EXACT-CONSENSUS DISTRIBUTED Q-LEARNING WITH FINITE-TIME GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study distributed Q-learning for networked Markov decision problems, where agents share a common environment, observe only local rewards, and communicate with their neighbors to learn the team-optimal Q-function. Since the local rewards differ across agents, existing methods rely on diminishing step-sizes to reach consensus, which sacrifices responsiveness over time. To overcome this limitation, we propose Exact-Consensus Q-learning (ECQ), which uses three variables: a local Bellman-learning variable, an agreement variable, and an auxiliary dynamic-consensus variable. By accumulating past disagreement, this auxiliary variable removes the steady-state consensus bias caused by reward heterogeneity. Under this scheme, the mean dynamics of ECQ are shown to converge to an equilibrium that exactly coincides with the centralized team-optimal Q-function. Moreover, using an exact switching-system representation, joint-spectral-radius stability, and martingale-difference noise analysis, we derive finite-time bounds that decompose the total error into a geometrically decaying transient and an $O(\sqrt{\alpha})$ residual. We also prove that the consensus error of ECQ vanishes geometrically without any residual under deterministic reward heterogeneity. Experiments show that ECQ reaches machine-precision consensus under deterministic local rewards, exhibits steady-state errors consistent with our theoretical results, and adapts robustly to sudden reward shifts.

1 INTRODUCTION

Many decision-making problems involve multiple agents operating in a shared environment, where each agent observes only partial information and communicates exclusively with its neighboring agents. Such problems are inherently distributed, requiring effective coordination and information sharing among agents to achieve a common objective. For this reason, consensus and cooperative control have been extensively studied to enable agents to reach an agreement through local communication (Vicsek et al., 1995; Jadbabaie et al., 2003; Olfati-Saber & Murray, 2004; Olfati-Saber, 2006; Olfati-Saber et al., 2007). The importance of information sharing has also been recognized in distributed reinforcement learning, where a team of agents learns a shared value function from reward information collected by individual agents (Lee et al., 2020). However, because agents receive different local rewards, their local learning processes tend to converge to different solutions. Related work in distributed optimization has addressed similar challenges using correction and tracking mechanisms to compensate for different local objectives (Shi et al., 2015; Nedic et al., 2017). In distributed reinforcement learning, $\mathcal{QD}$-learning combined Q-learning with a consensus mechanism using diminishing step-sizes and consensus gains. This enables the agents to eliminate disagreement and converge to the optimal value function of the network-averaged reward (Kar et al., 2013). However, diminishing step-sizes make the learning process progressively less responsive and require careful selection of the gain schedules. These limitations raise a natural question:

Can agents eliminate the consensus bias while maintaining responsiveness?

To answer this question, we propose Exact-Consensus Q-learning (ECQ), a distributed Q-learning method that eliminates the steady-state consensus bias caused by heterogeneous local rewards while retaining a constant step-size. The key idea is to separate local learning from network-wide agreement through three variables: a local Bellman-learning variable, an agreement variable, and an

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.

auxiliary dynamic-consensus variable. Unlike conventional consensus corrections, the auxiliary dynamic-consensus variable accumulates disagreement over time to compensate for persistent errors so that the consensus bias is removed without diminishing step-sizes. This prevents differences in local rewards from shifting the distributed equilibrium away from the centralized team-optimal solution. We further establish finite-time convergence guarantees for ECQ, providing a theoretical characterization of its convergence behavior. To this end, we represent the algorithm as an exact switching system and establish its contractivity by analyzing its joint spectral radius (JSR). By combining this property with bounds on policy-induced perturbations and sampling noise, we derive finite-time error bounds that explicitly characterize how quickly the iterates approach the optimal solution: the resulting bound on the total error consists of a geometrically decaying transient term and an $O(\sqrt{\alpha})$ steady-state residual. Furthermore, under deterministic reward heterogeneity, we show that the consensus error vanishes geometrically without any residual.

Our main contributions are summarized as follows:

1. We propose ECQ, a distributed Q-learning method that removes the consensus bias induced by heterogeneous local rewards under a constant step-size. We establish that the equilibrium of its mean dynamics coincides with the centralized team-optimal Q-function.
2. We establish a finite-time error bound on the total error for ECQ through a switching system representation, JSR stability, and a martingale-difference noise bound. Under deterministic reward heterogeneity, we further show that the consensus error of ECQ vanishes geometrically.
3. Experiments demonstrate that ECQ achieves machine-precision consensus under deterministic local rewards, maintains steady-state errors consistent with our theoretical bounds, and adapts robustly to sudden reward shifts.
4. We also provide a two-variable variant of ECQ that avoids an explicit auxiliary consensus state and uses an appropriate initialization condition in the appendix. We show that it achieves comparable performance while requiring one fewer variable per agent.

2 Related Work

Consensus and cooperative control have a long history in multi-agent systems (Vicsek et al., 1995; Jadbabaie et al., 2003; Olfati-Saber et al., 2007). In networked systems, consensus is commonly achieved through the graph Laplacian, which drives agents toward agreement with their neighbors (Olfati-Saber & Murray, 2004). However, with different local objectives, this mechanism alone with constant gain can leave persistent disagreement. A related idea has been studied in dynamic average consensus, where an auxiliary state accumulates the disagreement between neighbors to eliminate the steady-state error (Freeman et al., 2006; Yang et al., 2008). Further approaches were developed in distributed optimization to remove the bias caused by a constant step-size: EXTRA (Shi et al., 2015) introduced an additional correction term, while DIGing (Nedic et al., 2017) used gradient tracking. More recently, in the context of reinforcement learning, Wang et al. (2020) combined TD with gradient tracking for bias-free decentralized policy evaluation. Inspired by these approaches, we extend dynamic consensus to Q-learning, where disagreement dynamics are driven by local value updates from sampled transitions.

Cooperation and information sharing have also been studied in multi-agent reinforcement learning, including fully decentralized settings with networked agents (Zhang et al., 2018). Tan (1993) showed that information sharing can accelerate learning, but Claus & Boutilier (1998) found that it does not guarantee convergence to a team-optimal solution. $\mathcal{QD}$-learning (Kar et al., 2013) incorporated the information sharing directly into the update rule by combining local Q-learning updates with a consensus term. Under diminishing step-sizes, it converges almost surely to the optimal Q-function of the network-averaged reward MDP. Subsequent work established finite-sample guarantees for related distributed Q-learning schemes under linear function approximation (Heredia et al., 2020; Zeng et al., 2022) and in the tabular setting (Lim & Lee, 2025). However, diminishing step-sizes can limit adaptation in nonstationary environments (Sutton & Barto, 1998). The constant-step-size approach has been studied as an alternative, but it suffers from a persistent $O(\alpha)$ consensus error under doubly stochastic mixing (Lim & Lee, 2025). We therefore introduce an auxiliary state

that removes this steady-state consensus error while keeping the constant step-size, thus preserving adaptivity.

From the perspective of convergence analysis, related results have been established for distributed TD and primal-dual variants (Doan et al., 2019; Stanković et al., 2023; Lim & Lee, 2023; Lee et al., 2022; Cassano et al., 2021; Sun et al., 2020) and for decentralized batch RL (Zhang et al., 2021). Most of these works either focused on linear update dynamics for policy evaluation without action maximization, or studied batch/offline learning based on repeated regression. Moreover, a prior tracking-based method with finite-time guarantees was limited to linear policy evaluation (Wang et al., 2020). In contrast, we extend tracking to Q-learning, where the Bellman maximization makes the mean dynamics piecewise linear and non-smooth. Building on the switching-system view of single-agent Q-learning (Lee et al., 2023), we extend it to the networked setting and establish finite-time stability with a JSR argument.

3 PRELIMINARIES

Networked Markov Decision Problem. We consider a distributed learning setting where agents share the same state-transition dynamics, receive different local rewards, and communicate only with their neighbors.

Definition 3.1 (Networked MDP)**.** *A networked Markov decision problem (MDP) is defined by a tuple $(\mathcal{G}, \mathcal{S}, \mathcal{A}, \{P_a\}_{a\in\mathcal{A}}, \{r^i\}_{i=1}^N, \gamma)$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected connected communication graph with agent set $\mathcal{V} = \{1, \dots, N\}$, edge set $\mathcal{E}$, and neighbor sets $\mathcal{N}_i$. $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, P_a is the transition matrix under action a, r^i is the bounded local reward of agent i, and $\gamma \in (0, 1)$ is the discount factor. All agents share the same state-action-transition structure $(\mathcal{S}, \mathcal{A}, \{P_a\})$, while agent i observes only its local reward $r^i(s_k, a_k, s_{k+1})$. The team objective is to cooperatively learn the optimal Q-function associated with the team-average reward $\bar{r}(s, a, s') := \frac{1}{N}\sum_{i=1}^N r^i(s, a, s')$, using only local rewards and communication over $\mathcal{G}$.*

The communication structure is represented by the graph Laplacian $L \in \mathbb{R}^{N\times N}$, defined by $L_{ii} = |\mathcal{N}_i|$, $L_{ij} = -1$ if $j \in \mathcal{N}_i$, and $L_{ij} = 0$ otherwise. Under the assumption that the graph $\mathcal{G}$ is undirected and connected, its Laplacian satisfies

$$L = L^\top \succeq 0, \qquad L\mathbf{1}_N = 0, \qquad \mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}, \tag{1}$$

where $\mathbf{1}_N$ denotes the all-ones vector in $\mathbb{R}^N$.

Bellman Operators and Notation. Let d denote the state-action sampling distribution satisfying $d(s, a) > 0$ for every (s, a) to ensure sufficient exploration, where (s_k, a_k) are drawn i.i.d. from d and $s'_k \sim P_{a_k}(s_k, \cdot)$, generating a single transition shared by all agents. Define the expected local reward of agent i, the team-average reward, the stacked transition matrix, and the diagonal sampling distribution matrix

$$\begin{aligned} &R^i(s, a) := \sum_{s'} P_a(s, s')r^i(s, a, s'), \qquad R := \frac{1}{N}\sum_{i=1}^N R^i, \\ &P := \mathrm{col}(P_1, \dots, P_{|\mathcal{A}|}), \qquad D := \mathrm{diag}(d(s, a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}, \end{aligned} \tag{2}$$

where $\mathrm{col}(\cdot)$ denotes the vertical stacking of its arguments and $\mathrm{diag}(\cdot)$ denotes the diagonal matrix formed from its arguments. For a stochastic policy μ, let $\mu(s)$ denote the probability distribution over actions at state s. The corresponding policy-selection matrix is defined as

$$\Pi^\mu := \mathrm{col}(\mu(1)^\top \otimes e_1^\top, \dots, \mu(|\mathcal{S}|)^\top \otimes e_{|\mathcal{S}|}^\top) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \tag{3}$$

where $\otimes$ denotes the Kronecker product and e_s denotes the standard basis vectors corresponding to state s. Deterministic policies are represented in the same way using one-hot action vectors, and Θ denotes the finite set of deterministic stationary policies. For a deterministic stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, the team-average Q-function can be denoted as

$$Q^\pi(s, a) := \mathbb{E}\left[\sum_{k=0}^\infty \gamma^k \bar{r}(s_k, a_k, s_{k+1}) \middle| s_0 = s, a_0 = a, a_k = \pi(s_k) \quad \forall k \geq 1\right]$$

Algorithm 1 Exact-Consensus Q-learning (ECQ)

1: Choose a constant step-size $\alpha \in (0,1)$ and a mixing gain $\kappa > 0$.
2: Initialize $W_0^i, Q_0^i, Z_0^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for all agents $i = 1, \ldots, N$.
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: Sample (s_k, a_k, s'_k). Each agent i observes its local reward $r_{k+1}^i = r^i(s_k, a_k, s'_k)$.
5: **for** each agent i **do**
6: Update W_k^i, Q_k^i, and Z_k^i at (s_k, a_k) using equations 5–7, respectively.
7: Leave all other coordinates unchanged.
8: **end for**
9: **end for**

with the optimal Q-function $Q^*(s,a) := \max_{\pi \in \Theta} Q^\pi(s,a)$. For a Q-function Q, define its associated value function by

$$V_Q(s) := \max_a Q(s,a), \quad V_Q := \mathrm{col}(V_Q(1), \ldots, V_Q(|\mathcal{S}|)). \tag{4}$$

Using this, the Bellman optimality operator for the team-average reward can be written as $F(Q) := R + \gamma P V_Q$ with $Q^* = F(Q^*)$, where Q^* is the unique fixed point of F, and thus the unique optimal Q-function.

4 Exact-Consensus Q-learning

4.1 Exact-Consensus Q-learning Algorithm

In this paper, we focus on learning the optimal Q-function under the team-average reward. To address the steady-state error inherent in conventional direct consensus, we introduce ECQ (Algorithm 1), which eliminates the steady-state consensus bias caused by heterogeneous local rewards while maintaining a constant step-size. The key idea is to decouple local learning from network-wide agreement. To this end, each agent i maintains three variables: W^i for local learning, Q^i for network agreement, and Z^i for tracking accumulated disagreement.

The learning variable W^i incorporates the local reward through

$$W_{k+1}^i(s_k,a_k) = W_k^i(s_k,a_k) + \alpha \left(\underbrace{r_{k+1}^i + \gamma \max_{a' \in \mathcal{A}} Q_k^i(s'_k, a') - W_k^i(s_k,a_k)}_{\text{temporal-difference error}} - \underbrace{\sum_{j=1}^N L_{ij} W_k^j(s_k,a_k)}_{\text{consensus correction}} \right). \tag{5}$$

A key feature of the temporal-difference error is that its target is computed using Q_k^i instead of W_k^i. Since W^i is updated from local rewards, using it as the target would continuously reintroduce local reward variances into the learning recursion. Instead, Q^i provides a network-coordinated target, decoupling the local Bellman update from the agent-shared value. Furthermore, the consensus correction reduces differences in W^i across neighboring agents.

On the other hand, the variables Q^i and Z^i coordinate the agents. The corresponding updates are

$$Q_{k+1}^i(s_k,a_k) = Q_k^i(s_k,a_k) + \alpha\kappa \left(\underbrace{W_k^i(s_k,a_k) - Q_k^i(s_k,a_k)}_{\text{tracking correction}} - \underbrace{\sum_{j=1}^N L_{ij} Q_k^j(s_k,a_k)}_{\text{consensus correction}} - \underbrace{\sum_{j=1}^N L_{ij} Z_k^j(s_k,a_k)}_{\text{auxiliary correction}} \right), \tag{6}$$

$$Z_{k+1}^i(s_k,a_k) = Z_k^i(s_k,a_k) + \underbrace{\alpha\kappa \sum_{j=1}^N L_{ij} Q_k^j(s_k,a_k)}_{\text{disagreement}}. \tag{7}$$

In equation 6, the tracking correction drives Q^i toward the locally updated W^i, whereas the consensus correction reduces disagreement among neighboring agents. The auxiliary dynamic-consensus state Z^i accumulates this disagreement, continuously adjusting Q^i via its auxiliary correction until exact agreement is achieved. Consequently, this mechanism removes the consensus bias without requiring a diminishing step-size. The gain $\kappa > 0$ scales the consensus updates and is used later to establish convergence.

A two-variable variant, R-ECQ, avoids an explicit Z-state but requires a specific initialization condition. It is a separately designed recursion, not a trajectory-equivalent elimination of Z from ECQ. This simplified scheme and its convergence analysis are provided in Appendix A.12.

4.2 Equilibrium of the Mean Dynamics

For the convergence analysis, we reformulate Algorithm 1 in three steps: (i) deriving its stacked recursion, (ii) characterizing the equilibrium of the mean dynamics, and (iii) expressing the resulting error dynamics as a switching system.

We first express the algorithm in stacked form. Based on equations 1–7, we define the stacked variables as

$$\begin{aligned} \mathbf{R} &:= \operatorname{col}\left(R^1, \ldots, R^N\right),\ \mathbf{\Pi^{\pi}} := \operatorname{diag}\left(\Pi^{\pi^1}, \ldots, \Pi^{\pi^N}\right),\ \mathbf{V}_{\mathbf{Q}_k} := \operatorname{col}\left(V_{Q_k^1}, \ldots, V_{Q_k^N}\right), \\ \mathbf{W}_k &:= \operatorname{col}\left(W_k^1, \ldots, W_k^N\right),\ \mathbf{Q}_k := \operatorname{col}\left(Q_k^1, \ldots, Q_k^N\right),\ \mathbf{Z}_k := \operatorname{col}\left(Z_k^1, \ldots, Z_k^N\right), \end{aligned} \tag{8}$$

and the lifted matrices as

$$\begin{aligned} \mathbf{D} &:= I_N \otimes D \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}| \times N|\mathcal{S}||\mathcal{A}|}, \qquad \mathbf{P} := I_N \otimes P \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}| \times N|\mathcal{S}|}, \\ \mathbf{L} &:= L \otimes I_{|\mathcal{S}||\mathcal{A}|} \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}| \times N|\mathcal{S}||\mathcal{A}|}, \qquad \mathbf{S} := \mathbf{DL} = L \otimes D \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}| \times N|\mathcal{S}||\mathcal{A}|}, \end{aligned} \tag{9}$$

where I_n denotes the $n \times n$ identity matrix. Using these notations, Algorithm 1 can be written in the following stacked form:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \alpha\left[\mathbf{DR} + \gamma \mathbf{DPV}_{\mathbf{Q}_k} - \mathbf{DW}_k - \mathbf{SW}_k\right] + \alpha\omega_{k+1}, \tag{10}$$

$$\mathbf{Q}_{k+1} = \mathbf{Q}_k + \alpha\kappa\left[\mathbf{D}(\mathbf{W}_k - \mathbf{Q}_k) - \mathbf{SQ}_k - \mathbf{SZ}_k\right] + \alpha\kappa\nu_{k+1}, \tag{11}$$

$$\mathbf{Z}_{k+1} = \mathbf{Z}_k + \alpha\kappa\mathbf{SQ}_k + \alpha\kappa\zeta_{k+1}, \tag{12}$$

where each update is decomposed into its conditional mean and a zero-mean sampling noise. The definitions of the noise terms ω_{k+1}, ν_{k+1}, and ζ_{k+1}, along with the detailed derivation, are provided in Appendix A.1.

We then analyze the equilibrium of the mean dynamics obtained by omitting the noise terms in equations 10–12. The following proposition shows that the Q-component of every equilibrium coincides exactly with the centralized team-optimal Q-function, with the proof deferred to Appendix A.2.

Proposition 4.1. *Let Q^* be the optimal Q-function associated with the team-average reward, and let V^* be the corresponding optimal value function defined by $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$. Then, any equilibrium $(\mathbf{W}_\infty, \mathbf{Q}_\infty, \mathbf{Z}_\infty)$ of the mean dynamics satisfies*

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes Q^*, \quad (\mathbf{D}+\mathbf{S})\mathbf{W}_\infty = \mathbf{DR}+\gamma\mathbf{DP}(\mathbf{1}_N \otimes V^*), \quad \mathbf{SZ}_\infty = \mathbf{D}(\mathbf{W}_\infty - \mathbf{1}_N \otimes Q^*). \tag{13}$$

Furthermore, the Z-equilibrium is not unique: $\mathbf{Z}_\infty + \mathbf{1}_N \otimes v$ satisfies the third relation for any $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

4.3 Switching System Formulation

Based on the equilibrium derived in Proposition 4.1, we now transform the recursion into error coordinates and formulate the switching system defined below.

Definition 4.1 (Liberzon (2003))**.** *Given a finite family of matrices $\{A_\sigma\}_{\sigma \in \mathcal{M}}$ and vectors $\{b_\sigma\}_{\sigma \in \mathcal{M}}$ together with a switching signal $\sigma_k \in \mathcal{M}$, the system $X_{k+1} = A_{\sigma_k} X_k + b_{\sigma_k}$ is called a switching system, where $X_k \in \mathbb{R}^n$ is the state at iteration $k \in \{0, 1, \ldots\}$. In our case, σ_k is determined by the greedy policy.*

To reformulate the convergence to the equilibrium as the convergence of the error variables to zero, we define

$$u_k := \mathbf{W}_k - \mathbf{W}_\infty, \qquad x_k := \mathbf{Q}_k - \mathbf{Q}_\infty, \qquad y_k := \mathbf{Z}_k - \mathbf{Z}_\infty. \tag{14}$$

Expressing equations 10–12 in terms of the error variables using equation 1 and equation 13 yields

$$\begin{aligned} u_{k+1} =& u_k + \alpha \left[\gamma \mathbf{D}\mathbf{P} \left(\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V^*\right) - (\mathbf{D}+\mathbf{S})u_k\right] + \alpha\omega_{k+1}, \\ x_{k+1} =& x_k + \alpha\kappa \left[\mathbf{D}(u_k - x_k) - \mathbf{S}x_k - \mathbf{S}y_k\right] + \alpha\kappa\nu_{k+1}, \\ y_{k+1} =& y_k + \alpha\kappa\mathbf{S}x_k + \alpha\kappa\zeta_{k+1}. \end{aligned} \tag{15}$$

To express the error dynamics as a switching system, we handle the nonlinear term in equation 15, namely the Bellman maximization in $\mathbf{V}_{\mathbf{Q}_k}$. By applying a stochastic-policy linearization, we can represent this term using a policy-dependent linear map. Additionally, recall from Proposition 4.1 that the Z-error contains a common component shared across all agents that is invariant under the dynamics, and the third relation in equation 13 does not uniquely determine $\mathbf{Z}_\infty$ along this component. We therefore fix $\mathbf{Z}_\infty$ to have the same network sum as the initialization Z_0,

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{Z}_0 - \mathbf{Z}_\infty) = 0, \tag{16}$$

and eliminate the resulting redundant component by projecting y_k onto the disagreement subspace. Let $U \in \mathbb{R}^{(N-1)\times N}$ be a matrix with orthonormal rows spanning $\{\mathbf{1}_N\}^\perp$, and set

$$\mathbf{U} := U \otimes I_{|\mathcal{S}||\mathcal{A}|}, \qquad z_k := \mathbf{U}y_k. \tag{17}$$

Then, the projected error dynamics admit the switching linear representation

$$\begin{gathered} X_k := \mathrm{col}(u_k, x_k, z_k), \qquad \eta_{k+1} := \mathrm{col}(\omega_{k+1}, \kappa\nu_{k+1}, \kappa\mathbf{U}\zeta_{k+1}), \\ X_{k+1} = A_{\boldsymbol{\mu}_k,\perp} X_k + \alpha\eta_{k+1},\ A_{\boldsymbol{\mu},\perp} := \begin{bmatrix} I - \alpha(\mathbf{D}+\mathbf{S}) & \alpha\gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}} & 0 \\ \alpha\kappa\mathbf{D} & I - \alpha\kappa(\mathbf{D}+\mathbf{S}) & -\alpha\kappa\mathbf{S}\mathbf{U}^\top \\ 0 & \alpha\kappa\mathbf{U}\mathbf{S} & I \end{bmatrix}, \end{gathered} \tag{18}$$

with the joint stochastic policy $\boldsymbol{\mu}_k$. The detailed derivation is provided in Appendix A.3. This switching-system representation serves as the foundation for the finite-time analysis in the next section.

5 Finite-Time Convergence Guarantees

This section establishes the main finite-time guarantees for Algorithm 1. To this end, we first verify that the switching system is contractive by analyzing its joint spectral radius (JSR). We then combine this contraction property with bounds on policy-induced perturbations and sampling noise to derive finite-time bounds on total error and consensus error.

5.1 Joint Spectral Radius Analysis of the Switching System

As a basis for our analysis, we first review the definition of JSR.

Definition 5.1 (Jungers (2009))**.** *For a bounded set of matrices $\mathcal{H} \subset \mathbb{R}^{m\times m}$, its joint spectral radius (JSR) is defined as*

$$\rho(\mathcal{H}) := \lim_{k\to\infty} \sup_{A_1,\ldots,A_k \in \mathcal{H}} \|A_k \cdots A_1\|^{1/k},$$

which is independent of the choice of matrix norm. If $\rho(\mathcal{H}) < 1$, the noise-free switching system is exponentially stable under arbitrary switching sequences.

To apply the concept of JSR to our system, we revisit the matrix structure of the system in equation 18. While the policy $\boldsymbol{\mu}$ in equation 18 is stochastic and varies continuously, its selection matrix can be represented as a convex combination of deterministic selection matrices:

$$\boldsymbol{\Pi}^{\boldsymbol{\mu}} \in \mathrm{co}\left\{\boldsymbol{\Pi}^{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \Theta^N\right\},$$

where $\mathrm{co}(\cdot)$ denotes the convex hull. Then, since $A_{\boldsymbol{\mu},\perp}$ depends linearly on $\boldsymbol{\Pi}^{\boldsymbol{\mu}}$, any such matrix lies within the convex hull of its deterministic counterparts. Accordingly, we define the finite mode set

$$\mathcal{A}_{\alpha,\perp} := \left\{A_{\boldsymbol{\pi},\perp} : \boldsymbol{\pi} \in \Theta^N\right\},$$

and establish the following lemma, which guarantees that system equation 18 always evolves within the convex hull of this finite set $\mathcal{A}_{\alpha,\perp}$. The detailed proof is provided in Appendix A.4.

Lemma 5.1. *For every joint stochastic policy $\boldsymbol{\mu}$, $A_{\boldsymbol{\mu},\perp} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$. Consequently, $\rho\left(\mathrm{co}(\mathcal{A}_{\alpha,\perp})\right) = \rho(\mathcal{A}_{\alpha,\perp})$.*

Lemma 5.1 shows that it suffices to analyze only this finite set $\mathcal{A}_{\alpha,\perp}$, as the stability properties of $\mathcal{A}_{\alpha,\perp}$ naturally extend to its entire convex hull. That is, convexification leaves the JSR unchanged. Therefore, the analysis of equation 18 reduces to a single scalar condition, $\rho(\mathcal{A}_{\alpha,\perp}) < 1$, over a finite matrix family. The following proposition establishes this condition for a sufficiently small step-size, with the full proof provided in Appendix A.5.

Proposition 5.1. *There exists a finite threshold $\kappa^* > 0$ such that, for every fixed $\kappa > \kappa^*$, there exists $\bar{\alpha}(\kappa) > 0$ such that the projected deterministic switching family $\mathcal{A}_{\alpha,\perp}$ satisfies $\rho(\mathcal{A}_{\alpha,\perp}) < 1$ for every $0 < \alpha < \bar{\alpha}(\kappa)$.*

While Proposition 5.1 ensures contraction for small step-sizes, establishing a finite-time bound requires an explicit quantitative dependence of the contraction parameters on α. To capture this dependence, we state the following assumption and rewrite equation 18 as follows.

Assumption 5.1. *All problem parameters $(P, R^1, \ldots, R^N, \gamma, \mathbf{D}, \mathbf{L}, \mathbf{U}, \kappa)$, initial conditions, and Z_∞ satisfying equation 16 remain fixed independently of the step-size α.*

Note that Assumption 5.1 is not restrictive: it simply ensures the underlying environment is invariant under step-size variation. Consequently, each matrix $A_{\boldsymbol{\mu},\perp}$ in equation 18 satisfies

$$A_{\boldsymbol{\mu},\perp}(\alpha) = I + \alpha H_{\boldsymbol{\mu},\perp}, \qquad H_{\boldsymbol{\mu},\perp} := \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) & \gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}} & 0 \\ \kappa\mathbf{D} & -\kappa(\mathbf{D}+\mathbf{S}) & -\kappa\mathbf{S}\mathbf{U}^\top \\ 0 & \kappa\mathbf{U}\mathbf{S} & 0 \end{bmatrix}, \tag{19}$$

where $H_{\boldsymbol{\mu},\perp}$ is independent of α. This affine structure allows interpolation across step-sizes: for any $0 < \alpha \le \bar{\alpha}$,

$$A_{\boldsymbol{\mu},\perp}(\alpha) = \left(1 - \frac{\alpha}{\bar{\alpha}}\right) I + \frac{\alpha}{\bar{\alpha}} A_{\boldsymbol{\mu},\perp}(\bar{\alpha}).$$

Thus, each small-step matrix is a convex combination of I and $A_{\boldsymbol{\mu},\perp}(\bar{\alpha})$. Consequently, certifying contraction at a single upper bound $\bar{\alpha}$ guarantees control over all smaller step-sizes. The following lemma formalizes this result with an explicit α-dependent contraction rate. Its proof is deferred to Appendix A.6.

Lemma 5.2. *Under Assumption 5.1, suppose there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$. Then, there exist constants $c > 0$ and $K < \infty$, independent of α, such that for every $0 < \alpha \le \bar{\alpha}$:*

(i) $\rho(\mathcal{A}_{\alpha,\perp}) \le 1 - c\alpha < 1$.

(ii) *every product of matrices drawn from $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ satisfies*

$$\sup_{M_0,\ldots,M_{\ell-1} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})} \|M_{\ell-1} \cdots M_0\|_2 \le K(1-c\alpha)^\ell, \qquad \ell \ge 0. \tag{20}$$

5.2 Finite-Time Convergence Analysis

To establish the main results, we state two key lemmas: The first bounds the distance between mode matrices corresponding to different policies, while the second provides the bounds for the sampling noise η_{k+1} in equation 18, controlled by α-independent constants. The proofs are provided in Appendices A.7 and A.8.

Lemma 5.3. *Let*

$$L_\Pi := \sup_{\boldsymbol{\mu},\boldsymbol{\nu}} \left\|\mathbf{D}\mathbf{P}(\boldsymbol{\Pi}^{\boldsymbol{\mu}} - \boldsymbol{\Pi}^{\boldsymbol{\nu}})\right\|_2, \tag{21}$$

where the supremum is taken over all joint stochastic policies μ and ν. Then, L_Π is finite, and for all μ and ν,

$$\|A_{\boldsymbol{\mu},\perp} - A_{\boldsymbol{\nu},\perp}\|_2 \le \alpha\gamma L_\Pi.$$

Lemma 5.4. *Under Assumption 5.1, let Z_∞ be chosen to satisfy equation 16. Then, there exist α-independent constants $\sigma_0, \sigma_1 < \infty$ such that for all $k \ge 0$,*

$$\mathbb{E}[\eta_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}\left[\|\eta_{k+1}\|_2^2 \mid \mathcal{F}_k\right] \le \sigma_0^2 + \sigma_1^2 \|X_k\|_2^2,$$

where $\mathcal{F}_k = \sigma(\mathbf{W}_0, \mathbf{Q}_0, \mathbf{Z}_0, \{(s_t, a_t, s_t', r_{t+1}^1, \ldots, r_{t+1}^N) : 0 \le t \le k-1\})$ is the natural filtration.

Using Lemmas 5.2–5.4, we now state the main results. The following proposition establishes a uniform mean-square error bound without assuming a priori boundedness of the iterates. The proof is provided in Appendix A.9.

Proposition 5.2 (Uniform mean-square error bound)**.** *Under Assumption 5.1, suppose there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$. Let c, K be as in Lemma 5.2, σ_0, σ_1 as in Lemma 5.4, and L_Π as in equation 21. Define*

$$C_0 := 1 + 4K^2 + 4\left(\frac{\gamma L_\Pi K^2}{c}\right)^2, C_1 := 4\left(\frac{K^2}{c} + \frac{\gamma^2 L_\Pi^2 K^4}{c^3}\right), \alpha_0 := \min\left\{\bar{\alpha}, \frac{1}{2C_1\sigma_1^2}\right\}, \quad (22)$$

with the convention $\alpha_0 = \bar{\alpha}$ if $\sigma_1 = 0$. Consider Algorithm 1 initialized with a deterministic condition and Z_∞ satisfying equation 16. Then, for every $0 < \alpha \le \alpha_0$ and $k \ge 0$,

$$\sup_{0\le t\le k} \mathbb{E}\left[\|X_t\|_2^2\right] \le 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2. \quad (23)$$

In particular,

$$\sup_{0\le t\le k} \mathbb{E}[\|Q_t - \mathbf{1}_N \otimes Q^*\|_2^2] \le 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2.$$

Moreover, the following theorem refines Proposition 5.2 to provide an explicit bound on $\mathbb{E}[\|X_k\|_2]$ for every iteration k, separating the error bound into a geometrically decaying transient and a steady-state component of order $\sqrt{\alpha}$. The proof is given in Appendix A.10.

Theorem 5.1 (Finite-time error bound)**.** *Under the hypotheses and notation of Proposition 5.2, set $\beta := 1 - c\alpha/2$ for $0 < \alpha \le \alpha_0$. Then, for all $k \ge 0$ with the convention $k\beta^{k-1} = 0$ at $k = 0$,*

$$\begin{aligned}\mathbb{E}\left[\|X_k\|_2\right] \le & K\beta^k\|X_0\|_2 + \alpha\gamma K^2 L_\Pi k\beta^{k-1}\|X_0\|_2 \\ & + \left(1 + \frac{2\gamma K L_\Pi}{c}\right) K\sqrt{\frac{2\alpha}{c}}\left(\sigma_0^2 + \sigma_1^2\left(2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2\right)\right)^{1/2}.\end{aligned} \quad (24)$$

Consequently, $\mathbb{E}[\|Q_k - \mathbf{1}_N \otimes Q^*\|_2] \le \mathbb{E}[\|X_k\|_2]$. For a fixed initial condition, the steady-state error term in equation 24 scales as $O(\sqrt{\alpha})$ as α vanishes.

Note that Theorem 5.1 bounds the total error, which accounts for both the optimality error and the consensus error. Hence, the $O(\sqrt{\alpha})$ residual applies to both components. The following corollary shows that the consensus component of ECQ converges to zero under a mild structural condition on the local rewards, with the proof provided in Appendix A.11.

Corollary 5.1 (Exact consensus under deterministic local rewards)**.** *Under Assumption 5.1 and deterministic reward heterogeneity, let Z_∞ be chosen to satisfy equation 16, and suppose there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\bar{\alpha}})) < 1$. Then, there exist constants $\tilde{C}_0 < \infty$, $\tilde{c} > 0$, and $\tilde{\alpha}_0 \in (0, \bar{\alpha}]$, such that for every $0 < \alpha \le \tilde{\alpha}_0$ and $k \ge 0$, the consensus error satisfies*

$$\mathbb{E}\left[\max_i \left\|Q_k^i - \bar{Q}_k\right\|_\infty\right] \le \sqrt{2\tilde{C}_0}\left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k/2}\left\|\tilde{X}_0\right\|_2$$

with $\bar{Q}_k = \frac{1}{N}\sum_i Q_k^i$ denoting the average Q-function across agents, which decays geometrically as $k \to \infty$ without any α-dependent residual error.

6 Experiments

To evaluate ECQ in Algorithm 1, we customize a 5×5 stochastic gridworld environment with 8 agents. Each agent has a distinct private goal and observes only its own reward, while all agents share the same transition dynamics. The agents communicate over a fixed ring graph and aim to reach consensus on the team-optimal Q-function Q^*. Through this environment, we present experimental results demonstrating (i) the consensus errors, (ii) the step-size dependence of the error floors, and (iii) responsiveness to sudden reward changes. To this end, we compare ECQ against IQL (Tan, 1993), $\mathcal{QD}$-learning (Kar et al., 2013), and Mixing Q-learning (Lim & Lee, 2025). Further details on the experimental setup and additional empirical results are provided in Appendix A.13.

Figure 1a reports the consensus error $\max_i\|Q_k^i - \bar{Q}_k\|_\infty$, which measures the disagreement among the agents' local value estimates. Notably, ECQ is the only method that drives this error down to

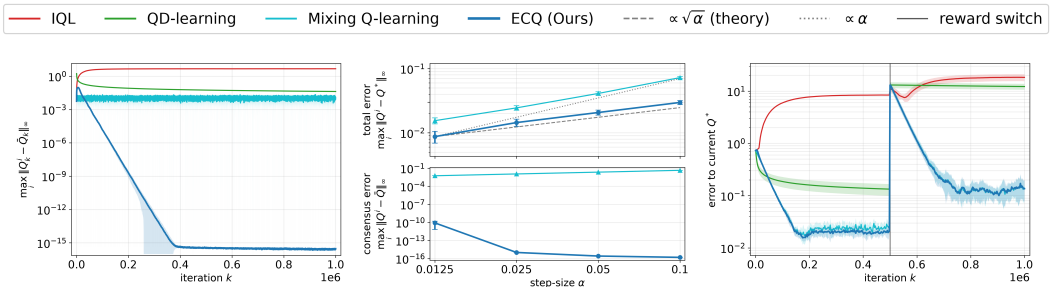


Figure 1: Experiments on gridworld: (a) Consensus error. (b) Step-size dependence of the error floors. (c) Re-tracking after a reward change.

10^{-16}, reaching machine-precision consensus. In contrast, mixing Q-learning settles onto a non-vanishing error floor without further improvement. Since IQL involves no communication, agents update their estimates independently, resulting in the largest consensus error. Within the reported budget, $\mathcal{QD}$-learning leaves a larger residual, a finite-budget effect rather than an established asymptotic floor: it converges almost surely to the common optimum (Kar et al., 2013), but its diminishing gains make the consensus error decay too slowly, staying orders of magnitude above ECQ throughout the run.

Figure 1b isolates the two error floors as a function of the constant step-size α, averaged over the final 10% of iterations. We compare ECQ exclusively against mixing Q-learning, as it is the sole baseline employing both a constant step-size and a consensus mechanism. In the top panel, we evaluate total error $\max_i \|Q_k^i - Q^*\|_\infty$. Fitting steady$(\alpha) = C\alpha^p$ to the experimental data for ECQ gives $\hat{p} = 0.590 \pm 0.100$, consistent with the $O(\sqrt{\alpha})$ upper bound in Theorem 5.1. Note that this worst-case bound neither implies a positive asymptotic error floor nor excludes smaller residuals in specific instances. In the bottom panel, we measure the consensus error $\max_i \|Q_k^i - \bar{Q}_k\|_\infty$. While Figure 1a shows how this error changes over time, this figure illustrates its steady-state floor across different step-sizes. Whereas mixing Q-learning is bounded by an error floor scaling as $O(\alpha)$ (Lim & Lee, 2025), ECQ consistently achieves machine-precision consensus regardless of α, which is aligned with the result of Corollary 5.1. Note that for $\alpha = 0.0125$, the higher consensus error of ECQ stems merely from an insufficient number of iterations relative to the small step size.

Figure 1c evaluates responsiveness under a mid-experiment reward switch, where the target shifts from Q_1^* to Q_2^* (see Appendix A.13). Once the switch happens, all algorithms experience a temporary spike in error. However, operating under constant step-sizes, ECQ and mixing Q-learning rapidly re-track the new optimum, reducing the error spikes within a similar number of iterations due to non-zero adaptation gains. In contrast, $\mathcal{QD}$-learning takes much longer to adapt because its decaying step-sizes have already become too small, leaving a large residual error. Without communication, IQL tracks only private optima and fails to adapt to the team-optimal target altogether. These results confirm that ECQ is highly robust to sudden changes in the reward function.

7 Conclusions

This paper introduced Exact-Consensus Q-learning (ECQ), a distributed Q-learning algorithm that removes the consensus bias under a constant step-size. ECQ decouples local learning from network agreement using three variables, aligning the mean dynamics equilibrium with the centralized optimum. We theoretically analyzed ECQ as a switching system, proving its contractivity in the JSR sense and deriving finite-time bounds: the total error bound consists of a decaying transient and an $O(\sqrt{\alpha})$ steady-state residual, whereas the consensus error vanishes geometrically without any residual under deterministic reward heterogeneity. Experiments confirmed that ECQ achieves machine-precision consensus under deterministic local rewards, aligns with our theoretical error bounds, and adapts robustly to reward changes. While this work focuses on tabular settings over fixed graphs, future research will explore time-varying topologies, function approximation, and larger-scale distributed control applications.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the development of the theoretical framework, assisting in the writing of proofs, and implementing the algorithms. We have not used generative AI tools for generating synthetic datasets or formulating the central hypotheses of this work, which originated with the authors. Other required-disclosure tasks (e.g., cleaning or reformatting datasets) are not applicable to this work. Additionally, we used generative AI tools for editing and creating experimental and plotting scripts, drafting and editing manuscripts to improve readability, summarizing related literature, and identifying relevant references. We have reviewed all AI-assisted work. All AI-suggested proof steps and mathematical claims were independently verified by the authors, and all AI-generated or AI-edited code was tested for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All modeling assumptions, such as the underlying networked MDP setting, constant step size conditions, initialization of Algorithm 2, invariant environment dynamics under step size variation, and deterministic reward heterogeneity, are explicitly stated in Section 3, Algorithms 1 and 2, Assumption 5.1, and Appendix A.11. Complete statements and self-contained proofs of all theoretical claims, including the corresponding analysis for ECQ and R-ECQ, are provided in Appendices A.1 through A.12. The proposed algorithms, ECQ and R-ECQ, are specified as pseudocode in Algorithms 1 and 2. Anonymized implementation code that reproduces all reported figures and tables is provided in the supplementary material as code.zip. The benchmark environments and data processing pipeline, including the number of random seeds, hyperparameter selection protocol, and baseline descriptions, are fully specified in Appendix A.13.

REFERENCES

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2009.07.008. URL https://www.sciencedirect.com/science/article/pii/S0005109809003549.

Lucas Cassano, Kun Yuan, and Ali H Sayed. Multiagent fully decentralized value function learning with linear convergence rates. *IEEE Transactions on Automatic Control*, 66(4):1497–1512, 2021.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998.

Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 1626–1635. PMLR, 2019.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

Randy A. Freeman, Peng Yang, and Kevin M. Lynch. Stability and convergence properties of dynamic average consensus estimators. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 338–343. IEEE, 2006.

Paulo Heredia, Hasan Ghadialy, and Shaoshuai Mou. Finite-sample analysis of distributed q-learning for multi-agent networks. In *2020 American Control Conference (ACC)*, pp. 3511–3516. IEEE, 2020.

Ali Jadbabaie, Jie Lin, and A Stephen Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on automatic control*, 48(6):988–1001, 2003.

Raphaël Jungers. *The joint spectral radius: theory and applications*, volume 385. Springer Science & Business Media, 2009.

Soummya Kar, José M. F. Moura, and H. Vincent Poor. Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

Donghwan Lee, Niao He, Parameswaran Kamalaruban, and Volkan Cevher. Optimization for reinforcement learning: From a single agent to cooperative agents. *IEEE Signal Processing Magazine*, 37(3):123–135, 2020.

Donghwan Lee, Do Wan Kim, and Jianghai Hu. Distributed off-policy temporal difference learning using primal-dual method. *IEEE Access*, 10:107077–107094, 2022.

Donghwan Lee, Jianghai Hu, and Niao He. A discrete-time switching system analysis of q-learning. *SIAM Journal on Control and Optimization*, 61(3):1861–1880, 2023.

Daniel Liberzon. *Switching in systems and control*, volume 190. Springer, 2003.

Han-Dong Lim and Donghwan Lee. A primal-dual perspective for distributed td-learning. *arXiv preprint arXiv:2310.00638*, 2023.

Han-Dong Lim and Donghwan Lee. A finite time analysis of distributed q-learning. *Reinforcement Learning Journal*, 6:165–200, 2025.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Reza Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on automatic control*, 51(3):401–420, 2006.

Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on automatic control*, 49(9):1520–1533, 2004.

Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

Walter Rudin. Principles of mathematical analysis. 2021.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Miloš S Stanković, Marko Beko, and Srdjan S Stanković. Distributed consensus-based multi-agent temporal-difference learning. *Automatica*, 151:110922, 2023.

Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4485–4495. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

Jean-Baptiste Hiriart Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms*. Springer-Verlag, 1993.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, and Jian Sun. Decentralized td tracking with linear function approximation and its finite-time analysis. *Advances in neural information processing systems*, 33:13762–13772, 2020.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Peng Yang, Randy A Freeman, and Kevin M Lynch. Multi-agent coordination by decentralized estimation and control. *IEEE Transactions on Automatic Control*, 53(11):2480–2496, 2008.

Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-time convergence rates of decentralized stochastic approximation with applications in multi-agent and multi-task learning. *IEEE Transactions on Automatic Control*, 68(5):2758–2773, 2022.

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.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940, 2021.

A Appendix

A.1 Decomposition of Stacked Updates and Noise Definitions

In this appendix, we detail the decomposition of update equations 10–12 into their conditional means and zero-mean sampling noises.

Let $e_k := e_{s_k,a_k} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the standard basis vector associated with the sampled state-action pair (s_k, a_k). We define the stacked sampled updates as $\hat{g}_{\cdot,k} := \mathrm{col}(\hat{g}^1_{\cdot,k}, \ldots, \hat{g}^N_{\cdot,k})$ with

$$\hat{g}^i_{W,k} := e_k \left[r^i_{k+1} + \gamma \max_{a' \in \mathcal{A}} Q^i_k(s'_k, a') - W^i_k(s_k, a_k) - \sum_j L_{ij} W^j_k(s_k, a_k) \right], \tag{25}$$

$$\hat{g}^i_{Q,k} := e_k \left[W^i_k(s_k, a_k) - Q^i_k(s_k, a_k) - \sum_j L_{ij} Q^j_k(s_k, a_k) - \sum_j L_{ij} Z^j_k(s_k, a_k) \right], \tag{26}$$

$$\hat{g}^i_{Z,k} := e_k \left[\sum_j L_{ij} Q^j_k(s_k, a_k) \right]. \tag{27}$$

Let $\mathcal{F}_k$ denote the natural filtration given by

$$\mathcal{F}_k := \sigma\left(\mathbf{W}_0, \mathbf{Q}_0, \mathbf{Z}_0, \left\{\left(s_t, a_t, s'_t, r^1_{t+1}, \ldots, r^N_{t+1}\right) : 0 \le t \le k-1\right\}\right), \qquad k \ge 0.$$

Conditioned on $\mathcal{F}_k$, the randomness in equations 25–27 originates from the current state-action pair, the next state transition, and the local rewards. Since $\mathbb{E}[e_k e_k^\top] = D$, taking the conditional expectation of the stacked sampled updates in equations 25–27 yields

$$\begin{aligned}
\mathbb{E}\left[\hat{g}_{W,k} \mid \mathcal{F}_k\right] &= \mathbf{DR} + \gamma \mathbf{DPV}_{\mathbf{Q}_k} - \mathbf{DW}_k - \mathbf{SW}_k, \\
\mathbb{E}\left[\hat{g}_{Q,k} \mid \mathcal{F}_k\right] &= \mathbf{D}\left(\mathbf{W}_k - \mathbf{Q}_k\right) - \mathbf{SQ}_k - \mathbf{SZ}_k, \\
\mathbb{E}\left[\hat{g}_{Z,k} \mid \mathcal{F}_k\right] &= \mathbf{SQ}_k.
\end{aligned}$$

These expected values correspond exactly to the conditional mean dynamics given in equations 10–12, with the mixing gain κ applied in the Q and Z-directions. Furthermore, defining the sampling noise as

$$\begin{aligned}
\omega_{k+1} &:= \hat{g}_{W,k} - \mathbb{E}\left[\hat{g}_{W,k} \mid \mathcal{F}_k\right], \\
\nu_{k+1} &:= \hat{g}_{Q,k} - \mathbb{E}\left[\hat{g}_{Q,k} \mid \mathcal{F}_k\right], \\
\zeta_{k+1} &:= \hat{g}_{Z,k} - \mathbb{E}\left[\hat{g}_{Z,k} \mid \mathcal{F}_k\right],
\end{aligned} \tag{28}$$

it readily follows that all three noise terms are conditionally zero-mean:

$$\mathbb{E}[\omega_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\nu_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0.$$

Therefore, equations 10–12 consist of the conditional mean terms together with zero-mean sampling noises. Omitting these noise terms yields the conditional mean dynamics used in the equilibrium analysis in Proposition 4.1.

A.2 Proof of Proposition 4.1

Proposition 4.1 Restated. Let Q^* be the optimal Q-function associated with the team-average reward, and let V^* be the corresponding optimal value function defined by $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$. Then, any equilibrium $(\mathbf{W}_\infty, \mathbf{Q}_\infty, \mathbf{Z}_\infty)$ of the mean dynamics satisfies

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes Q^*, \quad (\mathbf{D} + \mathbf{S})\mathbf{W}_\infty = \mathbf{D}\mathbf{R} + \gamma \mathbf{D}\mathbf{P}(\mathbf{1}_N \otimes V^*), \quad \mathbf{S}\mathbf{Z}_\infty = \mathbf{D}(\mathbf{W}_\infty - \mathbf{1}_N \otimes Q^*).$$

Furthermore, the Z-equilibrium is not unique: $\mathbf{Z}_\infty + \mathbf{1}_N \otimes v$ satisfies the third relation for any $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Proof. For $\mathbf{Q}_\infty$, we first consider the conditional mean dynamics of equation 12 with the noise term deleted. At equilibrium, this dynamics yield $\mathbf{Z}_\infty = \mathbf{Z}_\infty + \alpha\kappa\mathbf{S}\mathbf{Q}_\infty$, which directly implies

$$\mathbf{S}\mathbf{Q}_\infty = 0. \tag{29}$$

Given $D \succ 0$ in equation 2 and $\mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}$ in equation 1, equation 29 together with equation 9 yields

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes q \tag{30}$$

for some $q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Moreover, from the conditional mean dynamics of equation 11, we get

$$\mathbf{D}(\mathbf{W}_\infty - \mathbf{Q}_\infty) - \mathbf{S}\mathbf{Q}_\infty - \mathbf{S}\mathbf{Z}_\infty = 0 \tag{31}$$

at equilibrium. Since $\mathbf{S}\mathbf{Q}_\infty = 0$ from equation 29, multiplying both sides of equation 31 from the left by $\frac{1}{N}(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})$ gives

$$\begin{aligned}
&\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{D}(\mathbf{W}_\infty - \mathbf{Q}_\infty) - \frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{S}\mathbf{Z}_\infty \\
=&\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{D}(\mathbf{W}_\infty - \mathbf{Q}_\infty) \\
=&\frac{1}{N}\left(\mathbf{1}_N^\top \otimes D\right)(\mathbf{W}_\infty - \mathbf{Q}_\infty) \\
=&\frac{1}{N}(1 \otimes D)\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{W}_\infty - \mathbf{Q}_\infty) \\
=&D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{W}_\infty - \mathbf{Q}_\infty) \\
=&D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{W}_\infty - \mathbf{1}_N \otimes q) \\
=&D\left[\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{W}_\infty - \frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{1}_N \otimes q)\right] \\
=&D\left[\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{W}_\infty - \frac{1}{N}\left(\mathbf{1}_N^\top \mathbf{1}_N \otimes q\right)\right] \\
=&D\left[\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{W}_\infty - q\right],
\end{aligned}$$

where the first equality uses $\mathbf{1}_N^\top L = 0$ along with equation 9, the second, third, and seventh equalities utilize the mixed-product property of the Kronecker product with equation 9, and the fifth equality follows from equation 30. This leads to

$$\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{W}_\infty = q. \tag{32}$$

From the conditional mean dynamics of equation 10, we obtain

$$\mathbf{DR} + \gamma \mathbf{DPV}_{\mathbf{Q}_\infty} - \mathbf{DW}_\infty - \mathbf{SW}_\infty = 0 \tag{33}$$

at equilibrium. Multiplying both sides of the above equation from the left by $\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)$ yields

$$\begin{aligned}
&\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\left(\mathbf{DR} + \gamma \mathbf{DPV}_{\mathbf{Q}_\infty} - \mathbf{DW}_\infty - \mathbf{SW}_\infty\right) \\
=&\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{D}\left(\mathbf{R} + \gamma \mathbf{PV}_{\mathbf{Q}_\infty} - \mathbf{W}_\infty\right) \\
=&D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\left(\mathbf{R} + \gamma \mathbf{PV}_{\mathbf{Q}_\infty} - \mathbf{W}_\infty\right) \\
=&DR + D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\gamma \mathbf{PV}_{\mathbf{Q}_\infty} - D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{W}_\infty \\
=&DR + D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\gamma \mathbf{PV}_{\mathbf{Q}_\infty} - Dq \\
=&DR + \gamma D\frac{1}{N}\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\left(I_N \otimes P\right)\left(\mathbf{1}_N \otimes V_q\right) - Dq \\
=&DR + \gamma DPV_q - Dq \\
=&0,
\end{aligned}$$

where the first equality uses $\mathbf{1}_N^\top L = 0$ along with equation 9, the second and sixth equalities utilize the mixed-product property of the Kronecker product together with equation 9, and the third equality follows from the definition equation 2 and equation 8. Furthermore, the fourth equality holds by equation 32, and the fifth equality is due to equation 8, equation 9, and equation 30. Since $V_q(s) = \max_{a\in\mathcal{A}} q(s,a)$ from equation 4, the relation $q = R + \gamma P V_q$ represents the Bellman optimality equation associated with the team-average reward. Because the corresponding Bellman operator is a γ-contraction mapping with respect to $\|\cdot\|_\infty$, it admits a unique fixed point. As Q^* is defined as this unique fixed point, it follows that $q = Q^*$, and

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes Q^*, \tag{34}$$

yielding the first relation in equation 13.

For the remaining claims, substituting equation 34 into the conditional mean dynamics of equations 10–11 at equilibrium gives the remaining relations in equation 13. Specifically, combining equation 33 with equation 34 yields the second relation in equation 13. Moreover, from equation 31 together with equation 29 and equation 34, we have

$$\mathbf{SZ}_\infty = \mathbf{D}(\mathbf{W}_\infty - \mathbf{1}_N \otimes Q^*).$$

For any $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we obtain from the above equality that

$$\mathbf{S}(\mathbf{Z}_\infty + \mathbf{1}_N \otimes v) = \mathbf{SZ}_\infty + \mathbf{S}(\mathbf{1}_N \otimes v) = \mathbf{SZ}_\infty = \mathbf{D}(\mathbf{W}_\infty - \mathbf{1}_N \otimes Q^*),$$

where the second equality follows from $L\mathbf{1}_N = 0$ along with equation 9. Hence, $\mathbf{Z}_\infty$ is determined up to the addition of elements in $\mathrm{null}(\mathbf{S}) = \{\mathbf{1}_N \otimes v\}$, a subspace of dimension $|\mathcal{S}||\mathcal{A}|$. Note that $\mathbf{W}_\infty$ is uniquely determined by the second relation in equation 13, since the matrix $\mathbf{D} + \mathbf{S} = (I_N \otimes D) + (L \otimes D) = (I_N + L) \otimes D$ is nonsingular. □

A.3 Projected Switching-System Representation of the Error Dynamics

In this appendix, we provide a detailed derivation of the projected switching-system representation of the error dynamics given in equation 18. We first treat the nonlinear term in equation 15 using a policy-dependent linear map, choose an appropriate Z-equilibrium consistent with the initialization, and finally eliminate the redundant component of the Z-error.

Policy-Dependent Linearization of the Bellman Maximization Operator. First, we reformulate the nonlinear term, specifically the Bellman maximization difference, into an exact policy-dependent linear mapping:

Lemma A.1. *Let Q^* be the optimal Q-function associated with the team-average reward, and let V^* be the corresponding optimal value function. Then, for every $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, there exists a stochastic policy $\mu_Q : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, which can be chosen as a measurable function of Q, such that $V_Q - V^* = \Pi^{\mu_Q}(Q - Q^*)$.*

Proof. This argument follows a standard linearization of the greedy value map used in approximate dynamic programming (see, e.g., (Munos, 2007; Farahmand et al., 2010)).

Fix a state $s \in \mathcal{S}$ and denote $e = Q - Q^*$. For any $a \in \mathcal{A}$, it holds that

$$\min_{a' \in \mathcal{A}} e(s, a') \le e(s, a) \le \max_{a' \in \mathcal{A}} e(s, a'). \tag{35}$$

Since

$$V_Q(s) = \max_{a \in \mathcal{A}} \{Q^*(s, a) + e(s, a)\}, \qquad V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a),$$

adding $Q^*(s, a)$ to equation 35 and taking the maximum over $a \in \mathcal{A}$ yields

$$V^*(s) + \min_{a \in \mathcal{A}} e(s, a) \le V_Q(s) \le V^*(s) + \max_{a \in \mathcal{A}} e(s, a),$$

which implies

$$\min_{a \in \mathcal{A}} e(s, a) \le V_Q(s) - V^*(s) \le \max_{a \in \mathcal{A}} e(s, a).$$

Consequently, $V_Q(s) - V^*(s)$ belongs to the convex hull of $\{e(s, a) : a \in \mathcal{A}\}$. Therefore, there exists $\mu_Q(\cdot \mid s) \in \Delta_{|\mathcal{A}|}$ satisfying

$$V_Q(s) - V^*(s) = \sum_{a \in \mathcal{A}} \mu_Q(a \mid s) e(s, a).$$

Choosing this convex combination for each state s defines a stochastic policy μ_Q, which can be selected measurably with respect to Q using a fixed tie-breaking rule. Since $e = Q - Q^*$, we obtain

$$V_Q - V^* = \Pi^{\mu_Q}(Q - Q^*),$$

as desired. □

By applying Lemma A.1 agent-wise to the stacked policy $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N)$ alongside equation 8 and equation 14, we obtain a joint stochastic policy $\boldsymbol{\mu}_k$ such that $\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V^* = \boldsymbol{\Pi}^{\boldsymbol{\mu}_k} x_k$. Then, the error dynamics equation 15 can be written in the exact switching linear representation:

$$\xi_k := \mathrm{col}(u_k, x_k, y_k), \qquad \varepsilon_{k+1} := \mathrm{col}(\omega_{k+1}, \kappa\nu_{k+1}, \kappa\zeta_{k+1}),$$
$$\xi_{k+1} = A_{\boldsymbol{\mu}_k} \xi_k + \alpha \varepsilon_{k+1}, \quad A_{\boldsymbol{\mu}} := \begin{bmatrix} I - \alpha(\mathbf{D} + \mathbf{S}) & \alpha\gamma \mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}} & 0 \\ \alpha\kappa\mathbf{D} & I - \alpha\kappa(\mathbf{D} + \mathbf{S}) & -\alpha\kappa\mathbf{S} \\ 0 & \alpha\kappa\mathbf{S} & I \end{bmatrix}.$$

However, the family $\{A_{\boldsymbol{\mu}}\}$ cannot be strictly contractive on the entire ξ-space because the Z-error contains a non-decaying component. To see this, consider $c_v := \mathrm{col}(0, 0, \mathbf{1}_N \otimes v)$, which represents a common Z-error component shared across all agents. Since the mixed-product property of the Kronecker product yields

$$\mathbf{S}(\mathbf{1}_N \otimes v) = (L\mathbf{1}_N) \otimes Dv = 0, \tag{36}$$

we have

$$A_{\boldsymbol{\mu}} c_v = c_v$$

for every $\boldsymbol{\mu}$. Consequently, this common component remains invariant under the error dynamics. Moreover, it does not affect the (u, x)-updates, because y enters the x-equation only through $-\alpha\kappa\mathbf{S}y$ in equation 15, which evaluates to zero for c_v by equation 36. Hence, this direction is neutral: it neither decays nor affects other error components. Therefore, it must be eliminated before establishing a contraction estimate for the error dynamics.

Z-Equilibrium Selection. To eliminate this neutral direction without loss of information, we select an appropriate representative of the Z-equilibrium. In Proposition 4.1, the third relation in equation 13 does not uniquely determine $\mathbf{Z}_\infty$. Thus, a specific representative must be selected before defining the error variable $y_k = \mathbf{Z}_k - \mathbf{Z}_\infty$.

The appropriate representative is determined by a conservation property of the mean Z-dynamics. Specifically, applying $(\mathbf{1}_N^\top \otimes I)$ to equation 12 after dropping the noise term yields

$$\left(\mathbf{1}_N^\top \otimes I\right) \mathbf{Z}_{k+1} = \left(\mathbf{1}_N^\top \otimes I\right) \mathbf{Z}_k + \alpha\kappa \left(\mathbf{1}_N^\top \otimes I\right) \mathbf{S}\mathbf{Q}_k = \left(\mathbf{1}_N^\top \otimes I\right) \mathbf{Z}_k,$$

where the last equality follows from $(\mathbf{1}_N^\top \otimes I)\mathbf{S} = 0$ along with equation 9. Hence, the network sum $(\mathbf{1}_N^\top \otimes I)\mathbf{Z}_k$ remains invariant under the mean dynamics. We therefore choose an equilibrium $\mathbf{Z}_\infty$ that matches the network sum of the initialization:

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) (\mathbf{Z}_0 - \mathbf{Z}_\infty) = 0.$$

Note that no restriction is imposed on the initial values W_0^i, Q_0^i, or Z_0^i themselves. Together with the conservation law, this choice guarantees that the Z-error maintains a zero network sum for all k. Since the components of y_k across all agents sum to zero, y_k contains no common component shared among agents. Consequently, we can remove this common direction and represent y_k using only the remaining $N-1$ directions without loss of information.

Eliminating the Redundant Component of Z-Error. We now eliminate the neutral consensus component of the Z-error. Let $U \in \mathbb{R}^{(N-1)\times N}$ be a matrix with orthonormal rows spanning $\{\mathbf{1}_N\}^\perp$, and set

$$\mathbf{U} := U \otimes I_{|\mathcal{S}||\mathcal{A}|}, \qquad z_k := \mathbf{U} y_k.$$

The following lemma confirms that no information is lost during this reduction.

Lemma A.2. *Under equation 16, $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}) y_k = 0$ holds for all $k \geq 0$. Moreover, $\mathbf{U}^\top\mathbf{U}$ is the orthogonal projector onto $\{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, so that $y_k = \mathbf{U}^\top z_k$.*

Proof. To prove the first claim, we first multiply $\mathbf{S}$ on the left by $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})$ to obtain

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \mathbf{S} = \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) (L \otimes D) = \left(\mathbf{1}_N^\top L\right) \otimes D = 0, \tag{37}$$

where the first equality follows from equation 9, the second equality uses the mixed-product property of the Kronecker product, and the last equality holds because $\mathbf{1}_N^\top L = 0$. Moreover, summing $\hat{g}_{Z,k}^i$ in equation 27 over all $i \in \{1, \ldots, N\}$ gives

$$\sum_{i=1}^N \hat{g}_{Z,k}^i = e_k \left[\sum_{j=1}^N \left(\sum_{i=1}^N L_{ij}\right) Q_k^j(s_k, a_k)\right] = 0,$$

since each column sum of L is zero. Hence, $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})\hat{g}_{Z,k} = \sum_{i=1}^N \hat{g}_{Z,k}^i = 0$. By the definition of ζ_{k+1} in equation 28, the linearity of conditional expectation gives

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \zeta_{k+1} = \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \hat{g}_{Z,k} - \mathbb{E}\left[\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \hat{g}_{Z,k} \mid \mathcal{F}_k\right] = 0. \tag{38}$$

We now proceed by induction. For $k = 0$, equation 16 states that

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) y_0 = \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) (\mathbf{Z}_0 - \mathbf{Z}_\infty) = 0.$$

Suppose that $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}) y_k = 0$ for some $k \geq 0$. Multiplying the last equation of equation 15 on the left by $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})$ and applying equation 37 and equation 38, we obtain

$$\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) y_{k+1} = \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) y_k + \alpha\kappa \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \mathbf{S} x_k + \alpha\kappa \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) \zeta_{k+1} = 0.$$

By induction, the claim holds for all $k \geq 0$. Equivalently, y_k lies in the disagreement subspace

$$\mathcal{Y} := \{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \tag{39}$$

for every k.

We now prove the second claim. Since y_k is confined to $\mathcal{Y}$ in equation 39 by the first claim, which has dimension $(N-1)|\mathcal{S}||\mathcal{A}|$, it suffices to track the reduced coordinates $z_k = \mathbf{U}y_k$. The second claim shows that y_k can be exactly recovered from z_k.

Since U has orthonormal rows spanning $\{\mathbf{1}_N\}^\perp$ by construction,

$$UU^\top = I_{N-1} \quad \text{and} \quad U\mathbf{1}_N = 0. \tag{40}$$

Therefore, the normalized vector $\frac{1}{\sqrt{N}}\mathbf{1}_N^\top$ is orthogonal to every row of U, and adjoining it yields an orthogonal matrix

$$\widetilde{U} := \left[U^\top \frac{1}{\sqrt{N}}\mathbf{1}_N\right]^\top \in \mathbb{R}^{N\times N},$$

whose rows and columns are orthonormal. Since $\widetilde{U}^\top\widetilde{U} = U^\top U + \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top = I_N$, we obtain

$$U^\top U = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top, \tag{41}$$

which is the orthogonal projector onto $\{\mathbf{1}_N\}^\perp$. Specifically, it maps $\mathbf{1}_N$ to zero: $U^\top U\mathbf{1}_N = \mathbf{1}_N - \frac{1}{N}\mathbf{1}_N(\mathbf{1}_N^\top\mathbf{1}_N) = \mathbf{1}_N - \mathbf{1}_N = 0$, and leaves every $w \in \{\mathbf{1}_N\}^\perp$ invariant: $U^\top Uw = w - \frac{1}{N}\mathbf{1}_N(\mathbf{1}_N^\top w) = w$. Then, taking the Kronecker product with $I_{|\mathcal{S}||\mathcal{A}|}$, we get

$$\mathbf{U}^\top\mathbf{U} = \left(U^\top U\right) \otimes I_{|\mathcal{S}||\mathcal{A}|} = \left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right) \otimes I_{|\mathcal{S}||\mathcal{A}|}, \tag{42}$$

which likewise maps $\mathbf{1}_N \otimes v$ to zero for every $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\left(\mathbf{U}^\top\mathbf{U}\right)(\mathbf{1}_N \otimes v) = \left(\left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)\mathbf{1}_N\right) \otimes (I_{|\mathcal{S}||\mathcal{A}|}v) = 0, \tag{43}$$

and leaves every element of $\mathcal{Y}$ in equation 39 invariant: any $w \otimes v$ with $w \in \{\mathbf{1}_N\}^\perp$ satisfies $(\mathbf{U}^\top\mathbf{U})(w \otimes v) = ((I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top)w) \otimes (I_{|\mathcal{S}||\mathcal{A}|}v) = w \otimes v$. Hence, $\mathbf{U}^\top\mathbf{U}$ is the orthogonal projector onto $\{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Since $y_k \in \mathcal{Y}$ by the first claim, the projector acts as the identity on y_k, yielding

$$\mathbf{U}^\top z_k = \mathbf{U}^\top\mathbf{U}y_k = y_k$$

with equation 17. □

Hence, by Lemma A.2, y_k always lies in $\{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and z_k provides an equivalent lower-dimensional representation of y_k. In particular, the mapping $y_k \mapsto z_k$ is invertible on this subspace through $y_k = \mathbf{U}^\top z_k$. Therefore, passing to $X_k := \mathrm{col}(u_k, x_k, z_k) \in \mathbb{R}^{(3N-1)|\mathcal{S}||\mathcal{A}|}$ constitutes an exact change of coordinates rather than an approximation:

$$X_k := \mathrm{col}(u_k, x_k, z_k), \qquad \eta_{k+1} := \mathrm{col}(\omega_{k+1}, \kappa\nu_{k+1}, \kappa\mathbf{U}\zeta_{k+1}),$$

$$X_{k+1} = A_{\boldsymbol{\mu},\perp}X_k + \alpha\eta_{k+1}, \quad A_{\boldsymbol{\mu},\perp} := \begin{bmatrix} I - \alpha(\mathbf{D}+\mathbf{S}) & \alpha\gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}} & 0 \\ \alpha\kappa\mathbf{D} & I - \alpha\kappa(\mathbf{D}+\mathbf{S}) & -\alpha\kappa\mathbf{S}\mathbf{U}^\top \\ 0 & \alpha\kappa\mathbf{U}\mathbf{S} & I \end{bmatrix}.$$

Since $y_k = \mathbf{U}^\top z_k$, the convergence of X_k to the origin is equivalent to the convergence of the original error variables (u_k, x_k, y_k) to the origin.

A.4 Proof of Lemma 5.1

Lemma 5.1 Restated. For every joint stochastic policy $\boldsymbol{\mu}$, $A_{\boldsymbol{\mu},\perp} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$. Consequently, $\rho\left(\mathrm{co}(\mathcal{A}_{\alpha,\perp})\right) = \rho(\mathcal{A}_{\alpha,\perp})$.

Proof. For each agent i, the matrix representation Π^{μ^i} of a stochastic policy μ^i can be written as a convex combination of deterministic policy matrices. Extending this to N agents, the joint policy matrix $\boldsymbol{\Pi}^{\boldsymbol{\mu}}$ is likewise a convex combination of deterministic joint policy matrices $\boldsymbol{\Pi}^{\boldsymbol{\pi}}$ for $\boldsymbol{\pi} \in \Theta^N$. Since $A_{\boldsymbol{\mu},\perp}$ in equation 18 depends affinely on $\boldsymbol{\Pi}^{\boldsymbol{\mu}}$ through the block $\gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}}$, it preserves these convex combinations, yielding $A_{\boldsymbol{\mu},\perp} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$. By the classical convex-hull invariance of JSR (Jungers, 2009), we conclude that $\rho(\mathrm{co}(\mathcal{A}_{\alpha,\perp})) = \rho(\mathcal{A}_{\alpha,\perp})$. □

A.5 Proof of Proposition 5.1

To prove Proposition 5.1, we rewrite equation 18 as

$$A_{\boldsymbol{\pi},\perp} = I + \alpha M_{\boldsymbol{\pi},\perp}, \qquad M_{\boldsymbol{\pi},\perp} := \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) & \gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\pi}} & 0 \\ \kappa\mathbf{D} & -\kappa(\mathbf{D}+\mathbf{S}) & -\kappa\mathbf{S}\mathbf{U}^\top \\ 0 & \kappa\mathbf{U}\mathbf{S} & 0 \end{bmatrix} \tag{44}$$

for each deterministic joint policy $\boldsymbol{\pi} \in \Theta^N$. Moreover, we view the recursion in equation 18 as an Euler discretization of the continuous-time linear ODE:

$$X = \mathrm{col}(u, x, z), \qquad \dot{X} = M_{\boldsymbol{\pi},\perp} X. \tag{45}$$

The key structural insight is that equation 45 represents a feedback interconnection of two subsystems. Partitioning $M_{\boldsymbol{\pi},\perp}$ with respect to u and (x, z) yields:

$$\text{(Bellman-estimation)} \quad \dot{u} = -(\mathbf{D}+\mathbf{S})u + \gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\pi}} x, \tag{46}$$

$$\text{(Consensus-mixing)} \quad \dot{x} = \mathbf{D}u - (\mathbf{D}+\mathbf{S})x - \mathbf{S}\mathbf{U}^\top z, \tag{47}$$

$$\dot{z} = \mathbf{U}\mathbf{S}x. \tag{48}$$

Specifically, the Bellman-estimation subsystem in equation 46 takes the consensus estimate x as input and produces the local estimate u, whereas the consensus-mixing subsystem in equations 47–48 takes u as input and returns x. Interconnecting these two blocks in a closed loop exactly recovers the original system in equation 45, and separating the dynamics into these two blocks enables a two-time-scale analysis, which can be handled through a single combined Lyapunov norm. Note that in the consensus-mixing subsystem equations 47–48, the mixing gain κ has been dropped, since it only accelerates the subsystem while leaving its equilibrium map unchanged. This feedback interconnection is depicted in Figure 2. Lastly, we provide the definition of the upper right Dini directional derivative, which is used in the subsequent proofs.

Definition A.1 (Khalil & Grizzle (2002)). *For $\Phi : \mathbb{R}^n \to \mathbb{R}$ and $\psi, \chi \in \mathbb{R}^n$, the upper right Dini directional derivative of Φ at ψ in the direction χ is*

$$D^+\Phi(\psi;\chi) := \limsup_{h\to 0^+} \frac{\Phi(\psi + h\chi) - \Phi(\psi)}{h}. \tag{49}$$

If Φ is locally Lipschitz near ψ and f is locally Lipschitz, then $D^+\Phi(\psi; f(\psi))$ equals the upper right Dini derivative of Φ along the solution $x(t)$ of $\dot{x} = f(x)$ with $x(0) = \psi$. Indeed, $x(h) = \psi + hf(\psi) + o(h)$, and local Lipschitz continuity of Φ makes the function-value discrepancy $o(h)$. All norms and linear vector fields used below satisfy these conditions.

Under this framework, we establish the following four lemmas. The first proves that the Bellman-estimation block (the upper block of Figure 2) is contractive: its output u decreases whenever its norm is sufficiently large relative to its input x. The second shows that the consensus-mixing block (the lower block of Figure 2) is stable and that its steady-state gain is non-expansive. The third combines these two blocks: by accelerating the consensus-mixing block in equations 47–48 with a sufficiently large mixing gain κ, it constructs a unified, policy-independent Lyapunov norm that contracts in continuous time. The fourth transfers this continuous-time contraction property to the discrete-time Euler matrices $I + \alpha M_{\boldsymbol{\pi},\perp}$ from equation 44, thereby establishing $\rho(\mathcal{A}_{\alpha,\perp}) < 1$.

Lemma A.3. *Consider the Bellman-estimation subsystem, i.e., the upper block of the interconnection in Figure 2, with input x and output u:*

$$\dot{u} = f_u(u, x) := -(\mathbf{D}+\mathbf{S})u + \gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\pi}} x. \tag{50}$$

Then, for every deterministic joint policy $\boldsymbol{\pi} \in \Theta^N$, every $u \neq 0$, and all x such that $\|u\|_\infty \geq \gamma\|x\|_\infty$,

$$D^+\|\cdot\|_\infty(u; f_u(u,x)) \leq -d_{\min}(\|u\|_\infty - \gamma\|x\|_\infty), \tag{51}$$

where $D^+\|\cdot\|_\infty(u; f_u(u,x))$ follows Definition A.1 with choices $\Phi = \|\cdot\|_\infty$, $\psi = u$ and $\chi = f_u(u,x)$.

Proof. Since $\|u\|_\infty$ is defined as the maximum of coordinatewise absolute values, its upper right Dini derivative is governed by the active coordinates that attain the maximum, defined by $\mathcal{I}(u) := \{(i,s,a) : |u_i(s,a)| = \|u\|_\infty\}$. Specifically,

$$D^+\|\cdot\|_\infty(u; f_u(u,x))$$

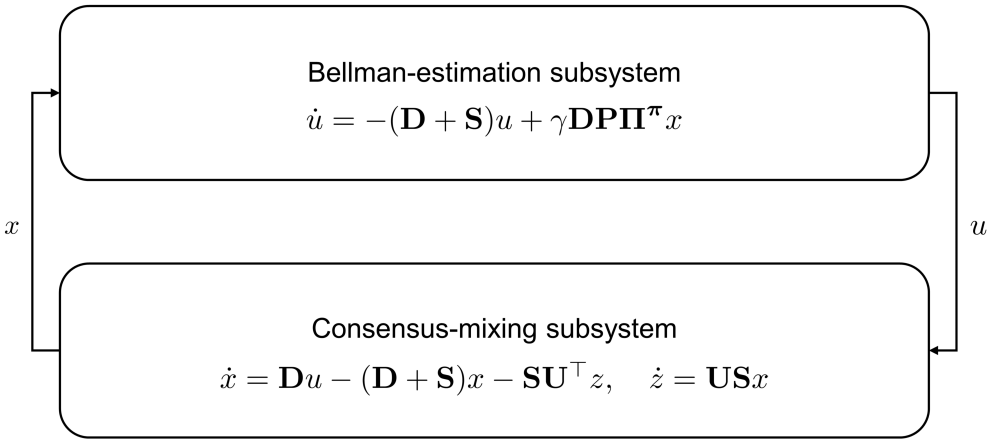


Figure 2: Interconnected-system view of ECQ without κ used in the proof of Proposition 5.1. The upper block takes the consensus estimate x as input to produce the local estimate u. The lower block blends neighboring estimates through the graph coupling $\mathbf{S}$, with state z enforcing exact agreement, and returns x. Its complete dynamics are multiplied by the mixing gain κ.

$$
\begin{aligned}
&= \limsup_{h\to 0^+} \frac{\|u+hf\|_\infty - \|u\|_\infty}{h} \\
&= \limsup_{h\to 0^+} \frac{\big\{\|u\|_\infty + h\max_{(i,s,a)\in\mathcal{I}(u)} \operatorname{sgn}(u_i(s,a))[f_u(u,x)]_i(s,a) + o(h)\big\} - \|u\|_\infty}{h} \\
&= \max_{(i,s,a)\in\mathcal{I}(u)} \operatorname{sgn}(u_i(s,a))[f_u(u,x)]_i(s,a), \qquad (52)
\end{aligned}
$$

where $\operatorname{sgn}(a)$ denotes the sign function, which yields $a/|a|$ for $a \neq 0$ and 0 for $a = 0$. Here, the second equality follows from a first-order Taylor expansion of the absolute values at the active coordinates in $\mathcal{I}(u)$, while the third equality holds since $o(h)/h \to 0$ as $h \to 0^+$.

Now, fixing any $(i,s,a) \in \mathcal{I}(u)$, the scalar component of the vector field equation 50 can be expressed as

$$
[f_u]_i(s,a) = -d(s,a)\left((1+|\mathcal{N}_i|)u_i(s,a) - \sum_{j\in\mathcal{N}_i} u_j(s,a)\right) + \gamma d(s,a)(\mathbf{P\Pi}^{\boldsymbol{\pi}}x)_i(s,a)
$$

according to equation 9. Multiplying both sides by $\operatorname{sgn}(u_i(s,a))$ yields

$$
\begin{aligned}
&\operatorname{sgn}(u_i)[f_u]_i \\
&= -d(s,a)(1+|\mathcal{N}_i|)\operatorname{sgn}(u_i)u_i + d(s,a)\operatorname{sgn}(u_i)\sum_{j\in\mathcal{N}_i} u_j + \gamma d(s,a)\operatorname{sgn}(u_i)(\mathbf{P\Pi}^{\boldsymbol{\pi}}x)_i \\
&= -d(s,a)(1+|\mathcal{N}_i|)|u_i| + d(s,a)\operatorname{sgn}(u_i)\sum_{j\in\mathcal{N}_i} u_j + \gamma d(s,a)\operatorname{sgn}(u_i)(\mathbf{P\Pi}^{\boldsymbol{\pi}}x)_i \\
&\le -d(s,a)(1+|\mathcal{N}_i|)|u_i| + d(s,a)\sum_{j\in\mathcal{N}_i}|u_j| + \gamma d(s,a)|(\mathbf{P\Pi}^{\boldsymbol{\pi}}x)_i| \\
&\le -d(s,a)(1+|\mathcal{N}_i|)|u_i| + d(s,a)\sum_{j\in\mathcal{N}_i}|u_j| + \gamma d(s,a)\|(\mathbf{P\Pi}^{\boldsymbol{\pi}}x)\|_\infty \\
&\le -d(s,a)(1+|\mathcal{N}_i|)\|u\|_\infty + d(s,a)|\mathcal{N}_i|\|u\|_\infty + \gamma d(s,a)\|x\|_\infty \\
&= -d(s,a)\big(\|u\|_\infty - \gamma\|x\|_\infty\big),
\end{aligned}
$$

where the second equality and the first inequality exploit the identity $\mathrm{sgn}(a)b \le |b|$ for any $a, b \in \mathbb{R}$. The third inequality follows because $|u_i(s,a)| = \|u\|_\infty$, $|u_j(s,a)| \le \|u\|_\infty$ for $j \in \mathcal{N}_i$, and $\|\mathbf{\Pi}\mathbf{\Pi}^{\boldsymbol{\pi}} x\|_\infty \le \|x\|_\infty$ with the row-stochasticity of $\mathbf{\Pi}\mathbf{\Pi}^{\boldsymbol{\pi}}$ for every deterministic joint policy. Combining this with equation 52 and using $\|u\|_\infty - \gamma\|x\|_\infty \ge 0$ together with $d(s,a) \ge d_{\min}$, we obtain equation 51, which completes the proof. □

Lemma A.4. *Let $v = \mathrm{col}(x, z)$ and consider the consensus-mixing subsystem, i.e., the lower block of the interconnection in Figure 2, with input u and output x:*

$$\dot{v} = f_v(v,u) := A_{\mathrm{mix}} v + B_{\mathrm{mix}} u, \qquad x = C_{\mathrm{mix}} v, \tag{53}$$

with

$$A_{\mathrm{mix}} = \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) & -\mathbf{S}\mathbf{U}^\top \\ \mathbf{U}\mathbf{S} & 0 \end{bmatrix}, \qquad B_{\mathrm{mix}} = \begin{bmatrix} \mathbf{D} \\ 0 \end{bmatrix}, \qquad C_{\mathrm{mix}} = [I \quad 0].$$

Then, A_{mix} is Hurwitz. Furthermore, defining

$$G := -A_{\mathrm{mix}}^{-1} B_{\mathrm{mix}}, \qquad J := \Big(\frac{1}{N}\mathbf{1}_N \mathbf{1}_N^\top\Big) \otimes I_{|\mathcal{S}||\mathcal{A}|},$$

yields $C_{\mathrm{mix}} G = J$, which implies $\|C_{\mathrm{mix}} G u\|_\infty \le \|u\|_\infty$ for every u.

Proof. To establish that A_{mix} is Hurwitz, we consider the quadratic Lyapunov function candidate

$$V_{\mathrm{mix}}(x,z) := \frac{1}{2} x^\top \mathbf{D}^{-1} x + \frac{1}{2} z^\top \mathbf{D}_\perp^{-1} z$$

with

$$\mathbf{D}_\perp := \mathbf{U}\mathbf{D}\mathbf{U}^\top = I_{N-1} \otimes D \succ 0.$$

Using

$$\mathbf{D}^{-1}\mathbf{S} = L \otimes I_{|\mathcal{S}||\mathcal{A}|} = \mathbf{L}, \qquad \mathbf{D}_\perp^{-1}\mathbf{U}\mathbf{S} = UL \otimes I_{|\mathcal{S}||\mathcal{A}|},$$

from equation 9 and $L = L^\top$ from equation 1, the time derivative of V_{mix} along the trajectories of equation 53 satisfies

$$\begin{aligned}
\dot{V}_{\mathrm{mix}} =& x^\top \mathbf{D}^{-1}\dot{x} + z^\top \mathbf{D}_\perp^{-1}\dot{z} \\
=& x^\top \mathbf{D}^{-1}\left(-(\mathbf{D}+\mathbf{S})x - \mathbf{S}\mathbf{U}^\top z + \mathbf{D}u\right) + z^\top \mathbf{D}_\perp^{-1}\mathbf{U}\mathbf{S}x \\
=& -x^\top (I+\mathbf{L})x - x^\top (LU^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})z + x^\top u + z^\top (UL \otimes I_{|\mathcal{S}||\mathcal{A}|})x \\
=& -x^\top (I+\mathbf{L})x - x^\top (LU^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})z + x^\top u + x^\top (LU^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})z \\
=& x^\top u - x^\top \left((I_N + L) \otimes I_{|\mathcal{S}||\mathcal{A}|}\right) x \\
\le& \|x\|_2 \|u\|_2 - \|x\|_2^2,
\end{aligned} \tag{54}$$

where the second equality follows from the subsystem dynamics equation 53, the fourth equality utilizes the fact that $z^\top (UL \otimes I_{|\mathcal{S}||\mathcal{A}|})x$ is the scalar value together with $L = L^\top$ from equation 1, and the last inequality follows from the Cauchy-Schwarz inequality together with the fact that $(I_N + L) \otimes I \succeq I$ due to $L \succeq 0$. Setting $u = 0$, equation 54 yields $\dot{V}_{\mathrm{mix}} \le 0$, where $\dot{V}_{\mathrm{mix}} = 0$ implies $x = 0$. For a trajectory to remain confined to the set where $x = 0$, the x-dynamics in equation 53 must satisfy

$$0 = \dot{x} = -\mathbf{S}\mathbf{U}^\top z, \quad \text{where } \mathbf{S}\mathbf{U}^\top = (LU^\top) \otimes D$$

by equation 9. Since $\mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}$ from equation 1 and $\mathrm{range}(U^\top) = \{\mathbf{1}_N\}^\perp$ by the definition of U, the matrix $LU^\top$ has full column rank. Furthermore, because $D \succ 0$ from equation 2, the matrix $\mathbf{S}\mathbf{U}^\top$ also has full column rank, which forces $z = 0$. Consequently, by LaSalle's invariance principle (Khalil & Grizzle, 2002), the origin $(x,z) = (0,0)$ is globally asymptotically stable. Since the subsystem is linear time-invariant, this implies that A_{mix} is Hurwitz.

Next, we establish the second statement. At an equilibrium where $\dot{v} = 0$ in equation 53, we have

$$\mathbf{U}\mathbf{S}x_\infty = 0, \tag{55}$$

$$\mathbf{D}u - (\mathbf{D}+\mathbf{S})x_\infty - \mathbf{S}\mathbf{U}^\top z_\infty = 0. \tag{56}$$

Since $\mathbf{S}x_\infty \in \mathrm{range}(\mathbf{S}) = \mathrm{range}(L) \otimes \mathrm{range}(D) = \{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\mathrm{range}(\mathbf{U}^\top) = \{\mathbf{1}_N\}^\perp \otimes \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and $\mathbf{U}^\top$ has full column rank, equation 55 yields $\mathbf{S}x_\infty = 0$. By the graph connectedness condition $\mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}$ from equation 1 along with equation 9, we obtain

$$x_\infty = \mathbf{1}_N \otimes q \tag{57}$$

for some $q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Multiplying equation 56 leftward by $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})$ gives

$$\begin{aligned}
&\left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{D}u - \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)(\mathbf{D}+\mathbf{S})x_\infty - \left(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}\right)\mathbf{S}\mathbf{U}^\top z_\infty \\
=&\left(\mathbf{1}_N^\top \otimes D\right)u - \left(\mathbf{1}_N^\top \otimes D\right)(\mathbf{1}_N \otimes q) \\
=&\left(\mathbf{1}_N^\top \otimes D\right)u - (N \otimes Dq) \\
=&D\sum_{i=1}^N u^i - NDq \\
=&0,
\end{aligned}$$

where the first equality follows from equation 57, $\mathbf{S}x_\infty = 0$, and the mixed-product property of the Kronecker product under equation 9. Then, the last equality directly leads to

$$q = \frac{1}{N}\sum_{i=1}^N u^i.$$

Therefore,

$$\begin{aligned}
\|x_\infty\|_\infty &= \left\|\mathbf{1}_N \otimes \frac{1}{N}\sum_{i=1}^N u^i\right\|_\infty \\
&= \left\|\frac{1}{N}\sum_{i=1}^N u^i\right\|_\infty \\
&= \max_{s,a}\left|\frac{1}{N}\sum_{i=1}^N u^i(s,a)\right| \\
&\le \max_{s,a}\frac{1}{N}\sum_{i=1}^N \left|u^i(s,a)\right| \\
&\le \|u\|_\infty.
\end{aligned} \tag{58}$$

Since A_{mix} is Hurwitz, it is nonsingular. At an equilibrium $\dot{v} = 0$, solving $0 = \dot{v} = A_{\mathrm{mix}}v + B_{\mathrm{mix}}u$ gives $v_\infty = -A_{\mathrm{mix}}^{-1}B_{\mathrm{mix}}u$, so that

$$x_\infty = C_{\mathrm{mix}}v_\infty = C_{\mathrm{mix}}(-A_{\mathrm{mix}}^{-1})B_{\mathrm{mix}}u = C_{\mathrm{mix}}Gu.$$

Recall that $Ju = \mathbf{1}_N \otimes \frac{1}{N}\sum_{i=1}^N u^i$ by definition. Combining this with equation 57 yields $C_{\mathrm{mix}}Gu = x_\infty = Ju$, from which it follows that $C_{\mathrm{mix}}G = J$. Moreover, equation 58 implies

$$\|C_{\mathrm{mix}}Gu\|_\infty = \|x_\infty\|_\infty \le \|u\|_\infty,$$

which completes the proof. □

Lemma A.5. *Under Lemmas A.3 and A.4, there exists a finite threshold $\kappa^* > 0$ such that, for every fixed $\kappa > \kappa^*$, there exists a norm W_c and a constant $\lambda_c > 0$, both independent of $\boldsymbol{\pi} \in \Theta^N$, satisfying*

$$D^+W_c\left(X; M_{\boldsymbol{\pi},\perp}X\right) \le -\lambda_c W_c(X), \qquad \forall X, \quad \forall \boldsymbol{\pi} \in \Theta^N, \tag{59}$$

where $M_{\boldsymbol{\pi},\perp}$ is defined as

$$M_{\boldsymbol{\pi},\perp} := \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) & \gamma\mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\pi}} & 0 \\ \kappa\mathbf{D} & -\kappa(\mathbf{D}+\mathbf{S}) & -\kappa\mathbf{S}\mathbf{U}^\top \\ 0 & \kappa\mathbf{U}\mathbf{S} & 0 \end{bmatrix}$$

from equation 44.

Proof. Recall from equation 45 that the continuous-time linear ODE can be represented as the feedback interconnection of the two subsystems shown in Figure 2. Making the mixing gain κ explicit, we rewrite the system as

$$\text{(Bellman-estimation)} \quad \dot{u} = -(\mathbf{D}+\mathbf{S})u + \gamma \mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\pi}} x \tag{60}$$

$$\text{(Consensus-mixing)} \quad \dot{v} = \kappa(A_{\text{mix}} v + B_{\text{mix}} u), \qquad x = C_{\text{mix}} v, \tag{61}$$

$$A_{\text{mix}} = \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) & -\mathbf{S}\mathbf{U}^\top \\ \mathbf{U}\mathbf{S} & 0 \end{bmatrix}, \; B_{\text{mix}} = \begin{bmatrix} \mathbf{D} \\ 0 \end{bmatrix}, \; C_{\text{mix}} = [I \quad 0].$$

Define the intermediate matrices

$$A_0 := -(\mathbf{D}+\mathbf{S}), \qquad B_{\boldsymbol{\pi}} := \gamma \mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\pi}}, \qquad F_{\boldsymbol{\pi}} := A_0 + B_{\boldsymbol{\pi}} J. \tag{62}$$

To analyze the feedback interconnection, we introduce the error variable relative to the steady-state equilibrium map

$$\epsilon := v - Gu. \tag{63}$$

Using this coordinate transformation, the continuous-time linear ODE $\dot{X} = M_{\boldsymbol{\pi},\perp} X$ in equation 45 can be rewritten as

$$\begin{aligned} \dot{u} =& A_0 u + B_{\boldsymbol{\pi}} x \\ =& A_0 u + B_{\boldsymbol{\pi}} C_{\text{mix}} v \\ =& A_0 u + B_{\boldsymbol{\pi}} C_{\text{mix}} (\epsilon + Gu) \\ =& A_0 u + B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon + B_{\boldsymbol{\pi}} J u \\ =& F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon, \end{aligned} \tag{64}$$

where the first equality uses equation 60, the second equality comes from $x = C_{\text{mix}} v$ from equation 61, the third equality follows from the error variable definition in equation 63, the fourth equality is due to $C_{\text{mix}} G = J$ from Lemma A.4, and the last equality utilizes the intermediate matrix definitions in equation 62. Similarly, the error dynamics $\dot{\epsilon}$ become

$$\begin{aligned} \dot{\epsilon} =& \dot{v} - G\dot{u} \\ =& \kappa(A_{\text{mix}} v + B_{\text{mix}} u) - G(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon) \\ =& \kappa(A_{\text{mix}} \epsilon + A_{\text{mix}} G u + B_{\text{mix}} u) - G(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon) \\ =& \kappa A_{\text{mix}} \epsilon - G F_{\boldsymbol{\pi}} u - G B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon, \end{aligned} \tag{65}$$

where the second equality comes from equation 61, together with the $\dot{u}$ relation in equation 64, the third equality uses the error variable definition in equation 63, and the last equality is due to $A_{\text{mix}} G + B_{\text{mix}} = 0$ from Lemma A.4.

Since A_{mix} is Hurwitz, there exists a unique symmetric matrix $P_m \succ 0$ solving the algebraic Lyapunov equation

$$A_{\text{mix}}^\top P_m + P_m A_{\text{mix}} = -I. \tag{66}$$

We define the error Lyapunov function and its rate constant as

$$W_m(\epsilon) := \sqrt{\epsilon^\top P_m \epsilon}, \qquad \lambda_m := \frac{1}{2\lambda_{\max}(P_m)} > 0, \tag{67}$$

and introduce the finite, policy-uniform constants

$$\begin{aligned} b_{12} &:= \max_{\boldsymbol{\pi} \in \Theta^N} \sup_{\epsilon \neq 0} \frac{\|B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon\|_\infty}{W_m(\epsilon)}, \\ b_{21} := \max_{\boldsymbol{\pi} \in \Theta^N} \sup_{u \neq 0} \frac{W_m(G F_{\boldsymbol{\pi}} u)}{\|u\|_\infty}, &\qquad b_{22} := \max_{\boldsymbol{\pi} \in \Theta^N} \sup_{\epsilon \neq 0} \frac{W_m(G B_{\boldsymbol{\pi}} C_{\text{mix}} \epsilon)}{W_m(\epsilon)}. \end{aligned} \tag{68}$$

Note that these constants in equation 68 are finite: W_m in equation 67 is a norm equivalent to $\|\cdot\|_2$, and each numerator is a fixed linear map applied to ϵ or u. Hence, each supremum represents an induced operator norm on a finite-dimensional space, which is therefore finite. Taking the outer maximum over the finite set Θ^N preserves this finiteness.

We first analyze the decoupled behavior of $\dot{u}$ in equation 64 when $\epsilon = 0$. In this case, combining $x = C_{\text{mix}} v$ from equation 61, $v = Gu$ from equation 63, and $C_{\text{mix}} G = J$ from Lemma A.4 yields

$x = Ju$. Since $\|Ju\|_\infty \le \|u\|_\infty$ from Lemma A.4 and $\gamma < 1$, the hypothesis $\gamma\|Ju\|_\infty \le \|u\|_\infty$ of Lemma A.3 holds, and applying equation 51 from Lemma A.3 with $x = Ju$ gives

$$D^+\|\cdot\|_\infty(u; F_{\boldsymbol{\pi}}u) \le -d_{\min}(\|u\|_\infty - \gamma\|Ju\|_\infty) \le -d_{\min}(1-\gamma)\|u\|_\infty.$$

Re-introducing the coupling term $B_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon$ and applying the subadditivity of the Dini derivative, $D^+\|\cdot\|_\infty(u; f+g) \le D^+\|\cdot\|_\infty(u; f) + \|g\|_\infty$, we get

$$\begin{aligned} D^+\|\cdot\|_\infty(u; \dot{u}) \le& D^+\|\cdot\|_\infty(u; F_{\boldsymbol{\pi}}u) + \|B_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon\|_\infty \\ \le& -d_{\min}(1-\gamma)\|u\|_\infty + \|B_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon\|_\infty \\ \le& -d_{\min}(1-\gamma)\|u\|_\infty + b_{12}W_m(\epsilon), \end{aligned} \tag{69}$$

where the last inequality follows from the definition of b_{12} in equation 68. Here, equation 69 is obtained for $u \neq 0$ via Lemma A.3; when $u = 0$, $\|u\|_\infty$ term vanishes and the bound reduces to $\|B_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon\|_\infty \le b_{12}W_m(\epsilon)$, which holds by the definition of b_{12}. Hence, equation 69 is valid for all u.

Next, we analyze $\dot{\epsilon}$ in equation 65 by splitting its dynamics into three terms. First, considering the trajectory along $A_{\mathrm{mix}}\epsilon$, we have

$$\begin{aligned} \frac{d}{dt}(\epsilon^\top P_m\epsilon) =& \dot{\epsilon}^\top P_m\epsilon + \epsilon^\top P_m\dot{\epsilon} \\ =& (A_{\mathrm{mix}}\epsilon)^\top P_m\epsilon + \epsilon^\top P_m(A_{\mathrm{mix}}\epsilon) \\ =& \epsilon^\top(A_{\mathrm{mix}}^\top P_m + P_mA_{\mathrm{mix}})\epsilon \\ =& -\epsilon^\top\epsilon \\ =& -\|\epsilon\|_2^2, \end{aligned}$$

where the fourth equality follows from equation 66. Differentiating $W_m(\epsilon)$ and applying the Rayleigh–Ritz theorem $\epsilon^\top P_m\epsilon \le \lambda_{\max}(P_m)\|\epsilon\|_2^2$, we obtain

$$\frac{d}{dt}W_m = \frac{d}{dt}\sqrt{\epsilon^\top P_m\epsilon} = \frac{1}{2\sqrt{\epsilon^\top P_m\epsilon}}\frac{d}{dt}\left(\epsilon^\top P_m\epsilon\right) = -\frac{\|\epsilon\|_2^2}{2W_m} \le -\frac{W_m}{2\lambda_{\max}(P_m)} = -\lambda_m W_m,$$

where the last equality follows from the definition of λ_m in equation 67. Since $W_m(\epsilon)$ is smooth for $\epsilon \neq 0$, its Dini derivative coincides with the ordinary derivative. Scaling by $\kappa > 0$ via positive homogeneity gives

$$D^+W_m\big(\epsilon; \kappa A_{\mathrm{mix}}\epsilon\big) = \kappa D^+W_m\big(\epsilon; A_{\mathrm{mix}}\epsilon\big) \le -\kappa\lambda_m W_m(\epsilon). \tag{70}$$

For the remaining two terms in equation 65, we use the inequality $D^+W_m(\epsilon; g) \le W_m(g)$, which follows from the triangle inequality: $D^+W_m(\epsilon; g) = \limsup_{h\to 0^+}(W_m(\epsilon + hg) - W_m(\epsilon))/h \le \limsup_{h\to 0^+}(W_m(\epsilon) + hW_m(g) - W_m(\epsilon))/h = W_m(g)$. Applying this property alongside the definitions of b_{21} and b_{22} in equation 68, we obtain

$$D^+W_m(\epsilon; -GF_{\boldsymbol{\pi}}u) \le W_m(-GF_{\boldsymbol{\pi}}u) \le b_{21}\|u\|_\infty,$$

$$D^+W_m(\epsilon; -GB_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon) \le W_m(-GB_{\boldsymbol{\pi}}C_{\mathrm{mix}}\epsilon) \le b_{22}W_m(\epsilon).$$

Then, combining equation 70 with the above inequalities yields the bound on the Dini derivative of $W_m(\epsilon)$:

$$D^+W_m(\epsilon; \dot{\epsilon}) \le b_{21}\|u\|_\infty - (\kappa\lambda_m - b_{22})W_m(\epsilon). \tag{71}$$

Now, we construct a unified Lyapunov norm to prove the contraction of the combined system with the mixing gain κ. To this end, we combine $\|u\|_\infty$ and $W_m(\epsilon)$ using a positive weight $\theta > 0$:

$$W_c := \|u\|_\infty + \theta W_m(\epsilon). \tag{72}$$

Then, summing equation 69 and equation 71 with the weight θ yields

$$\begin{aligned} D^+W_c(X; M_{\boldsymbol{\pi},\perp}X) \le& -d_{\min}(1-\gamma)\|u\|_\infty + b_{12}W_m(\epsilon) + \theta\left(b_{21}\|u\|_\infty - (\kappa\lambda_m - b_{22})W_m(\epsilon)\right) \\ =& -\left(d_{\min}(1-\gamma) - \theta b_{21}\right)\|u\|_\infty - \left(\theta(\kappa\lambda_m - b_{22}) - b_{12}\right)W_m(\epsilon). \end{aligned}$$

To ensure that the right-hand side yields a contraction, both coefficient terms must be strictly positive. For the first term, if $b_{21} > 0$, we require $\theta < (d_{\min}(1-\gamma))/b_{21}$ (note that $b_{21} \ge 0$ by definition,

and $b_{21} = 0$ trivially satisfies the bound). Furthermore, setting $\kappa > b_{22}/\lambda_m$ yields $\kappa\lambda_m - b_{22} > 0$, which implies $\theta > b_{12}/(\kappa\lambda_m - b_{22})$. Therefore, we obtain the valid range for θ as

$$\frac{b_{12}}{\kappa\lambda_m - b_{22}} < \theta < \frac{d_{\min}(1-\gamma)}{b_{21}}.$$

A valid θ exists within these bounds if and only if the upper bound strictly exceeds the lower bound, i.e.,

$$\frac{b_{12}}{\kappa\lambda_m - b_{22}} < \frac{d_{\min}(1-\gamma)}{b_{21}} \quad \Longrightarrow \quad \kappa > \frac{b_{22} + b_{12}b_{21}/\left(d_{\min}(1-\gamma)\right)}{\lambda_m} =: \kappa^*.$$

Thus, for any chosen $\kappa > \kappa^*$, there exists a suitable weight $\theta > 0$ such that

$$D^+W_c\left(X; M_{\boldsymbol{\pi},\perp}X\right) \le -\lambda_c W_c(X),$$

where $\lambda_c := \min\{d_{\min}(1-\gamma) - \theta b_{21}, \kappa\lambda_m - b_{22} - b_{12}/\theta\} > 0$.

Finally, because the state transformation $(u, v) \mapsto (u, v - Gu)$ is linear and invertible with inverse $(u, e) \mapsto (u, e + Gu)$, W_c in equation 72 defines a valid norm on the transformed state space. Since all gain constants b_{12}, b_{21}, and b_{22} are uniform over the finite policy space Θ^N, the constructed norm W_c and decay rate λ_c are completely policy-independent, completing the proof of equation 59. □

Lemma A.6. *Suppose a norm W_c and a constant $\lambda_c > 0$ satisfy Lemma A.5 for the finite family $\{M_{\boldsymbol{\pi},\perp} : \boldsymbol{\pi} \in \Theta^N\}$, where $M_{\boldsymbol{\pi},\perp}$ is defined as in equation 44. Then, there exists $\bar{\alpha} > 0$ such that, for all $0 < \alpha < \bar{\alpha}$,*

$$W_c\left((I + \alpha M_{\boldsymbol{\pi},\perp})X\right) \le \left(1 - \frac{\lambda_c\alpha}{2}\right) W_c(X), \qquad \forall X, \quad \forall \boldsymbol{\pi} \in \Theta^N. \tag{73}$$

Consequently,

$$\rho(\mathcal{A}_{\alpha,\perp}) < 1. \tag{74}$$

Proof. We begin by establishing the one-step discrete contraction in equation 73. Since the state space is unbounded, finding a uniform step-size bound $\bar{\alpha} > 0$ directly over all $X \in \mathbb{R}^n$ is intractable. To overcome this, we exploit the positive homogeneity of the norm W_c and the linearity of $M_{\boldsymbol{\pi},\perp}$, which allow us to restrict our attention to the unit sphere

$$\mathbb{S}_c := \{X : W_c(X) = 1\}.$$

Fix a mode $M_{\boldsymbol{\pi},\perp}$. For any $Y \in \mathbb{S}_c$ and $\alpha > 0$, define the average rate of change

$$s_{\boldsymbol{\pi}}(Y, \alpha) := \frac{W_c\left(Y + \alpha M_{\boldsymbol{\pi},\perp}Y\right) - W_c(Y)}{\alpha}.$$

Since $M_{\boldsymbol{\pi},\perp}$ is linear, the map $\alpha \mapsto Y + \alpha M_{\boldsymbol{\pi},\perp}Y$ is affine in α. Moreover, as W_c is a norm, it is convex, implying that the composition $g(\alpha) := W_c(Y + \alpha M_{\boldsymbol{\pi},\perp}Y)$ is also a convex function of α. A fundamental property of convex functions (Urruty & Lemaréchal, 1993) guarantees that $s_{\boldsymbol{\pi}}(Y, \alpha) = (g(\alpha) - g(0))/\alpha$ is nondecreasing as $\alpha > 0$ increases. Conversely, as $\alpha \to 0^+$, $s_{\boldsymbol{\pi}}(Y, \alpha)$ decreases monotonically to the right Dini derivative $D^+W_c(Y; M_{\boldsymbol{\pi},\perp}Y)$.

By the continuous-time contraction property in equation 59, we have $D^+W_c(Y; M_{\boldsymbol{\pi},\perp}Y) \le -\lambda_c < -\lambda_c/2$ for every $Y \in \mathbb{S}_c$. Furthermore, since W_c is continuous and $M_{\boldsymbol{\pi},\perp}$ is linear, $s_{\boldsymbol{\pi}}(\cdot, \alpha)$ is continuous on $\mathbb{S}_c$ for each fixed $\alpha > 0$. Consequently, the sets

$$\mathcal{O}_{\boldsymbol{\pi}}(\alpha) := \left\{Y \in \mathbb{S}_c : s_{\boldsymbol{\pi}}(Y, \alpha) < -\frac{\lambda_c}{2}\right\}$$

are open in $\mathbb{S}_c$. By the monotonicity of $s_{\boldsymbol{\pi}}$, these sets expand as $\alpha \to 0^+$ and together cover the entire unit sphere $\mathbb{S}_c$. Because $\mathbb{S}_c$ is compact, this open cover $\{\mathcal{O}_{\boldsymbol{\pi}}(\alpha)\}_{\alpha>0}$ admits a finite subcover $\mathcal{O}_{\boldsymbol{\pi}}(\alpha_1), \ldots, \mathcal{O}_{\boldsymbol{\pi}}(\alpha_n)$. Since these sets are nested, setting $\bar{\alpha}_{\boldsymbol{\pi}} := \min\{\alpha_1, \ldots, \alpha_n\} > 0$ ensures that a single set covers the unit sphere, i.e., $\mathcal{O}_{\boldsymbol{\pi}}(\bar{\alpha}_{\boldsymbol{\pi}}) = \mathbb{S}_c$. By monotonicity, this bound holds for all $0 < \alpha \le \bar{\alpha}_{\boldsymbol{\pi}}$, yielding

$$s_{\boldsymbol{\pi}}(Y, \alpha) \le -\frac{\lambda_c}{2}, \qquad \forall Y \in \mathbb{S}_c, \quad 0 < \alpha \le \bar{\alpha}_{\boldsymbol{\pi}}.$$

Since the policy set Θ^N is finite, taking the minimum over all joint policies gives a strictly positive uniform threshold $\bar{\alpha} := \min_{\boldsymbol{\pi} \in \Theta^N} \bar{\alpha}_{\boldsymbol{\pi}} > 0$. Therefore, substituting the definition of $s_{\boldsymbol{\pi}}(Y, \alpha)$ gives

$$s_{\boldsymbol{\pi}}(Y, \alpha) \leq -\frac{\lambda_c}{2} \quad \Longrightarrow \quad W_c\left(Y + \alpha M_{\boldsymbol{\pi},\perp} Y\right) \leq 1 - \frac{\lambda_c \alpha}{2}, \tag{75}$$

for all $Y \in \mathbb{S}_c$, $\boldsymbol{\pi} \in \Theta^N$, and $0 < \alpha \leq \bar{\alpha}$.

Finally, we scale this result from $\mathbb{S}_c$ to all $X \neq 0$. For any $X \neq 0$, set $Y := X/W_c(X) \in \mathbb{S}_c$. By the positive homogeneity of W_c and linearity of $M_{\boldsymbol{\pi},\perp}$, multiplying both sides of equation 75 by $W_c(X)$ yields

$$W_c\left((I + \alpha M_{\boldsymbol{\pi},\perp})X\right) = W_c(X) W_c\left(Y + \alpha M_{\boldsymbol{\pi},\perp} Y\right) \leq \left(1 - \frac{\lambda_c \alpha}{2}\right) W_c(X),$$

which establishes equation 73. For $X = 0$, the inequality holds trivially as $0 \leq 0$. This completes the proof of the first statement.

It remains to show that $\rho(\mathcal{A}_{\alpha,\perp}) < 1$. For any sequence of joint policies of length k, repeated application of equation 73 gives

$$W_c(A_{\boldsymbol{\pi}_{k-1},\perp} \cdots A_{\boldsymbol{\pi}_0,\perp} X) \leq \left(1 - \frac{\lambda_c \alpha}{2}\right)^k W_c(X).$$

Taking the supremum over all policy sequences of length k and evaluating the limit in the JSR definition from Definition 5.1 yields

$$\rho(\mathcal{A}_{\alpha,\perp}) \leq 1 - \frac{\lambda_c \alpha}{2} < 1.$$

□

We now complete the proof of Proposition 5.1.

Proposition 5.1 Restated. There exists a finite threshold $\kappa^* > 0$ such that, for every fixed $\kappa > \kappa^*$, there exists $\bar{\alpha}(\kappa) > 0$ such that the projected deterministic switching family $\mathcal{A}_{\alpha,\perp}$ satisfies $\rho(\mathcal{A}_{\alpha,\perp}) < 1$ for every $0 < \alpha < \bar{\alpha}(\kappa)$.

Proof. From equation 18, the associated continuous-time dynamics without κ are given by

$$\dot{u} = -(\mathbf{D} + \mathbf{S})u + \gamma \mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\pi}} x, \quad \dot{x} = \mathbf{D}u - (\mathbf{D} + \mathbf{S})x - \mathbf{S}\mathbf{U}^\top z, \quad \dot{z} = \mathbf{U}\mathbf{S}x,$$

which forms the feedback interconnection of the Bellman-estimation subsystem and the consensus-mixing subsystem analyzed in Lemmas A.3 and A.4. Lemma A.5 provides a finite threshold κ^* such that for every fixed $\kappa > \kappa^*$, there exists a policy-independent norm W_c with a decay rate $\lambda_c > 0$ satisfying equation 59 that is common to all generators $M_{\boldsymbol{\pi},\perp}$ for $\boldsymbol{\pi} \in \Theta^N$. Since this policy family is finite, Lemma A.6 yields a step-size threshold $\bar{\alpha}(\kappa) > 0$ such that, for every $0 < \alpha < \bar{\alpha}(\kappa)$, each matrix in $\mathcal{A}_{\alpha,\perp}$ acts as a strict contraction in W_c:

$$W_c(A_{\boldsymbol{\pi},\perp} X) \leq \left(1 - \frac{\lambda_c \alpha}{2}\right) W_c(X), \qquad \boldsymbol{\pi} \in \Theta^N.$$

Taking products over arbitrary policy sequences and applying the definition of JSR as in Lemma A.6 yields the desired bound equation 74. This completes the proof. □

A.6 Proof of Lemma 5.2

Lemma 5.2 Restated. Under Assumption 5.1, suppose there exists $\bar{\alpha} \in (0, 1)$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$. Then, there exist constants $c > 0$ and $K < \infty$, independent of α, such that for every $0 < \alpha \leq \bar{\alpha}$:

(i) $\rho(\mathcal{A}_{\alpha,\perp}) \leq 1 - c\alpha < 1$.

(ii) every product of matrices drawn from $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ satisfies

$$\sup_{M_0, \ldots, M_{\ell-1} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})} \|M_{\ell-1} \cdots M_0\|_2 \leq K(1 - c\alpha)^\ell, \qquad \ell \geq 0.$$

Proof. We first prove claim (i). While $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$ governs the asymptotic growth rate of matrix products, it does not directly guarantee a one-step contraction under the standard Euclidean norm $\|\cdot\|_2$. To establish a one-step contraction for every certified matrix, we construct an induced norm that incorporates all future product behaviors. Select $\bar{\beta}$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < \bar{\beta} < 1$, and define

$$p(X) := \sup_{\ell \geq 0} \bar{\beta}^{-\ell} \sup_{M_0,\ldots,M_{\ell-1} \in \mathcal{A}_{\bar{\alpha},\perp}} \|M_{\ell-1} \cdots M_0 X\|_2$$

for the finite family $\mathcal{A}_{\bar{\alpha},\perp}$, where the product of length zero denotes the identity matrix I. By the definition of the JSR (Jungers, 2009), $p(X)$ is finite for all X. Furthermore, the length-zero term ensures $p(X) \geq \|X\|_2$, and the homogeneity together with the triangle inequality confirms that $p(\cdot)$ is indeed a valid norm.

For any $A_{\pi,\perp}(\bar{\alpha})$ and any length-ℓ product $M_{\ell-1} \cdots M_0$, the composite $M_{\ell-1} \cdots M_0 \, A_{\pi,\perp}(\bar{\alpha})$ forms a product of length $\ell + 1$, which $p(\cdot)$ weights by $\bar{\beta}^{-(\ell+1)}$. Splitting $\bar{\beta}^{-\ell} = \bar{\beta} \cdot \bar{\beta}^{-(\ell+1)}$ and taking the supremum over ℓ yields

$$p(A_{\pi,\perp}(\bar{\alpha})X) = \sup_{\ell \geq 0} \bar{\beta} \cdot \bar{\beta}^{-(\ell+1)} \sup_{M_0,\ldots,M_{\ell-1} \in \mathcal{A}_{\bar{\alpha},\perp}} \|M_{\ell-1} \cdots M_0 A_{\pi,\perp}(\bar{\alpha})X\|_2 \leq \bar{\beta} p(X),$$

where the inequality follows from $p(X)$ being the supremum over all $\ell \geq 0$. By the convexity of the norm $p(\cdot)$, this contraction extends directly to any matrix $B := \sum_i \lambda_i A_{\pi_i,\perp}(\bar{\alpha}) \in \mathrm{co}(\mathcal{A}_{\bar{\alpha},\perp})$:

$$p(BX) \leq \sum_i \lambda_i p(A_{\pi_i,\perp}(\bar{\alpha})X) \leq \bar{\beta} p(X),$$

demonstrating that the contraction holds across the entire convex hull $\mathrm{co}(\mathcal{A}_{\bar{\alpha},\perp})$.

To transfer this property to smaller step-sizes $\alpha \in (0, \bar{\alpha}]$, we utilize the affine decomposition in equation 19. By setting $t := \alpha/\bar{\alpha} \in (0, 1]$, we can express $A_{\pi,\perp}(\alpha) = (1-t)I + tA_{\pi,\perp}(\bar{\alpha})$. Then, for any matrix $M \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$, it can be written as $M = \sum_i \lambda_i A_{\pi_i,\perp}(\alpha) = (1-t)I + t\sum_i \lambda_i A_{\pi_i,\perp}(\bar{\alpha}) = (1-t)I + tB$ for some $B \in \mathrm{co}(\mathcal{A}_{\bar{\alpha},\perp})$. Since the identity matrix I preserves $p(X)$ and B contracts it, we have

$$p(MX) \leq (1-t)p(X) + tp(BX) \leq \left\{(1-t) + t\bar{\beta}\right\} p(X) = \left(1 - \frac{(1-\bar{\beta})\alpha}{\bar{\alpha}}\right) p(X).$$

Defining $c := (1-\bar{\beta})/\bar{\alpha}$ yields the uniform one-step contraction

$$p(MX) \leq (1 - c\alpha)p(X), \qquad \forall X, \forall M \in \mathrm{co}(\mathcal{A}_{\alpha,\perp}),$$

where $c > 0$ depends solely on $\bar{\alpha}$ and $\bar{\beta}$, making it independent of α. Finally, iterating the above inequality over a product of length ℓ yields $p(M_{\ell-1} \cdots M_0 X) \leq (1 - c\alpha)^\ell p(X)$. Taking the ℓ-th root and letting $\ell \to \infty$ bounds the JSR of $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ by $1 - c\alpha$. Since Lemma 5.1 establishes $\rho\big(\mathrm{co}(\mathcal{A}_{\alpha,\perp})\big) = \rho(\mathcal{A}_{\alpha,\perp})$, we conclude that

$$\rho(\mathcal{A}_{\alpha,\perp}) = \rho\big(\mathrm{co}(\mathcal{A}_{\alpha,\perp})\big) < 1 - c\alpha < 1,$$

which completes the proof of claim (i).

For claim (ii), by the Norm Equivalence Theorem on finite-dimensional spaces (Rudin, 2021), there exist positive constants $a, b > 0$ such that $\|X\|_2 \leq ap(X)$ and $p(X) \leq b\|X\|_2$ for all X. Note that a and b depend only on the fixed norm $p(\cdot)$ and are therefore independent of α. Applying these inequalities to the iterated contraction bound gives

$$\|M_{\ell-1} \cdots M_0 X\|_2 \leq a\, p(M_{\ell-1} \cdots M_0 X) \leq ab\,(1 - c\alpha)^\ell \|X\|_2.$$

Setting $K := ab$ establishes equation 20, which completes the proof of claim (ii). □

A.7 Proof of Lemma 5.3

Lemma 5.3 Restated. Let

$$L_\Pi := \sup_{\boldsymbol{\mu},\boldsymbol{\nu}} \|\mathbf{DP}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}})\|_2,$$

where the supremum is taken over all joint stochastic policies μ and ν. Then, L_Π is finite, and for all μ and ν,

$$\|A_{\boldsymbol{\mu},\perp} - A_{\boldsymbol{\nu},\perp}\|_2 \leq \alpha\gamma L_\Pi.$$

Proof. Since the set of joint stochastic policies is compact and the mapping $(\boldsymbol{\mu}, \boldsymbol{\nu}) \mapsto \|\mathbf{DP}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}})\|_2$ is continuous, the Weierstrass extreme value theorem (Rudin, 2021) guarantees that the supremum is achieved and finite.

Moreover, taking the difference of the two mode matrices in equation 18 reveals that all blocks vanish except for the (u, x)-block:

$$A_{\boldsymbol{\mu},\perp} - A_{\boldsymbol{\nu},\perp} = \begin{bmatrix} 0 & \alpha\gamma\mathbf{DP}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}}) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Taking the Euclidean operator norm and applying the definition of L_Π completes the proof. □

A.8 Proof of Lemma 5.4

Lemma 5.4 Restated. Under Assumption 5.1, let Z_∞ be chosen to satisfy equation 16. Then, there exist α-independent constants $\sigma_0, \sigma_1 < \infty$ such that for all $k \geq 0$,

$$\mathbb{E}[\eta_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}\left[\|\eta_{k+1}\|_2^2 \mid \mathcal{F}_k\right] \leq \sigma_0^2 + \sigma_1^2 \|X_k\|_2^2,$$

where $\mathcal{F}_k = \sigma(\mathbf{W}_0, \mathbf{Q}_0, \mathbf{Z}_0, \{(s_t, a_t, s_t', r_{t+1}^1, \ldots, r_{t+1}^N) : 0 \leq t \leq k-1\})$ is the natural filtration.

Proof. By construction of equation 28, ω_{k+1}, ν_{k+1}, and ζ_{k+1} have zero conditional mean. Since $\mathbf{U}$ is deterministic, $\mathbb{E}[\mathbf{U}\zeta_{k+1} \mid \mathcal{F}_k] = \mathbf{U}\mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0$, which implies $\mathbb{E}[\eta_{k+1} \mid \mathcal{F}_k] = 0$. This establishes the first claim.

For the second claim, we decompose the iterates around their limits using equation 14, Proposition 4.1, and Lemma A.2:

$$\mathbf{W}_k = \mathbf{W}_\infty + u_k, \qquad \mathbf{Q}_k = \mathbf{1}_N \otimes Q^* + x_k, \qquad \mathbf{Z}_k = \mathbf{Z}_\infty + \mathbf{U}^\top z_k \tag{76}$$

Since $UU^\top = I_{N-1}$ with equation 40, we have $\mathbf{U}\mathbf{U}^\top = (UU^\top) \otimes I = I$, yielding $\|\mathbf{U}^\top\|_2 = 1$. Using the relation $\|\cdot\|_\infty \leq \|\cdot\|_2$, equation 76, the definition of X_k in equation 18, and setting

$$K_0 := \max\{\|\mathbf{W}_\infty\|_\infty, \|\mathbf{1}_N \otimes Q^*\|_\infty, \|\mathbf{Z}_\infty\|_\infty\}, \tag{77}$$

each block-norm satisfies

$$\begin{aligned} \|W_k^i\|_\infty \leq& \|\mathbf{W}_\infty\|_\infty + \|u_k\|_2 \leq K_0 + \|X_k\|_2, \\ \|Q_k^i\|_\infty \leq& \|\mathbf{1}_N \otimes Q^*\|_\infty + \|x_k\|_2 \leq K_0 + \|X_k\|_2, \\ \|Z_k^i\|_\infty \leq& \|\mathbf{Z}_\infty\|_\infty + \|\mathbf{U}^\top z_k\|_2 \leq K_0 + \|X_k\|_2. \end{aligned}$$

Then, setting $K_k := K_0 + \|X_k\|_2$, we obtain

$$\max_i\{\|W_k^i\|_\infty, \|Q_k^i\|_\infty, \|Z_k^i\|_\infty\} \leq K_k. \tag{78}$$

Defining $R_{\max} := \sup_{i,s,a,s'} |r^i(s,a,s')| < \infty$, the term $\hat{g}_{W,k}^i$ from equation 25 satisfies, almost surely,

$$\begin{aligned} \left|\hat{g}_{W,k}^i(s_k, a_k)\right| =& \left| r_{k+1}^i + \gamma \max_{a' \in \mathcal{A}} Q_k^i(s_k', a') - W_k^i(s_k, a_k) - \sum_j L_{ij} W_k^j(s_k, a_k) \right| \\ \leq& R_{\max} + \gamma \max_{a' \in \mathcal{A}} \left|Q_k^i(s_k', a')\right| + \left|W_k^i(s_k, a_k)\right| + \left| \sum_j L_{ij} W_k^j(s_k, a_k) \right| \\ \leq& R_{\max} + \gamma K_k + K_k + \left| \sum_j L_{ij} W_k^j(s_k, a_k) \right| \\ =& R_{\max} + \gamma K_k + K_k + \left| |\mathcal{N}_i| W_k^i(s_k, a_k) - \sum_{j \in \mathcal{N}_i} W_k^j(s_k, a_k) \right| \end{aligned}$$

$$
\begin{aligned}
&\leq R_{\max} + \gamma K_k + K_k + |\mathcal{N}_i||W_k^i(s_k, a_k)| + \sum_{j \in \mathcal{N}_i} |W_k^j(s_k, a_k)| \\
&\leq R_{\max} + \gamma K_k + K_k + 2|\mathcal{N}_i| K_k \\
&\leq R_{\max} + (\gamma + 1 + 2 \max_i |\mathcal{N}_i|)(K_0 + \|X_k\|_2), \qquad (79)
\end{aligned}
$$

where equation 78 is applied in the second and fourth inequalities, the definition of the graph Laplacian L is used in the second equality, and the last inequality follows from the definition of K_k. By a similar argument, we obtain from equations 26–27 that

$$
\left|\hat{g}_{Q,k}^i(s_k, a_k)\right| \leq (2 + 4 \max_i |\mathcal{N}_i|)(K_0 + \|X_k\|_2), \quad \left|\hat{g}_{Z,k}^i(s_k, a_k)\right| \leq 2 \max_i |\mathcal{N}_i|(K_0 + \|X_k\|_2).
$$

For simplicity, define

$$
a_W = \gamma + 1 + 2 \max_i |\mathcal{N}_i|, \qquad a_Q = 2 + 4 \max_i |\mathcal{N}_i|, \qquad a_Z = 2 \max_i |\mathcal{N}_i| \qquad (80)
$$

and

$$
b_W = R_{\max} + a_W K_0, \qquad b_Q = a_Q K_0, \qquad b_Z = a_Z K_0. \qquad (81)
$$

Since each sampled update has only one possibly nonzero coordinate per agent with $\|e_k\|_2 = 1$, we have

$$
\|\hat{g}_{W,k}^i\|_2^2 = \|e_k\, \hat{g}_{W,k}^i(s_k, a_k)\|_2^2 = |\hat{g}_{W,k}^i(s_k, a_k)|^2,
$$

and the same identity holds for $\hat{g}_{Q,k}^i$ and $\hat{g}_{Z,k}^i$. Stacking over all N agents and applying equation 79 gives

$$
\|\hat{g}_{W,k}\|_2^2 = \sum_{i=1}^N \|\hat{g}_{W,k}^i\|_2^2 = \sum_{i=1}^N |\hat{g}_{W,k}^i(s_k, a_k)|^2 \leq N \left(b_W + a_W \|X_k\|_2\right)^2. \qquad (82)
$$

Analogously, we obtain

$$
\|\hat{g}_{Q,k}\|_2^2 \leq N \left(b_Q + a_Q \|X_k\|_2\right)^2, \qquad \|\hat{g}_{Z,k}\|_2^2 \leq N \left(b_Z + a_Z \|X_k\|_2\right)^2. \qquad (83)
$$

Then, from the definition of the noise vector in equation 18, we have

$$
\begin{aligned}
\|\eta_{k+1}\|_2^2 &= \left\| \begin{bmatrix} \omega_{k+1} \\ \kappa \nu_{k+1} \\ \kappa \mathbf{U} \zeta_{k+1} \end{bmatrix} \right\|_2^2 \\
&= \|\omega_{k+1}\|_2^2 + \kappa^2 \|\nu_{k+1}\|_2^2 + \kappa^2 \|\mathbf{U}\zeta_{k+1}\|_2^2 \\
&= \|\omega_{k+1}\|_2^2 + \kappa^2 \|\nu_{k+1}\|_2^2 + \kappa^2 (\mathbf{U}\zeta_{k+1})^\top (\mathbf{U}\zeta_{k+1}) \\
&\leq \|\omega_{k+1}\|_2^2 + \kappa^2 \|\nu_{k+1}\|_2^2 + \kappa^2 \|\mathbf{U}\|_2^2 \|\zeta_{k+1}\|_2^2 \\
&= \|\omega_{k+1}\|_2^2 + \kappa^2 \|\nu_{k+1}\|_2^2 + \kappa^2 \|\zeta_{k+1}\|_2^2,
\end{aligned}
$$

where the last equality follows from $\|\mathbf{U}\|_2 = 1$. By the definition of the variance and equation 28,

$$
\begin{aligned}
\mathbb{E}[\|\omega_{k+1}\|_2^2 \mid \mathcal{F}_k] &= \mathbb{E}[\|\widehat{g}_{W,k} - \mathbb{E}[\hat{g}_{W,k} \mid \mathcal{F}_k]\|_2^2 \mid \mathcal{F}_k] \\
&= \mathbb{E}[\|\widehat{g}_{W,k}\|_2^2 \mid \mathcal{F}_k] - \|\mathbb{E}[\hat{g}_{W,k} \mid \mathcal{F}_k]\|_2^2 \\
&\leq \mathbb{E}[\|\widehat{g}_{W,k}\|_2^2 \mid \mathcal{F}_k],
\end{aligned}
$$

and analogous bounds hold for ν_{k+1} and ζ_{k+1}. Hence,

$$
\begin{aligned}
&\mathbb{E}[\|\eta_{k+1}\|_2^2 \mid \mathcal{F}_k] \\
&\leq \mathbb{E}\left[\|\widehat{g}_{W,k}\|_2^2 + \kappa^2 \|\widehat{g}_{Q,k}\|_2^2 + \kappa^2 \|\widehat{g}_{Z,k}\|_2^2 \mid \mathcal{F}_k\right] \\
&\leq N \left((b_W + a_W \|X_k\|_2)^2 + \kappa^2 (b_Q + a_Q \|X_k\|_2)^2 + \kappa^2 (b_Z + a_Z \|X_k\|_2)^2\right) \\
&\leq 2N \left(b_W^2 + \kappa^2 b_Q^2 + \kappa^2 b_Z^2\right) + 2N \left(a_W^2 + \kappa^2 a_Q^2 + \kappa^2 a_Z^2\right) \|X_k\|_2^2,
\end{aligned}
$$

where the second inequality applies the bounds equations 82–83, and the last inequality uses $2mn \leq m^2 + n^2$ for all $m, n \in \mathbb{R}$. This proves the second claim with

$$
\sigma_0^2 := 2N \left(b_W^2 + \kappa^2 b_Q^2 + \kappa^2 b_Z^2\right), \qquad \sigma_1^2 := 2N \left(a_W^2 + \kappa^2 a_Q^2 + \kappa^2 a_Z^2\right).
$$

Note that σ_0 and σ_1 are independent of α, as they depend solely on $\gamma, \max_i |\mathcal{N}_i|, R_{\max}, N, K_0$ and the fixed gain κ via equation 77 and equations 80–81, none of which involves the step-size. □

A.9 Proof of Proposition 5.2

Proposition 5.2 Restated. Under Assumption 5.1, suppose there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$. Let c, K be as in Lemma 5.2, σ_0, σ_1 as in Lemma 5.4, and L_Π as in equation 21. Define

$$C_0 := 1 + 4K^2 + 4\left(\frac{\gamma L_\Pi K^2}{c}\right)^2, \quad C_1 := 4\left(\frac{K^2}{c} + \frac{\gamma^2 L_\Pi^2 K^4}{c^3}\right), \quad \alpha_0 := \min\left\{\bar{\alpha}, \frac{1}{2C_1\sigma_1^2}\right\},$$

with the convention $\alpha_0 = \bar{\alpha}$ if $\sigma_1 = 0$. Consider Algorithm 1 initialized with a deterministic condition and Z_∞ satisfying equation 16. Then, for every $0 < \alpha \le \alpha_0$ and $k \ge 0$,

$$\sup_{0\le t\le k} \mathbb{E}\left[\|X_t\|_2^2\right] \le 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2.$$

In particular,

$$\sup_{0\le t\le k} \mathbb{E}[\|Q_t - \mathbf{1}_N \otimes Q^*\|_2^2] \le 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2.$$

Proof. Fix $0 < \alpha \le \bar{\alpha}$. By Lemma 5.2, any product of matrices from $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ satisfies equation 20. Since Lemma 5.1 guarantees that $A_{\boldsymbol{\mu}_k,\perp} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$ for every k, the matrix product along the trajectory, $A_{\boldsymbol{\mu}_{\ell-1},\perp}\cdots A_{\boldsymbol{\mu}_0,\perp}$, is a specific instance of these bounded products. Consequently, equation 20 directly applies along the actual trajectory.

To analyze X_k, our strategy is to compare it with a simpler auxiliary process that uses a single fixed policy and then bound their gap. Specifically, for a fixed joint stochastic policy $\bar{\boldsymbol{\mu}}$, we introduce the reference filter

$$Y_{k+1} = A_{\bar{\boldsymbol{\mu}},\perp}Y_k + \alpha\eta_{k+1}, \qquad Y_0 = X_0, \tag{84}$$

which is driven by the same noise as X_k but maintains a constant policy $\bar{\boldsymbol{\mu}}$. We then decompose Y_k into its deterministic transient $\bar{Y}_k$ and noise response Ξ_k by writing $Y_k = \bar{Y}_k + \Xi_k$, where

$$\bar{Y}_{k+1} = A_{\bar{\boldsymbol{\mu}},\perp}\bar{Y}_k, \qquad \bar{Y}_0 = X_0, \tag{85}$$

$$\Xi_{k+1} = A_{\bar{\boldsymbol{\mu}},\perp}\Xi_k + \alpha\eta_{k+1}, \qquad \Xi_0 = 0. \tag{86}$$

Applying equation 85 recursively together with equation 20 yields

$$\|\bar{Y}_k\|_2 \le K(1-c\alpha)^k\|X_0\|_2. \tag{87}$$

Similarly, unrolling equation 86 recursively gives

$$\Xi_k = \alpha\sum_{t=0}^{k-1} A_{\bar{\boldsymbol{\mu}},\perp}^{k-1-t}\eta_{t+1}.$$

Taking the mean-square norm of the above equation and applying Lemma 5.4 alongside equation 20, we obtain

$$\mathbb{E}\left[\|\Xi_k\|_2^2\right] \le \alpha^2K^2\sum_{t=0}^{k-1}(1-c\alpha)^{2(k-1-t)}\left(\sigma_0^2 + \sigma_1^2\mathbb{E}\left[\|X_t\|_2^2\right]\right) \tag{88}$$

for $k \ge 1$. Let

$$M_k := \sup_{0\le t\le k}\mathbb{E}\left[\|X_t\|_2^2\right] \tag{89}$$

denote the running mean-square bound. Summing the geometric series and using $1-(1-c\alpha)^2 = c\alpha(2-c\alpha) \ge c\alpha$ with $c > 0$ and $\alpha \in (0,1)$, we get

$$\mathbb{E}\left[\|\Xi_k\|_2^2\right] \le \frac{\alpha K^2}{c}\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right). \tag{90}$$

It remains to bound the gap $\Psi_k := X_k - Y_k$ between the true iterate and the reference filter. Subtracting equation 18 from equation 84 cancels the noise term, yielding the error recursion

$$\Psi_{k+1} = A_{\boldsymbol{\mu}_k,\perp}\Psi_k + (A_{\boldsymbol{\mu}_k,\perp} - A_{\bar{\boldsymbol{\mu}},\perp})Y_k, \qquad \Psi_0 = 0.$$

That is, the gap is driven solely by the policy mismatch $A_{\boldsymbol{\mu}_k,\perp} - A_{\bar{\boldsymbol{\mu}},\perp}$ acting on Y_k. Since Y_k decomposes into $\bar{Y}_k$ and Ξ_k, we track their individual effects by splitting $\Psi_k = U_k + V_k$, where

$$U_{k+1} = A_{\boldsymbol{\mu}_k,\perp} U_k + (A_{\boldsymbol{\mu}_k,\perp} - A_{\bar{\boldsymbol{\mu}},\perp})\bar{Y}_k, \qquad U_0 = 0,$$
$$V_{k+1} = A_{\boldsymbol{\mu}_k,\perp} V_k + (A_{\boldsymbol{\mu}_k,\perp} - A_{\bar{\boldsymbol{\mu}},\perp})\Xi_k, \qquad V_0 = 0.$$

For U_k, combining Lemma 5.3 with equation 20 and equation 87 gives

$$\begin{aligned}\|U_k\|_2 \leq& \sum_{i=0}^{k-1} K(1-c\alpha)^{k-1-i}\alpha\gamma L_\Pi K(1-c\alpha)^i \|X_0\|_2 \\ =&\alpha\gamma L_\Pi K^2 k(1-c\alpha)^{k-1}\|X_0\|_2 \qquad (91)\\ \leq&\frac{\gamma L_\Pi K^2}{c}\|X_0\|_2, \qquad (92)\end{aligned}$$

where the last inequality uses $\sup_{k\geq 1} \alpha k(1-c\alpha)^{k-1} \leq 1/c$ derived from $k(1-ca)^{k-1} \leq \sum_{i=0}^{k-1}(1-ca)^i \leq \frac{1}{ca}$. Repeating the construction for V_k with Ξ_i in place of $\bar{Y}_i$, we have

$$\begin{aligned}\|V_k\|_2^2 \leq&(\alpha\gamma L_\Pi K)^2 \left(\sum_{i=0}^{k-1}(1-c\alpha)^{k-1-i}\|\Xi_i\|_2\right)^2 \qquad (93)\\ \leq&(\alpha\gamma L_\Pi K)^2 \left(\sum_{i=0}^{k-1}(1-c\alpha)^{k-1-i}\right)\left(\sum_{i=0}^{k-1}(1-c\alpha)^{k-1-i}\|\Xi_i\|_2^2\right),\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. By taking expectations,

$$\begin{aligned}\mathbb{E}\left[\|V_k\|_2^2\right] \leq&(\alpha\gamma L_\Pi K)^2 \left(\sum_{i=0}^{k-1}(1-c\alpha)^{k-1-i}\right)\left(\sum_{i=0}^{k-1}(1-c\alpha)^{k-1-i}\mathbb{E}\left[\|\Xi_i\|_2^2\right]\right) \qquad (94)\\ \leq&\frac{\gamma^2 L_\Pi^2 K^2}{c^2}\sup_{0\leq i\leq k-1}\mathbb{E}[\|\Xi_i\|_2^2]\\ \leq&\frac{\alpha\gamma^2 L_\Pi^2 K^4}{c^3}\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right) \qquad (95)\end{aligned}$$

via equation 90. Finally, we combine the four terms. Since $X_k = Y_k + \Psi_k = \bar{Y}_k + \Xi_k + U_k + V_k$, applying Cauchy-Schwarz via $\|a+b+c+d\|_2^2 \leq 4(\|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 + \|d\|_2^2)$ yields

$$\|X_k\|_2^2 \leq 4\|\bar{Y}_k\|_2^2 + 4\|\Xi_k\|_2^2 + 4\|U_k\|_2^2 + 4\|V_k\|_2^2.$$

Taking expectations and substituting equations 87,90, 92, and 95, we first use $(1-c\alpha)^{2t} \leq 1$ and $M_{t-1} \leq M_{k-1}$ for $1 \leq t \leq k$. Taking the supremum over $0 \leq t \leq k$, including $t = 0$ since $K \geq 1$, then gives, for $k \geq 1$,

$$\begin{aligned}M_k \leq&4K^2\|X_0\|_2^2 + \frac{4\alpha K^2}{c}\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right)\\ &+ \frac{4\gamma^2 L_\Pi^2 K^4}{c^2}\|X_0\|_2^2 + \frac{4\alpha\gamma^2 L_\Pi^2 K^4}{c^3}\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right)\\ =&\left(4K^2 + \frac{4\gamma^2 L_\Pi^2 K^4}{c^2}\right)\|X_0\|_2^2 + \alpha\left(\frac{4K^2}{c} + \frac{4\gamma^2 L_\Pi^2 K^4}{c^3}\right)\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right)\\ \leq&C_0\|X_0\|_2^2 + C_1\alpha\left(\sigma_0^2 + \sigma_1^2 M_{k-1}\right),\end{aligned}$$

where M_k is defined in equation 89 and C_0, C_1 are defined as in equation 22. For $0 < \alpha \leq \alpha_0$ with equation 22, choosing $\alpha \leq \alpha_0 \leq 1/(2C_1\sigma_1^2)$ ensures $C_1\alpha\sigma_1^2 \leq 1/2$ with the stated convention when $\sigma_1 = 0$. Since $M_{k-1} \leq M_k$, we obtain

$$M_k \leq C_0\|X_0\|_2^2 + C_1\alpha\sigma_0^2 + \frac{1}{2}M_k \quad\Longrightarrow\quad M_k \leq 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2.$$

The base case $k = 0$ holds trivially since $C_0 \geq 1$. This completes the proof of equation 23. The stated Q-error bound follows immediately, as the second block of X_t in equation 18 corresponds to $Q_t - \mathbf{1}_N \otimes Q^*$. □

A.10 Proof of Theorem 5.1

Theorem 5.1 Restated. Under the hypotheses and notation of Proposition 5.2, set $\beta := 1 - c\alpha/2$ for $0 < \alpha \le \alpha_0$. Then, for all $k \ge 0$ with the convention $k\beta^{k-1} = 0$ at $k = 0$,

$$\mathbb{E}\left[\|X_k\|_2\right] \le K\beta^k\|X_0\|_2 + \alpha\gamma K^2 L_\Pi k\beta^{k-1}\|X_0\|_2 + \left(1 + \frac{2\gamma K L_\Pi}{c}\right) K\sqrt{\frac{2\alpha}{c}}\left(\sigma_0^2 + \sigma_1^2\left(2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2\right)\right)^{1/2}.$$

Consequently, $\mathbb{E}[\|Q_k - \mathbf{1}_N \otimes Q^*\|_2] \le \mathbb{E}[\|X_k\|_2]$. For a fixed initial condition, the steady-state error term in equation 24 scales as $O(\sqrt{\alpha})$ as α vanishes.

Proof. By Lemma 5.2, every product generated by matrices in $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ is bounded by $K(1 - c\alpha)^\ell \le K\beta^\ell$, since $1 - c\alpha < \beta < 1$. We reuse the reference-filter decomposition $X_k = \bar{Y}_k + \Xi_k + U_k + V_k$ from Appendix A.9, replacing $1 - c\alpha$ with the refined product bound $K\beta^\ell$. Then, from equation 87 and equation 91,

$$\|\bar{Y}_k\|_2 \le K\beta^k\|X_0\|_2, \qquad \|U_k\|_2 \le \alpha\gamma L_\Pi K^2 k\beta^{k-1}\|X_0\|_2.$$

Moreover, define $M_\alpha := 2C_0\|X_0\|_2^2 + 2C_1\alpha\sigma_0^2$. Applying Jensen's inequality and the geometric series summation to equation 88 and equation 93 yields

$$\mathbb{E}\left[\|\Xi_k\|_2\right] \le \left(\alpha^2 K^2 \sum_{t=0}^{k-1} \beta^{2(k-1-t)}\left(\sigma_0^2 + \sigma_1^2 M_\alpha\right)\right)^{1/2} \le \frac{\alpha K}{\sqrt{1-\beta^2}}\left(\sigma_0^2 + \sigma_1^2 M_\alpha\right)^{1/2},$$

and

$$\mathbb{E}[\|V_k\|_2] \le \alpha\gamma L_\Pi K \sum_{i=0}^{k-1} \beta^{k-1-i}\mathbb{E}\left[\|\Xi_i\|_2\right] \le \frac{\alpha\gamma L_\Pi K}{1-\beta} \cdot \frac{\alpha K}{\sqrt{1-\beta^2}}\left(\sigma_0^2 + \sigma_1^2 M_\alpha\right)^{1/2}.$$

Summing these four estimates and using

$$1 - \beta = \frac{c\alpha}{2}, \qquad 1 - \beta^2 = (1-\beta)(1+\beta) \ge \frac{c\alpha}{2}$$

establishes equation 24. The output-Q error bound follows immediately from $\|Q_k - \mathbf{1}_N \otimes Q^*\|_2 \le \|X_k\|_2$, and the $O(\sqrt{\alpha})$ scaling holds because all constants in the steady-state residual are independent of α for a fixed initial condition. □

A.11 Exact consensus of the disagreement component in ECQ

Theorem 5.1 bounds the total error, which decomposes into the optimality error and the consensus error:

$$\underbrace{Q_k^i - Q^*}_{\text{total error}} = \underbrace{(\bar{Q}_k - Q^*)}_{\text{optimality error}} + \underbrace{(Q_k^i - \bar{Q}_k)}_{\text{consensus error}} = [\mathbf{J}x_k]_i + [(I - \mathbf{J})x_k]_i, \tag{96}$$

which follows from equation 14, Proposition 4.1, and the definition $\mathbf{J} := (\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top) \otimes I_{|\mathcal{S}||\mathcal{A}|}$. Here, $[v]_i$ denotes the ith agent block of a stacked vector v. Then, the $O(\sqrt{\alpha})$ residual from Theorem 5.1 affects both error components, leaving their individual contributions indistinguishable. Therefore, we additionally show that, under the sampling model of Algorithm 1 and a mild structural condition on the local rewards, the consensus error carries no residual and converges to zero geometrically. Specifically, we assume deterministic reward heterogeneity, where local rewards are decomposed as

$$r^i(s, a, s') = R^i_{\mathrm{det}}(s, a) + n(s, a, s'), \qquad \forall i, (s, a, s'), \tag{97}$$

with bounded R^i_{det} and a common noise term n. Equivalently, the reward differences between agents, $r^i - r^j = R^i_{\mathrm{det}} - R^j_{\mathrm{det}}$, depend solely and deterministically on (s, a), while any dependence on s' or common noise can be shared across all agents.

Switching System Formulation. To prove exact consensus, we show that the disagreement entirely vanishes. To derive the disagreement recursion, we apply $\mathbf{U}$ to the error dynamics equation 15, since $\mathbf{U}$ acts as a projection that isolates the disagreement components by definition. We define the reduced disagreement state as

$$\tilde{u}_k := \mathbf{U}u_k, \qquad \tilde{x}_k := \mathbf{U}x_k = \mathbf{U}Q_k, \qquad z_k = \mathbf{U}y_k, \tag{98}$$

using equations 14, 17, Proposition 4.1, and the property $(\mathbf{U}^\top\mathbf{U})(\mathbf{1}_N \otimes v) = 0$ for every $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ from equation 43. To derive the $\tilde{u}$-equation, we apply $\mathbf{U}$ to equation 15. For state-value vectors, define $\mathbf{U}_\mathcal{S} := U \otimes I_{|\mathcal{S}|}$, so that $\mathbf{U}_\mathcal{S}(\mathbf{1}_N \otimes v) = 0$ for every $v \in \mathbb{R}^{|\mathcal{S}|}$ by equation 40. Since $\mathbf{UDP} = (I_{N-1} \otimes DP)\mathbf{U}_\mathcal{S}$, the common value term of the form $1_N \otimes v$ vanishes under $\mathbf{UDP}$, leaving $\mathbf{UDPV}_{\mathbf{Q}_k}$. To linearize the Bellman maximization in $\mathbf{V}_{\mathbf{Q}_k}$, we apply Lemma A.1 agent-wise with $\bar{Q}_k$ in place of Q^*: there exists a joint stochastic policy $\tilde{\boldsymbol{\mu}}_k$ such that $\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V_{\bar{Q}_k} = \mathbf{\Pi}^{\tilde{\boldsymbol{\mu}}_k}(Q_k - \mathbf{1}_N \otimes \bar{Q}_k)$. Since $1_N \otimes V_{\bar{Q}_k}$ is likewise a common value term annihilated by $\mathbf{UDP}$, we obtain $\mathbf{UDPV}_{\mathbf{Q}_k} = \mathbf{UDP\Pi}^{\tilde{\boldsymbol{\mu}}_k}(Q_k - \mathbf{1}_N \otimes \bar{Q}_k)$. Moreover, using equations 42, 96, and 98, we have $Q_k - \mathbf{1}_N \otimes \bar{Q}_k = (I - \mathbf{J})x_k = \mathbf{U}^\top\tilde{x}_k$. Hence, we obtain

$$\mathbf{UDV}_{\mathbf{Q}_k} = \mathbf{UDP\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top\tilde{x}_k. \tag{99}$$

By defining

$$\mathbf{D}_\perp := I_{N-1} \otimes D, \qquad \mathbf{S}_\perp := (ULU^\top) \otimes D \tag{100}$$

such that $\mathbf{UD} = \mathbf{D}_\perp\mathbf{U}$ and $\mathbf{US} = \mathbf{S}_\perp\mathbf{U}$, we get

$$\tilde{u}_{k+1} = \tilde{u}_k + \alpha[\gamma\mathbf{UDP\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top\tilde{x}_k - (\mathbf{D}_\perp + \mathbf{S}_\perp)\tilde{u}_k] + \alpha\mathbf{U}\omega_{k+1}.$$

For the $\tilde{x}$-equation with equation 15, since $\mathbf{US}y_k = \mathbf{S}_\perp\mathbf{U}y_k = \mathbf{S}_\perp z_k$ by definition in equations 98 and 100, we have

$$\tilde{x}_{k+1} = \tilde{x}_k + \alpha\kappa\left[\mathbf{D}_\perp(\tilde{u}_k - \tilde{x}_k) - \mathbf{S}_\perp\tilde{x}_k - \mathbf{S}_\perp z_k\right] + \alpha\kappa\mathbf{U}\nu_{k+1}.$$

The $\tilde{z}$-equation can be driven by multiplying equation 15 by $\mathbf{U}$ and using $\mathbf{US} = \mathbf{S}_\perp\mathbf{U}$. This yields

$$\tilde{X}_k := \mathrm{col}(\tilde{u}_k, \tilde{x}_k, z_k), \qquad \tilde{\eta}_{k+1} := \mathrm{col}(\mathbf{U}\omega_{k+1},\ \kappa\mathbf{U}\nu_{k+1},\ \kappa\mathbf{U}\zeta_{k+1}),$$
$$\tilde{X}_{k+1} = \tilde{A}_{\tilde{\boldsymbol{\mu}}_k}\tilde{X}_k + \alpha\tilde{\eta}_{k+1},\ \tilde{A}_{\boldsymbol{\mu}} := \begin{bmatrix} I - \alpha(\mathbf{D}_\perp + \mathbf{S}_\perp) & \alpha\gamma\,\mathbf{UDP\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top & 0 \\ \alpha\kappa\mathbf{D}_\perp & I - \alpha\kappa(\mathbf{D}_\perp + \mathbf{S}_\perp) & -\alpha\kappa\mathbf{S}_\perp \\ 0 & \alpha\kappa\mathbf{S}_\perp & I \end{bmatrix}, \tag{101}$$

Analytical Framework and Key Lemmas. Since the policy enters $\tilde{A}_{\boldsymbol{\mu}}$ in equation 101 affinely through $\alpha\gamma\,\mathbf{UDP\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top$, we can apply Lemma 5.1 to obtain $\tilde{A}_{\boldsymbol{\mu}} \in \mathrm{co}(\tilde{\mathcal{A}}_\alpha)$, where

$$\tilde{\mathcal{A}}_\alpha := \left\{\tilde{A}_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \Theta^N\right\}. \tag{102}$$

Moreover, $\tilde{A}_{\boldsymbol{\mu}}(\alpha)$ can be written as $\tilde{A}_{\boldsymbol{\mu}}(\alpha) = I + \alpha\tilde{H}_{\boldsymbol{\mu}}$, where $\tilde{H}_{\boldsymbol{\mu}}$ independent of α. Lemma 5.3 is also applicable here because subtracting equation 101 for two distinct policies leaves only the affine policy term. Therefore, we only need to establish the remaining required lemmas using methods similar to those used in proving the finite-time guarantees on the total error for ECQ.

First, to establish the contraction of the disagreement family in a manner similar to Proposition 5.1, we write $\tilde{A}_{\boldsymbol{\pi}} = I + \alpha\tilde{M}_{\boldsymbol{\pi}}$ with

$$\tilde{M}_{\boldsymbol{\pi}} := \begin{bmatrix} -(\mathbf{D}_\perp + \mathbf{S}_\perp) & \gamma\mathbf{UDP\Pi}^{\boldsymbol{\pi}}\mathbf{U}^\top & 0 \\ \kappa\mathbf{D}_\perp & -\kappa(\mathbf{D}_\perp + \mathbf{S}_\perp) & -\kappa\mathbf{S}_\perp \\ 0 & \kappa\mathbf{S}_\perp & 0 \end{bmatrix}, \tag{103}$$

and view the continuous-time system $\dot{\tilde{X}} = \tilde{M}_{\boldsymbol{\pi}}\tilde{X}$ analogously to Figure 2, with κ removed from the lower block. The feedback interconnection consists of the Bellman-estimation block represented as

$$\dot{\tilde{u}} = \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{x}, \qquad \tilde{A}_0 := -\left(\mathbf{D}_\perp + \mathbf{S}_\perp\right), \quad \tilde{B}_0 := \gamma\mathbf{UDP\Pi}^{\boldsymbol{\pi}}\mathbf{U}^\top, \tag{104}$$

and the consensus-mixing block with state $\tilde{v} := \mathrm{col}(\tilde{x}, z)$ given by

$$\dot{\tilde{v}} = \tilde{A}_{\mathrm{mix}}\tilde{v} + \tilde{B}_{\mathrm{mix}}\tilde{u}, \qquad \tilde{x} = \tilde{C}_{\mathrm{mix}}\tilde{v},$$

$$\tilde{A}_{\text{mix}} := \begin{bmatrix} -(\mathbf{D}_\perp + \mathbf{S}_\perp) & -\mathbf{S}_\perp \\ \mathbf{S}_\perp & 0 \end{bmatrix}, \qquad \tilde{B}_{\text{mix}} := \begin{bmatrix} \mathbf{D}_\perp \\ 0 \end{bmatrix}, \qquad \tilde{C}_{\text{mix}} := [I \quad 0]\,. \tag{105}$$

Moreover, from the definition of $\mathbf{D}_\perp$ and $\mathbf{S}_\perp$ in equation 100, we get the following properties:

$$\begin{aligned} &\mathbf{U}^\top \mathbf{D}_\perp = \mathbf{D}\mathbf{U}^\top, \qquad &&\mathbf{U}^\top \mathbf{S}_\perp = \mathbf{S}\mathbf{U}^\top, \\ &\mathbf{S}_\perp = \mathbf{S}_\perp^\top \succ 0, \qquad &&\mathbf{D}_\perp^{-1}\mathbf{S}_\perp = (ULU^\top) \otimes I_{|\mathcal{S}||\mathcal{A}|} = (\mathbf{D}_\perp^{-1}\mathbf{S}_\perp)^\top. \end{aligned} \tag{106}$$

Lemma A.7. *For every deterministic joint policy $\boldsymbol{\pi} \in \Theta^N$, every $u \neq 0$, and all x,*

$$D^+\Phi_U\left(\tilde{u}; \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{x}\right) \leq -d_{\min}\left\|\mathbf{U}^\top\tilde{u}\right\|_\infty + \left\|\mathbf{U}^\top\tilde{B}_0\tilde{x}\right\|_\infty,$$

where $\Phi_U(\tilde{u}) := \|\mathbf{U}^\top\tilde{u}\|_\infty$, and $D^+\Phi_U(\tilde{u}; \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{x})$ follows Definition A.1 by setting $\Phi = \Phi_U$, $\psi = \tilde{u}$ and $\chi = \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{x}$.

Proof. Combining equation 104 and equation 106, we obtain $\mathbf{U}^\top\tilde{A}_0\tilde{u} = -\mathbf{U}^\top(\mathbf{D}_\perp + \mathbf{S}_\perp)\tilde{u} = -(\mathbf{D} + \mathbf{S})\mathbf{U}^\top\tilde{u}$. Since $\mathbf{U}^\top$ is a linear operator, we have $D^+\Phi_U(\tilde{u}; \tilde{A}_0\tilde{u}) = D^+\|\cdot\|_\infty(\mathbf{U}^\top\tilde{u}; \mathbf{U}^\top\tilde{A}_0\tilde{u})$. Here, using equation 106, the directional vector simplifies to $-(\mathbf{D}+\mathbf{S})\mathbf{U}^\top\tilde{u} = f_u(\mathbf{U}^\top\tilde{u}, 0)$, where $f_u(\cdot)$ defined in equation 50. Hence, by Lemma A.3 with $x = 0$, it holds that

$$D^+\Phi_U\left(\tilde{u}; \tilde{A}_0\tilde{u}\right) \leq -d_{\min}\left\|\mathbf{U}^\top\tilde{u}\right\|_\infty.$$

Applying the subadditivity of the Dini derivative, $D^+\Phi_U(\tilde{u}; f + g) \leq D^+\Phi_U(\tilde{u}; f) + \|\mathbf{U}^\top g\|_\infty$, yields the desired conclusion by incorporating the coupling term. □

Lemma A.8. *Under equation 105, $\tilde{A}_{\text{mix}}$ is Hurwitz, and $\tilde{G} := -\tilde{A}_{\text{mix}}^{-1}\tilde{B}_{\text{mix}}$ satisfies*

$$\tilde{G}\tilde{u} = \operatorname{col}\left(0,\ \mathbf{S}_\perp^{-1}\mathbf{D}_\perp\tilde{u}\right), \qquad \textit{implying that} \quad \tilde{C}_{\text{mix}}\tilde{G} = 0.$$

Proof. Consider the quadratic Lyapunov function candidate $\tilde{V}(\tilde{x}, z) := \frac{1}{2}\tilde{x}^\top\mathbf{D}_\perp^{-1}\tilde{x} + \frac{1}{2}z^\top\mathbf{D}_\perp^{-1}z$. Using equations 105–106, the time derivative along the trajectories satisfies

$$\begin{aligned} \dot{\tilde{V}} &= -\tilde{x}^\top\left(\left(I_{N-1} + ULU^\top\right) \otimes I\right)\tilde{x} - \tilde{x}^\top\left(ULU^\top \otimes I\right)z + z^\top\left(ULU^\top \otimes I\right)\tilde{x} + \tilde{x}^\top\tilde{u} \\ &= -\tilde{x}^\top\left(\left(I_{N-1} + ULU^\top\right) \otimes I\right)\tilde{x} + \tilde{x}^\top\tilde{u}, \end{aligned}$$

where the two cross terms cancel in the second equality because they are scalar transposes of each other and $ULU^\top$ is symmetric. Setting $\tilde{u} = 0$ yields $\dot{\tilde{V}} \leq -\|\tilde{x}\|_2^2$, since $(I_{N-1} + ULU^\top) \otimes I \succeq I$ and $L \succeq 0$. For the unforced system $\tilde{u} = 0$, the zero-derivative set is

$$\mathcal{E} := \{(\tilde{x}, z) : \dot{\tilde{V}}(\tilde{x}, z) = 0\} = \{(\tilde{x}, z) : \tilde{x} = 0\}.$$

The equality $\tilde{x} = 0$ at a single time does not imply $\dot{\tilde{x}} = 0$. Instead, consider a trajectory that remains in $\mathcal{E}$. Along such a trajectory, $\tilde{x}(t) \equiv 0$, and equation 105 gives

$$0 = \dot{\tilde{x}}(t) = -(\mathbf{D}_\perp + \mathbf{S}_\perp)\tilde{x}(t) - \mathbf{S}_\perp z(t) = -\mathbf{S}_\perp z(t).$$

Since $\mathbf{S}_\perp \succ 0$, we obtain $z(t) = 0$. Hence, the largest invariant subset of $\mathcal{E}$ is the origin. Moreover, $\tilde{V}$ is positive definite and radially unbounded, and its sublevel sets are compact and positively invariant. LaSalle's invariance principle (Khalil & Grizzle, 2002) therefore yields global asymptotic stability of the unforced linear system, which implies that $\tilde{A}_{\text{mix}}$ is Hurwitz.

At an equilibrium where $\dot{\tilde{v}} = 0$, the second row of $\tilde{A}_{\text{mix}}$ in equation 105 yields $\mathbf{S}_\perp\tilde{x}_\infty = 0$, i.e.,

$$\tilde{x}_\infty = 0.$$

Then, the first row of $\tilde{A}_{\text{mix}}$ gives

$$z_\infty = \mathbf{S}_\perp^{-1}\mathbf{D}_\perp\tilde{u}.$$

Alternatively, setting $0 = \dot{v} = \tilde{A}_{\text{mix}}v + \tilde{B}_{\text{mix}}u$ yields $\tilde{v}_\infty = -\tilde{A}_{\text{mix}}^{-1}\tilde{B}_{\text{mix}}\tilde{u} = \tilde{G}\tilde{u}$, so that

$$\tilde{v}_\infty = \operatorname{col}\left(0,\ \mathbf{S}_\perp^{-1}\mathbf{D}_\perp\tilde{u}\right) = \tilde{G}\tilde{u}, \qquad \tilde{C}_{\text{mix}}\tilde{G}\tilde{u} = 0.$$

Since these results hold for all $\tilde{u}$, this completes the proof. □

Lemma A.9. *There exists a finite threshold $\tilde{\kappa}^* > 0$ such that, for every fixed $\kappa > \tilde{\kappa}^*$, there exists a norm $\tilde{W}_c$ and a constant $\tilde{\lambda}_c > 0$, both independent of $\boldsymbol{\pi} \in \Theta^N$, satisfying*

$$D^+\tilde{W}_c\left(\tilde{X}; \tilde{M}_{\boldsymbol{\pi}}\tilde{X}\right) \le -\tilde{\lambda}_c\tilde{W}_c(\tilde{X}), \qquad \forall \tilde{X}, \quad \forall \boldsymbol{\pi} \in \Theta^N,$$

where $D^+\tilde{W}_c(\tilde{X}; \tilde{M}_{\boldsymbol{\pi}}\tilde{X})$ follows Definition A.1 with choices $\Phi = \tilde{W}_c$, $\psi = \tilde{X}$ and $\chi = \tilde{M}_{\boldsymbol{\pi}}\tilde{X}$, and $\tilde{M}_{\boldsymbol{\pi}}$ is defined as in equation 103.

Proof. We follow a procedure similar to that in Lemma A.5, with $(A_0,\ B_{\boldsymbol{\pi}},\ A_{\mathrm{mix}},\ B_{\mathrm{mix}},\ C_{\mathrm{mix}},\ G,\ \|u\|_\infty)$ replaced by $(\tilde{A}_0, \tilde{B}_0, \tilde{A}_{\mathrm{mix}}, \tilde{B}_{\mathrm{mix}}, \tilde{C}_{\mathrm{mix}}, \tilde{G}, \|\mathbf{U}^\top\tilde{u}\|_\infty)$. Define $\tilde{\epsilon} := \tilde{v} - \tilde{G}\tilde{u}$. Since $\tilde{C}_{\mathrm{mix}}\tilde{G} = 0$ by Lemma A.8, we have $\tilde{x} = \tilde{C}_{\mathrm{mix}}\tilde{\epsilon}$ from equation 105. Performing calculations similar to those in equations 64–65 yields

$$\dot{\tilde{u}} = \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{C}_{\mathrm{mix}}\tilde{\epsilon}, \qquad \dot{\tilde{\epsilon}} = \kappa\tilde{A}_{\mathrm{mix}}\tilde{\epsilon} - \tilde{G}\tilde{A}_0\tilde{u} - \tilde{G}\tilde{B}_0\tilde{C}_{\mathrm{mix}}\tilde{\epsilon}, \tag{107}$$

where we have used $\tilde{C}_{\mathrm{mix}}\tilde{G} = 0$ from Lemma A.8 and $\tilde{A}_{\mathrm{mix}}\tilde{G} + \tilde{B}_{\mathrm{mix}} = 0$ by definition. Thus, in comparison with equations 64–65, the counterpart of $F_{\boldsymbol{\pi}} = A_0 + B_{\boldsymbol{\pi}}\mathbf{J}$ is $\tilde{A}_0$, which is policy independent.

Let $\tilde{P}_m \succ 0$ be the unique solution to $\tilde{A}_{\mathrm{mix}}^\top\tilde{P}_m + \tilde{P}_m\tilde{A}_{\mathrm{mix}} = -I$. Define $\tilde{W}_m(\tilde{\epsilon}) := \sqrt{\tilde{\epsilon}^\top\tilde{P}_m\tilde{\epsilon}}$, $\tilde{\lambda}_m := 1/(2\lambda_{\max}(\tilde{P}_m))$, and

$$\tilde{b}_{12} := \max_{\boldsymbol{\pi}\in\Theta^N}\sup_{\tilde{\epsilon}\neq 0}\frac{\left\|\mathbf{U}^\top\tilde{B}_0\tilde{C}_{\mathrm{mix}}\tilde{\epsilon}\right\|_\infty}{\tilde{W}_m(\tilde{\epsilon})},$$

$$\tilde{b}_{21} := \sup_{\tilde{u}\neq 0}\frac{\tilde{W}_m(\tilde{G}\tilde{A}_0\tilde{u})}{\|\mathbf{U}^\top\tilde{u}\|_\infty}, \qquad \tilde{b}_{22} := \max_{\boldsymbol{\pi}\in\Theta^N}\sup_{\tilde{\epsilon}\neq 0}\frac{\tilde{W}_m(\tilde{G}\tilde{B}_0\tilde{C}_{\mathrm{mix}}\tilde{\epsilon})}{\tilde{W}_m(\tilde{\epsilon})}.$$

Applying Lemma A.7 to the first equation of equation 107, we get

$$D^+\Phi_U(\tilde{u}; \dot{\tilde{u}}) \le -d_{\min}\left\|\mathbf{U}^\top\tilde{u}\right\|_\infty + \tilde{b}_{12}\tilde{W}_m(\tilde{\epsilon}),$$

which holds for all $\tilde{u}$; for $\tilde{u} = 0$, it reduces to the definition of $\tilde{b}_{12}$. Moreover, applying the three-term split to the second equation of equation 107 as in equation 71 gives

$$D^+\tilde{W}_m(\tilde{\epsilon}; \dot{\tilde{\epsilon}}) \le \tilde{b}_{21}\left\|\mathbf{U}^\top\tilde{u}\right\|_\infty - \left(\kappa\tilde{\lambda}_m - \tilde{b}_{22}\right)\tilde{W}_m(\tilde{\epsilon}).$$

Defining $\tilde{W}_c := \|\mathbf{U}^\top\tilde{u}\|_\infty + \theta\tilde{W}_m(\tilde{\epsilon})$ for $\theta > 0$, we obtain

$$D^+\tilde{W}_c\left(\tilde{X}; \tilde{M}_{\boldsymbol{\pi}}\tilde{X}\right) \le -\left(d_{\min} - \theta\tilde{b}_{21}\right)\left\|\mathbf{U}^\top\tilde{u}\right\|_\infty - \left(\theta\left(\kappa\tilde{\lambda}_m - \tilde{b}_{22}\right) - \tilde{b}_{12}\right)\tilde{W}_m(\tilde{\epsilon}).$$

A weighting factor $\theta > 0$ that makes both coefficients strictly positive exists if and only if $\tilde{b}_{12}/(\kappa\tilde{\lambda}_m - \tilde{b}_{22}) < d_{\min}/\tilde{b}_{21}$, which is equivalent to

$$\kappa > \tilde{\kappa}^* := \frac{\tilde{b}_{22} + (\tilde{b}_{12}\tilde{b}_{21})/d_{\min}}{\tilde{\lambda}_m}.$$

In this case, setting $\tilde{\lambda}_c := \min\{d_{\min} - \theta\tilde{b}_{21}, \kappa\tilde{\lambda}_m - \tilde{b}_{22} - \tilde{b}_{12}/\theta\} > 0$ ensures the contraction $D^+\tilde{W}_c \le -\tilde{\lambda}_c\tilde{W}_c$.

It is straightforward to verify that $\tilde{W}_c$ defines a valid norm on the $\tilde{X}$-space because the linear mapping $(\tilde{u}, \tilde{v}) \mapsto (\tilde{u}, \tilde{v} - \tilde{G}\tilde{u})$ is invertible. Furthermore, the pair $(\tilde{W}_c, \tilde{\lambda}_c)$ is independent of $\boldsymbol{\pi}$ since all gain constants are policy-independent. □

Lemma A.10. *Let $\tilde{\kappa}^*$ be as in Lemma A.9. For every fixed $\kappa > \tilde{\kappa}^*$ there is $\tilde{\alpha}(\kappa) > 0$ such that the projected deterministic switching family $\tilde{\mathcal{A}}_\alpha$ with the definition of equation 102 satisfies $\rho(\tilde{\mathcal{A}}_\alpha) < 1$ for every $0 < \alpha < \tilde{\alpha}(\kappa)$. Consequently, for any $\bar{\alpha} \in (0, 1)$ with $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\bar{\alpha}})) < 1$, there exist α-independent constants $\tilde{c} > 0$ and $\tilde{K} < \infty$ such that for all $0 < \alpha \le \bar{\alpha}$,*

$$\sup_{M_0,\ldots,M_{\ell-1}\in\mathrm{co}(\tilde{\mathcal{A}}_\alpha)}\|M_{\ell-1}\cdots M_0\|_2 \le \tilde{K}(1 - \tilde{c}\alpha)^\ell, \qquad \ell \ge 0. \tag{108}$$

Proof. By Lemma A.9, for every fixed $\kappa > \tilde{\kappa}^*$, the finite family $\{\tilde{M}_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \Theta^N\}$ admits a policy-independent pair $(\tilde{W}_c, \tilde{\lambda}_c)$. Since Lemma A.6 relies solely on the existence of such a pair and the affine structure $\tilde{A}_{\boldsymbol{\pi}} = I + \alpha\tilde{M}_{\boldsymbol{\pi}}$, applying it yields an $\tilde{\alpha}(\kappa) > 0$ such that $\tilde{W}_c(\tilde{A}_{\boldsymbol{\pi}}\tilde{X}) \leq (1 - \tilde{\lambda}_c\alpha/2)\tilde{W}_c(\tilde{X})$ for all $\boldsymbol{\pi} \in \Theta^N$ and $0 < \alpha < \tilde{\alpha}(\kappa)$. The product bound equation 108 then follows directly from Lemma 5.2, whose proof requires only $\tilde{A}_{\boldsymbol{\mu}}(\alpha) = I + \alpha\tilde{H}_{\boldsymbol{\mu}}$ with $\tilde{H}_{\boldsymbol{\mu}}$ independent of α, alongside $\tilde{A}_{\boldsymbol{\mu}} \in \mathrm{co}(\tilde{\mathcal{A}}_\alpha)$ and $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\alpha,\perp})) = \rho(\tilde{\mathcal{A}}_{\alpha,\perp})$ from Lemma 5.1. □

Lemma A.11. *Under Assumption 5.1, let Z_∞ be chosen to satisfy equation 16. Then, there exists an α-independent constant $\tilde{\sigma}_1 < \infty$ such that for all $k \geq 0$,*

$$\mathbb{E}\left[\tilde{\eta}_{k+1} \mid \mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left[\|\tilde{\eta}_{k+1}\|_2^2 \mid \mathcal{F}_k\right] \leq \tilde{\sigma}_1^2 \left\|\tilde{X}_k\right\|_2^2,$$

where $\mathcal{F}_k = \sigma(\mathbf{W}_0, \mathbf{Q}_0, \mathbf{Z}_0, \{(s_t, a_t, s'_t, r^1_{t+1}, \ldots, r^N_{t+1}) : 0 \leq t \leq k-1\})$ is the natural filtration.

Proof. The first claim follows from Lemma 5.4 together with the fact that $\mathbf{U}$ is deterministic.

For the second claim, we show that the norm of $\mathbf{U}\hat{g}_{W,k}$, $\mathbf{U}\hat{g}_{Q,k}$, and $\mathbf{U}\hat{g}_{Z,k}$ with equations 25–27 are bounded by a constant multiple of $\tilde{X}_k$ without an additive constant. From the second equality of Proposition 4.1, we have $(I_N + L)W_\infty(s,a) = R(s,a) + \gamma\mathbf{1}_N(P_aV^*)(s)$, which implies

$$U(I_N + L)W_\infty(s,a) = UR_{\mathrm{det}}(s,a)$$

by using equation 40 and $r_{k+1} = R_{\mathrm{det}}(s_k,a_k) + n(s_k,a_k,s'_k)\mathbf{1}_N$ from equation 97. Since $U(I_N + L) = U + UL = U + ULU^\top U = (I_{N-1} + ULU^\top)U$ together with equation 41 and $L\mathbf{1}_N = 0$ from equation 1, we obtain

$$\begin{aligned} U(I_N + L)W_k(s_k,a_k) =& U(I_N + L)\left(u_k(s_k,a_k) + W_\infty(s_k,a_k)\right) \\ =& (I_{N-1} + ULU^\top)\tilde{u}_k(s_k,a_k) + UR_{\mathrm{det}}(s,a) \end{aligned} \tag{109}$$

using equation 14 and equation 98. Multiplying $\hat{g}_{W,k}$ in equation 25 by $\mathbf{U}$ and setting $\delta_k := r_{k+1} + \gamma V_{Q_k}(s'_k) - W_k(s_k,a_k) - LW_k(s_k,a_k)$, we get, almost surely,

$$\begin{aligned} \|\mathbf{U}\hat{g}_{W,k}\|_2 =& \|(U \otimes I_{|\mathcal{S}||\mathcal{A}|})(I_N \otimes e_k)\delta_k\|_2 \\ =& \|(U \otimes e_k)\delta_k\|_2 \\ =& \|(I_{N-1} \otimes e_k)(U \otimes 1)\delta_k\|_2 \\ =& \|U\delta_k\|_2 \\ =& \left\|U\left(r_{k+1} + \gamma V_{Q_k}(s'_k) - W_k(s_k,a_k) - LW_k(s_k,a_k)\right)\right\|_2 \\ =& \left\|UR_{\mathrm{det}}(s_k,a_k) + U\left(\gamma V_{Q_k}(s'_k) - W_k(s_k,a_k) - LW_k(s_k,a_k)\right)\right\|_2 \\ =& \left\|UR_{\mathrm{det}}(s_k,a_k) + \gamma\left[\mathbf{U}_{\mathcal{S}}\boldsymbol{\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top\tilde{x}_k\right](s'_k) + U\left(-W_k(s_k,a_k) - LW_k(s_k,a_k)\right)\right\|_2 \\ =& \left\|-\left(I_{N-1} + ULU^\top\right)\tilde{u}_k(s_k,a_k) + \gamma\left[\mathbf{U}_{\mathcal{S}}\boldsymbol{\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top\tilde{x}_k\right](s'_k)\right\|_2 \\ \leq& \left\|I_{N-1} + ULU^\top\right\|_2 \|\tilde{u}_k(s_k,a_k)\|_2 + \gamma\|\mathbf{U}_{\mathcal{S}}\boldsymbol{\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top\|_2 \|\tilde{x}_k\|_2 \\ \leq& (1 + \|L\|_2)\|\tilde{u}_k\|_2 + \gamma\|\tilde{x}_k\|_2 \\ \leq& (1 + 2\max_i |\mathcal{N}_i|)\|\tilde{u}_k\|_2 + \gamma\|\tilde{x}_k\|_2 \\ \leq& (1 + 2\max_i |\mathcal{N}_i| + \gamma)\left\|\tilde{X}_k\right\|_2, \end{aligned} \tag{110}$$

where the first equality uses $\hat{g}_{W,k} = (I_N \otimes e_k)\delta_k$, the third equality follows from the mixed-product property of the Kronecker product, and the fourth equality is due to $\|e_k\|_2 = 1$. The sixth equality utilizes $r_{k+1} = R_{\mathrm{det}}(s_k,a_k) + n(s_k,a_k,s'_k)\mathbf{1}_N$ from equation 97 and $U\mathbf{1}_N = 0$ from equation 40, the seventh equality follows from equations 98–99 with the definition $\mathbf{U}_{\mathcal{S}} := U \otimes I_{|\mathcal{S}|}$, the eighth equality comes from equation 109, and the first inequality uses the triangle inequality and the fact that extracting a state block cannot increase the Euclidean norm. Moreover, the second inequality uses $\|\mathbf{U}_{\mathcal{S}}\|_2 = \|\mathbf{U}^\top\|_2 = 1$ and $\|\boldsymbol{\Pi}^{\tilde{\boldsymbol{\mu}}_k}\|_2 \leq 1$, and the third inequality follows from the definition of L.

For $\hat{g}_{Q,k}$, we have, almost surely,

$$\|\mathbf{U}\hat{g}_{Q,k}\|_2$$

$$
\begin{aligned}
&= \|U\left(W_k(s_k,a_k) - Q_k(s_k,a_k) - LQ_k(s_k,a_k) - LZ_k(s_k,a_k)\right)\|_2 \\
&= \|UW_\infty(s_k,a_k) + \tilde{u}_k(s_k,a_k) - \tilde{x}_k(s_k,a_k) \\
&\quad - ULU^\top \tilde{x}_k(s_k,a_k) - ULZ_\infty(s_k,a_k) - ULU^\top z_k(s_k,a_k)\|_2 \\
&= \|\tilde{u}_k(s_k,a_k) - \tilde{x}_k(s_k,a_k) - ULU^\top \tilde{x}_k(s_k,a_k) - ULU^\top z_k(s_k,a_k)\|_2 \\
&\le \|\tilde{u}_k(s_k,a_k)\|_2 + \|\tilde{x}_k(s_k,a_k)\|_2 + \|L\|_2\|\tilde{x}_k(s_k,a_k)\|_2 + \|L\|_2\|z_k(s_k,a_k)\|_2 \\
&\le (2 + 4\max_i |\mathcal{N}_i|)\left\|\tilde{X}_k\right\|_2,
\end{aligned}
$$

where the second equality uses equations 14, 98, and $UL = ULU^\top U$ along with equation 41 and $L\mathbf{1}_N = 0$ from equation 1, the third equality follows from $LZ_\infty = W_\infty - Q^*\mathbf{1}_N$ from equation 13 together with $U\mathbf{1}_N = 0$ from equation 40, the first inequality applies the triangle inequality with $\|\mathbf{U}\|_2 = 1$, and the last inequality follows from the definition of $\tilde{L}$.

Similarly, $\hat{g}_{Z,k}$ satisfies, almost surely,

$$
\begin{aligned}
\|\mathbf{U}\hat{g}_{Z,k}\|_2 &= \|ULQ_k(s_k,a_k)\|_2 \\
&= \left\|ULU^\top \tilde{x}_k(s_k,a_k)\right\|_2 \\
&\le \|L\|_2\|\tilde{x}_k(s_k,a_k)\|_2 \\
&\le 2\max_i |\mathcal{N}_i|\left\|\tilde{X}_k\right\|_2,
\end{aligned}
$$

where the second equality uses equation 98 and $UL = ULU^\top U$ along with equation 41 and $L\mathbf{1}_N = 0$ from equation 1, the first inequality uses $\|\mathbf{U}\|_2 = 1$, and the last inequality follows from the definition of L.

Finally, combining these bounds with $\mathbb{E}[\|\tilde{\eta}_{k+1}\|_2^2 \mid \mathcal{F}_k] \le \mathbb{E}[\|\mathbf{U}\widehat{g}_{W,k}\|_2^2 + \kappa^2\|\mathbf{U}\widehat{g}_{Q,k}\|_2^2 + \kappa^2\|\mathbf{U}\widehat{g}_{Z,k}\|_2^2 \mid \mathcal{F}_k]$ via the procedure similar to that in Lemma 5.4, we obtain $\mathbb{E}[\|\tilde{\eta}_{k+1}\|_2^2 \mid \mathcal{F}_k] \le \tilde{\sigma}_1^2\|\tilde{X}_k\|_2^2$, where

$$
\tilde{\sigma}_1^2 := (1 + 2\max_i |\mathcal{N}_i| + \gamma)^2 + \kappa^2(2 + 4\max_i |\mathcal{N}_i|)^2 + \kappa^2(2\max_i |\mathcal{N}_i|)^2.
$$

This completes the proof. □

Consensus Error Convergence. Now, we show that the consensus error contains no residual term and converges to zero geometrically.

Corollary 5.1 Restated. Under Assumption 5.1 and deterministic reward heterogeneity, let Z_∞ be chosen to satisfy equation 16, and suppose there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\bar{\alpha}})) < 1$. Then, there exist constants $\tilde{C}_0 < \infty$, $\tilde{c} > 0$, and $\tilde{\alpha}_0 \in (0, \bar{\alpha}]$, such that for every $0 < \alpha \le \tilde{\alpha}_0$ and $k \ge 0$, the consensus error satisfies

$$
\mathbb{E}\left[\max_i \left\|Q_k^i - \bar{Q}_k\right\|_\infty\right] \le \sqrt{2\tilde{C}_0}\left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k/2}\left\|\tilde{X}_0\right\|_2
$$

with $\bar{Q}_k = \frac{1}{N}\sum_i Q_k^i$ denoting the average Q-function across agents, which decays geometrically as $k \to \infty$ without any α-dependent residual error.

Proof. Let $\tilde{c}, \tilde{K}$ be as in Lemma A.10, $\tilde{\sigma}_1$ as in Lemma A.11, and L_Π as in equation 21. Define the constants

$$
\tilde{C}_0 := 1 + 4\tilde{K}^2 + 64\frac{\gamma^2 L_\Pi^2 \tilde{K}^4}{\tilde{c}^2}, \quad \tilde{C}_1 := 16\left(\frac{\tilde{K}^2}{\tilde{c}} + 4\frac{\gamma^2 L_\Pi^2 \tilde{K}^4}{\tilde{c}^3}\right), \quad \tilde{\alpha}_0 := \min\left\{\bar{\alpha}, \frac{1}{2\tilde{C}_1\tilde{\sigma}_1^2}\right\},
$$

with the convention $\tilde{\alpha}_0 = \bar{\alpha}$ if $\tilde{\sigma}_1 = 0$. Fix $0 < \alpha \le \tilde{\alpha}_0$ and the joint stochastic policy $\bar{\boldsymbol{\mu}}$. As in the proof of Proposition 5.2, define

$$
\begin{aligned}
&Y_{k+1} = \tilde{A}_{\bar{\boldsymbol{\mu}}}Y_k + \alpha\tilde{\eta}_{k+1}, \qquad \bar{Y}_{k+1} = \tilde{A}_{\bar{\boldsymbol{\mu}}}\bar{Y}_k, \qquad \Xi_{k+1} = \tilde{A}_{\bar{\boldsymbol{\mu}}}\Xi_k + \alpha\tilde{\eta}_{k+1}, \\
&U_{k+1} = \tilde{A}_{\tilde{\boldsymbol{\mu}}_k}U_k + (\tilde{A}_{\tilde{\boldsymbol{\mu}}_k} - \tilde{A}_{\bar{\boldsymbol{\mu}}})\bar{Y}_k, \qquad V_{k+1} = \tilde{A}_{\tilde{\boldsymbol{\mu}}_k}V_k + (\tilde{A}_{\tilde{\boldsymbol{\mu}}_k} - \tilde{A}_{\bar{\boldsymbol{\mu}}})\Xi_k,
\end{aligned}
$$

with

$$Y_0 = \tilde{X}_0, \qquad \bar{Y}_0 = \tilde{X}_0, \qquad \Xi_0 = 0, \qquad U_0 = 0, \qquad V_0 = 0,$$

so that $\tilde{X}_k = \bar{Y}_k + \Xi_k + U_k + V_k$. Applying Lemmas 5.3, A.10, and A.11, we obtain the following inequalities, corresponding to equations 87, 88, 91, and 94:

$$\left\|\bar{Y}_k\right\|_2 \leq \tilde{K}(1-\tilde{c}\alpha)^k \left\|\tilde{X}_0\right\|_2, \tag{111}$$

$$\mathbb{E}\left[\|\Xi_k\|_2^2\right] \leq \alpha^2 \tilde{K}^2 \tilde{\sigma}_1^2 \sum_{t=0}^{k-1} (1-\tilde{c}\alpha)^{2(k-1-t)} \mathbb{E}\left[\left\|\tilde{X}_t\right\|_2^2\right], \tag{112}$$

$$\|U_k\|_2 \leq \alpha\gamma L_\Pi \tilde{K}^2 k (1-\tilde{c}\alpha)^{k-1} \left\|\tilde{X}_0\right\|_2, \tag{113}$$

$$\mathbb{E}\left[\|V_k\|_2^2\right] \leq \frac{\alpha\gamma^2 L_\Pi^2 \tilde{K}^2}{\tilde{c}} \sum_{i=0}^{k-1} (1-\tilde{c}\alpha)^{k-1-i} \mathbb{E}\left[\|\Xi_i\|_2^2\right]. \tag{114}$$

The only modification with respect to the proof of Proposition 5.2 is that $\sigma_0^2 + \sigma_1^2 \mathbb{E}\|X_t\|_2^2$ is replaced by $\tilde{\sigma}_1^2 \mathbb{E}\|\tilde{X}_t\|_2^2$ in equation 112.

Now, we prove by induction that $\mathbb{E}[\|\tilde{X}_t\|_2^2] \leq 2\tilde{C}_0(1-\tilde{c}\alpha/2)^t \|\tilde{X}_0\|_2^2$ for all $t \geq 0$. The base case $t = 0$ holds trivially since $\tilde{C}_0 \geq 1$. Assume that the bound holds for all $t \leq k-1$. Note that $(1-\tilde{c}\alpha)^2 \leq 1 - \tilde{c}\alpha \leq 1 - \tilde{c}\alpha/2$ and

$$\frac{1-\tilde{c}\alpha}{1-\tilde{c}\alpha/2} = 1 - \frac{\tilde{c}\alpha/2}{1-\tilde{c}\alpha/2} \leq 1 - \frac{\tilde{c}\alpha}{2}, \qquad 1 - \frac{\tilde{c}\alpha}{2} \geq \frac{1}{2}, \tag{115}$$

where the second inequality holds because $\tilde{c}\alpha \leq \tilde{c}\bar{\alpha} < 1$ by the construction of $\tilde{c}$ in Lemma A.10. Moreover, we utilize the elementary bounds

$$\begin{aligned}
\sum_{t=0}^{k-1} (1-\tilde{c}\alpha)^{k-1-t} \left(1 - \frac{\tilde{c}\alpha}{2}\right)^t \leq & \frac{(1-\tilde{c}\alpha/2)^{k-1}}{1-(1-\tilde{c}\alpha)/(1-\tilde{c}\alpha/2)} \\
\leq & \frac{(1-\tilde{c}\alpha/2)^{k-1}}{\tilde{c}\alpha/2} \\
\leq & \frac{4\,(1-\tilde{c}\alpha/2)^k}{\tilde{c}\alpha},
\end{aligned} \tag{116}$$

where the second inequality follows from equation 115, and the third inequality holds since $(1-\tilde{c}\alpha/2)^{k-1} \leq 2(1-\tilde{c}\alpha/2)^k$ by equation 115. Similarly,

$$\begin{aligned}
k(1-\tilde{c}\alpha)^{k-1} = & k \left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k-1} \left(\frac{1-\tilde{c}\alpha}{1-\tilde{c}\alpha/2}\right)^{k-1} \\
\leq & k \left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k-1} \left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k-1} \\
\leq & \frac{(1-\tilde{c}\alpha/2)^{k-1}}{\tilde{c}\alpha/2} \\
\leq & \frac{4(1-\tilde{c}\alpha/2)^k}{\tilde{c}\alpha},
\end{aligned} \tag{117}$$

where the first inequality follows from equation 115, the second inequality holds because $k(1-\tilde{c}\alpha/2)^{k-1} \leq \sum_{i=0}^{k-1}(1-\tilde{c}\alpha/2)^i \leq 1/(\tilde{c}\alpha/2)$, and the last inequality follows from equation 116. Substituting the inductive hypothesis into equation 112 and applying $(1-\tilde{c}\alpha)^2 \leq 1 - \tilde{c}\alpha$ together with equation 116 yields

$$\begin{aligned}
\mathbb{E}\left[\|\Xi_k\|_2^2\right] \leq & 2\alpha^2 \tilde{K}^2 \tilde{\sigma}_1^2 \tilde{C}_0 \|\tilde{X}_0\|_2^2 \sum_{t=0}^{k-1} (1-\tilde{c}\alpha)^{k-1-t} \left(1 - \frac{\tilde{c}\alpha}{2}\right)^t \\
\leq & \frac{8\alpha \tilde{K}^2 \tilde{\sigma}_1^2}{\tilde{c}} \tilde{C}_0 \left(1 - \frac{\tilde{c}\alpha}{2}\right)^k \|\tilde{X}_0\|_2^2.
\end{aligned} \tag{118}$$

From equation 113 and equation 117, we obtain $\|U_k\|_2 \le \alpha\gamma L_\Pi \tilde{K}^2 \cdot \frac{4(1-\tilde{c}\alpha/2)^k}{\tilde{c}\alpha}\|\tilde{X}_0\|_2$, and therefore

$$\|U_k\|_2^2 \le \frac{16\gamma^2 L_\Pi^2 \tilde{K}^4}{\tilde{c}^2}\left(1-\frac{\tilde{c}\alpha}{2}\right)^{2k}\left\|\tilde{X}_0\right\|_2^2 \le \frac{16\gamma^2 L_\Pi^2 \tilde{K}^4}{\tilde{c}^2}\left(1-\frac{\tilde{c}\alpha}{2}\right)^{k}\left\|\tilde{X}_0\right\|_2^2.$$

Applying equation 114, equation 118 for indices $i \le k-1$, and equation 116, we get

$$\begin{aligned}\mathbb{E}\left[\|V_k\|_2^2\right] \le& \frac{\alpha\gamma^2 L_\Pi^2 \tilde{K}^2}{\tilde{c}} \cdot \frac{8\alpha\tilde{K}^2\tilde{\sigma}_1^2}{\tilde{c}}\tilde{C}_0\|\tilde{X}_0\|_2^2 \sum_{i=0}^{k-1}(1-\tilde{c}\alpha)^{k-1-i}\left(1-\frac{\tilde{c}\alpha}{2}\right)^i \\ \le& \frac{32\alpha\gamma^2 L_\Pi^2 \tilde{K}^4\tilde{\sigma}_1^2}{\tilde{c}^3}\tilde{C}_0\left(1-\frac{\tilde{c}\alpha}{2}\right)^k\left\|\tilde{X}_0\right\|_2^2.\end{aligned}$$

Finally, from equation 111, we have $\|\bar{Y}_k\|_2^2 \le \tilde{K}^2(1-\tilde{c}\alpha)^{2k}\|\tilde{X}_0\|_2^2 \le \tilde{K}^2(1-\tilde{c}\alpha/2)^k\|\tilde{X}_0\|_2^2$. Applying Cauchy-Schwarz via $\|a+b+c+d\|_2^2 \le 4(\|a\|_2^2+\|b\|_2^2+\|c\|_2^2+\|d\|_2^2)$ and combining the four terms, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\tilde{X}_k\|_2^2\right] \le& \left(4\tilde{K}^2 + \frac{64\gamma^2 L_\Pi^2\tilde{K}^4}{\tilde{c}^2}\right)\left(1-\frac{\tilde{c}\alpha}{2}\right)^k\left\|\tilde{X}_0\right\|_2^2 \\ &+ \alpha\tilde{\sigma}_1^2\left(\frac{32\tilde{K}^2}{\tilde{c}} + \frac{128\gamma^2 L_\Pi^2\tilde{K}^4}{\tilde{c}^3}\right)\tilde{C}_0\left(1-\frac{\tilde{c}\alpha}{2}\right)^k\left\|\tilde{X}_0\right\|_2^2 \\ \le& \left(\tilde{C}_0 + 2\alpha\tilde{\sigma}_1^2\tilde{C}_0\tilde{C}_1\right)\left(1-\frac{\tilde{c}\alpha}{2}\right)^k\left\|\tilde{X}_0\right\|_2^2.\end{aligned}$$

For $0 < \alpha \le \tilde{\alpha}_0 \le 1/(2\tilde{C}_1\tilde{\sigma}_1^2)$, the second coefficient is bounded above by $\tilde{C}_0$. Thus, $\mathbb{E}[\|\tilde{X}_k\|_2^2] \le 2\tilde{C}_0(1-\tilde{c}\alpha/2)^k\|\tilde{X}_0\|_2^2$, which closes the induction and proves

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \le 2\tilde{C}_0\left(1-\frac{\tilde{c}\alpha}{2}\right)^k\left\|\tilde{X}_0\right\|_2^2.$$

The bound on the consensus error follows from $\max_i \|Q_k^i - \bar{Q}_k\|_\infty \le \|(I-\mathbf{J})x_k\|_2 = \|\tilde{x}_k\|_2 \le \|\tilde{X}_k\|_2$ using equation 42, equation 96, $\|\mathbf{U}^\top\|_2 = 1$, and Jensen's inequality. □

Corollary 5.1 gives a geometrically vanishing consensus error, and Theorem 5.1 gives an $O(\sqrt{\alpha})$ upper bound on the limit superior of the expected total error for admissible step-sizes under deterministic reward heterogeneity. These two results have different status. The consensus bound vanishes, so it pins the limiting consensus error to zero. The total error bound does not vanish, so it establishes neither a matching lower bound nor the existence of a limit nor a positive error floor; a total error of $O(\alpha)$ or smaller in specific instances is not excluded. Residual total-error fluctuations can arise from next-state sampling in the Bellman target even with deterministic rewards and from reward sampling when present.

The comparison with Lim & Lee (2025) concerns consensus only. Their algorithm lacks the integral action of Z, and their upper bound on the consensus error retains an $O(\alpha)$ residual term; ours is residual-free. This is a comparison of guarantees, not of true residuals: since their bound does not vanish, it neither proves nor rules out exact consensus for their algorithm. No comparison of total error is implies as both analyses provide only upper bounds for that quantity.

A.12 Reduced version of ECQ

In this appendix, we introduce a two-variable variant of ECQ. R-ECQ retains W as the local learning variable and Q as the network agreement variable but uses different Q-dynamics rather than a trajectory-equivalent algebraic elimination of the auxiliary state Z. Unlike Algorithm 1, the reduced recursion requires a specific initialization condition.

Algorithm 2 Reduced Exact-Consensus Q-learning (R-ECQ)

1: Choose a constant step-size $\alpha \in (0,1)$ and a mixing gain $\kappa > 0$.
2: Initialize $W_0^i, Q_0^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for all agents $i = 1, \dots, N$ so that $\sum_{i=1}^N (Q_0^i - W_0^i) = 0$.
3: **for** $k = 0, 1, 2, \dots$ **do**
4: Sample (s_k, a_k, s_k'). Each agent i observes its local reward $r_{k+1}^i = r^i(s_k, a_k, s_k')$.
5: **for** each agent i **do**
6: Update W_k^i and Q_k^i at (s_k, a_k) using equations 5 and 119, respectively.
7: Leave all other coordinates unchanged.
8: **end for**
9: **end for**

Reduced Exact-Consensus Q-learning Algorithm. In Algorithm 2, the learning variable W^i follows the same update rule as in equation 5. In contrast, the update for Q^i is given by

$$\begin{aligned} Q_{k+1}^i(s_k, a_k) =& Q_k^i(s_k, a_k) \\ &+ \alpha \left(\underbrace{r_{k+1}^i + \gamma \max_{a' \in \mathcal{A}} Q_k^i(s_k', a') - W_k^i(s_k, a_k) - \sum_{j=1}^N L_{ij} W_k^j(s_k, a_k)}_{\text{innovation shared with the } W\text{-update}} \right) \\ &\underbrace{- \alpha\kappa \sum_{j=1}^N L_{ij} Q_k^j(s_k, a_k)}_{\text{consensus correction}} . \end{aligned} \tag{119}$$

Here, the second term in equation 119 is identical to the innovation term used in the update of W^i. Therefore, both W^i and Q^i share the same local learning signal for each sampled state-action pair. The only additional term in the Q^i-update is $-\alpha\kappa \sum_{j=1}^N L_{ij} Q_k^j(s_k, a_k)$, which reduces disagreement among neighboring agents' Q-values. The gain $\kappa > 0$ scales the consensus update and is used later to establish convergence.

These update equations define a two-variable algorithm without an explicit Z-state. Subtracting the Q^i-update equation 119 from the W^i-update equation 5 yields

$$(W_{k+1}^i - Q_{k+1}^i) = \left(W_k^i - Q_k^i\right) + \alpha\kappa e_k e_k^\top \sum_{j=1}^N L_{ij} Q_k^j. \tag{120}$$

Consequently, the gap $W_k^i - Q_k^i$ evolves purely through the consensus disagreement in Q^i. The Laplacian term $\sum_{j=1}^N L_{ij} Q_k^j$ accumulates in this gap over time, so the gap serves as an integral of disagreement, analogous to the auxiliary state Z^i in Algorithm 1. This analogy does not imply that the W- and Q-trajectories of the two algorithms coincide.

In Algorithm 2, the initialization condition constrains only the network average of the initial gap; individual initial gaps may be nonzero. Summing equation 120 over all agents and applying $\mathbf{1}_N^\top L = 0$ from equation 1 yields

$$\sum_{i=1}^N (W_{k+1}^i - Q_{k+1}^i) = \sum_{i=1}^N (W_k^i - Q_k^i), \tag{121}$$

and setting $\sum_{i=1}^N \left(W_0^i - Q_0^i\right) = 0$ guarantees that $\sum_{i=1}^N \left(W_k^i - Q_k^i\right) = 0$ for all k. Thus, the network average of the gap $\frac{1}{N} \sum_{i=1}^N \left(W_k^i - Q_k^i\right)$ remains zero throughout the recursion, ensuring that the gap carries only the disagreement information across the agents. In general, the gap retains the initial disagreement $W_0^i - Q_0^i$ in addition to the accumulated consensus increments. The stronger choice $W_0^i = Q_0^i$ for every agent removes this initial term.

Equilibrium and Switching-System Formulation. In terms of the stacked variables from Section 4.2, Algorithm 2 is expressed in stacked form as

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \alpha\left(\mathbf{DR} + \gamma\mathbf{DPV}_{\mathbf{Q}_k} - \mathbf{DW}_k - \mathbf{SW}_k\right) + \alpha\omega_{k+1}, \tag{122}$$

$$\mathbf{Q}_{k+1} = \mathbf{Q}_k + \alpha\left(\mathbf{DR} + \gamma\mathbf{DPV}_{\mathbf{Q}_k} - \mathbf{DW}_k - \mathbf{SW}_k - \kappa\mathbf{SQ}_k\right) + \alpha\nu_{k+1}. \tag{123}$$

As in Appendix A.1, the noise terms represent the sampled updates minus their conditional expectations, satisfying

$$\mathbb{E}[\omega_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\nu_{k+1} \mid \mathcal{F}_k] = 0.$$

Since the noise decomposition follows the ECQ case identically, it is omitted here.

We next characterize the equilibrium of the reduced mean dynamics. The proposition below establishes that for any equilibrium of R-ECQ in the invariant subspace specified by its initialization condition, its Q-component matches the centralized team-optimal Q-function exactly.

Proposition A.1. *Let Q^* be the optimal Q-function associated with the team-average reward, and let V^* be the corresponding optimal value function defined by $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$. Then, any equilibrium $(\mathbf{W}_\infty, \mathbf{Q}_\infty)$ of the mean dynamics restricted to $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})(\mathbf{Q} - \mathbf{W}) = 0$ satisfies*

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes Q^*, \qquad (\mathbf{D} + \mathbf{S})\mathbf{W}_\infty = \mathbf{DR} + \gamma\mathbf{DP}(\mathbf{1}_N \otimes V^*).$$

Proof. At equilibrium, subtracting the conditional mean dynamics of equation 123 from those of equation 122 yields

$$(\mathbf{W}_\infty - \mathbf{Q}_\infty) = (\mathbf{W}_\infty - \mathbf{Q}_\infty) + \kappa\mathbf{SQ}_\infty,$$

which implies $\mathbf{SQ}_\infty = 0$. Since $D \succ 0$ from equation 2 and $\mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}$ from equation 1, this ensures that

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes q \tag{124}$$

for some $q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. By the stated invariant-subspace restriction, $\sum_{i=1}^N (W_\infty^i - Q_\infty^i) = 0$. Combining this equality with equation 124 gives

$$\frac{1}{N}\sum_{i=1}^N W_\infty^i = \frac{1}{N}\sum_{i=1}^N Q_\infty^i = q \tag{125}$$

at equilibrium. Moreover, from the conditional mean dynamics of equation 122, we get

$$\mathbf{DR} + \gamma\mathbf{DPV}_{\mathbf{Q}_\infty} - \mathbf{DW}_\infty - \mathbf{SW}_\infty = 0. \tag{126}$$

Expanding this equation for each agent gives

$$DR^i + \gamma DPV_{Q_\infty^i} - DW_\infty^i - D\sum_{j=1}^N L_{ij}W_\infty^j = 0, \qquad i = 1, \ldots, N.$$

Summing this equation over all agents and dividing by N, we obtain

$$D\left(\frac{1}{N}\sum_{i=1}^N R^i\right) + \gamma DP\left(\frac{1}{N}\sum_{i=1}^N V_{Q_\infty^i}\right) - D\left(\frac{1}{N}\sum_{i=1}^N W_\infty^i\right) - D\left(\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N L_{ij}W_\infty^j\right) = 0.$$

Since $\sum_i L_{ij} = 0$ by the definition of the graph Laplacian L, we can eliminate the consensus term and therefore

$$DR + \gamma DP\left(\frac{1}{N}\sum_{i=1}^N V_{Q_\infty^i}\right) - D\left(\frac{1}{N}\sum_{i=1}^N W_\infty^i\right) = 0$$

with equation 2. Substituting $Q_\infty^i = q$ for every i from equation 124 and applying equation 125, we have

$$DR + \gamma DPV_q - Dq = 0,$$

which leads to $q = R + \gamma PV_q$ by removing D. Therefore,

$$\mathbf{Q}_\infty = \mathbf{1}_N \otimes Q^*. \tag{127}$$

Thus, on the stated invariant subspace, R-ECQ has the same team-optimal Q-equilibrium as Algorithm 1, despite using different Q-dynamics. Finally, combining equation 126 with equation 127 yields the remaining claim statement, completing the proof. □

Based on the equilibrium derived in Proposition A.1, we now transform the recursion into error coordinates to establish convergence. Define

$$u_k := \mathbf{W}_k - \mathbf{W}_\infty, \qquad x_k := \mathbf{Q}_k - \mathbf{1}_N \otimes Q^*, \qquad h_k := x_k - u_k. \tag{128}$$

Subtracting the corresponding equilibrium from equations 122–123 yields

$$\begin{aligned} u_{k+1} =& u_k + \alpha\left[\gamma \mathbf{DP}\left(\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V^*\right) - (\mathbf{D}+\mathbf{S})u_k\right] + \alpha\omega_{k+1}, \\ x_{k+1} =& x_k + \alpha\left[\gamma \mathbf{DP}\left(\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V^*\right) - (\mathbf{D}+\mathbf{S})u_k - \kappa\mathbf{S}x_k\right] + \alpha\nu_{k+1}. \end{aligned} \tag{129}$$

Combining the above equations, we get

$$\begin{aligned} h_{k+1} =& x_{k+1} - u_{k+1} \\ =& x_k - u_k - \alpha\kappa\mathbf{S}x_k + \alpha\nu_{k+1} - \alpha\omega_{k+1} \\ =& h_k - \kappa\alpha\mathbf{S}x_k + \alpha(\nu_{k+1} - \omega_{k+1}). \end{aligned} \tag{130}$$

As in Appendix A.3, we also verify whether there is any redundant component in Algorithm 2. From equation 128, multiplying $(\mathbf{1}_N^\top \otimes I)$ by h_k yields

$$(\mathbf{1}_N^\top \otimes I)h_k = \sum_{i=1}^N (Q_k^i - W_k^i) - NQ^* + \sum_{i=1}^N W_\infty^i.$$

Here, the initialization condition guarantees $\sum_{i=1}^N \left(W_k^i - Q_k^i\right) = 0$ for all k via equation 121. Moreover, multiplying the second relation in Proposition A.1 by $(\mathbf{1}_N^\top \otimes I_{|\mathcal{S}||\mathcal{A}|})$ gives $\sum_{i=1}^N W_\infty^i = N(R+\gamma PV^*) = NQ^*$ since $\mathbf{1}_N^\top L = 0$ along with equation 9 and $R+\gamma PV^* = Q^*$ by the Bellman optimality equation. Therefore, $(\mathbf{1}_N^\top \otimes I)h_k = 0$, implying that h_k contains no common component shared among agents. Consequently, we can remove this common direction and represent h_k using only the remaining $N-1$ directions without loss of information. Using the projection matrix $\mathbf{U}$ from equation 17, we project h_k onto this subspace as

$$z_k := \mathbf{U}h_k, \qquad h_k = \mathbf{U}^\top z_k. \tag{131}$$

Left-multiplying equation 130 by $\mathbf{U}$ and using $x_k = u_k + \mathbf{U}^\top z_k$ yields

$$z_{k+1} = z_k - \alpha\kappa\mathbf{U}\mathbf{S}u_k - \alpha\kappa\mathbf{U}\mathbf{S}\mathbf{U}^\top z_k + \alpha\mathbf{U}(\nu_{k+1} - \omega_{k+1}). \tag{132}$$

Meanwhile, by Lemma A.1, there exists a joint stochastic policy μ_k satisfying $\mathbf{V}_{\mathbf{Q}_k} - \mathbf{1}_N \otimes V^* = \mathbf{\Pi}^{\boldsymbol{\mu}_k} x_k$. Substituting this relation along with $x_k = u_k + \mathbf{U}^\top z_k$ into equation 129 gives

$$u_{k+1} = \left[I + \alpha\left(\gamma\mathbf{DP}\mathbf{\Pi}^{\boldsymbol{\mu}_k} - \mathbf{D} - \mathbf{S}\right)\right]u_k + \alpha\gamma\mathbf{DP}\mathbf{\Pi}^{\boldsymbol{\mu}_k}\mathbf{U}^\top z_k + \alpha\omega_{k+1}. \tag{133}$$

Combining equations 132–133, we obtain the projected switching-system representation

$$\begin{gathered} X_k := \mathrm{col}(u_k, z_k), \qquad \eta_{k+1} := \mathrm{col}(\omega_{k+1}, \mathbf{U}(\nu_{k+1} - \omega_{k+1})), \\ X_{k+1} = A_{\boldsymbol{\mu}_k,\perp} X_k + \alpha\eta_{k+1},\ A_{\boldsymbol{\mu},\perp} := \begin{bmatrix} I + \alpha\left(\gamma\mathbf{DP}\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{D} - \mathbf{S}\right) & \alpha\gamma\mathbf{DP}\mathbf{\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top \\ -\alpha\kappa\mathbf{U}\mathbf{S} & I - \alpha\kappa\mathbf{U}\mathbf{S}\mathbf{U}^\top \end{bmatrix}. \end{gathered} \tag{134}$$

For deterministic joint policies, we define the finite mode set

$$\mathcal{A}_{\alpha,\perp} := \left\{A_{\boldsymbol{\pi},\perp} : \boldsymbol{\pi} \in \Theta^N\right\}. \tag{135}$$

As with Algorithm 1, this forms the foundation for the subsequent stability and finite-time convergence analysis.

Analytical Framework and Key Lemmas. The remaining analysis follows the same structure as the finite-time analysis of Algorithm 1. We therefore present only the reduced counterparts of the essential results.

Lemma A.12. *For every joint stochastic policy $\boldsymbol{\mu}$, $A_{\boldsymbol{\mu},\perp} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})$ with the definition of equations 134–135. Consequently, $\rho\left(\mathrm{co}(\mathcal{A}_{\alpha,\perp})\right) = \rho(\mathcal{A}_{\alpha,\perp})$.*

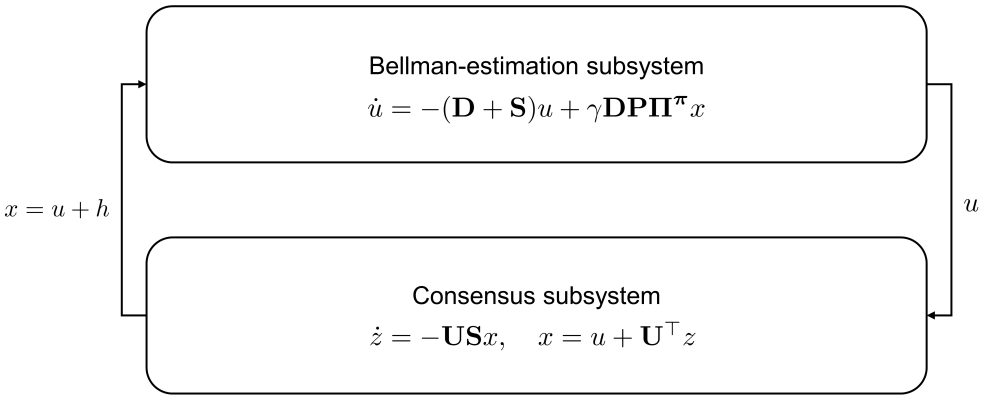


Figure 3: Interconnected-system view of R-ECQ without κ used in the proof of Lemma A.15. The upper block takes the consensus estimate x as input and produces the local estimate u. The lower block accumulates the disagreement in the projected state z, which is added back through $x = u + \mathbf{U}^\top z$ to form the feedback signal. Its complete dynamics are multiplied by the mixing gain κ.

Proof. The argument is identical to that of Lemma 5.1: a stochastic joint policy is a convex combination of the deterministic joint policies, and the policy enters $A_{\boldsymbol{\mu},\perp}$ affinely through the block $I+\alpha\left(\gamma\mathbf{DP\Pi}^{\boldsymbol{\mu}} - \mathbf{D} - \mathbf{S}\right)$ and $\alpha\gamma\mathbf{DP\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top$. Since an affine map preserves convex combinations, this yields $A_{\boldsymbol{\mu},\perp} \in \mathrm{co}(A_{\alpha,\perp})$. Finally, by the classical convex-hull invariance of the JSR (Jungers, 2009), $\rho(\mathrm{co}(A_{\alpha,\perp})) = \rho(A_{\alpha,\perp})$. □

To show that $\rho(\mathcal{A}_{\alpha,\perp}) < 1$, we follow a similar procedure to that in Appendix A.5. To this end, we rewrite equation 134 as

$$A_{\boldsymbol{\pi},\perp} = I + \alpha M_{\boldsymbol{\pi},\perp}, \qquad M_{\boldsymbol{\pi},\perp} := \begin{bmatrix} -(\mathbf{D}+\mathbf{S}) + \gamma\mathbf{DP\Pi}^{\boldsymbol{\pi}} & \gamma\mathbf{DP\Pi}^{\boldsymbol{\pi}}\mathbf{U}^\top \\ -\kappa\mathbf{US} & -\kappa\mathbf{USU}^\top \end{bmatrix} \tag{136}$$

for each deterministic joint policy $\boldsymbol{\pi} \in \Theta^N$. Moreover, as depicted in Figure 3, we view the recursion in equation 134 as an Euler discretization of the continuous-time linear ODE, which can be interpreted as two feedback-interconnected subsystems. Based on this formulation, we establish the following two lemmas by leveraging Lemma A.3 and Lemma A.6. This is justified because both ECQ and R-ECQ share the same upper block update equation, while Lemma A.6 relies solely on the existence of policy-independent (W_c, λ_c) and the affine structure $A_{\boldsymbol{\pi},\perp} = I + \alpha M_{\boldsymbol{\pi},\perp}$.

Lemma A.13. *Consider the consensus subsystem, i.e., the lower block of the interconnection in Figure 3, with input u and output x:*

$$\dot{z} = A_z z + B_z u, \qquad x = C_z z + u, \tag{137}$$

with $A_z := -\mathbf{USU}^\top$, $B_z := -\mathbf{US}$, $C_z := \mathbf{U}^\top$. Then, A_z is Hurwitz. Furthermore, defining

$$G_z := -A_z^{-1}B_z = -\mathbf{U}, \qquad J := \left(\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right) \otimes I_{|\mathcal{S}||\mathcal{A}|},$$

the steady-state input-output map is $u \mapsto x_\infty = Ju$, i.e., $C_z G_z + I = J$, which implies $\|x_\infty\|_\infty \leq \|u\|_\infty$ for every u.

Proof. To establish that A_z is Hurwitz, we express it as $A_z = -\mathbf{USU}^\top = -(U \otimes I_{|\mathcal{S}||\mathcal{A}|})(L \otimes D)(U^\top \otimes I_{|\mathcal{S}||\mathcal{A}|}) = -(ULU^\top) \otimes D$ using equation 9 and equation 17. Note that $L = L^\top \succeq 0$

with $\mathrm{null}(L) = \mathrm{span}\{\mathbf{1}_N\}$ by equation 1. Since the rows of U form an orthonormal basis for $\{\mathbf{1}_N\}^\perp$, $ULU^\top$ is the positive-definite restriction of L to $\{\mathbf{1}_N\}^\perp$, i.e., $x^\top Lx > 0$ for every nonzero $x \perp \mathbf{1}_N$, yielding $ULU^\top \succ 0$. Together with $D \succ 0$ being in diagonal form in equation 2, both $ULU^\top$ and D are symmetric positive definite. Consequently, their Kronecker product $(ULU^\top) \otimes D$ is symmetric positive definite, making $A_z = -(ULU^\top) \otimes D$ symmetric negative definite. Because A_z is symmetric, all its eigenvalues are real and strictly negative; therefore A_z is Hurwitz.

Next, we prove the second statement. Since A_z is Hurwitz, it is nonsingular. At an equilibrium $\dot{z} = 0$, solving $0 = \dot{z} = A_z z + B_z u$ gives $z_\infty = -A_z^{-1} B_z u$, so that

$$x_\infty = C_z z_\infty + u = C_z(-A_z^{-1})B_z u + u = (C_z G_z + I)u. \tag{138}$$

By the definitions of C_z and G_z, along with the properties of $\mathbf{U}$ from equation 42, we have

$$C_z G_z + I = -\mathbf{U}^\top \mathbf{U} + I = J. \tag{139}$$

Consequently, equation 138 implies

$$\|x_\infty\|_\infty = \|Ju\|_\infty \le \|u\|_\infty,$$

which completes the proof. □

Lemma A.14. *There exists a finite $\kappa^* > 0$ such that, for every fixed $\kappa > \kappa^*$, there exist a norm W_c and a constant $\lambda_c > 0$, both independent of $\boldsymbol{\pi} \in \Theta^N$, satisfying $D^+ W_c(X; M_{\boldsymbol{\pi},\perp} X) \le -\lambda_c W_c(X)$ for all X and $\boldsymbol{\pi} \in \Theta^N$, where $M_{\boldsymbol{\pi},\perp}$ is defined as in equation 136.*

Proof. We follow the same line of reasoning as in Lemma A.5. Using the definition in equation 62, we introduce the error variable relative to the steady-state equilibrium map as

$$\epsilon := z - G_z u = z + \mathbf{U}u.$$

Recall that $x = u + \mathbf{U}^\top z$ from equation 137. Substituting $z = \epsilon - \mathbf{U}u$ and applying $\mathbf{U}^\top(-\mathbf{U})u = (J - I)u$ from equation 139, we obtain

$$x = Ju + \mathbf{U}^\top \epsilon, \tag{140}$$

which yields

$$\dot{u} = A_0 u + B_{\boldsymbol{\pi}} x = F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon. \tag{141}$$

Moreover,

$$\begin{aligned}
\dot{\epsilon} =& \dot{z} + \mathbf{U}\dot{u} \\
=& \kappa(A_z z + B_z u) + \mathbf{U}(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon) \\
=& \kappa(A_z \epsilon - A_z \mathbf{U}u + B_z u) + \mathbf{U}(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon) \\
=& \kappa(A_z \epsilon + \mathbf{U}\mathbf{S}(\mathbf{U}^\top \mathbf{U})u - \mathbf{U}\mathbf{S}u) + \mathbf{U}(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon) \\
=& \kappa(A_z \epsilon - \mathbf{U}\mathbf{S}Ju) + \mathbf{U}(F_{\boldsymbol{\pi}} u + B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon) \\
=& \kappa A_z \epsilon + \mathbf{U}F_{\boldsymbol{\pi}} u + \mathbf{U}B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon,
\end{aligned}$$

where the second equality follows from equation 137 with the mixing gain κ and equation 141, the third equality uses $z = \epsilon - \mathbf{U}u$, and the fourth equality utilizes the definitions of A_z and B_z. The fifth equality applies $\mathbf{U}^\top(-\mathbf{U})u = (J - I)u$ from equation 139, and the last equality holds because $L\mathbf{1}_N = 0$ along with equation 9 and the definition of J. This yields a system similar to that in Lemma A.5, with C_{mix} and G replaced by $\mathbf{U}^\top$ and G_z, respectively.

Since A_z is Hurwitz, there exists a unique symmetric matrix $P_z \succ 0$ solving the algebraic Lyapunov equation $A_z^\top P_z + P_z A_z = -I$. We define

$$W_m(\epsilon) := \sqrt{\epsilon^\top P_z \epsilon}, \qquad \lambda_m := \frac{1}{2\lambda_{\max}(P_z)} > 0, \qquad b_{12} := \max_{\boldsymbol{\pi}\in\Theta^N} \sup_{\epsilon \neq 0} \frac{\|B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon\|_\infty}{W_m(\epsilon)},$$

$$b_{21} := \max_{\boldsymbol{\pi}\in\Theta^N} \sup_{u \neq 0} \frac{W_m(\mathbf{U}F_{\boldsymbol{\pi}} u)}{\|u\|_\infty}, \qquad b_{22} := \max_{\boldsymbol{\pi}\in\Theta^N} \sup_{\epsilon \neq 0} \frac{W_m(\mathbf{U}B_{\boldsymbol{\pi}} \mathbf{U}^\top \epsilon)}{W_m(\epsilon)}.$$

By applying arguments similar to those in Lemma A.5 and noting that $x = Ju$ when $\epsilon = 0$ from equation 140, we retrieve the same inequality as in equation 69. The three-term split of $\dot{\epsilon}$ likewise yields results identical to those in Lemma A.5. Following the procedure of Lemma A.5, we obtain the exact same threshold κ^*, with the resulting λ_c derived in precisely the same manner as in Lemma A.5. This completes the proof. □

Lemma A.15. *There exists a finite threshold $\kappa^* > 0$ such that, for every fixed $\kappa > \kappa^*$, there exists $\bar{\alpha}(\kappa) > 0$ such that the projected deterministic switching family $\mathcal{A}_{\alpha,\perp}$ with the definition of equation 135 satisfies $\rho(\mathcal{A}_{\alpha,\perp}) < 1$ for every $0 < \alpha < \bar{\alpha}(\kappa)$.*

Proof. We follow the proof argument of Proposition 5.1: the Bellman-estimation subsystem and the consensus subsystem are analyzed in Lemmas A.3 and A.13, the existence of a policy-independent pair (W_c, λ_c) is established in Lemma A.14, and the transfer of the continuous-time contraction property to the discrete-time Euler matrices $I + \alpha M_{\boldsymbol{\pi},\perp}$ is guaranteed by Lemma A.6. □

Lemma A.16. *Under Assumption 5.1, suppose that there exists $\bar{\alpha} \in (0,1)$ such that $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$ with the definition of equation 135. Then, there exist constants $c > 0$ and $K < \infty$, independent of α, such that for every $0 < \alpha \le \bar{\alpha}$:*

(i) $\rho(\mathcal{A}_{\alpha,\perp}) \le 1 - c\alpha < 1$.

(ii) *every product of matrices drawn from $\mathrm{co}(\mathcal{A}_{\alpha,\perp})$ satisfies*

$$\sup_{M_0,\dots,M_{\ell-1} \in \mathrm{co}(\mathcal{A}_{\alpha,\perp})} \|M_{\ell-1} \cdots M_0\|_2 \le K(1 - c\alpha)^\ell, \quad \ell \ge 0.$$

Proof. The proof follows the same argument as Lemma 5.2, where the only difference lies in the form of the matrix $A_{\boldsymbol{\mu},\perp}(\alpha) = I + \alpha H_{\boldsymbol{\mu},\perp}$ with equation 134 and

$$H_{\boldsymbol{\mu},\perp} := \begin{bmatrix} \gamma \mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}} - \mathbf{D} - \mathbf{S} & \gamma \mathbf{D}\mathbf{P}\boldsymbol{\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top \\ -\kappa \mathbf{U}\mathbf{S} & -\kappa \mathbf{U}\mathbf{S}\mathbf{U}^\top \end{bmatrix}.$$

This modification does not affect the proof, because the internal structure of $H_{\boldsymbol{\mu},\perp}$ is never required. The proof relies solely on each mode being affine in α and equal to the identity at $\alpha = 0$. □

Lemma A.17. *Let*

$$L_\Pi := \sup_{\boldsymbol{\mu},\boldsymbol{\nu}} \left\| \mathbf{D}\mathbf{P}(\boldsymbol{\Pi}^{\boldsymbol{\mu}} - \boldsymbol{\Pi}^{\boldsymbol{\nu}})[I \ \mathbf{U}^\top] \right\|_2,$$

where the supremum is taken over all joint stochastic policies μ and ν. Then, L_Π is finite, and for all μ and ν,

$$\|A_{\mu,\perp} - A_{\nu,\perp}\|_2 \le \alpha\gamma L_\Pi$$

with the definition of equation 134.

Proof. We apply the same technique as in Lemma 5.3. When subtracting equation 134 for two policies $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, all policy-independent terms cancel out, and L_Π remains finite due to the compactness of the policy space. The only difference lies in the difference matrix:

$$A_{\boldsymbol{\mu},\perp} - A_{\boldsymbol{\nu},\perp} = \alpha\gamma \begin{bmatrix} \mathbf{D}\mathbf{P}(\boldsymbol{\Pi}^{\boldsymbol{\mu}} - \boldsymbol{\Pi}^{\boldsymbol{\nu}})[I \ \mathbf{U}^\top] \\ 0 \end{bmatrix}.$$

Taking $\|\cdot\|_2$ gives $\|A_{\mu,\perp} - A_{\nu,\perp}\|_2 \le \alpha\gamma L_\Pi$. □

Lemma A.18. *Under Assumption 5.1, assume that the initialization of Algorithm 2 satisfies $(\mathbf{1}_N^\top \otimes I)(\mathbf{Q}_0 - \mathbf{W}_0) = 0$. Then, there exist α-independent constants $\sigma_0, \sigma_1 < \infty$ such that for all $k \ge 0$,*

$$\mathbb{E}[\eta_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}\big[\|\eta_{k+1}\|_2^2 \mid \mathcal{F}_k\big] \le \sigma_0^2 + \sigma_1^2 \|X_k\|_2^2,$$

where $\mathcal{F}_k = \sigma\left(\mathbf{W}_0, \mathbf{Q}_0, \{(s_t, a_t, s'_t, r_{t+1}^1, \dots, r_{t+1}^N) : 0 \le t \le k-1\}\right)$ is the natural filtration and η is defined in equation 134.

Proof. We follow the proof argument of Lemma 5.4. By construction, ω_{k+1} and ν_{k+1} have zero conditional mean. Since $\mathbf{U}$ is deterministic, it follows that $\mathbb{E}[\mathbf{U}(\nu_{k+1} - \omega_{k+1}) \mid \mathcal{F}_k] = 0$ as well. Therefore, $\mathbb{E}[\eta_{k+1} \mid \mathcal{F}_k] = 0$, which establishes the first claim.

For the second claim, we decompose the iterates around their limits using equation 128 and equation 131:

$$\mathbf{W}_k = \mathbf{W}_\infty + u_k, \qquad \mathbf{Q}_k = \mathbf{1}_N \otimes Q^* + u_k + \mathbf{U}^\top z_k.$$

Since $UU^\top = I_{N-1}$ with equation 40, we have $\mathbf{U}\mathbf{U}^\top = (UU^\top) \otimes I = I$, yielding $\|\mathbf{U}^\top\|_2 = 1$. Setting

$$K_0 := \max\{\|\mathbf{W}_\infty\|_\infty, \|\mathbf{1}_N \otimes Q^*\|_\infty\},$$

the block norm of X_k in equation 134 satisfies

$$\max_i \left\{\|W_k^i\|_\infty, \|Q_k^i\|_\infty\right\} \le K_0 + \sqrt{2}\|X_k\|_2.$$

By defining $R_{\max} := \sup_{i,s,a,s'} |r^i(s,a,s')| < \infty$ and the sampled updates

$$\begin{aligned}
\hat{h}^i_{W,k} &:= e_k \left[r^i_{k+1} + \gamma \max_{a' \in \mathcal{A}} Q^i_k(s'_k, a') - W^i_k(s_k, a_k) - \sum_j L_{ij} W^j_k(s_k, a_k) \right], \\
\hat{h}^i_{Q,k} &:= e_k \Bigg[r^i_{k+1} + \gamma \max_{a' \in \mathcal{A}} Q^i_k(s'_k, a') - W^i_k(s_k, a_k) \\
&\qquad - \sum_j L_{ij} W^j_k(s_k, a_k) - \kappa \sum_j L_{ij} Q^j_k(s_k, a_k) \Bigg],
\end{aligned} \tag{142}$$

we observe that $\hat{h}^i_{W,k}$ and $\hat{h}^i_{Q,k}$ satisfy, almost surely,

$$|\hat{h}^i_{W,k}(s_k, a_k)| \le R_{\max} + \left(\gamma + 1 + 2\max_i |\mathcal{N}_i|\right)(K_0 + \sqrt{2}\|X_k\|_2),$$

$$|\hat{h}^i_{Q,k}(s_k, a_k)| \le R_{\max} + \left(\gamma + 1 + 2(1+\kappa)\max_i |\mathcal{N}_i|\right)(K_0 + \sqrt{2}\|X_k\|_2).$$

For simplicity, define

$$a_W = \sqrt{2}(\gamma + 1 + 2\max_i |\mathcal{N}_i|), \qquad a_Q = \sqrt{2}(\gamma + 1 + 2(1+\kappa)\max_i |\mathcal{N}_i|)$$

and

$$b_W = R_{\max} + (\gamma + 1 + 2\max_i |\mathcal{N}_i|)K_0, \qquad b_Q = R_{\max} + (\gamma + 1 + 2(1+\kappa)\max_i |\mathcal{N}_i|)K_0.$$

Since each sampled update has only one possibly nonzero coordinate per agent with $\|e_k\|_2 = 1$, it holds that

$$\|\hat{h}^i_{W,k}\|_2^2 = |\hat{h}^i_{W,k}(s_k, a_k)|^2, \qquad \|\hat{h}^i_{Q,k}\|_2^2 = |\hat{h}^i_{Q,k}(s_k, a_k)|^2.$$

Stacking these terms over all N agents yields

$$\|\hat{h}_{W,k}\|_2^2 = \sum_{i=1}^N \|\hat{h}^i_{W,k}\|_2^2 = \sum_{i=1}^N |\hat{h}^i_{W,k}(s_k, a_k)|^2 \le N\left(b_W + a_W\|X_k\|_2\right)^2,$$

$$\|\hat{h}_{Q,k}\|_2^2 = \sum_{i=1}^N \|\hat{h}^i_{Q,k}\|_2^2 = \sum_{i=1}^N |\hat{h}^i_{Q,k}(s_k, a_k)|^2 \le N\left(b_Q + a_Q\|X_k\|_2\right)^2.$$

From equation 134, using $\|\mathbf{U}\|_2 = 1$ along with $(m+n)^2 \le 2m^2 + 2n^2$ for all $m, n \in \mathbb{R}$, we obtain

$$\|\eta_{k+1}\|_2^2 \le \|\omega_{k+1}\|_2^2 + \|\mathbf{U}\|_2^2\|\nu_{k+1} - \omega_{k+1}\|_2^2 \le 3\|\omega_{k+1}\|_2^2 + 2\|\nu_{k+1}\|_2^2.$$

By the definition of the variance, we have $\mathbb{E}[\|\omega_{k+1}\|_2^2 \mid \mathcal{F}_k] \le \mathbb{E}[\|\widehat{h}_{W,k}\|_2^2 \mid \mathcal{F}_k]$ and $\mathbb{E}[\|\nu_{k+1}\|_2^2 \mid \mathcal{F}_k] \le \mathbb{E}[\|\widehat{h}_{Q,k}\|_2^2 \mid \mathcal{F}_k]$, which implies

$$\begin{aligned}
\mathbb{E}[\|\eta_{k+1}\|_2^2 \mid \mathcal{F}_k] &\le 3N(b_W + a_W\|X_k\|_2)^2 + 2N(b_Q + a_Q\|X_k\|_2)^2 \\
&\le 2N\left(3b_W^2 + 2b_Q^2\right) + 2N\left(3a_W^2 + 2a_Q^2\right)\|X_k\|_2^2.
\end{aligned}$$

This completes the proof of the second bound with

$$\sigma_0^2 := 2N(3b_W^2 + 2b_Q^2), \qquad \sigma_1^2 := 2N(3a_W^2 + 2a_Q^2),$$

both of which are independent of α, as in Lemma 5.4. □

Finite-Time Convergence Analysis. We now state the following corollaries, which establish a uniform mean-square error bound and an explicit bound on $\mathbb{E}[\|X_k\|_2]$ for every iteration k.

Corollary A.1 (Mean-square bound for Algorithm 2)**.** *Under Assumption 5.1, assume that the initialization of Algorithm 2 satisfies* $(\mathbf{1}_N^\top \otimes I)(\mathbf{Q}_0 - \mathbf{W}_0) = 0$ *and that there exists* $\bar{\alpha} \in (0,1)$ *such that* $\rho(\mathcal{A}_{\bar{\alpha},\perp}) < 1$ *with the definition of equation 135. Let* c, K *be as in Lemma A.16,* σ_0, σ_1 *as in Lemma A.18,* L_Π *as in Lemma A.17, and* C_0, C_1, α_0 *as in equation 22. Then, for every* $0 < \alpha \le \alpha_0$ *and* $k \ge 0$*,*

$$\sup_{0 \le t \le k} \mathbb{E}\left[\|Q_t - \mathbf{1}_N \otimes Q^*\|_2^2\right] \le 4C_0\|X_0\|_2^2 + 4C_1\alpha\sigma_0^2.$$

Proof. We adapt the proof argument of Proposition 5.2. The proof of Proposition 5.2 relies on Lemmas 5.1–5.4; replacing these with Lemmas A.12, A.16, A.17, and A.18 leaves every estimate unchanged and yields the exact same bound on X_k.

The only modification is in the conclusion lines in the Q-error bound. Since $x_t = u_t + \mathbf{U}^\top z_t$ from equation 128 and equation 131, and $\|\mathbf{U}^\top\|_2 = 1$, we have

$$\|\mathbf{Q}_t - \mathbf{1}_N \otimes Q^*\|_2^2 = \|x_t\|_2^2 \le \left(\|u_t\|_2 + \|z_t\|_2\right)^2 \le 2\left(\|u_t\|_2^2 + \|z_t\|_2^2\right) = 2\|X_t\|_2^2,$$

which doubles the resulting constants. □

Corollary A.2 (Finite-time bound for Algorithm 2)**.** *Under the hypotheses and notation of Corollary A.1, set* $\beta := 1 - c\alpha/2$ *for* $0 < \alpha \le \alpha_0$*. Then, for all* $k \ge 0$ *with the convention* $k\beta^{k-1} = 0$ *at* $k = 0$*,* X_k *in equation 134 satisfies the same finite-time bound as in equation 24. Consequently,*

$$\mathbb{E}[\|\mathbf{Q}_k - \mathbf{1}_N \otimes Q^*\|_2] \le \sqrt{2}\mathbb{E}[\|X_k\|_2].$$

Proof. We adapt the proof argument of Theorem 5.1. The proof of Theorem 5.1 relies on Lemma 5.2 and Proposition 5.2; replacing these with Lemma A.16 and Corollary A.1 leaves every estimate unchanged and yields the exact same bound on X_k.

The only modification in the conclusion lies in the Q-error bound. Since $x_t = u_t + \mathbf{U}^\top z_t$ from equation 128 and equation 131, and $\|\mathbf{U}^\top\|_2 = 1$, we have

$$\|\mathbf{Q}_k - \mathbf{1}_N \otimes Q^*\|_2 \le \|u_k\|_2 + \|z_k\|_2 \le \sqrt{2}\|X_k\|_2,$$

which accounts for the additional factor of $\sqrt{2}$ in the stated bound. □

Exact consensus of the disagreement component in R-ECQ. Similar to Theorem 5.1, Corollary A.2 bounds the total error of R-ECQ, which decomposes according to equation 96. We show that, under the sampling model of Algorithm 2 and the deterministic reward heterogeneity condition in equation 97, the consensus error of R-ECQ likewise carries no residual term. The argument parallels that in Appendix A.11.

Switching System Formulation for Exact Consensus Analysis. Unlike ECQ, where the disagreement of Q_k is represented by a single block $\tilde{x}_k$ as in equation 98, R-ECQ removes x_k from the state space via $x_k = u_k + \mathbf{U}^\top z_k$ using equation 128 and equation 131. Consequently, the disagreement of Q_k is expressed as the sum

$$\mathbf{U}x_k = \mathbf{U}u_k + z_k = \tilde{u}_k + z_k.$$

We define the reduced disagreement state as

$$\tilde{u}_k := \mathbf{U}u_k, \qquad \tilde{x}_k := \mathbf{U}x_k = \tilde{u}_k + z_k. \tag{143}$$

Applying Lemma A.1 agent-wise with $\bar{Q}_k$ in place of Q^* yields the same relation as in equation 99. Moreover, using equations 42, 96, and 143, we have

$$Q_k - \mathbf{1}_N \otimes \bar{Q}_k = (I - \mathbf{J})x_k = \mathbf{U}^\top \tilde{x}_k. \tag{144}$$

We then obtain

$$\mathbf{UDP}V_{Q_k} = \mathbf{UDP\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top(\tilde{u}_k + z_k).$$

Applying $\mathbf{U}$ to the u-recursion in equation 129 and using $\mathbf{U}(\mathbf{D}+\mathbf{S}) = (\mathbf{D}_\perp + \mathbf{S}_\perp)\mathbf{U}$ from equation 100, we get

$$\tilde{u}_{k+1} = \tilde{u}_k + \alpha\left[\gamma \mathbf{U}\mathbf{D}\mathbf{P}\mathbf{\Pi}^{\tilde{\boldsymbol{\mu}}_k}\mathbf{U}^\top(\tilde{u}_k + z_k) - (\mathbf{D}_\perp + \mathbf{S}_\perp)\tilde{u}_k\right] + \alpha\mathbf{U}\omega_{k+1}.$$

The z-recursion in equation 132 is already in disagreement coordinates; using $\mathbf{U}\mathbf{S}u_k = \mathbf{S}_\perp\tilde{u}_k$ and $\mathbf{U}\mathbf{S}\mathbf{U}^\top = \mathbf{S}_\perp$, it becomes

$$z_{k+1} = z_k - \alpha\kappa\mathbf{S}_\perp\tilde{u}_k - \alpha\kappa\mathbf{S}_\perp z_k + \alpha\mathbf{U}(\nu_{k+1} - \omega_{k+1}).$$

Hence, the system can be written as

$$\tilde{X}_k := \mathrm{col}(\tilde{u}_k, z_k), \qquad \tilde{\eta}_{k+1} := \mathrm{col}\left(\mathbf{U}\omega_{k+1}, \mathbf{U}(\nu_{k+1} - \omega_{k+1})\right),$$
$$\tilde{X}_{k+1} = \tilde{A}_{\tilde{\boldsymbol{\mu}}_k}\tilde{X}_k + \alpha\tilde{\eta}_{k+1}, \tag{145}$$
$$\tilde{A}_{\boldsymbol{\mu}} := \begin{bmatrix} I + \alpha\left(\gamma\mathbf{U}\mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top - \mathbf{D}_\perp - \mathbf{S}_\perp\right) & \alpha\gamma\mathbf{U}\mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\mu}}\mathbf{U}^\top \\ -\alpha\kappa\mathbf{S}_\perp & I - \alpha\kappa\mathbf{S}_\perp \end{bmatrix}.$$

Analytical Framework and Key Lemmas for Exact Consensus Analysis. Since the policy enters $\tilde{A}_{\boldsymbol{\mu}}$ in equation 101 affinely, we can apply Lemma A.12 to obtain $\tilde{A}_{\boldsymbol{\mu}} \in \mathrm{co}(\tilde{\mathcal{A}}_\alpha)$, where $\tilde{\mathcal{A}}_\alpha$ is defined as in equation 102, and write $\tilde{A}_{\boldsymbol{\mu}}(\alpha) = I + \alpha\tilde{H}_{\boldsymbol{\mu}}$ with $\tilde{H}_{\boldsymbol{\mu}}$ independent of α. Since subtracting two modes yields

$$\tilde{A}_{\boldsymbol{\mu}} - \tilde{A}_{\boldsymbol{\nu}} = \alpha\gamma\begin{bmatrix}\mathbf{U}\mathbf{D}\mathbf{P}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}})\mathbf{U}^\top[I\ I] \\ 0\end{bmatrix},$$

by applying Lemma A.17, we define

$$\tilde{L}_\Pi := \sup_{\boldsymbol{\mu},\boldsymbol{\nu}} \left\|\mathbf{U}\mathbf{D}\mathbf{P}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}})\mathbf{U}^\top[I\ I]\right\|_2 \le \sqrt{2}\sup_{\boldsymbol{\mu},\boldsymbol{\nu}}\left\|\mathbf{D}\mathbf{P}(\mathbf{\Pi}^{\boldsymbol{\mu}} - \mathbf{\Pi}^{\boldsymbol{\nu}})\right\|_2 < \infty. \tag{146}$$

We write $\tilde{A}_{\boldsymbol{\pi}} = I + \alpha\tilde{M}_{\boldsymbol{\pi}}$ with

$$\tilde{M}_{\boldsymbol{\pi}} := \begin{bmatrix}\gamma\mathbf{U}\mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\pi}}\mathbf{U}^\top - \mathbf{D}_\perp - \mathbf{S}_\perp & \gamma\mathbf{U}\mathbf{D}\mathbf{P}\mathbf{\Pi}^{\boldsymbol{\pi}}\mathbf{U}^\top \\ -\kappa\mathbf{S}_\perp & -\kappa\mathbf{S}_\perp\end{bmatrix}, \tag{147}$$

and view the continuous-time system $\dot{\tilde{X}} = \tilde{M}_{\boldsymbol{\pi}}\tilde{X}$ analogously to Figure 3, with κ removed from the lower block. The feedback interconnection consists of the Bellman-estimation block represented as equation 104 and the consensus block given by

$$\dot{z} = \tilde{A}_z z + \tilde{B}_z\tilde{u}, \qquad \tilde{x} = \tilde{C}_z z + \tilde{u}, \qquad \tilde{A}_z := -\mathbf{S}_\perp, \qquad \tilde{B}_z := -\mathbf{S}_\perp, \qquad \tilde{C}_z := I. \tag{148}$$

Since the Bellman block is shared with ECQ, Lemma A.7 applies directly.

Lemma A.19. *Under equation 148, $\tilde{A}_z$ is Hurwitz, and $\tilde{G}_z := -\tilde{A}_z^{-1}\tilde{B}_z = -I$. Consequently, the steady-state input-output map is $\tilde{u} \mapsto \tilde{x}_\infty = (\tilde{C}_z\tilde{G}_z + I)\tilde{u} = 0$.*

Proof. By a method similar to that in Lemma A.13, $\tilde{A}_z$ is trivially Hurwitz. Moreover, by definition, we have $\tilde{G}_z = -(-\mathbf{S}_\perp)^{-1}(-\mathbf{S}_\perp) = -I$, which yields $\tilde{C}_z\tilde{G}_z + I = -I + I = 0$, thereby completing the proof. □

Lemma A.20. *There exists a finite threshold $\tilde{\kappa}^* > 0$ such that, for every fixed $\kappa > \tilde{\kappa}^*$, there exists a norm $\tilde{W}_c$ and a constant $\tilde{\lambda}_c > 0$, both independent of $\boldsymbol{\pi} \in \Theta^N$, satisfying $D^+\tilde{W}_c(\tilde{X}; \tilde{M}_{\boldsymbol{\pi}}\tilde{X}) \le -\tilde{\lambda}_c\tilde{W}_c(\tilde{X})$ for all $\tilde{X}$ and all $\boldsymbol{\pi} \in \Theta^N$, with $\tilde{M}_{\boldsymbol{\pi}}$ as in equation 147.*

Proof. We follow a procedure similar to that in Lemma A.14, with $(\tilde{A}_{\mathrm{mix}}, \tilde{B}_{\mathrm{mix}}, \tilde{C}_{\mathrm{mix}}, \tilde{G})$ replaced by $(\tilde{A}_z, \tilde{B}_z, \tilde{C}_z, \tilde{G}_z)$. Define

$$\tilde{\epsilon} := z - \tilde{G}_z\tilde{u} = z + \tilde{u} = \tilde{x}.$$

Substituting this relation into equation 104 and equation 148 yields

$$\dot{\tilde{u}} = \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{\epsilon}, \qquad \dot{\tilde{\epsilon}} = \dot{z} + \dot{\tilde{u}} = \kappa\tilde{A}_z\tilde{\epsilon} + \tilde{A}_0\tilde{u} + \tilde{B}_0\tilde{\epsilon}.$$

Let $\tilde{P}_m \succ 0$ be the unique solution to $\tilde{A}_z^\top \tilde{P}_m + \tilde{P}_m \tilde{A}_z = -I$. Using $\tilde{W}_m(\tilde{\epsilon})$ and $\tilde{\lambda}_m$ as defined in Lemma A.9, we define

$$\tilde{b}_{12} := \max_{\boldsymbol{\pi}\in\Theta^N} \sup_{\tilde{\epsilon}\neq 0} \frac{\left\|\mathbf{U}^\top \tilde{B}_0 \tilde{\epsilon}\right\|_\infty}{\tilde{W}_m(\tilde{\epsilon})}, \qquad \tilde{b}_{21} := \sup_{\tilde{u}\neq 0} \frac{\tilde{W}_m(\tilde{A}_0 \tilde{u})}{\|\mathbf{U}^\top \tilde{u}\|_\infty}, \qquad \tilde{b}_{22} := \max_{\boldsymbol{\pi}\in\Theta^N} \sup_{\tilde{\epsilon}\neq 0} \frac{\tilde{W}_m(\tilde{B}_0 \tilde{\epsilon})}{\tilde{W}_m(\tilde{\epsilon})}.$$

Defining $\tilde{W}_c := \|\mathbf{U}^\top \tilde{u}\|_\infty + \theta \tilde{W}_m(\tilde{\epsilon})$, the remainder of the proof of Lemma A.9 applies directly, yielding the same threshold $\tilde{\kappa}^* = (\tilde{b}_{22} + \tilde{b}_{12}\tilde{b}_{21}/d_{\min})/\tilde{\lambda}_m$ and rate $\tilde{\lambda}_c = \min\{d_{\min} - \theta\tilde{b}_{21},\ \kappa\tilde{\lambda}_m - \tilde{b}_{22} - \tilde{b}_{12}/\theta\} > 0$. Finally, $\tilde{W}_c$ defines a valid norm on the state space since the linear transformation $(\tilde{u}, z) \mapsto (\tilde{u}, z + \tilde{u})$ is invertible, and the pair $(\tilde{W}_c, \tilde{\lambda}_c)$ is policy-independent as all gain constants are independent of $\boldsymbol{\pi}$. □

Lemma A.21. *Let $\tilde{\kappa}^*$ be as in Lemma A.20. For every fixed $\kappa > \tilde{\kappa}^*$ there is $\tilde{\alpha}(\kappa) > 0$ such that the projected deterministic switching family $\tilde{\mathcal{A}}_\alpha$ with the definition of equation 102 satisfies $\rho(\tilde{\mathcal{A}}_\alpha) < 1$ for every $0 < \alpha < \tilde{\alpha}(\kappa)$. Consequently, for any $\bar{\alpha} \in (0, 1)$ with $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\bar{\alpha}})) < 1$, there exist α-independent constants $\tilde{c} > 0$, $\tilde{K} < \infty$ such that equation 108 holds for all $0 < \alpha \leq \bar{\alpha}$.*

Proof. The proof is identical to that of Lemma A.10, with Lemma A.20 supplying the policy-independent pair $(\tilde{W}_c, \tilde{\lambda}_c)$ for the finite family $\{\tilde{M}_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \Theta^N\}$ in equation 147. Furthermore, Lemmas A.6 and A.16 rely solely on the existence of this pair, the affine structure $\tilde{A}_{\boldsymbol{\pi}} = I + \alpha \tilde{M}_{\boldsymbol{\pi}}$, and $\tilde{A}_{\boldsymbol{\mu}} \in \mathrm{co}(\tilde{\mathcal{A}}_\alpha)$. □

Lemma A.22. *Under Assumption 5.1, assume that the initialization of Algorithm 2 satisfies $(\mathbf{1}_N^\top \otimes I)(\mathbf{Q}_0 - \mathbf{W}_0) = 0$. Then, there exists an α-independent constant $\tilde{\sigma}_1 < \infty$ such that for all $k \geq 0$,*

$$\mathbb{E}[\tilde{\eta}_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}\left[\|\tilde{\eta}_{k+1}\|_2^2 \mid \mathcal{F}_k\right] \leq \tilde{\sigma}_1^2 \|\tilde{X}_k\|_2^2,$$

where $\mathcal{F}_k = \sigma\left(\mathbf{W}_0, \mathbf{Q}_0, \{(s_t, a_t, s_t', r_{t+1}^1, \ldots, r_{t+1}^N) : 0 \leq t \leq k-1\}\right)$ is the natural filtration and η is defined in equation 145.

Proof. The first claim follows from Lemma 5.4 together with the fact that $\mathbf{U}$ is deterministic.

For the second claim, we show that the norms of $\mathbf{U}\hat{h}_{W,k}$ and $\mathbf{U}\hat{h}_{Q,k}$ with equation 142 are bounded by a constant multiple of $\|\tilde{X}_k\|_2$ without an additive constant. Since the W-update is shared with ECQ and the second relation of Proposition A.1 coincides with that of Proposition 4.1, the computation in equation 110 applies directly, with the single modification that the target disagreement is now evaluated at $\tilde{x}_k = \tilde{u}_k + z_k$ via equation 143. Since $\|\tilde{x}_k\|_2 \leq \|\tilde{u}_k\|_2 + \|z_k\|_2 \leq \sqrt{2}\|\tilde{X}_k\|_2$, we have, almost surely,

$$\left\|\mathbf{U}\hat{h}_{W,k}\right\|_2 \leq \left(1 + 2\max_i |\mathcal{N}_i|\right) \|\tilde{u}_k\|_2 + \gamma\|\tilde{x}_k\|_2 \leq \left(1 + 2\max_i |\mathcal{N}_i| + \sqrt{2}\gamma\right) \left\|\tilde{X}_k\right\|_2.$$

For $\hat{h}_{Q,k}$, the bracket term in equation 142 equals that of $\hat{h}_{W,k}$ minus $\kappa L Q_k(s_k, a_k)$. Using $UL = ULU^\top U$ along with equation 41, and $UQ_k(s, a) = \tilde{x}_k(s, a)$ from equation 144 and $U\mathbf{1}_N = 0$ from equation 40, we obtain, almost surely,

$$\left\|\mathbf{U}\hat{h}_{Q,k}\right\|_2 \leq \left\|\mathbf{U}\hat{h}_{W,k}\right\|_2 + \kappa\|L\|_2\|\tilde{x}_k\|_2 \leq \left(1 + 2\max_i |\mathcal{N}_i|(1 + \sqrt{2}\kappa) + \sqrt{2}\gamma\right) \left\|\tilde{X}_k\right\|_2.$$

By equation 145, $\|\mathbf{U}\|_2 = 1$, and the inequality $(m+n)^2 \leq 2m^2 + 2n^2$ for all $m, n \in \mathbb{R}$, we get

$$\|\tilde{\eta}_{k+1}\|_2^2 \leq \|\mathbf{U}\omega_{k+1}\|_2^2 + 2\|\mathbf{U}\nu_{k+1}\|_2^2 + 2\|\mathbf{U}\omega_{k+1}\|_2^2 = 3\|\mathbf{U}\omega_{k+1}\|_2^2 + 2\|\mathbf{U}\nu_{k+1}\|_2^2.$$

As in Lemma A.18, we have $\mathbb{E}[\|\mathbf{U}\omega_{k+1}\|_2^2 \mid \mathcal{F}_k] \leq \mathbb{E}[\|\mathbf{U}\hat{h}_{W,k}\|_2^2 \mid \mathcal{F}_k]$ and similarly for ν_{k+1} and $\hat{h}_{Q,k}$. Combining these bounds yields $\mathbb{E}[\|\tilde{\eta}_{k+1}\|_2^2 \mid \mathcal{F}_k] \leq \tilde{\sigma}_1^2 \|\tilde{X}_k\|_2^2$, where

$$\tilde{\sigma}_1^2 := 3\left(1 + 2\max_i |\mathcal{N}_i| + \sqrt{2}\gamma\right)^2 + 2\left(1 + 2\max_i |\mathcal{N}_i| \left(1 + \sqrt{2}\kappa\right)^2 + \sqrt{2}\gamma\right)^2,$$

which yields the desired conclusion. □

Consensus Error Convergence. Now, we show that the consensus error contains no residual term and converges to zero geometrically.

Corollary A.3. *Under Assumption 5.1, deterministic reward heterogeneity, and the initialization* $(\mathbf{1}_N^\top \otimes I)(\mathbf{Q}_0 - \mathbf{W}_0) = 0$*, suppose there exists* $\bar{\alpha} \in (0,1)$ *such that* $\rho(\mathrm{co}(\tilde{\mathcal{A}}_{\bar{\alpha}})) < 1$ *with* $\tilde{\mathcal{A}}_\alpha$*. Then, there exist constants* $\tilde{C}_0 < \infty$*,* $\tilde{c} > 0$*, and* $\tilde{\alpha}_0 \in (0, \bar{\alpha}]$ *such that, for every* $0 < \alpha \le \tilde{\alpha}_0$ *and* $k \ge 0$*, the consensus error of R-ECQ satisfies*

$$\mathbb{E}\left[\max_i \left\|Q_k^i - \bar{Q}_k\right\|_\infty\right] \le 2\sqrt{\tilde{C}_0}\left(1 - \frac{\tilde{c}\alpha}{2}\right)^{k/2} \left\|\tilde{X}_0\right\|_2,$$

which decays geometrically as $k \to \infty$ *without any* α*-dependent residual error.*

Proof. Let $\tilde{c}, \tilde{K}$ be as in Lemma A.21, $\tilde{\sigma}_1$ as in Lemma A.22, and $\tilde{L}_\Pi$ as in equation 146. Define $\tilde{C}_0$, $\tilde{C}_1$, and $\tilde{\alpha}_0$ as in the proof of Corollary 5.1, with L_Π replaced by $\tilde{L}_\Pi$. The reference-filter decomposition $\tilde{X}_k = \bar{Y}_k + \Xi_k + U_k + V_k$ and the inductive argument in that proof rely solely on the product bound equation 108, the mode-difference bound $\|\tilde{A}_{\boldsymbol{\mu}} - \tilde{A}_{\boldsymbol{\nu}}\|_2 \le \alpha\gamma\tilde{L}_\Pi$, and the noise bound in Lemma A.22. Since all three conditions hold here, the same argument yields

$$\mathbb{E}\left[\left\|\tilde{X}_k\right\|_2^2\right] \le 2\tilde{C}_0\left(1 - \frac{\tilde{c}\alpha}{2}\right)^k \left\|\tilde{X}_0\right\|_2^2.$$

The only modification appears in the final step. Using equation 96, $(I - \mathbf{J})x_k = \mathbf{U}^\top\mathbf{U}x_k$ with equation 42, $\|\mathbf{U}^\top\|_2 = 1$, and equation 143, we obtain

$$\max_i \left\|Q_k^i - \bar{Q}_k\right\|_\infty \le \|(I - \mathbf{J})x_k\|_2 = \left\|\mathbf{U}^\top\tilde{x}_k\right\|_2 = \|\tilde{u}_k + z_k\|_2 \le \sqrt{2}\left\|\tilde{X}_k\right\|_2,$$

which accounts for the additional factor of $\sqrt{2}$. Applying Jensen's inequality completes the proof. □

A.13 Experimental Setup and Additional Experiments

To evaluate ECQ (Algorithm 1) and R-ECQ (Algorithm 2), we consider two environments: a Garnet MDP (Bhatnagar et al., 2009) and a customized 5×5 stochastic gridworld. In both settings, agents share identical transition dynamics but observe distinct, local reward functions that are deterministic in (s, a). Their collective objective is to reach consensus on the team-optimal Q-function Q^*, allowing us to test whether the proposed methods can learn a unified solution despite heterogeneous individual rewards.

In the Garnet MDP, we set $|\mathcal{S}| = 10$ states and $|\mathcal{A}| = 4$ actions. For each state-action pair, three next states are chosen uniformly at random with Dirichlet-sampled transition probabilities. Each agent's reward function is constructed by adding a small agent-specific perturbation to a shared random base reward, introducing distinct individual preferences. Unless specified otherwise, experiments run for $K = 5 \times 10^5$ iterations on a ring topology. We also test path, star, and complete graphs to analyze network connectivity effects (see Figure 4).

For the gridworld benchmark, agents navigate a 5×5 grid using four cardinal actions (up, down, left, and right). Actions succeed with probability 0.9; with probability 0.1, the agent drifts to a randomly selected adjacent cell, with transitions constrained by the grid boundaries. Each agent is assigned a unique goal cell, receiving a higher reward at its goal and a small penalty elsewhere. Experiments in this environment run for $K = 10^6$ iterations over a fixed ring topology.

In both environments, we set the number of agents to $N = 8$ and the discount factor to $\gamma = 0.9$. Both environments are illustrated in Figures 5 and 6, respectively. All experimental curves report the mean across 10 random seeds, with shaded regions representing ± 1 standard deviation. We compare ECQ against the following baseline algorithms:

1. IQL (Tan, 1993): Each agent executes standard tabular Q-learning using only its local reward with no inter-agent communication, treating all other agents as part of the environment.

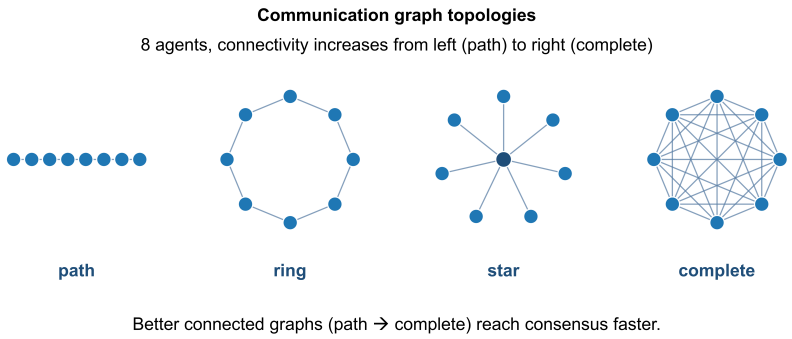


Figure 4: Overview of the communication graph topologies. Denser graphs achieve consensus faster (Olfati-Saber & Murray, 2004).

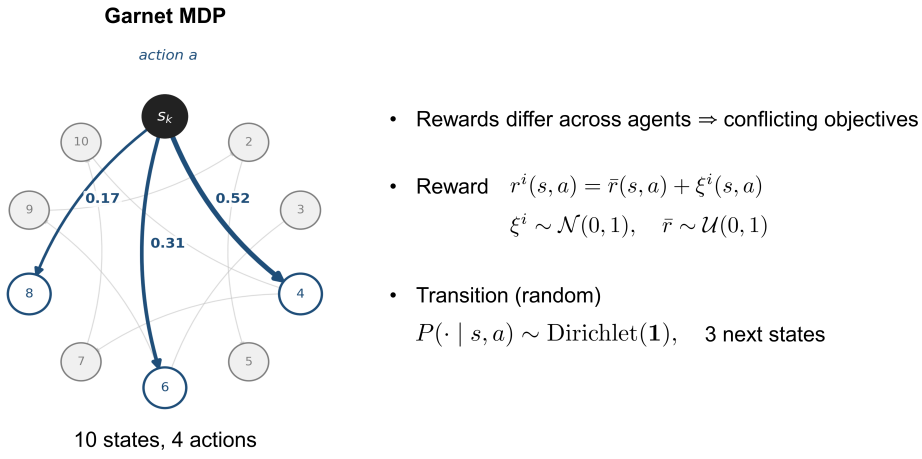


Figure 5: Overview of the garnet MDP environment.

2. $\mathcal{QD}$-learning (Kar et al., 2013): A distributed Q-learning scheme based on the consensus-plus-innovations framework, in which each agent mixes its neighbors' estimates with its own local update under decaying step-sizes.

3. Mixing Q-learning (Lim & Lee, 2025): Each agent performs a constant-step-size local Q-learning update and replaces its estimate with a convex combination of its neighbors' estimates using a doubly stochastic mixing matrix (corresponding to Algorithm 2 in Lim & Lee (2025)).

4. ECQ ablation: Variant of ECQ (Algorithm 1) with the auxiliary variable Z removed, such that agents only compensate for immediate local disagreement without accumulating disagreement history over time.

5. Centralized Q-learning (oracle) (Watkins & Dayan, 1992): A centralized learner executing standard Q-learning on the mean team reward $\bar{r} = \frac{1}{N}\sum_{i=1}^{N} r^i$. This serves as an ideal upper bound with zero consensus error by construction.

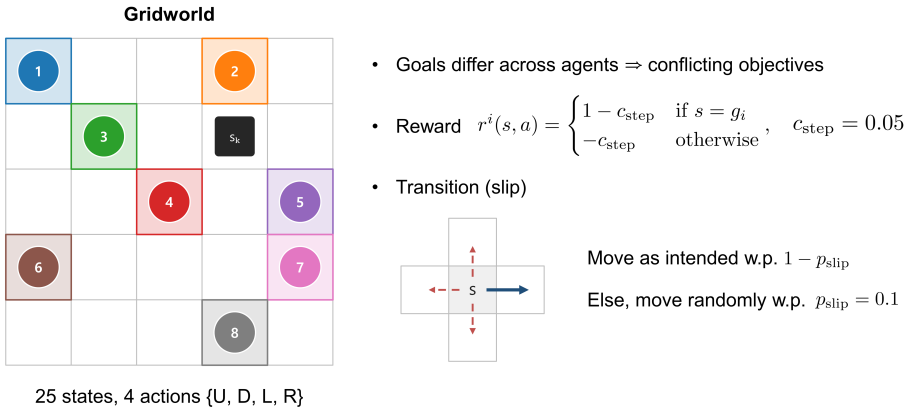


Figure 6: Overview of the 5×5 stochastic gridworld environment.

Table 1: Baseline hyperparameter search configurations. Each baseline is optimized via grid search, with performance evaluated using the mean error over the final 10% of iterations. By contrast, ECQ and R-ECQ use a fixed step-size of $\alpha = 0.05$ without any hyperparameter tuning.

Method	Parameter	Search grid	Selected
IQL	step-size α	$\{0.0125, 0.025, 0.05, 0.1, 0.2\}$	0.0125
$\mathcal{QD}$-learning	a (innovation gain)	$\{0.5, 1.0, 2.0, 5.0\}$	1.0
	b (consensus gain)	$\{0.2, 0.4, 0.8\}$	0.4
	τ_2 (decay rate)	$\{0.5, 0.6, 0.75\}$	0.6
Mixing Q-learning	step-size α	$\{0.0125, 0.025, 0.05, 0.1, 0.2\}$	0.025
ECQ / R-ECQ / oracle	step-size α	*not tuned*	0.05

Unless specified otherwise, ECQ and R-ECQ operate under a fixed step-size of $\alpha = 0.05$ across all tests without tuning. Conversely, constant- and decaying-step-size baselines are optimized via grid search, with top-performing hyperparameter settings detailed in Table 1. Moreover, we set $\kappa = 1$ for both ECQ and R-ECQ. While Theorem 5.1 and Corollary A.2 require a sufficiently large κ, the theoretical threshold κ^* is highly conservative due to worst-case assumptions. In practice, we choose a smaller value of $\kappa = 1$ for evaluating quantitative performance.

Figure 7 shows the consensus error $\max_i \|Q_k^i - \bar{Q}_k\|_\infty$, where $\bar{Q}_k = \frac{1}{N}\sum_i Q_k^i$, across four communication topologies ranging from complete to path graphs. Here, centralized Q-learning is omitted because its single Q-function yields zero disagreement by construction. Across the tested topologies, the ECQ and R-ECQ curves decay toward machine precision within the reported iteration budgets, with R-ECQ doing so faster than ECQ. Since network topology controls the speed of convergence rather than the final error floor, consensus is fastest on the highly connected complete graph and slows down as connectivity drops. In contrast, mixing Q-learning and the ECQ ablation settles on positive error floors, $\mathcal{QD}$-learning reduces disagreement gradually but retains a larger residual within the finite iteration budget, and IQL fails to reach agreement altogether due to the lack of communication. Note that the observed residual does not establish a positive asymptotic consensus error for $\mathcal{QD}$-learning: under the assumptions and diminishing-gain conditions of Kar et al. (2013), it converges almost surely to the common optimum, so the consensus error does vanish - only its diminishing gains make this decay too slow to matter within a finite budget, leaving it orders of magnitude above ECQ throughout the run.

Figure 8 shows the total error $\max_i \|Q_k^i - Q^*\|_\infty$ across the same communication topologies. Under every topology, ECQ and R-ECQ converge to the steady-state error floor of the centralized

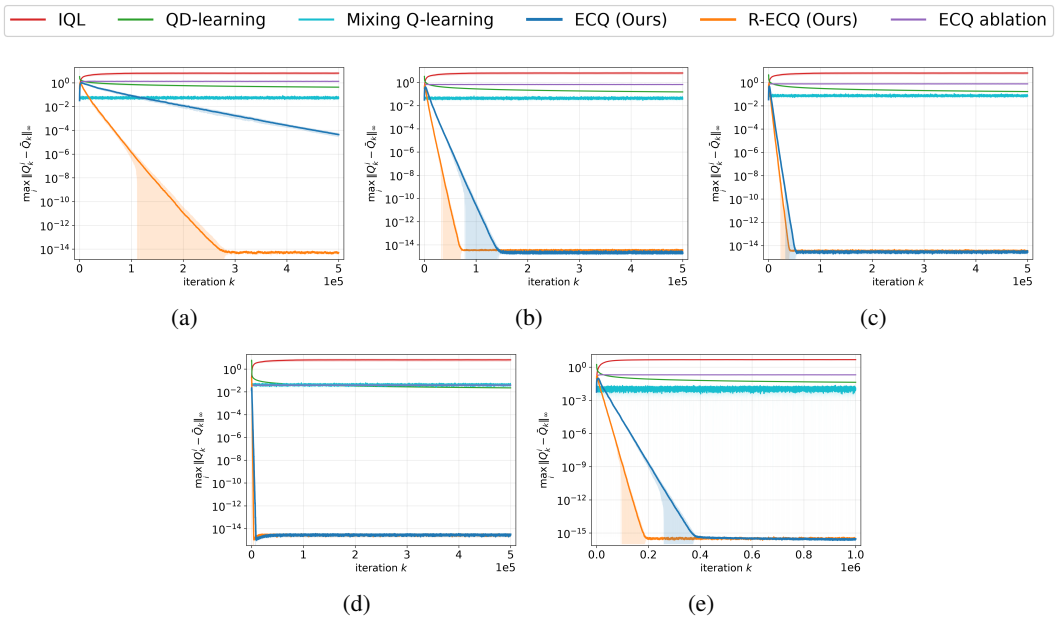


Figure 7: Consensus error comparison: (a)–(d) Garnet MDP under path, ring, star, and complete topologies, respectively; (e) stochastic gridworld environment.

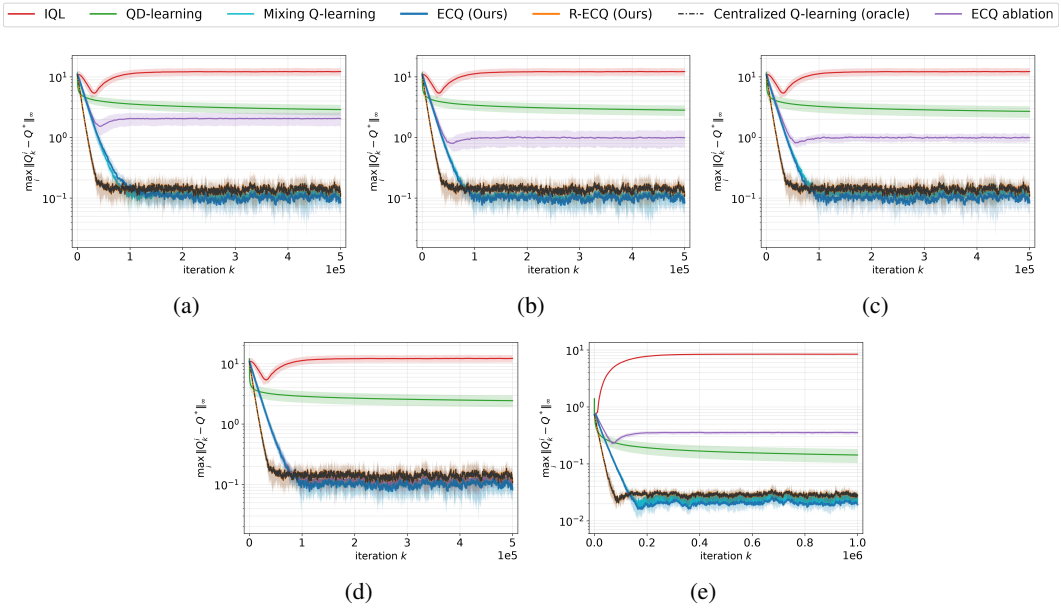
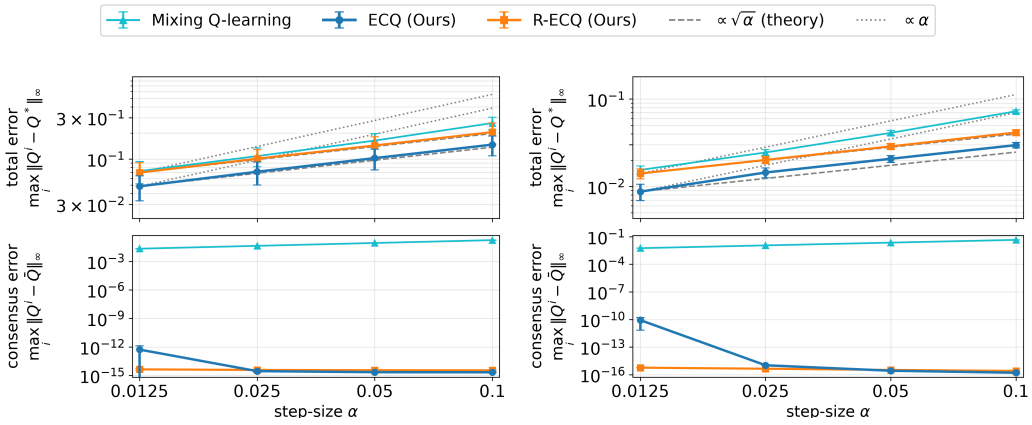


Figure 8: Convergence to team-optimal Q^*: (a)–(d) Garnet MDP under path, ring, star, and complete topologies, respectively; (e) stochastic gridworld environment.

Q-learning without accuracy penalty. Since R-ECQ's average update becomes equivalent to centralized Q-learning on the mean team reward once agreement is established, it closely matches the oracle. Mixing Q-learning settles at a comparable error level because optimality error is dominated by the persistent noise of a constant step-size rather than by the consensus dynamics. Within the reported iteration budget, the other baselines do not reach this error level: non-communicating IQL converges toward individual private optima and deviates from Q^*, $\mathcal{QD}$-learning with decaying gains



(a) (b)

Figure 9: Step-size dependence of the error floors: (a) Garnet MDP and (b) stochastic gridworld under a ring topology.

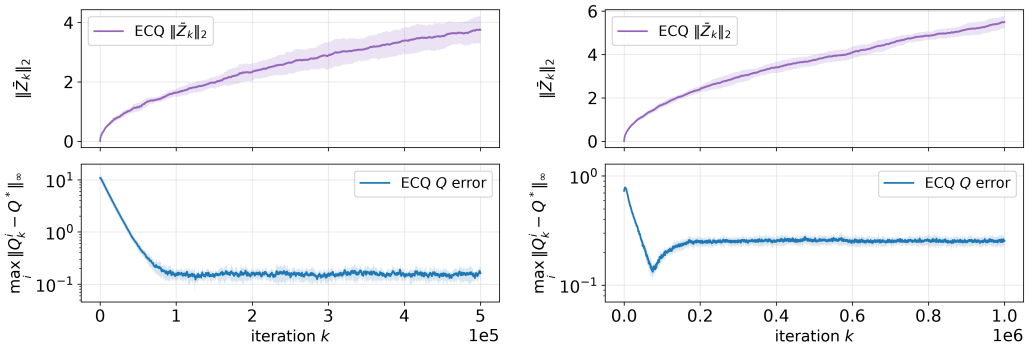
decreases slowly and ends with a larger finite-budget error, and the ECQ ablation settles at a higher error floor whose magnitude depends directly on the network topology.

Figure 9 isolates the dependence of the two error floors on the constant step-size α. This step-size sweep uses the ring topology; it does not establish topology-independent sharp asymptotic scaling, and the constants in the theoretical upper bounds can depend on the graph. We benchmark ECQ and R-ECQ exclusively against mixing Q-learning, which is the sole baseline incorporating both a constant step-size and a consensus mechanism. In the top panel, we evaluate the total error $\max_i \|Q_k^i - Q^*\|_\infty$. By fitting steady$(\alpha) = C\alpha^p$ to the experimental data for ECQ and R-ECQ yields exponents of $\hat{p} = 0.537 \pm 0.067$ and $\hat{p} = 0.518 \pm 0.038$ in the Garnet MDP, and $\hat{p} = 0.590 \pm 0.100$ and $\hat{p} = 0.522 \pm 0.053$ in the gridworld, respectively. These fitted exponents are empirical and compatible with the $O(\sqrt{\alpha})$ upper bounds in Theorem 5.1 and Corollary A.2. Being upper bounds, however, they establish neither a matching $\Theta(\sqrt{\alpha})$ order, a limiting error, nor a positive floor, and do not rule out $O(\alpha)$ or smaller error in particular instances. The residual total error is consistent with such bounds: rather than stemming solely from team-average reward noise, it reflects next-state sampling in the Bellman target as well as reward sampling when present. Having characterized the total error, we now turn to the consensus error, on which ECQ and R-ECQ are further distinguished. In the bottom panel, we measure the consensus error $\max_i \|Q_k^i - \bar{Q}_k\|_\infty$. While mixing Q-learning is bounded by an error floor scaling as $O(\alpha)$ (Lim & Lee, 2025), ECQ and R-ECQ achieve machine-precision consensus in sufficiently long runs for the tested stable step-sizes, consistent with the geometric-decay guarantees for admissible α in Corollaries 5.1 and A.3. Note that for $\alpha = 0.0125$, the higher consensus error of ECQ stems merely from an insufficient number of iterations relative to the small step size. However, R-ECQ reaches its error floor faster for small α by applying consensus correction directly to the $W - Q$ gap rather than through the accumulated Z.

Figure 10 evaluates the robustness of ECQ under imperfect communication settings. In this experiment, synthetic noise is applied exclusively to the Z-communication channel on a ring topology. Specifically, we inject i.i.d. Gaussian noise with a standard deviation of 0.3 exclusively into the Z-update, while keeping the Q-channel noise-free. Such a setting is reasonable for modeling communication-specific errors, since local observations and computations can remain reliable while the communication link may be affected by noise, quantization, or transmission disturbances. Here, $\bar{Z}_k = \frac{1}{N}\sum_i Z_k^i$ denotes the average value of the Z state across all agents. Figure 10 shows a small late-iteration team-optimality error even as $\bar{Z}_k$ drifts.

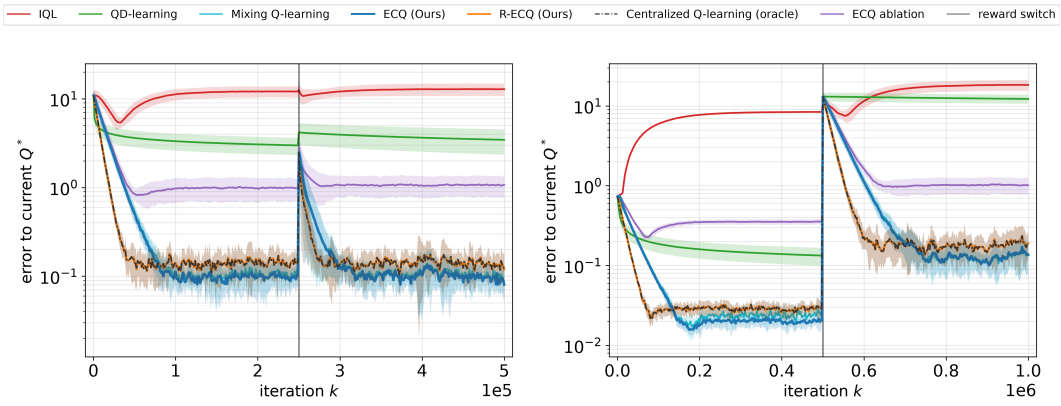
This behavior is consistent with the structure of the auxiliary coupling. Decomposing a perturbation of the stacked auxiliary state as

$$\delta\mathbf{Z}_k = \mathbf{J}\delta\mathbf{Z}_k + (I - \mathbf{J})\delta\mathbf{Z}_k, \qquad \mathbf{SJ} = 0.$$



(a) (b)

Figure 10: Convergence in Q under communication noise: (a) Garnet MDP and (b) stochastic gridworld under a ring topology.



(a) (b)

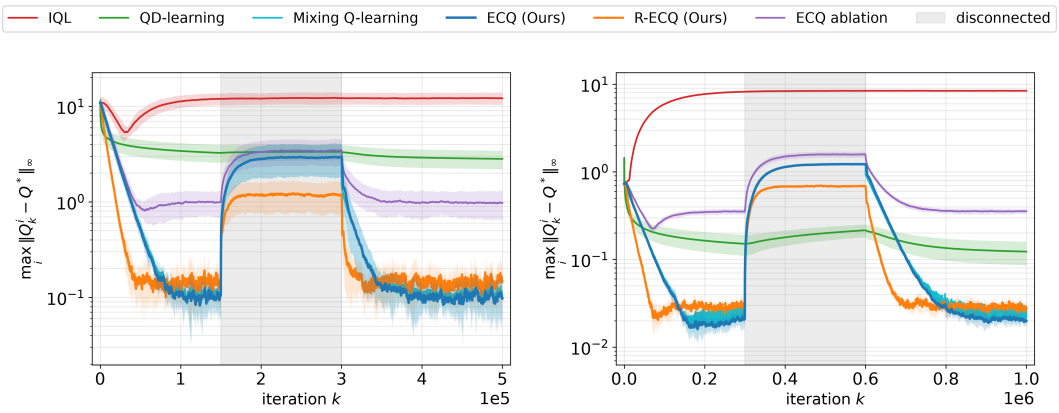
Figure 11: Re-tracking after a reward change: (a) Garnet MDP and (b) stochastic gridworld under a ring topology.

Its contribution through the auxiliary-state coupling in the conditional-mean Q-update is

$$-\alpha\kappa\mathbf{S}\delta\mathbf{Z}_k = -\alpha\kappa\mathbf{S}(I - \mathbf{J})\delta\mathbf{Z}_k.$$

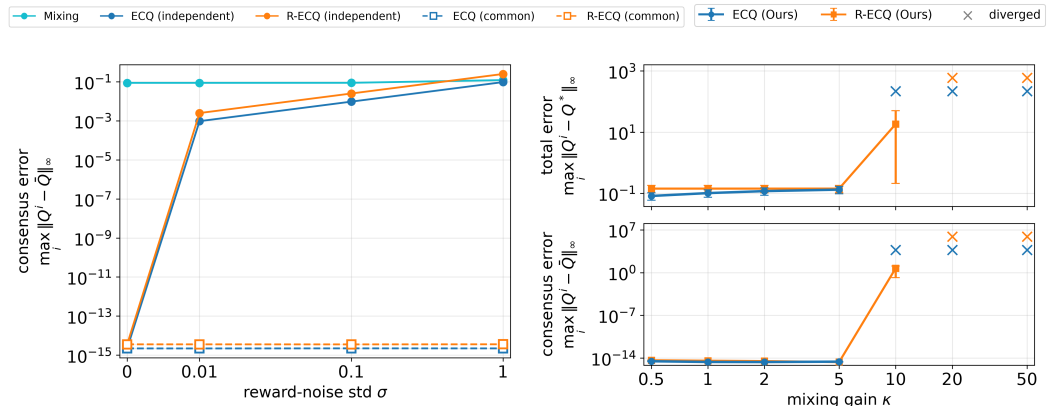
Thus, a perturbation entirely in the consensus direction is canceled, so the drift of $\bar{Z}_k$ does not corrupt Q, whereas agent-specific disagreement noise can. Consequently, the nonuniqueness of the auxiliary equilibrium in Proposition 4.1, a consensus-direction property as in dynamic average consensus (Yang et al., 2008), explains the tolerance to $\bar{Z}_k$ drift but does not by itself imply robustness to arbitrary Z-channel noise. Accordingly, Figure 10 is an empirical robustness observation rather than a proof of convergence under general Z-channel noise. The finite-time theorems analyze the original sampling-noise model, and extending them to this additional communication disturbance requires a separate bound for the perturbed recursion.

We further evaluate network robustness under two types of disruptions on a ring graph: a sudden reward change (Figure 11) and a temporary communication disconnection (Figure 12). Although we present results on a ring topology, the overall trend remains the same across different graphs. Figure 11 considers a mid-training reward switch, where the target changes from Q_1^* to Q_2^* using a new random realization of the original Garnet reward procedure. The same procedure is applied to both environments; in the gridworld, the switched reward departs from the standard goal-based structure to evaluate re-tracking performance under an abrupt target change rather than a shift in goal locations. Figure 12 models a temporary network disconnection, where the graph is split into two isolated subgraphs before being restored. Centralized Q-learning is omitted here, as a single-agent



(a) (b)

Figure 12: Recovery after temporary disconnection: (a) Garnet MDP and (b) stochastic gridworld under a ring topology.



(a) (b)

Figure 13: (a) Effect of reward noise on consensus error floor in the Garnet MDP under a ring topology. (b) Effect of the mixing gain κ on the error floors in the Garnet MDP under a ring topology.

architecture cannot represent inter-agent disconnection. In both scenarios, ECQ and R-ECQ rapidly re-track the new optimum after the change, where R-ECQ exhibits a faster recovery. Notably, in Figure 11, R-ECQ tracks the new target almost as quickly as centralized Q-learning. This occurs because, once agreement is established, R-ECQ's average update becomes equivalent to centralized updates on the mean team reward. Mixing Q-learning achieves a similar steady-state error because its performance is limited by constant step-size noise rather than consensus dynamics. Conversely, other baselines lacking this adaptivity: the ECQ ablation settles at a higher error, $\mathcal{QD}$-learning adapts too slowly, and IQL remains stuck due to its lack of communication. Overall, these results demonstrate that ECQ and R-ECQ reliably adapt to both environment shifts and network outages.

Figure 13 evaluates the sensitivity to the key parameters introduced in this paper: the reward noise n in equation 97 and the mixing gain κ in Algorithms 1 and 2. Note that both ablation studies are conducted solely on the Garnet MDP, as their primary goal is to illustrate how the error floors depend on the reward-noise structure and the mixing gain.

The left side of Figure 13 evaluates the scope of the exact-consensus claim by introducing per-agent reward noise $n_k^i \sim \mathcal{N}(0, \sigma^2)$ and sweeping $\sigma \in \{0, 0.01, 0.1, 1\}$ at the fixed step-size $\alpha = 0.05$. Two noise models are compared: *independent* noise, drawn separately for each agent, which makes the inter-agent difference $r^i - r^j$ stochastic and thus violates the deterministic-heterogeneity

condition in equation 97; and *common* noise, a single scalar added identically to all agents, which leaves $r^i - r^j$ deterministic and preserves the reward-difference structure of equation 97. At $\sigma = 0$, both ECQ and R-ECQ reach a consensus floor near machine precision, whereas mixing Q-learning stays at its $O(\alpha)$ floor. However, under *independent* noise, the consensus floor of both ECQ and R-ECQ increases with σ, indicating that exact consensus is not preserved when the deterministic reward heterogeneity condition is violated. Conversely, under *common* noise, the consensus error floor of ECQ and R-ECQ remains at machine precision for all values of σ, coinciding with the $\sigma = 0$ case. For $\sigma > 0$, Gaussian reward noise is unbounded, so these cases lie outside the stated bounded-reward assumptions of Corollaries 5.1 and A.3. Their observed behavior is qualitatively consistent with the deterministic-heterogeneity cancellation mechanism, not a direct consequence of the stated guarantees. Although common noise cancels under the disagreement projection, extending the guarantees to this unbounded-noise model requires a separate justification of the relevant noise and moment bounds.

The right side of Figure 13 evaluates the sensitivity to κ by sweeping $\kappa \in \{0.5, 1, 2, 5, 10, 20, 50\}$ at a fixed step-size $\alpha = 0.05$, reporting the final total error $\max_i \|Q_k^i - Q^*\|_\infty$ and consensus error $\max_i \|Q_k^i - \bar{Q}_k\|_\infty$. For $\kappa \in \{0.5, 1, 2, 5\}$, both ECQ and R-ECQ remain stable: the consensus error reaches machine precision independently of κ, while the total error floor increases only mildly. However, beyond this range, ECQ diverges at $\kappa = 10$, while R-ECQ degrades sharply at $\kappa = 10$ and diverges by $\kappa = 20$. These observations concern the sampled iterates, not only their conditional-mean dynamics. Proposition 5.1 and Lemma A.15 give sufficient conditions for stability of the mean-dynamics switching family. The stochastic guarantee additionally requires $0 < \alpha \leq \alpha_0$, where α_0 is defined in equation 22; in particular, when $\sigma_1 > 0$, it also requires $\alpha \leq (2C_1\sigma_1^2)^{-1}$. These conditions enter Proposition 5.2 for ECQ and Corollary A.1 with the corresponding constants for R-ECQ. The sweep does not evaluate the JSR or these sufficient step-size bounds, so the observed divergence does not identify which bound is exceeded. For a fixed κ satisfying the relevant mean-dynamics stability assumptions, the theory guarantees stability only when α is also small enough to meet the stochastic restriction. Thus, the curves demonstrate sensitivity to the chosen pair (κ, α), rather than determine a theoretical stability boundary. The choice $\kappa = 1$ performed stably in the reported runs at $\alpha = 0.05$; this observation does not certify its inclusion in a theoretical admissible region or quantify the conservatism of κ^*.