mathbb{E}_{t-2\tau}|y_t^k|^2,
\end{aligned}$$

when $2\tau U_\delta^2 U_\omega^3 \zeta \le \tau U_\delta U_\omega^2$, and $32\tau^2 U_\omega^5 U_\delta^2 \zeta^2 \le 8\tau^2 U_\omega^3 U_\delta^2$.

Lemma K.4 aligns the error terms indexed by j to a common reference time t, which is required for a consistent Lyapunov analysis. The bounds follow from the triangle inequality applied along the trajectory from j to t; since $j \ge t-\tau$, the difference between x_j^k (resp., y_j^k) and x_t^k (resp., y_t^k) accumulates at most τ one-step perturbations. Each perturbation is upper bounded using the update dynamics, yielding deviations that scale linearly in τ, while $\mathbb{E}_{t-2\tau}\|x_t^k\|$ and $\mathbb{E}_{t-2\tau}|y_t^k|$ serve as anchored reference terms.

Proof. For any $t-1 \ge j \ge t-\tau$, we have

$$\begin{aligned}
&\|x_j^k\|^2 \le 2\|x_j^k - x_t^k\|^2 + 2\|x_t^k\|^2\\
=&2\|\mathbf{B}_t\omega_t^k - \mathbf{B}_j\omega_j^k\|^2 + 2\|x_t^k\|^2\\
\le&4\|\mathbf{B}_t\omega_t^k - \mathbf{B}_t\omega_j^k\|^2 + 4\|\mathbf{B}_t\omega_j^k - \mathbf{B}_j\omega_j^k\|^2 + 2\|x_t^k\|^2\\
\le&4\|\omega_t^k - \omega_j^k\|^2 + 4U_\omega\|\mathbf{B}_t - \mathbf{B}_j\|^2 + 2\|x_t^k\|^2\\
\le&4\tau^2 U_\delta^2\beta^2 + 4\tau^2 U_\omega(U_\delta U_\omega \zeta + 2U_\delta^2 U_\omega^2 \zeta^2)^2 + 2\|x_t^k\|^2 \quad \text{(by Lemma K.1)}.
\end{aligned}$$

Then, taking expectation, we get

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|x_j^k\|^2 \le 4\tau^2 U_\delta^2\beta^2 + 8\tau^2 U_\omega^3 U_\delta^2 \zeta^2 + 32\tau^2 U_\omega^5 U_\delta^2 \zeta^4 + 2\mathbb{E}_{t-2\tau}\|x_t^k\|^2\\
\le&4\tau^2 U_\delta^2\beta^2 + 16\tau^2 U_\omega^3 U_\delta^2 \zeta^2 + 2\mathbb{E}_{t-2\tau}\|x_t^k\|^2.
\end{aligned}$$

The last inequality holds since $32\tau^2 U_\omega^5 U_\delta^2 \zeta^2 \le 8\tau^2 U_\omega^3 U_\delta^2$.

Analogously, we get,

$$\begin{aligned}
&\mathbb{E}_{t-2\tau}\|x_j^k\| \le \tau U_\delta\beta + \tau U_\delta U_\omega^2\zeta + 2\tau U_\delta^2 U_\omega^3 \zeta^2 + \mathbb{E}_{t-2\tau}\|x_t^k\|\\
\le&\tau U_\delta\beta + 2\tau U_\delta U_\omega^2\zeta + \mathbb{E}_{t-2\tau}\|x_t^k\|.
\end{aligned}$$

The last inequality holds since $2\tau U_\delta^2 U_\omega^3 \zeta \le \tau U_\delta U_\omega^2$.

Similarly, for $|y_j^k|$, we have,

$$\begin{aligned}
&|y_j^k|^2 \le 2|y_j^k - y_t^k|^2 + 2|y_t^k|^2\\
\le&2\tau\sum_{i=j}^{t-1}|y_{i+1}^k - y_i^k|^2 + 2|y_t^k|^2\\
=&2\tau\sum_{i=j}^{t-1}|\eta_{i+1}^k - \eta_i^k|^2 + |y_t^k|\\
=&2\tau\gamma^2\sum_{i=j}^{t-1}\left|\frac{1}{L}\sum_{\ell=0}^{L-1} r_{t,\ell}^k - \eta_t^k\right|^2 + 2|y_t^k|^2\\
\le&8\tau^2 U_r^2\gamma^2 + 2|y_t^k|^2.
\end{aligned}$$

Taking expectation, we get

$$\mathbb{E}_{t-2\tau}|y_j^k|^2 \le 8\tau^2 U_r^2\gamma^2 + 2\mathbb{E}_{t-2\tau}|y_t^k|^2.$$

Analogously, we get,

$$\mathbb{E}_{t-2\tau}|y_j^k| \leq 2\tau U_r \gamma + \mathbb{E}_{t-2\tau}|y_t^k|.$$

□

Lemma K.5. *For $k \in [K]$, $t \geq 0$, $L > 0$, and any state $s_{t,0}^k$, we have*

$$\mathbb{E}\left[\|\mathbf{b}_t^k\|\right] \leq \frac{U_\delta}{\sqrt{L}}, \tag{148}$$

$$\mathbb{E}\|\mathbf{b}_t^k\|^2 \leq \frac{U_\delta^2}{L}, \tag{149}$$

$$\mathbb{E}\|\widetilde{b}_t^k - \bar{b}^k\| \leq \frac{1}{\sqrt{L}}\sqrt{16U_r^2 + 32U_r^2\frac{m}{1-\rho}}, \tag{150}$$

$$\mathbb{E}\|\widetilde{b}_t^k - \bar{b}^k\|^2 \leq \frac{1}{L}\left(16U_r^2 + 32U_r^2\frac{m}{1-\rho}\right), \tag{151}$$

$$\mathbb{E}\left\|\widetilde{A}_t^k - A^k\right\| \leq \frac{1}{\sqrt{L}}\sqrt{16 + 32\frac{m}{1-\rho}}, \tag{152}$$

$$\mathbb{E}\left\|\widetilde{A}_t^k - A^k\right\|^2 \leq \frac{1}{L}\left(16 + 32\frac{m}{1-\rho}\right), \tag{153}$$

where $U_\delta := 2U_r + 2U_\omega$, and the expectation is taken over the Markovian sampling trajectories.

Proof. We know that $V(s) = \phi(s)^\top z^{k,*} = \mathbb{E}_{\pi\otimes P^k}[R(s,a) - J^k + V(s')]$, which implies that for any s,

$$\mathbb{E}_{(a,s')\sim\pi\otimes P^k}[R(s,a) - J^k + (\phi(s') - \phi(s))^\top z^{k,*} \mid s] = 0. \tag{154}$$

Recall from Eq. (16) that $\mathbf{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1} g_\ell^k$, where

$$g_\ell^k := \left[(r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}^*\omega^{k,*}\right]\phi(s_{t,\ell}^k).$$

Each g_ℓ^k is a function of $(s_{t,\ell}^k, a_{t,\ell}^k, s_{t,\ell+1}^k)$, hence measurable with respect to $v_{0:(t,\ell)}$. By the Bellman equation Eq. (154) applied pointwise at $s_{t,\ell}^k$ (recall $z^{k,*} = \mathbf{B}^*\omega^{k,*}$),

$$\mathbb{E}\big[g_\ell^k \mid v_{0:(t,\ell-1)}\big] = \phi(s_{t,\ell}^k)\,\mathbb{E}\big[(r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top z^{k,*} \,\big|\, s_{t,\ell}^k\big] = 0,$$

so $\{g_\ell^k\}_{\ell=0}^{L-1}$ is a martingale difference sequence with respect to $\{v_{0:(t,\ell)}\}_\ell$, and $\|g_\ell^k\| \leq U_\delta$ since $\left|(r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top z^{k,*}\right| \leq U_\delta$ and $\|\phi(s_{t,\ell}^k)\| \leq 1$. In particular, for any $\ell < \ell'$, since g_ℓ^k is measurable w.r.t. $v_{0:(t,\ell)}$ and $v_{0:(t,\ell)} \subseteq v_{0:(t,\ell'-1)}$, the tower property gives

$$\mathbb{E}\langle g_\ell^k, g_{\ell'}^k\rangle = \mathbb{E}\big[\langle g_\ell^k, \mathbb{E}[g_{\ell'}^k \mid v_{0:(t,\ell'-1)}]\rangle\big] = 0,$$

so all cross terms vanish. Therefore,

$$\mathbb{E}\|\mathbf{b}_t^k\|^2 = \frac{1}{L^2}\mathbb{E}\Big\|\sum_{\ell=0}^{L-1} g_\ell^k\Big\|^2 = \frac{1}{L^2}\sum_{\ell=0}^{L-1}\mathbb{E}\|g_\ell^k\|^2 \leq \frac{1}{L^2}\cdot L\cdot U_\delta^2 = \frac{U_\delta^2}{L},$$

which proves Eq. (149). Finally, by Jensen's inequality, $\mathbb{E}\|\mathbf{b}_t^k\| \leq \sqrt{\mathbb{E}\|\mathbf{b}_t^k\|^2} \leq \frac{U_\delta}{\sqrt{L}}$, which proves Eq. (148).

Now we prove Eq. (151). For ease of exposition, we omit the explicit specification of the underlying distributions in the expectation when it is clear from the context. Recall from Eq. (14) that $\widetilde{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k)$ and $\bar{b}^k = \mathbb{E}_{s\sim\mu^k,a\sim\pi}[(R(s,a) - J^k)\phi(s)]$. Denoting $f_\ell^k := (r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k) - \bar{b}^k$, we establish two facts.

(i) Boundedness. Since $|r_{t,\ell}^k - J^k| \leq 2U_r$, $\|\phi(\cdot)\| \leq 1$, and $\|\bar{b}^k\| = \|\mathbb{E}_{s\sim\mu^k,a\sim\pi}[(R(s,a) - J^k)\phi(s)]\| \leq 2U_r$, we have $\|f_\ell^k\| \leq 2U_r + 2U_r = 4U_r$, hence $\mathbb{E}\|f_\ell^k\|^2 \leq 16U_r^2$.

(ii) Mixing. Since $\bar{b}^k$ is a μ^k-expectation while the samples follow the Markov chain, $\mathbb{E}[f_\ell^k] \neq 0$ in general, so the cross terms are controlled by mixing. The stationary mean vanishes, $\mathbb{E}_{s\sim\mu^k, a\sim\pi}[f_\ell^k] = \bar{b}^k - \bar{b}^k = 0$. Fix $\ell < \ell'$. By the Markov property, f_ℓ^k and $f_{\ell'}^k$ are conditionally independent given $s_{t,\ell+1}^k$, so

$$\mathbb{E}\langle f_\ell^k, f_{\ell'}^k\rangle = \mathbb{E}\big\langle \mathbb{E}[f_\ell^k \mid s_{t,\ell+1}^k],\ \mathbb{E}[f_{\ell'}^k \mid s_{t,\ell+1}^k]\big\rangle.$$

Since $\mathbb{E}_{s\sim\mu^k, a\sim\pi}[f_{\ell'}^k] = 0$ and $\|(R(s,a) - J^k)\phi(s)\| \le 2U_r$, using $\|\mathbb{E}_\nu[X] - \mathbb{E}_{\mu^k}[X]\| \le 2\|X\|_\infty d_{TV}(\nu, \mu^k)$ and Assumption 2 we obtain

$$\big\|\mathbb{E}[f_{\ell'}^k \mid s_{t,\ell+1}^k]\big\| \le 2 \cdot 2U_r \cdot d_{TV}\big(\mathbb{P}^k(s_{t,\ell'}^k \mid s_{t,\ell+1}^k),\ \mu^k\big) \le 4U_r\, m\rho^{\ell'-\ell-1}.$$

Combined with $\|\mathbb{E}[f_\ell^k \mid s_{t,\ell+1}^k]\| \le \mathbb{E}[\|f_\ell^k\| \mid s_{t,\ell+1}^k] \le 4U_r$, Cauchy–Schwarz gives

$$\big|\mathbb{E}\langle f_\ell^k, f_{\ell'}^k\rangle\big| \le 4U_r \cdot 4U_r\, m\rho^{\ell'-\ell-1} = 16U_r^2\, m\rho^{\ell'-\ell-1}.$$

Combining (i) and (ii),

$$\begin{aligned}
\mathbb{E}\|\widetilde{b}_t^k - \bar{b}^k\|^2 &= \frac{1}{L^2}\mathbb{E}\Big\|\sum_{\ell=0}^{L-1} f_\ell^k\Big\|^2 \\
&\le \frac{1}{L^2}\left[\sum_{\ell=0}^{L-1}\mathbb{E}\|f_\ell^k\|^2 + 2\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\big|\mathbb{E}\langle f_\ell^k,\, f_{\ell'}^k\rangle\big|\right] \\
&\le \frac{1}{L^2}\left(16U_r^2 L + 32U_r^2 m\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\rho^{\ell'-\ell-1}\right) \\
&\le \frac{1}{L}\left(16U_r^2 + 32U_r^2\frac{m}{1-\rho}\right).
\end{aligned}$$

For Eq. (150), by Jensen's inequality, we get

$$\mathbb{E}\|\widetilde{b}_t^k - \bar{b}^k\| \le \sqrt{\mathbb{E}\|\widetilde{b}_t^k - \bar{b}^k\|^2} \le \frac{1}{\sqrt{L}}\sqrt{16U_r^2 + 32U_r^2\frac{m}{1-\rho}}.$$

Similarly, for Eq. (153), recall from Eq. (16) that $\widetilde{A}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top$ and $A^k = \mathbb{E}_{s\sim\mu^k,\, a\sim\pi,\, s'\sim P^k}[\phi(s)(\phi(s') - \phi(s))^\top]$. Denoting $h_\ell^k := \phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top - A^k$, the same two facts hold by the identical argument applied to $\phi(s)(\phi(s') - \phi(s))^\top$, whose norm is at most 2 and whose μ^k-mean is A^k: *(i)* $\|h_\ell^k\| \le \|\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top\| + \|A^k\| \le 2 + 2 = 4$, hence $\mathbb{E}\|h_\ell^k\|^2 \le 16$; and *(ii)* h_ℓ^k has zero stationary mean $\mathbb{E}_{s\sim\mu^k}[h_\ell^k] = 0$; by the same argument as in (ii) above, for $\ell < \ell'$, $\|\mathbb{E}[h_{\ell'}^k \mid s_{t,\ell+1}^k]\| \le 2 \cdot 2\, m\rho^{\ell'-\ell-1} = 4m\rho^{\ell'-\ell-1}$, so $|\mathbb{E}\langle h_\ell^k, h_{\ell'}^k\rangle| \le \|\mathbb{E}[h_\ell^k \mid s_{t,\ell+1}^k]\| \cdot 4m\rho^{\ell'-\ell-1} \le 16\, m\rho^{\ell'-\ell-1}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left\|\widetilde{A}_t^k - A^k\right\|^2 &= \frac{1}{L^2}\mathbb{E}\Big\|\sum_{\ell=0}^{L-1} h_\ell^k\Big\|^2 \\
&\le \frac{1}{L^2}\left[\sum_{\ell=0}^{L-1}\mathbb{E}\|h_\ell^k\|^2 + 2\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\big|\mathbb{E}\langle h_\ell^k,\, h_{\ell'}^k\rangle\big|\right] \\
&\le \frac{1}{L^2}\left(16L + 32m\sum_{\ell=0}^{L-1}\sum_{\ell'=\ell+1}^{L-1}\rho^{\ell'-\ell-1}\right) \le \frac{1}{L}\left(16 + 32\frac{m}{1-\rho}\right).
\end{aligned}$$

By Jensen's inequality, Eq. (152) follows:

$$\mathbb{E}\left\|\widetilde{A}_t^k - A^k\right\| \le \sqrt{\mathbb{E}\left\|\widetilde{A}_t^k - A^k\right\|^2} \le \frac{1}{\sqrt{L}}\sqrt{16 + 32\frac{m}{1-\rho}}.$$

□

Lemma K.6. *Suppose that Assumption 2. Fix $\tau > 0$. For any $t \geq \tau$, $L > 0$, and $k \in [K]$, it holds that*

$$\|\mathbb{E}_{t-\tau}\big[\widetilde{b}_t^k - \bar{b}^k\big]\| \leq \frac{4U_r m\rho^{\tau L}}{L(1-\rho)}$$
$$\|\mathbb{E}_{t-\tau}\big[\widetilde{A}_t^k - A^k\big]\| \leq \frac{4m\rho^{\tau L}}{L(1-\rho)}$$
$$\|\mathbb{E}_{t-\tau}[\mathbf{b}_t^k]\| \leq \frac{2U_\delta m\rho^{\tau L}}{L(1-\rho)}.$$

Proof. Recall from Eq. (16) that

$$\mathbf{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1} g_\ell^k,$$

where $g_\ell^k = \big((r_{t,\ell}^k - J^k) + (\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top \mathbf{B}^* \omega^{k,*}\big)\phi(s_{t,\ell}^k)$. By the Bellman equation Eq. (154), $\mathbb{E}_{s\sim\mu^k}[g_\ell^k] = 0$ and $\|g_\ell^k\| \leq U_\delta$. Since $s_{t,\ell}^k$ is $\tau L + \ell$ steps ahead of $s_{t-\tau}^k$, the total-variation bound $\|\mathbb{E}_\nu[X] - \mathbb{E}_{\mu^k}[X]\| \leq 2\|X\|_\infty d_{TV}(\nu, \mu^k)$ and Assumption 2 give

$$\|\mathbb{E}_{t-\tau}[g_\ell^k]\| = \|\mathbb{E}_{t-\tau}[g_\ell^k] - \mathbb{E}_{s\sim\mu^k}[g_\ell^k]\| \leq 2\|g_\ell^k\|_\infty \, d_{TV}\big(\mathbb{P}^k(s_{t,\ell}^k \mid s_{t-\tau}^k), \mu^k\big) \leq 2U_\delta \, m\rho^{\tau L+\ell}.$$

Averaging over the block and using $\sum_{\ell=0}^{L-1} \rho^{\tau L+\ell} \leq \rho^{\tau L}/(1-\rho)$,

$$\|\mathbb{E}_{t-\tau}[\mathbf{b}_t^k]\| = \Big\|\tfrac{1}{L}\sum_{\ell=0}^{L-1} \mathbb{E}_{t-\tau}[g_\ell^k]\Big\| \leq \frac{1}{L}\sum_{\ell=0}^{L-1} 2U_\delta \, m\rho^{\tau L+\ell} \leq \frac{2U_\delta m\rho^{\tau L}}{L(1-\rho)}.$$

Recall from Eq. (14) that

$$\widetilde{b}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1}(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k), \quad \bar{b}^k = \mathbb{E}_{s\sim\mu^k, a\sim\pi}\big[(R(s,a) - J^k)\phi(s)\big].$$

The same argument applied to $(r_{t,\ell}^k - J^k)\phi(s_{t,\ell}^k)$, which is bounded by $2U_r$ with stationary mean $\bar{b}^k$, gives

$$\|\mathbb{E}_{t-\tau}\big[\widetilde{b}_t^k - \bar{b}^k\big]\| \leq \frac{1}{L}\sum_{\ell=0}^{L-1} 4U_r \, m\rho^{\tau L+\ell} \leq \frac{4U_r m\rho^{\tau L}}{L(1-\rho)}.$$

Similarly, for $\widetilde{A}_t^k = \frac{1}{L}\sum_{\ell=0}^{L-1} \phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top$, the same argument applied to $\phi(s_{t,\ell}^k)(\phi(s_{t,\ell+1}^k) - \phi(s_{t,\ell}^k))^\top$, which is bounded by 2 with stationary mean A^k, gives

$$\|\mathbb{E}_{t-\tau}\big[\widetilde{A}_t^k - A^k\big]\| \leq \frac{1}{L}\sum_{\ell=0}^{L-1} 4 \, m\rho^{\tau L+\ell} \leq \frac{4m\rho^{\tau L}}{L(1-\rho)}.$$

□

L Single- versus Two-Timescale Stepsizes

Our stepsizes are constant over the run, so their ratios do not change with t. A two-timescale scheme instead gives the head and the subspace different decay exponents. On the instances in synthetic MDPs at $K = L = 16$, we compare

$$\beta = 0.08, \quad \zeta = 0.05, \quad \gamma = 0.05 \qquad \text{against} \qquad \beta_t = \frac{200}{t+2}, \quad \zeta_t = \frac{5}{(t+2)^{5/6}}, \quad \gamma_t = \frac{0.5}{t+2},$$

from the same initialization. In Figure 4, the two-timescale run is the faster of the two until $t \approx 8\times10^3$ and then flattens as ζ_t decays; at $T = 6 \times 10^4$ its errors are $\bar{X} = 7.2 \times 10^{-3}$ and $M = 7.4 \times 10^{-3}$ against 2.1×10^{-3} and 1.0×10^{-3} for the constant stepsizes.

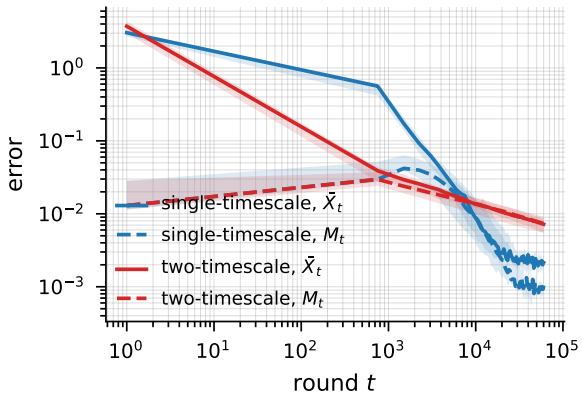


Figure 4: Single- versus two-timescale stepsizes at $K = L = 16$, median and interquartile range over eight seeds.

M Experiments on Gym Environments

M.1 Prediction under a fixed policy

In this section, we evaluate the performance of **pFedAR-TD** for approximating value functions under a fixed policy. We compare our proposed approach, **pFedAR-TD**, against the baselines introduced in Sections 3: **Single TD** and **FedTD-Uniform**. **Single TD** is the standard single-agent TD learning; **FedTD-Uniform** periodically averages the parameters of all agents and outputs a universal value function for all agents across different environments. The following are the specifications of how to evaluate the performance of value function approximation:

Parameters	Agent 0	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
link length 1	2.2	1.5	2	3	2.1	1.5
link length 2	1	1	0.8	1	1	1

Table 2: Acrobot agent configuration.

A pre-trained single-agent TD model is employed as the fixed policy. A fixed initial state is selected and applied consistently across all methods. Value function convergence is monitored by periodically evaluating this state via a forward pass, which ensures the preselected state appears in every training episode. The value of the preselected state is recorded during training and serves as a metric to quantify

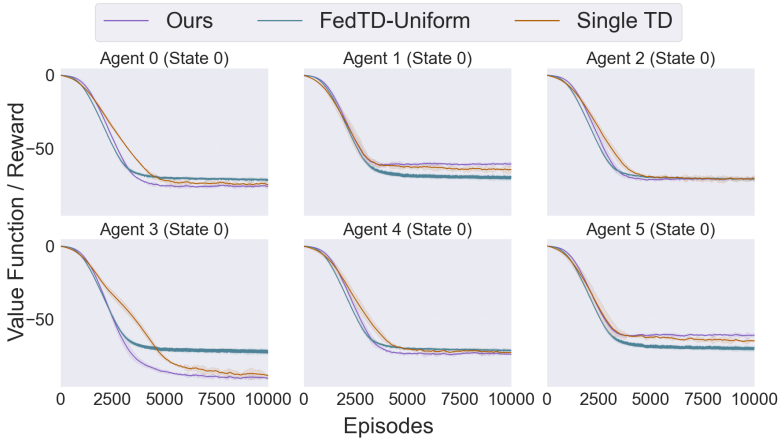


Figure 5: Value function convergence comparison on Acrobot.

the convergence rate toward the optimal value function for each approach. For the experiments, we utilize three distinct seeds and a synchronization rate of $E = 50$ as shared hyperparameters. The specific configuration of the Acrobot environment for each agent is detailed in Table 2.

We analyze Figure 5 from two perspectives. **(1) Convergence Speed: pFedAR-TD** exhibits significant acceleration over **Single TD** across most agents. **(2) Approximation Accuracy: FedTD-Uniform** converges to a suboptimal value in most configurations due to its lack of adaptability to heterogeneous environments, whereas **pFedAR-TD** maintains consistency with the optimal baseline.

Evaluation of value functions is relatively simple when the policy is fixed and the corresponding true values are determined. Consequently, we consider control problems, where policies are learned and updated throughout training, leading to increased uncertainty in learning.

M.2 Application to control problems

We extend **pFedAR-TD** to the actor-critic framework **pFed-AC** by utilizing **pFedAR-TD** as the critic component and incorporating an additional actor component. The action heads are also distinct across agents, i.e., agents do not share the parameters of those layers.

M.2.1 Model Design

The architecture is illustrated in Figure 6.

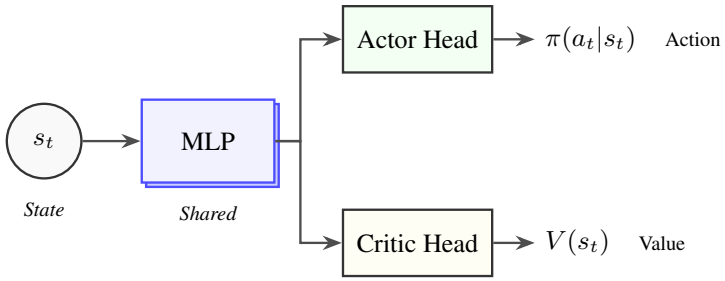


Figure 6: Model architecture.

M.2.2 Environment Configurations

We find that settings in prior work (Jin et al., 2022; Xiong et al., 2025) are not sufficiently heterogeneous. We introduce stronger heterogeneity via a mirrored environment where actions are reversed, making a shared policy ineffective. We use three random seeds and a synchronization interval $E = 50$. Environment parameters are listed in Tables 3 and 4.

Table 3: CartPole agent configuration.

Parameters	Agent 0	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
pole length	0.5	1	0.5	1	0.5	1
gravity	9.8	15	9.8	15	9.8	15
force	10	10	10	10	10	10
start position	$\frac{\pi}{6}$	$-\frac{\pi}{6}$	$\frac{\pi}{6}$	$-\frac{\pi}{6}$	$\frac{\pi}{6}$	$-\frac{\pi}{6}$
env	normal	mirrored	normal	mirrored	normal	mirrored

Table 4: Acrobot agent configuration.

Parameters	Agent 0	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
link length 1	1	1	1	1	1	1
link length 2	1	1	1	1	1	1
env	normal	mirrored	normal	mirrored	normal	mirrored

M.2.3 Results

(1) Convergence Speed: pFed-AC outperforms **FedAC-Uniform** and **Single AC** across most agents. Despite a marginal lag in early exploration, it surpasses baselines in mid-training and maintains

this advantage until convergence. **(2) Reward: pFed-AC** achieves superior asymptotic reward; in Acrobot, shared feature synchronization enables convergence closer to the theoretical optimum. **(3) Training Stability: pFed-AC** consistently narrows its confidence intervals, achieving the smallest variance across seeds in both environments. See Figures 7 and 8.

M.2.4 More Experiment Results

To further validate **pFed-AC** beyond these environments, we conduct an additional experiment on the `Hopper-v5` locomotion task using PPO (Schulman et al., 2017), one of the most widely adopted algorithms in the actor-critic family. Interestingly, the results in Figure 9 demonstrate that **pFed-PPO** consistently outperforms the baselines.

All experiments are conducted on a MacBook Air with an Apple M2 chip and 16 GB of unified memory.

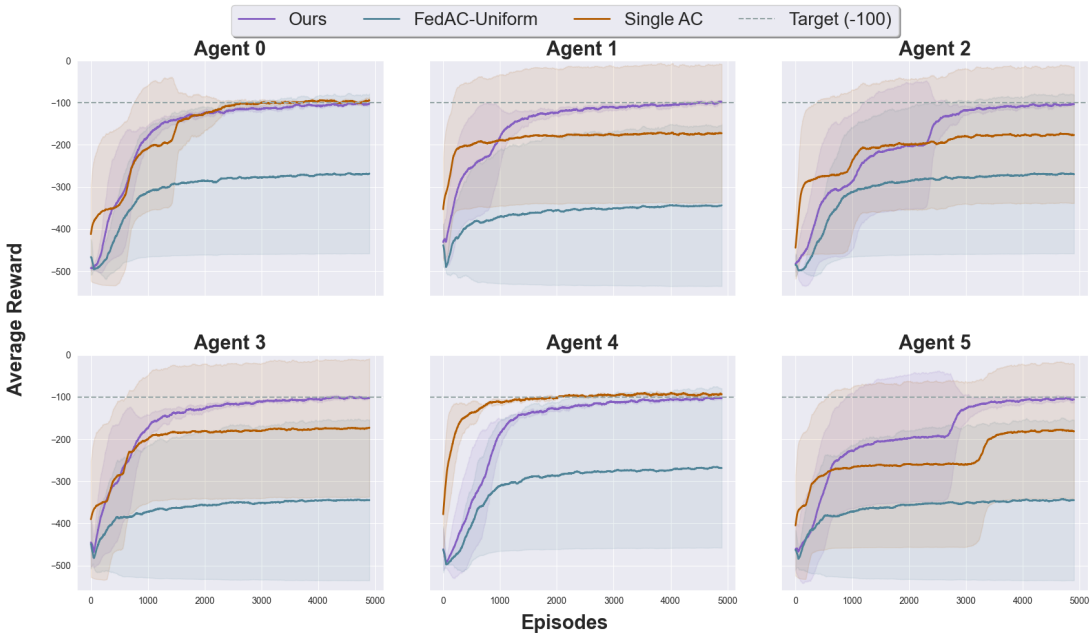


Figure 7: Policy Learning Performance Comparison on Acrobot



Figure 8: Policy Learning Performance Comparison on Cartpole

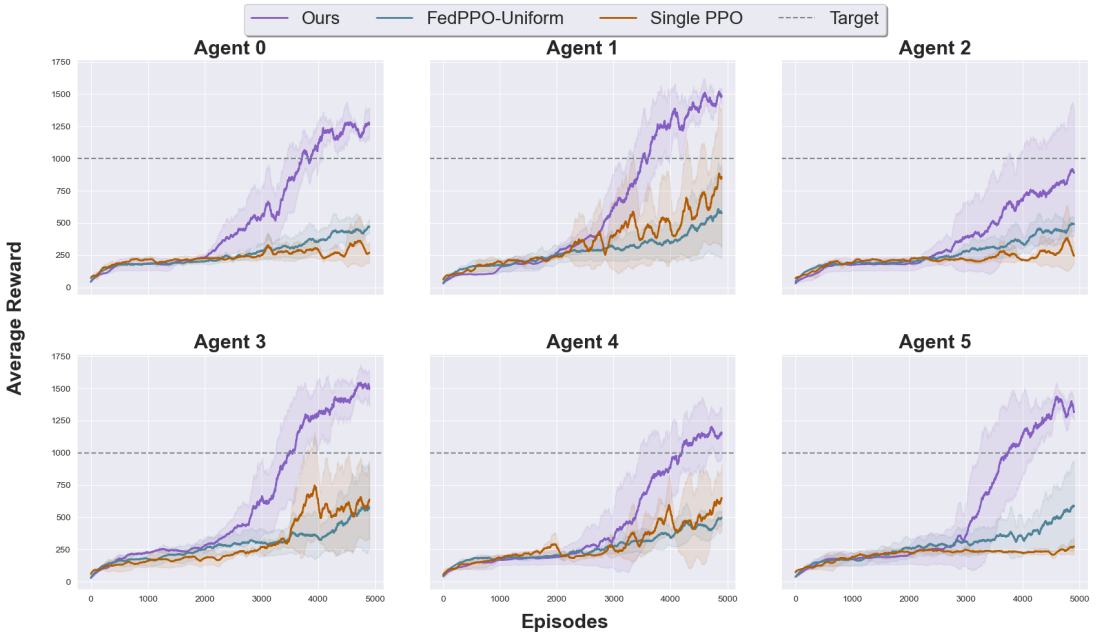


Figure 9: Policy Learning Performance Comparison on Hopper